COMPUTATIONAL TOOLS FOR ENDANGERED LANGUAGE DOCUMENTATION

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Antonios Anastasopoulos

David Chiang, Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana

April 2019

COMPUTATIONAL TOOLS FOR ENDANGERED LANGUAGE DOCUMENTATION

Abstract

by

Antonios Anastasopoulos

The traditional method for documenting a language involves the collection of audio or video sources, which are then annotated at multiple granularity levels by trained linguists. This is a painstaking and time-consuming process, which could benefit from machine learning techniques at almost all stages. However, most existing machine learning methods are being developed for high-resource languages and rely on abundant data, rendering them unsuitable for such applications.

At the same time, for many low-resource and endangered languages speech data is easier to obtain than textual data, particularly since most of the world's languages are unwritten. Nevertheless, it is relatively easy to provide written or spoken translations for audio sources, as speakers of a minority language are often bilingual and literate in a high-resource language.

This work is aimed at solving certain problems that arise in the documentation process of an endangered language, due to the minimal annotated resources that are available at this stage. This dissertation mainly focuses on spoken corpora of endangered and low-resource languages with limited translation annotations, tackling problems that cover every layer of linguistic annotation:

- speech-to-translation alignment: we present an unsupervised method for discovering word or phrase boundaries in the audio signal and aligning the discovered segments with translation words.

- speech transcription: we developed two novel neural methods for creating a phoneme or grapheme level transcription of the audio, also utilizing any available translations.
- speech translation: our novel multitask neural model jointly produces a transcription and a (free) translation of an audio segment.
- morphological analysis: producing a layer of annotation that provides word-level or morpheme-level information. In this work we focus on grammatical (part-of-speech) tagging on an endangered language.

Building off limited or no annotations, our methods are capable of producing helpful suggestions for word or phrase boundaries, as well as transcriptions, translations, or grammatical tags. As a result, our work provides the machine learning methods that could form the backbone of a modern linguistic annotation toolkit, one that could have the potential to significantly accelerate the language documentation process.

CONTENTS

FIGURES

TABLES

ACKNOWLEDGMENTS

After finishing my undergraduate degree in Greece and a brief research internship in Italy, I had set my mind on a career in research, so my next step would need to be a PhD. Hopeful and determined, I applied to numerous schools in the United States, only to be rejected by all of them. But to my good fortune, my aspirations didn't end there, as my application was salvaged from the "reject" pile by David Chiang, in the midst of transferring from the University of Southern California to Notre Dame. To this day, I cannot believe that I was the recipient of such a generous dose of serendipity, and I am immensely grateful that David decided to take a chance on an unlikely and seemingly inauspicious electrical engineering undergraduate from the National Technical University of Athens.

As providence would have it, David turned out to be the best advisor I could have hoped for, and I am nothing short of honored to have been his student. He is a pillar of intellect and ethics and an incredibly knowledgeable and kind mentor, who supported me in exploring both my own ideas and academia as a whole, guiding me towards ambitions and consequential goals. His principled disposition and dedication to his students and family are a constant inspiration for me, exemplifying everything I aspire to someday become.

In addition, I am very grateful to my committee members, Graham Neubig, Walter Scheirer, and Tim Weninger for their comments and feedback in my PhD proposal and the final drafts of this dissertation. I am also thankful to Sharon Goldwater and Adam Lopez for their mentorship during the time that I spent visiting the University of Edinburgh. They too were an integral part of my journey, and a continuation of that turn of good favor that pulled me unsuspectingly from the West of the West Coast to the West of the Midwest.

I am also more than indebted to my immediate academic siblings: Tomer Levinboim,

Arturo Argueta, and Kenton Murray. Tomer helped me immensely, leading by example, with the transition from student to researcher, with navigating conferences and academia in general, and with conducting effective research. Arturo and Kenton where there from the very beginning of this ND journey 5 years ago, from picking me up the first time I set foot on campus, to Taco Tuesdays every week, to graduating together. And beyond being just excellent researchers and colleagues, each helped me acclimate to a new country, and each became a dear friend.

In addition, I have had the opportunity to work with or alongside many incredible and generous individuals: from Notre Dame, Toan Nguyen, Justin DeBenedetto, Brian DuSell, Paige Rodeghero, and Alison Lui; from Melbourne, Long Duong, Trevor Cohn, and Steven Bird; in Edinburgh, Sameer Bansal and Spandana Gela; from Grenoble, Marcely Zanon-Boito and Laurent Besacier; from Ioannina, Eleni Zimianiti and Marika Lekakou; from Barcelona, Josep Quer; and at Google NYC, Chris Alberti, Manaal Faruqui, Shankar Kumar, and Hank Liao.

I would also like to extend my gratitude to Joyce Yeats, who, on top of keeping the whole ND CSE graduate department running, was always there to ensure that I didn't miss a deadline, to help navigate the painful bureaucracy, and to keep me and the other graduate students well fed on countless occasions.

Finally, I could not have possibly reached the end of this journey without the support of my family back in Greece: my father, sister, and aunts and uncles, as well as my good friends, Foteini, Christos, Nikolas, Valentinos, and others. It was often those hour-long Skype calls that kept me going during the South Bend winters. And last but not least, I would like to dedicate this dissertation to my late mother Dionysia, to my sister Stavroula, and to my fiance Gabriela, as a token of my gratitude for their love, support, and constant belief in me.

CHAPTER 1

INTRODUCTION

Throughout human history, knowledge and culture have, for the most part, been passed from generation to generation through oral tradition. In fact, language and culture are so inseparably connected, that the loss of the former often marks the end of the latter. According to recent conservative estimates from UNESCO's Atlas Of The World's Languages In Danger" (UNESCO, 2010), more than 43% of the world's languages are endangered or vulnerable to extinction.

The loss of a language is deemed to have a staggering number of negative consequences. Importantly, according to Lee and Van Way (2016), it leads the loss of cultural or ethnic identity (Tsunoda, 2017), the loss of knowledge of prehistory by losing the only means of reconstructing words about a culture's past (Evans, 2011), the loss of linguistic diversity (Hale, 1992) and of part of the sum of human knowledge (Crystal, 2000), including traditional ecological knowledge. Estimates of the rate of language extinctions vary from the worst-case scenario of about 90% of the world's languages disappearing within 100 years (Krauss, 2007), to the more moderate, but not less catastrophic rate of a language disappearing very three months (Campbell et al., 2013). Acknowledging the value and importance of these languages, not only to their respective communities, but to humanity as a whole, significant efforts have been channelled towards their documentation and preservation.

Language preservation can be loosely defined as all efforts aimed towards maintaining or creating a pool of active, native speakers of a language, such that it will ensure the continuing usage of the language. A language starts to become *endangered* when transgen-

erational transmission starts to decrease, i.e. the language is not passed down to the young generations. Other factors that lead to endangerment can be the lack of official (state) support for the language, political turmoil, or association of the language with a lower social class. Naturally, in the majority of cases, preservation efforts consist of collecting the knowledge of –typically– elder fluent speakers and teaching the language to children, in order to create a new generation of fluent speakers.

Systematic language documentation, on the other hand, aims to codify the rules that describe a language, that is, its grammar, as well as to study its use among a speech community. Himmelmann (1998) defines language documentation as "a comprehensive and representative sample of communicative events as natural as possible," which is geared more towards the scientific understanding of the language. However, despite their fundamentally different goals, documentation can be and often is an integral part of the process of language preservation; if anything, a proper codification of the language is crucial for creating educational material.

In order to achieve proper documentation, however, a comprehensive record of the language is needed, which can be then used to not only study and preserve the language, but also to aid any revitalization efforts. Traditionally, this record is collected by field linguists who study the phonetics, phonology, morphology, syntax, and so forth, of the language. In order to make a collection of speech data usable for future studies of the language, the minimum requirement is something resembling Interlinear Glossed Text (IGT), which includes a *transcription* with a phonetic or standard orthography, *morphological analysis* and *glosses* which provide sub-word and word-level information, and finally a *free translation* in a more high-resource language, which captures the semantics of an utterance. An example from Russian, that follows the Leipzig Glossing Rules (Bickel et al., 2008) is shown here:

In the recent decades, modern technology in the form of tools, webpages, and apps is also increasingly used in the documentation process. In Chapter 2, we provide an in-

My	s	Marko	poexa-l-i	avtobus-om	v	Peredelkino
1PL	COM	Marko	go-PST-PL	bus-INS	All	Peredelkino
we	with	Marko	go-PST-PL	bus-by	to	Peredelkino

"Marko and I went to Perdelkino by bus "

depth discussion of the involvement of technology and computational methods in language documentation.

The traditional paradigm of a linguist working manually, or with minimal computational assistance, is still the prevalent one. However, our modern, data-driven era opens up the path to an alternative paradigm, originally proposed by Liberman (2006) and Bird (2010). In the traditional documentation scenario, a linguist can ask questions aimed at clarifying a specific phenomenon of a language. In our "alternative" documentation scenario, the assumption is that the answers to such questions will arise from the volume of the data and can be discovered by combining statistical methods with linguistic knowledge.

Although somewhat unorthodox, this "quantity over quality" approach will be our general pre-supposed framework throughout this dissertation, for two main reasons. First, abundance of data in a language could lead to the development of technologies that in turn could be combined with the traditional approach, allowing a linguist to produce larger quantities of high quality annotations. Second, this scenario is applicable in the case of severely endangered languages that stand little chance of preservation. Instead of lamentably futile attempts in revitalization or preservation, one could focus instead on collecting as (ideally interpretable) data as possible. Unfortunately, the case for severely endangered languages, i.e. with a handful of elderly native speakers, is that there is simply not enough time left to take the "traditional" documentation approach. Instead the linguist could first ensure that enough data are collected, and only later analyze them.

The first aspect of documentation that can be significantly scaled using modern technologies is data collection. In fact, new mobile and web-based technologies are being de-

veloped to facilitate collection of spoken samples in endangered languages, as well as translations (Bird et al., 2014a). The collection of translations ensures the interpretability of the resource even if the language eventually falls out of use, and the quantity of the data collected aims to counter the loss of quality that stems from the lack of a specific elicitation process with the linguist's participation. Recent examples of such parallel speech collection efforts focused on endangered languages are already underway (Adda et al., 2016; Blachon et al., 2016).

After data collection, the next step of the documentation process, and the most time consuming one, is analysis. For example, it is estimated that it takes a trained linguist about an hour to phonetically transcribe a minute of speech (Thi-Ngoc-Diep Do and Castelli, 2014). Therefore, we propose to develop computational methods that could automate parts of the documentation process. Since the aforementioned data collection scenario provides not only speech in an endangered language, but also translations in a high-resource language, we propose computational methods that combine approaches from the fields of automatic speech recognition and machine translation.

The reason that our approach requires techniques from the speech processing field is non-trivial. Out of the 7,097 living languages currently listed in Ethnologue (Lewis et al., 2009), only 3,909 have a developed writing system. In many of these instances, it might still be the case that a writing system exists but is not widely adopted by the community. The remaining 3,188 languages most likely exist only in spoken form without a standardized writing system. This presents further obstacles to the systematic documentation of these languages.

In this dissertation, I describe our research efforts towards the goal of enhancing the language documentation pipeline with machine learning. The main theme of this work is that the techniques take advantage of translations in a higher-resource language. Overall, this work provides a set of computational methods that could be employed at each sub-task of a language documentation pipeline, potentially speeding up the entire process.

1.1 Document Structure

This dissertation contains the following chapters:

- In Chapter 2 we provide a description of the language documentation process and an overview of the current state of relevant computational methods. We also summarize our contributions and briefly introduce the language corpora that we have used throughout this work.

- In Chapter 3 we present a novel unsupervised method for speech-to-translation alignment, and show how such alignments can be used for spoken term discovery in unannotated corpora. We also present a case study showing that a transcription interface that provides speech-to-translation alignments leads to better crowdsourced mismatched transcriptions. This chapter is based on an EMNLP'16, a Comput-EL 2, and a SCNLP'17 paper (Anastasopoulos et al., 2017; Anastasopoulos and Chiang, 2017; Anastasopoulos et al., 2016).

- In Chapter 4 we present three neural architectures that leverage translations for speech transcriptions. *Multi-source* models can be used when translations are available at test time. *DLUPI* models that learn using translations as privileged information, could be used when translations are only available during training, achieving comparable performance. *Tied multitask* models, finally, train a transcription and a translation model jointly and are able to produce both at test time. This chapter is based on a NAACL'18 and an Interspeech'18 paper (Anastasopoulos and Chiang, 2018a,b), as well as unpublished ongoing work.

- In Chapter 5 we describe how we collected a new resource on Griko, an endangered language, along with Italian translations. Furthermore, we combined a semi-supervised part-of-speech tagging method with cross lingual projections in order to provide word-level grammatical tags for our corpus. We also discuss how we employed active learning to facilitate faster annotation of the test set by our linguist collaborators.

- In Chapter 6 we summarize our contributions, and discuss possible directions for future work, in light of the needs of the linguistics community for computational assistance in the documentation process.

1.2 Thesis Statement

Annotation of linguistic resources is done in multiple layers and in several annotating passes. An example of tiered annotation is shown in Figure 1.1, where each tier provides different information. We expand the standard IGT formulation to include the corresponding audio, also time-aligning it with the corresponding annotation tiers. The different tiers include transcription (in a working orthography, or a phonetic-level International Phonetic Alphabet one), glossing, morphological analysis, and a free translation.

Since the data collection framework under which we operate provides audio in an endangered language and its translation in a high-resource one, the main thesis of this work can be summarized as follows:

> **Thesis:** Machine learning techniques that leverage translations can be applied in every layer of tiered linguistic annotation, accelerating the language documentation process.

el	ɣa.to		se	’sjen.to	en	la	a.’fo.βra
el	gato		se	sientò	en	la	afobra
the.Masc	cat.Masc		3SG	sit.3SG.PST	on	the.F	mat

‘The cat sat on the mat’

Figure 1.1. Visualization of tiered annotation of a speech utterance, equivalent to Interlinear Gloss Text. The alignments with the speech utterance are often annotated through a tool like PRAAT or ELAN. The tiers of this example correspond to phonetic transcription, orthographic transcription, glossing with morphological information, and a free translation.

CHAPTER 2

TECHNOLOGY IN LANGUAGE DOCUMENTATION

This chapter focuses on the ways that Natural Language Processing techniques have been used as part of the language documentation pipeline. We break down the pipeline into discrete subtasks, providing a concrete formulation whenever possible. We also attempt to provide a cohesive overview of the previous work on each subtask, with a particular focus on low-resource language or language documentation scenarios.

Woodbury (2003) defines language documentation as "comprehensive and transparent records supporting wide ranging scientific investigations of the language." Although the specific process of documenting a language is neither rigid nor clearly defined, common practice nonetheless follows a sequence of the following general steps, as outlined by Bird and Chiang (2012):

1. Collect (record) data, as in a series of communicative events.
2. Transcribe and translate (as much as possible of) the recordings.
3. Perform basic morphosyntactic analysis in order to create morphological glosses and/or a lexicon.
4. Elicit further paradigms that will allow the study of specific phenomena and/or reveal underlying patterns.
5. Prepare a grammar of the language i.e. descriptive reports that outline how the language is structured.

2.1 Data Collection

Even at the early stages of descriptive linguistics in the 19th century, text collection was the first major component of the documentation process. With the development of digital

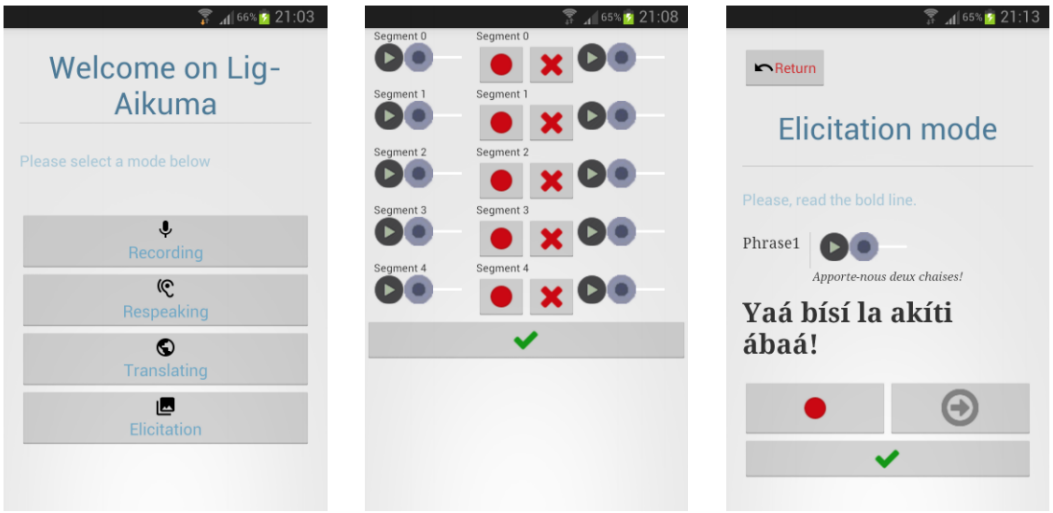


Figure 2.1. Screenshots of the LIG-Aikuma app taken from (Gauthier et al., 2016), showing the home view, summary view, and elicitation mode view.

tools in the 20th century, the process evolved to recording of audio, which allows for further study of the phonology of the language. Recording of videos is a newer trend, better situating the discourse and physical context, while enabling the combination of documentary linguistics with fields such as sociology and anthropology.

With the explosion of digital and mobile technologies in the 21st century, there has been a further shift in the data collection process as well as the general practice of documentary linguistics (Birch et al., 2013). For example, even cheap smartphones allow for much easier collection of audio and video than the heavy specialised equipment that was needed in the previous decades.

With the primary examples of SayMore (Hatton, 2013) and Aikuma (Bird et al., 2014b), academics have developed mobile apps that facilitate easy collection of audio, as well as recordings of oral re-speaking and translation. Furthermore, they allow for basic metadata collection and management. Aikuma has already been used to collect bilingual audio in remote indigenous communities, from Papua New Guinea (Bird et al., 2013) to Brazil and Nepal (Bird et al., 2014a), working on the Tembé, Nhengatu, and Kagate languages among

others. An extension of Aikuma, LIG-Aikuma (Gauthier et al., 2016), was also used in the field, collecting over 80 hours of speech on three languages from the Congo-Brazzaville area. Examples of the interface of LIG-Aikuma app are displayed in Figure 2.1.

2.2 Transcription and Annotation Tools

After data collection, linguists typically use a dedicated annotation software to aid the annotation process. Popular examples are ELAN[1] (Wittenburg et al., 2006), PRAAT[2] (Boersma et al., 2002), or FLEx.[3] ELAN is widely used for annotation of multimodal data, such as data that include video recordings. PRAAT is geared more towards phonetics and fine-grained annotation of audio, while FLEx focuses more on building lexica and interlinearized texts. Examples of the ELAN and Praat interfaces are shown in Figure 2.2.

Although still very popular among the linguistics community, the aforementioned tools require specialized training and are often platform specific. Web-based tools, instead, could avoid any platform-specific restrictions. One such prototype is Aikuma-NG (Bettinson and Bird, 2017), which was developed to work in conjunction with the Aikuma mobile app. It delivers a feature set similar to the desktop software, allowing for the annotation of audio collected with the Aikuma app.

2.2.1 Transcription

The transcription of the collected audio is both essential and one of the most time-consuming processes of the documentation pipeline. However, even for a single endangered language, the volume of the data would be prohibitive of simply manual annotation: it is estimated that the equivalent of 10 million words, or $1,000$ hours of speech should

[1] `https://tla.mpi.nl/tools/tla-tools/elan/`

[2] `http://www.fon.hum.uva.nl/praat/`

[3] `https://software.sil.org/fieldworks/`

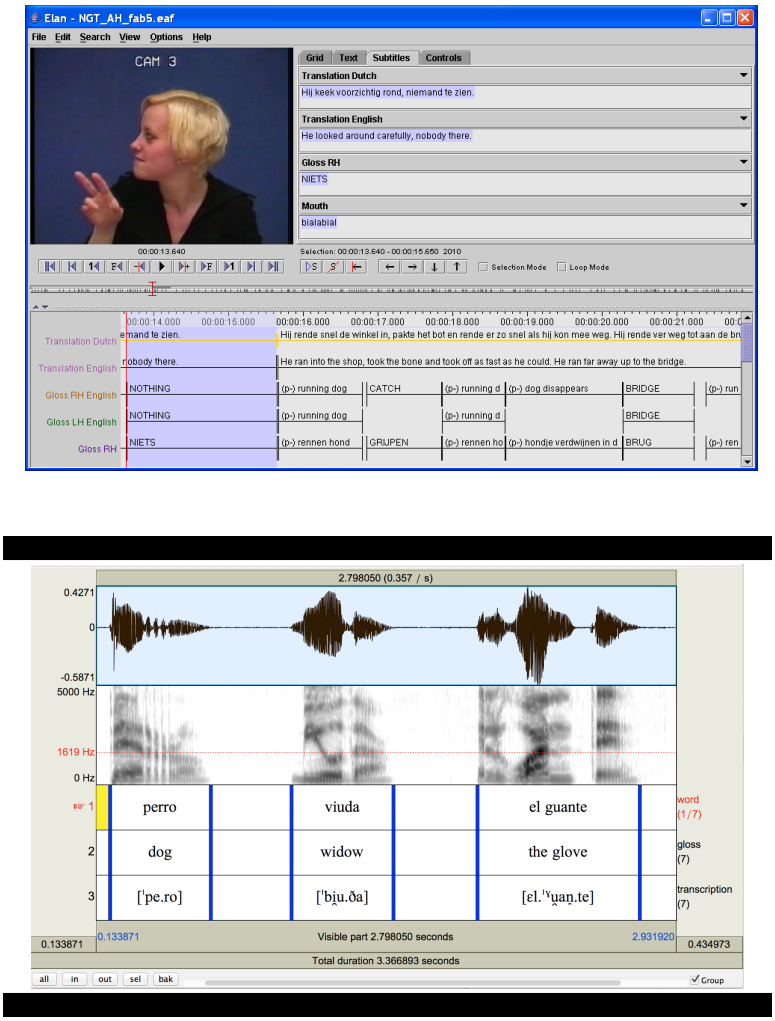


Figure 2.2. The annotation interfaces of the ELAN (top) and PRAAT (bottom) toolkits.

be transcribed and translated, in order to support extensive investigations of the language (Liberman, 2006).

An additional hurdle arises from the lack of official orthography for several of the endangered languages, or even just from the plain need to study the phonology of a language. In both cases, a phonemic transcription is needed. When no standard or working orthography is available, the International Phonetic Alphabet (IPA) is usually used. However, some estimates (Thi-Ngoc-Diep Do and Castelli, 2014) claim that it could take up to one hour for a trained linguist to transcribe the phonemes of one minute of speech.

Clearly, as noted by Thieberger (2017), automatic speech transcription systems have the potential to greatly aid the "time-intensive task of transcription [...] building transcripts for many more hours of recordings than has previously been possible."

Automatic Speech Recognition (ASR) emerged very early as one of the first tasks that could be performed by computers. Initially, HMM-based methods with limited vocabularies were developed (Jelinek, 1976; Rabiner and Juang, 1993). Large-vocabulary speech recognition was enabled by the collection of larger annotated datasets and by the development of toolkits such as `htk` (Woodland et al., 1994) and `kaldi` (Povey et al., 2011). Such models rely on vast amounts of data in order to train an *acoustic model*, which produces a phone lattice from the audio signal. Then, large *phonetic dictionaries* and *language models* are combined, in order to score the paths of the phone lattice and produce the final word-level transcription.

More recently, end-to-end neural models have been proposed, that alleviate the dependence on phonetic dictionaries and thus the need for separate components. The most popular approaches (Amodei et al., 2016; Graves et al., 2013; Hannun et al., 2014; Maas et al., 2015) rely either on Connectionist Temporal Classification (CTC) (Graves et al., 2006) or on attention-based models (Bahdanau et al., 2015b; Chan et al., 2016; Chorowski et al., 2014). Recently, such deep neural systems have achieved state-of-the-art results (Amodei et al., 2016; Hannun et al., 2014). However, training such systems requires orders-of-

magnitude more data (thousands of transcribed hours of speech) than what is available for a language documentation setting, where often less than 10 transcribed hours are available.

In fact, it is not unlikely that no transcribed data are available in the language. In this zero-resource setting, the goal is to recognize phoneme-like units in the audio, assuming no prior knowledge of the phonology of the language. Originally centered around the tasks of phonetic and lexical discovery, the field of zero-resource speech was framed as the study of models of early language acquisition. Nevertheless, it is well suited for the case of endangered language documentation where no data might be available.

The field was pioneered by Roy et al. (2006); Roy and Pentland (2002), and it eventually evolved around Dynamic Time Warping (DTW) based methods (Jansen et al., 2010; Kamper et al., 2015, 2016). Several unsupervised techniques were initially proposed (McInnes and Goldwater, 2011; Park and Glass, 2008; Siu et al., 2011; Varadarajan et al., 2008), and the field was furthered by the Zero-Resource speech challenges (Dunbar et al., 2017; Jansen et al., 2013; Versteegh et al., 2015). Recently though, new methods that use variational inference (Ondel et al., 2016), variational autoencoders (Ebbers et al., 2017), or language grounding on images (Harwath and Glass, 2017; Harwath et al., 2018) have been shown to outperform the DTW-based approaches. The zero-resource systems produce crude transcriptions, which in turn could be used for other downstream tasks. Examples include speech topic identification (Kesiraju et al., 2017; Liu et al., 2017), situation frame detection (Wiesner et al., 2018), word discovery and segmentation (Boito et al., 2018, 2017; Glarner et al., 2017), or even building speech recognition systems from completely untranscribed data (Burget et al., 2016; Scharenborg et al., 2018b), also using multimodal signals like images (Scharenborg et al., 2018a).

Endangered language data fall well within the realm of low-resource settings, so any work on *low-resource speech recognition* indirectly tackles the same problem. This includes, for example, the recent Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages (Srivastava et al., 2018). A common approach is

pre-training the system on a high-resource language and fine-tuning it to the low-resource one (Bansal et al., 2018b; Dalmia et al., 2018a; Scharenborg et al., 2017), as well as using multilingual or universal models (Dalmia et al., 2018b; Li et al., 2018).

Work that specifically focuses on the applications of speech recognition technology for language documentation is more sparse. Adams et al. (2018) trained systems for the tonal languages Yongning Na and Chatino, using up to 224 and 50 minutes of training data, respectively. Jimerson et al. (2018); Jimerson and Prud'hommeaux (2018) built systems for North American indigenous languages like Seneca, while Scharenborg et al. (2018b) built systems for the central African language Mboshi. Finally, a big collaborative project by CoEDL and Google built systems for 16 endangered languages from the Asia-Pacific region (Foley et al., 2018), creating a pipeline of tools named ELPIS.

Our work explores whether automatic speech transcription in language documentation scenarios can be aided by the availability of translations. We investigated low-resource speech transcription scenarios on Spanish, Ainu, and Mboshi. We showed that when transcriptions are available during both training and inference time, our proposed *multisource* models can produce more accurate character or phoneme-level transcriptions (Anastasopoulos and Chiang, 2018a). Furthermore, we showed that even if we only have translations at training time but not at test time, we can still take advantage of them during training (using the learning under privileged information paradigm) and achieve comparable performance.

Automatic phonemic or character transcription models could be integrated with documentation tools in order to provide, if not completely accurate transcriptions, a "rough draft" that the user could edit, speeding up the process. It is worth noting that tools like Praat are usually equipped with some basic functionalities designed to aid the documentation process, such as voice activity detection. However, none of the existing tools take advantage of recent advances in statistical speech and language processing which could further enhance the documentation process, hence updated tools are much needed. A re-

cent example of such a prototype toolkit is Persephone (Michaud et al., 2018), an offline tool that incorporates models for automatic phonemic transcription with a linguistic annotation back-end (Neubig et al., 2018).

2.2.2 Speech Translation

In language documentation, data is really usable only if it is interpretable. The interpretability of any transcribed data collections is ensured by providing free or word-level translations in a more high-resource language. Again, the sheer volume of the data that needs to be translated requires automatic solutions, rather than relying on a bilingual community member or the linguist to produce manual translations.

Similar with speech recognition, *Machine Translation* (MT) was one of the first applications envisioned for computers. The first statistical attempts at IBM (Brown et al., 1988, 1990) were followed by statistical phrase-based (Chiang, 2005; Koehn et al., 2003) and syntax-based systems (Yamada and Knight, 2001), which have now been surpassed by neural systems (Vaswani et al., 2017; Wu et al., 2016) that take advantage of massive amounts of parallel text data, as well as monolingual data with techniques such as back-translation (Sennrich et al., 2016).

In documentation scenarios, however, we deal with the more nuanced problem of *speech translation*. A linguist can and does produce translations of spoken utterances, even if a standardized orthographic transcription of these utterances does not exist.

The speech translation problem has been traditionally approached by using the output of an ASR system as input to a MT system. For example, Ney (1999) and Matusov et al. (2005) use ASR output lattices as input to translation models, integrating speech recognition uncertainty into the translation model. Recent work has focused more on modelling speech translation without explicit access to transcriptions. Duong et al. (2016) introduced a sequence-to-sequence model for speech translation without transcriptions but only evaluated on alignment, while one of our own contributions (Anastasopoulos et al.,

2016) presented an unsupervised alignment method for speech-to-translation alignment. In another of our contributions, Bansal et al. (2017) used an unsupervised term discovery system (Jansen et al., 2010) to cluster recurring audio segments into pseudowords and translate speech using a bag-of-words model. Bérard et al. (2016) translated synthesized speech data using a model similar to the Listen Attend and Spell speech recognition model (Chan et al., 2016). A larger-scale study (Bérard et al., 2018) used an end-to-end neural system system for translating audio books between French and English. On a different line of work, Boito et al. (2017) used the attentions of a sequence-to-sequence model for word discovery.

Weiss et al. (2017) used sequence-to-sequence models to transcribe Spanish speech and translate it in English, by jointly training the two tasks in a multitask scenario where the decoders share the encoder. They use a large corpus for training an 8-layer-deep model on roughly 163 hours of data, using the Spanish Fisher and CALLHOME conversational speech corpora. However, training such a large model on endangered language datasets would be infeasible.

Our contribution is inspired by the work of Weiss et al. (2017), but it adapts the model to our extremely low-resource settings and further expands it to incorporate "common sense" notions. Our *tied multitask models* are able to produce both the transcription and the translation of a speech utterance, while they are encouraged to obey the notion of transitivity, leading to better performance.

The subsequent IWSLT shared task on end-to-end speech translation (Jan et al., 2018) further explored the field with architecture search and comparing end-to-end models to pipeline approaches (Inaguma et al., 2018; Liu et al., 2018; Matusov et al., 2018; Sulubacak et al., 2018; Zenkel et al., 2018). Another notable submission focused on data filtering for the task (Di Gangi et al., 2018) leading to further improvements.

Furthermore, Bansal et al. (2018a,b) found that pre-training a speech translation system on a high-resource language and fine-tuning in the low-resource language of interest results

in improved performance. Kano et al. (2018) explored curriculum learning, while Jia et al. (2018) used speech synthesis systems to fully take advantage of monolingual data, in a back-translation-inspired semi-supervised approach.

2.2.3 Morphosyntactic Analysis

The previous subsections dealt with the basic set of annotations that render a collected data collection interpretable: transcriptions, and translations. After collecting these, linguistic research requires additional levels of annotation that highlight specific phenomena.

One of the hardest and most important tasks that the linguist has to complete is the morphological analysis of the language. That is, discovering the basic units that form words and influence meaning, and possibly glossing and assigning grammatical categories to them. Syntax, studying and discovering the rules the govern the structure and formation of sentences, is the next step. The computational linguistics field has thoroughly studied the potential of automating these tasks, leading to subfields that study (automatic) morphological segmentation, (automatic) Part-of-Speech tagging, and (automatic) syntactic or dependency parsing.

Naturally, most of the breakthroughs in those subfields have taken place in high-resource settings and languages with millions of speakers.

Segmentation *Word segmentation* is the task of segmenting an unsegmented stream of symbols, such as phonemes or characters, into delimited sequences corresponding to actual words or word-like units in the language. For a significant number of languages (e.g. those without an established orthography, or for languages with orthography that does not include explicit word boundaries like Chinese) this low-level task is non-trivial. "Words" can be generally defined as basic syntactic units that do not always coincide with phonological or orthographic words, as per the Universal Dependencies project (Kirov et al., 2018; Nivre et al., 2016).

The task is is often framed as the task that children implicitly solve, in an unsupervised way, when learning a language. Thus, unsupervised Bayesian approaches using Dirichlet Processes (Goldwater et al., 2009) or Pitman-Yor Processes (Mochihashi et al., 2009) have been exceptional baselines and the state-of-the-art for many years. Extending these approaches with external linguistic knowledge using Adaptor Grammars and annotated parse trees (Sirts and Goldwater, 2013) or small dictionaries (Eskander et al., 2016) leads to better performance, but of course requires the involvement of the linguist in a semi-supervised way.

Recently proposed segmentation models use neural encoder-decoder models as autoencoders (Chung et al., 2016), also enforcing memory limitations on the model (Elsner and Shain, 2017), in order to match the way that human memory limitations guide lexical acquisition, by applying pressure to discover compressed representations (i.e. words). Another recently proposed segmentation approach uses segmental language models (Kawakami et al., 2018) providing a neural "equivalent" to the Bayesian approaches. Previously, the attention weights of encoder-decoder translation models had been used to inform the word segmentation (Boito et al., 2018; Duong et al., 2016).

Godard et al. (2018) built word segmentation models with adaptor grammars with a specific focus on rendering useful for a linguist during the documentation process. They performed experiments on Myene and Mboshi, two Bantu languages. Kann et al. (2018) focused on neural segmentation approaches for indigenous Mexican polysynthetic languages.

It is worth noting that despite all the advances on the segmentation task, simple models like the HMM-based Morfessor FlatCat (Grönroos et al., 2014) or Byte-Pair Encoding (Sennrich et al., 2016) are often preferred, especially when word segmentation is merely a preprocessing step for another downstream task, or if large amount of monolingual data are available.

Tagging and Parsing *Part-of-Speech (POS) tagging* is a very well studied problem; probabilistic models like Hidden Markov Models and Conditional Random Fields (CRF) were initially proposed (Lafferty et al., 2001; Toutanova et al., 2003), with neural network approaches taking over in recent years (Huang et al., 2015; Mikolov et al., 2010).

The use of parallel data for projecting POS tag information across languages was introduced by Yarowsky and Ngai (2001), and further improved at a large scale by Das and Petrov (2011) who used graph-based label propagation to expand the coverage of labeled tokens. Täckström et al. (2013) used high-quality alignments to construct type and token level dictionaries. In the neural realm, Zhang et al. (2016) used only a few word translations in order to train cross-lingual word embeddings, using them in an unsupervised setting. Fang and Cohn (2017), on the other hand, used parallel dictionaries of 20k entries along with 20 annotated sentences. Plank and Agić (2018), finally, utilized joint training on the high and low resource datasets.

Most of the previous approaches are rarely tested on under-represented languages, with research on POS tagging for endangered languages being sporadic. In Ptaszynski and Momouchi (2012), for example, an HMM-based POS tagger for the extremely endangered Ainu language was presented, based on dictionaries, 12 narratives (yukar), using one annotated story (200 words) for evaluation. To our knowledge, no other previous work has extensively tested several approaches on an actual endangered language.

The lack of high quality annotated data led to approaches that attempt to use monolingual resources in a semi-supervised setting. Notably, Garrette and Baldridge (2013) used about 200 annotated sentences along with monolingual corpora improving the accuracy of an HMM-based model. They tested their model on two low-resource African languages, Kinyarwanda and Malagasy and they found that in this time-constrained scenario type-level annotation leads to slightly higher improvements than token-level annotation, increasing the accuracy of their taggers to slightly less than 80%. Similar conclusions were reached in Garrette et al. (2013): 4 hours of annotation are more wisely spent if annotating at the type-

level, provided there exist additional raw monolingual data. This line of work adequately addressed the question of what labeled data are preferable when there is (exceptionally) restricted access to annotators.

Our contribution (Anastasopoulos et al., 2018) corroborated the evidence from the Garrette and Baldridge (2013) work, but extended it to take advantage of cross-lingual information. In particular, since the resource we collected included translations of the endangered language text, we were able to use cross-lingual projection a la Täckström et al. (2013) and include them in the training data. We also explored active learning and managed to speed up the process of annotating the test set of our corpus.

Syntactic or Dependency Parsing is the task of analyzing a sentence into its constituencies, resulting into a structure (a parse tree) that describes the syntactic or semantic relation among the words. Research on this field has been primarily driven by the Universal Dependencies project (Nivre et al., 2016) and the associated CoNLL shared tasks on dependency parsing (Zeman et al., 2018). The Universal Dependencies project has collected treebanks from several languages, while using consistent syntactic representations across them. As a result, it currently includes more than 100 treebanks on more than 70 languages. Naturally, very few of them are endangered or in the process of documentation.

Our work does not delve into parsing. Nonetheless, the Universal Dependencies project provides an invaluable resource, which has been used to investigate low-resource parsing. A common approach is knowledge transfer from a high-resource language to a low-resource one (Cotterell and Heigold, 2017; Malaviya et al., 2018). This can be achieved by joint training (Ammar et al., 2016), annotation projection (Agić et al., 2016; Ponti et al., 2018; Täckström et al., 2012), or zero-shot transfer (Ahmad et al., 2018).

Similar approaches could in theory be employed at the last step of the documentation process, but they would have to be further extended to handle the extreme scarcity and noise of documentation data.

2.3 Contributions Summary

This work assumes that we are provided with audio resources in an unknown language, along with translations, similar to the data collected through apps like Aikuma. We aim to provide automated solutions for each of the steps that are crucial to understanding an unknown language.

A first step towards this goal would be to automatically align spoken segments with their translations at the word or phrase level. This task has been implicitly or explicitly performed by the linguists using traditional tools like Praat, whenever they create phrase-, word-, or phoneme-level annotation boundaries.

Our work in this area is collected in Chapter 3. Our first contribution, detailed in §3.2, produces such speech-to-translation alignments. Furthermore, in two more contributions, we evaluated whether such alignments are potentially useful for other downstream tasks. We investigated the use of alignment information for collecting mismatched transcriptions (§3.3), showing that they are indeed beneficial. Also, as we discuss in §3.4, we leveraged speech-to-translation alignments for keyword spotting on unannotated resources. A synopsis of these works is outlined here:

- **An Unsupervised Probability Model for Speech-to-Translation Alignment of Low-Resource Languages (Anastasopoulos et al., 2016) :** We presented a speech-to-translation alignment model that combines Dyer et al.'s reparameterization of IBM Model 2 (`fast_align`) and k-means clustering using Dynamic Time Warping as a distance measure. The two components are trained jointly using expectation-maximization. In extremely low-resource scenarios, this model performs significantly better than both a strong naive baseline as well as our previous approach based on the attentions of a neural model.

- **A case study on using speech-to-translation alignments for language documentation (Anastasopoulos and Chiang, 2017):** We investigated whether augmenting an utterance with a translation and speech-to-translation alignment information can aid in producing better crowdsourced *mismatched* transcriptions, that is, transcriptions by speakers who do not speak the language. These transcriptions could in turn be valuable for training speech recognition systems. We showed that they can indeed be beneficial through a small-scale case study as a proof-of-concept. We also presented a simple phonetically aware string averaging technique that combines the

collected mismatched transcriptions into a transcription of higher quality.

- **Spoken Term Discovery for Language Documentation using Translations (Anastasopoulos et al., 2017):** We presented a method for partially labeling unannotated speech with translations in a scenario where we have access to limited translated speech. We modified an unsupervised speech-to-translation alignment model and obtained prototype speech segments that match the translation words, which were in turn used to discover terms in the unlabelled data. We evaluated our method on a Spanish-English speech translation corpus and on two endangered languages corpora, Arapaho and Ainu, demonstrating its appropriateness and applicability in an actual very-low-resource scenario.

Other than alignment, proper documentation needs to include a transcription of the speech utterance. Moreover, in order for the meaning of a segment to be understandable, it is usually followed by a free translation. Thus, producing these two layers of annotation is the natural next step of our work. We list our contributions on the transcription and translation tasks here, and they are laid out in detail in Chapter 4:

- **Leveraging translations for speech transcription in low-resource settings (Anastasopoulos and Chiang, 2018a):** We explored whether having access to translations allows us to improve transcription accuracy in extremely low-resource scenarios. This scenario is applicable to the data collection process that we described earlier, and improving the transcription quality has the potential to significantly reduce the time required for fully annotating the collected resources. We find that in most cases the multi-source approach combined with a shared attention mechanism significantly reduces the Character Error Rate of the transcriptions.
- **Translations as Privileged Information for Low-Resource Speech Transcription (unpublished):** We explore the recently proposed Learning Under Privileged Information for deep neural models (DLUPI) framework (Lambert et al., 2018). We adapt it to the speech transcription task and enhance it with attention mechanism in order to receive fine-grained privileged information from translations. We show that in low-resource settings we can achieve performance comparable to the best multi-source models, despite not having access to translations at inference time. At the same time the DLUPI models surpass single-source baselines that do not use translations at all.
- **Tied Multitask Learning for Neural Speech Translation (Anastasopoulos and Chiang, 2018b):** We explored multitask models for neural translation of speech, augmenting them in order to reflect two intuitive notions. First, we introduced a model where the second task decoder receives information from the decoder of the first task, since higher-level intermediate representations should provide useful information. Second, we applied regularization that encourages *transitivity* and *invertibility*. We show that the application of these notions on jointly trained models improves

performance on the tasks of low-resource speech transcription and translation. It also leads to better performance when using attention information for word discovery over unsegmented input.

The final layers of annotation usually analyze more complex phenomena, like syntax and morphology, building upon the previous annotation layers. Towards this end, we collected a new parallel corpus for an endangered language, Griko, and developed a Part-of-Speech (POS) tagger that takes advantage of cross-lingual information. This work is presented in Chapter 5:

- **POS-tagging on an Endangered Language: a parallel Griko-Italian resource (Anastasopoulos et al., 2018):** Most work on part-of-speech (POS) tagging is focused on high resource languages, or examines low-resource and active learning settings through simulated studies. We evaluated POS tagging techniques on an actual endangered language, Griko. We collected and released a resource that contains 114 narratives in Griko, along with sentence-level translations in Italian. The resource also provides gold POS annotations for the test set. Based on a previously collected small corpus, we investigated several traditional methods, as well as methods that take advantage of monolingual data or project cross-lingual POS tags. We showed that the combination of a semi-supervised method with cross-lingual transfer is more appropriate for this extremely challenging setting, with the best tagger achieving an accuracy of 72.9%. With an applied active learning scheme, which we used to collect sentence-level annotations over the test set, we achieved improvements of more than 21 percentage points on POS-tagging accuracy.

2.4 Datasets

This section describes the corpora on which we have evaluated our contributions. We particularly focus on the endangered or extremely low-resource language corpora that we used, providing a brief overview of these languages.

Spanish-English Spanish is obviously neither an endangered nor a low-resource language, but we pretend that it is one, by not making use of any Spanish resources like additional transcribed speech or pronunciation lexicons. We use the Spanish CALLHOME corpus (LDC96S35) and the Fisher corpus (LDC2010T04), which consist of telephone con-

versations between Spanish native speakers based in the US and their relatives abroad, together with English translations produced by Post et al. (2013) and silver standard speech-to-translation alignments the we produced (Duong et al., 2016). Both datasets are split into utterances based on the dialogue turns. This results in 17,532 Spanish utterances for the CALLHOME corpus, and 143,355 utterances for the Fisher corpus.

Mboshi-French Mboshi (Bantu C25 in the Guthrie classification) is a language spoken in Congo-Brazzaville by about 110,000 speakers, without standard orthography. We use a corpus (Godard et al., 2017) of 5,517 parallel utterances (about 4.4 hours of audio) collected from three native speakers using the LIG-Aikuma app for the BULB project (Adda et al., 2016). The corpus provides non-standard grapheme transcriptions (close to the language phonology) and word segmentation produced by linguists, as well as French translations.

Ainu-English Hokkaido Ainu is the sole surviving member of the Ainu language family and is generally considered a language isolate. As of 2007, only ten native speakers were alive. The Glossed Audio Corpus of Ainu Folklore provides 24 narratives (about 5 hours of audio), transcribed at the utterance level, glossed, and translated in Japanese and English.[4]

Arapaho-English Arapaho is an Algonquian language with about 1,000 native speakers, mostly in Wyoming. We use 8 narratives published at The Arapaho Language Project,[5] which provides the narratives' audio along with English translations, among other language learning resources. This is a corpus that is not aligned at the utterance level, but only at the narrative level.

[4] `http://ainucorpus.ninjal.ac.jp/corpus/en/`

[5] `http://www.colorado.edu/csilw/alp/index.html`

Griko-Italian Griko is a Greek dialect spoken in south Italy, believed to be the last living trace of the ancient Greek elements that once formed Magna Graecia. It is only partially intelligible with modern Greek. In 2010, less than 20,000 people were registered as native speakers, and about 50,000 were registered as L2 speakers. The Griko-Italian corpus[6] consists of about 20 minutes of speech in Griko, an endangered minority dialect of Greek spoken in south Italy, along with text translations into Italian (Lekakou et al., 2013). The corpus consists of 330 mostly prompted utterances by nine native speakers. All utterances were manually annotated and transcribed by a trained linguist and bilingual speaker of both languages, who produced the Griko transcriptions and Italian glosses. We created full translations into Italian and manually aligned the translations with the Griko transcriptions. We then combined the two alignments (speech-to-transcription and transcription-to-translation) to produce speech-to-translation alignments.

Furthermore, we compiled an additional parallel text resource of 114 Griko narratives (along with Italian translations), taken from a website.[7] We further elaborate on this work in section §5.1.

[6] `http://griko.project.uoi.gr`

[7] http://wwww.ciuricepedi.it

CHAPTER 3

SPEECH-TO-TRANSLATION ALIGNMENT AND WORD DISCOVERY

In this chapter we focus on the problem of aligning speech segments to translation words. After briefly providing the necessary background, we describe a generative alignment model trained with expectation-maximization (§3.2). In addition, we show that such alignments can be useful for two tasks related to language documentation: collecting mismatched transcriptions (§3.3), and keyword spotting for labelling untranscribed data (§3.4).

3.1 Background

In this section, we present an overview of several influential publications which presented the basis of research on Alignment for Machine Translation, as well as Term Discovery.

3.1.1 Word Alignment

Given a set of parallel sentences in two languages, the task of finding a correspondence between the words of each language is the task of word alignment (Brown et al., 1993; Koehn, 2010). Word alignments have been traditionally used as a tool for phrase extraction, a crucial preprocessing step for phrase-based MT models.

IBM Models The IBM translation models (Brown et al., 1993) are the most commonly used word alignment models, aiming to model the distribution $p(\mathbf{e} \mid \mathbf{f})$ for an English sentence $\mathbf{e} = e_1 \cdots e_l$, given a French sentence $\mathbf{f} = f_1 \cdots e_m$. They all introduce a hidden

variable $\mathbf{a} = a_1 \cdots a_l$ that gives the position of the French word to which each English word is aligned.

The general form of IBM Models 1 and 2 is

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = p(l) \prod_{i=1}^{l} t(e_i \mid f_{a_i})\, \delta(a_i \mid i, l, m)$$

where $t(e \mid f)$ is the probability of translating French word f to English word e, and $\delta(a_i = j \mid i, l, m)$ is the probability of aligning the i-th English word with the j-th French word (also often referred to as *distortion*).

In Model 1, δ is uniform; in Model 2, it is a categorical distribution. Dyer et al. (2013) proposed a reparameterization of Model 2, known as `fast_align`:

$$h(i, j, l, m) = -\left|\frac{i}{l} - \frac{j}{m}\right|$$

$$\delta(a_i \mid i, l, m) = \begin{cases} p_0 & a_i = 0 \\ (1 - p_0)\frac{\exp \lambda h(i, a_i, l, m)}{Z_\lambda(i, l, m)} & a_i > 0 \end{cases}$$

where the null alignment probability p_0 and precision $\lambda \geq 0$ are hyperparameters optimized by grid search. As $\lambda \to 0$, the distribution gets closer to the distribution of IBM Model 1, and as λ gets larger, the model prefers monotone word alignments more strongly.

IBM Models 3-5 are more complex, also introducing the notion of *fertility*, excplicitly modelling the probability of many-to-one target-to-source alignments.

Other Alignment Models Stahlberg et al. (2012) presented a modification of IBM Model 3, named `Model 3P`, designed specifically for phone-to-word alignment. They introduced an intermediate step so that the generative process of the model model incorporates the length (in phonemes) of the target phonetic sequence. They tested it on Spanish phones to English words alignments on the BTEC corpus improving over GIZA++ on both Alignment Error Rates as well on the extrinsic task of word segmentation.

In addition, `pialign` (Neubig et al., 2011) is an unsupervised model for joint phrase alignment and extraction, based on inversion transduction grammars. It has been shown to work well at the character level (Neubig et al., 2012), and it extends naturally to work on phones.

Alignment Evaluation Given a collection of annotated parallel sentences with *gold alignments*, the performance of an alignment method can be evaluated by computing Precision (P), Recall (R), and F-score (F_s) over the gold alignments. If $\mathbf{a}^*$ are the gold alignments, and $\mathbf{a}$ are produced alignments, then:

$$P = \frac{|\mathbf{a}^* \cap \mathbf{a}|}{|\mathbf{a}|} \qquad R = \frac{|\mathbf{a}^* \cap \mathbf{a}|}{|\mathbf{a}^*|} \qquad F_s = 2\frac{P \times R}{P + R}$$

Another metric that is used in the literature is Alignment Error Rate (AER) (Mihalcea and Pedersen, 2003) which incorporates uncertainty over both the gold standard and the produced alignments, such that $\mathbf{a}_S$ denote certain alignments and $\mathbf{a}_P$ denote probable alignments, is computed as:

$$AER = 1 - \frac{|\mathbf{a}_P \cap \mathbf{a}_S^*| + |\mathbf{a}_P \cap \mathbf{a}_P^*|}{|\mathbf{a}_P| + |\mathbf{a}_S^*|}.$$

3.1.2 Forced Alignment

Given an audio file containing speech, and the corresponding transcript, computing a forced alignment is the process of determining, for each fragment of the transcript, the time interval (in the audio file) containing the spoken text of the fragment. It can be performed at either phoneme level, word level, utterance or dialogue turn level, or even at a document level.

Training an acoustic model traditionally required training examples often annotated at the phonetic level, or at the level of context-dependent subphones. Manually annotating examples at that level is an incredibly time consuming task, although it was done for a few

corpora like TIMIT (Garofolo et al., 1993). Traditionally, for larger speech collections, and also given a lexicon with phonetic pronunciations of words, the transcription word sequence is used to constrain an optimal alignment between an existing speech model and the new speech data, providing labels that are then used for training the acoustic model.

3.1.3 Term Discovery, DTW, and DBA

Spoken Term Discovery is the task of finding spoken terms (words, phrases) in a collection of audio resources, usually in an unsupervised setting, where we have no access to transcriptions. *Unsupervised Term Discovery* (UTD) or keyword spotting has been studied extensively through the Zero-Resource Speech Challenges (Dunbar et al., 2017; Versteegh et al., 2015), and various approaches (Jansen et al., 2010; Muscariello et al., 2009; Park and Glass, 2008; Ten Bosch and Cranen, 2007; Zhang and Glass, 2010) have been tried. Most of them rely on segmental DTW (described below) to identify repeated trajectories in the speech signal.

Other approaches have been recently proposed too; Kamper et al. (2016) try to discover word segmentation and a pronunciation lexicon in a zero-resource setting, combining DTW with acoustic embeddings; their methods operate in a very low-vocabulary setting. Finally, Ondel et al. (2016) proposed a bayesian approach for acoustic unit discovery, using variational inference.

DTW and DBA Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) is a dynamic programming method for measuring distance between two temporal sequences of variable length, as well as computing an alignment based on this distance. Given two sequences ϕ, ϕ' of length m and m' respectively, DTW constructs an $m \times m'$ matrix w. The warping path can be found by evaluating the following recurrence:

$$w_{i,j} = d(\phi_i, \phi'_j) + \min\{w_{i-1,j}, w_{i-1,j-1}, w_{i,j-1}\}$$

where d is a distance measure. The cost of the warping path is often length-normalized so that it lies between zero and one:

$$\mathrm{DTW}(\phi, \phi') = \frac{w_{m,m'}}{m + m'}.$$

DTW Barycenter Averaging (DBA) (Petitjean et al., 2011) is an iterative approximate method that attempts to find a centroid of a set of sequences, minimizing the sum of squared DTW distances.

In the original definition, given a set of sequences, DBA chooses one sequence randomly to be a "skeleton." Then, at each iteration, DBA computes the DTW between the skeleton and every sequence in the set, aligning each of the skeleton's points with points in all the sequences. The skeleton is then refined using the found alignments, by updating each frame in the skeleton to the mean of all the frames aligned to it. Note that, in our implementation of DBA, in order to avoid picking a skeleton that is too short or too long, we randomly choose one of the sequences with median (or close to median) length.

3.2 An Unsupervised Probability Model for Speech-to-Translation Alignments

Abstract: We proposed a generative model for alignment between speech frames and translation words, that combines Dyer et al.'s reparameterization of IBM Model 2 (`fast_align`) and k-means clustering using Dynamic Time Warping as a distance measure. The two components are trained jointly using expectation-maximization. In an extremely low-resource scenario, our model performs significantly better than both a neural model and a strong baseline.

3.2.1 Introduction

IBM alignment models have been very popular for aligning parallel corpora in Machine Translation. We combine IBM Model 2 with a k-means clustering approach in order to allow alignment between speech frames and translation words. The clustering component uses Dynamic Time Warping as a distance measure, and the whole generative model is trained using hard expectation-maximization. Our model outperforms two strong baselines on all datasets, particularly improving precision.

3.2.2 Model

We use a generative model from a source-language speech segment consisting of feature frames $\boldsymbol{\phi} = \phi_1 \cdots \phi_m$ to a target-language segment consisting of words $\mathbf{e} = e_1 \ldots e_l$. We chose to model $p(\mathbf{e} \mid \boldsymbol{\phi})$ rather than $p(\boldsymbol{\phi} \mid \mathbf{e})$ because it makes it easier to incorporate DTW. In addition to the target-language sentence $\mathbf{e}$, our model hypothesizes a sequence $\mathbf{f} = f_1 \cdots f_l$ of source-language clusters (intuitively, source-language words), and spans (a_i, b_i) of the source signal that each target word e_i is aligned to. Thus, the clusters $\mathbf{f} = f_1 \cdots f_l$ and the spans $\mathbf{a} = a_1, \ldots, a_l$ and $\mathbf{b} = b_1, \ldots, b_l$ are the hidden variables of the model:

$$p(\mathbf{e} \mid \boldsymbol{\phi}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{f}} p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi}).$$

The model generates $\mathbf{e}, \mathbf{a}, \mathbf{b}$, and $\mathbf{f}$ from $\boldsymbol{\phi}$ as follows.

1. Choose l, the number of target words, with uniform probability. (Technically, this assumes a maximum target sentence length, which we can just set to be very high.)
2. For each target word position $i = 1, \ldots, l$:
   (a) Choose a cluster f_i.
   (b) Choose a span of source frames (a_i, b_i) for e_i to be aligned to.
   (c) Generate a target word e_i from f_i.

Accordingly, we decompose $p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi})$ into several submodels:

$$p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi}) = p(l) \prod_{i=1}^{l} u(f_i) \times s(a_i, b_i \mid f_i, \boldsymbol{\phi}) \times \delta(a_i, b_i \mid i, l, |\boldsymbol{\phi}|) \times t(e_i \mid f_i).$$

Note that submodels δ and s both generate spans (corresponding to step 2b), making the model deficient. We could make the model sum to one by replacing $u(f_i)s(a_i, b_i \mid f_i, \boldsymbol{\phi})$ with $s(f_i \mid a_i, b_i, \boldsymbol{\phi})$, and this was in fact our original idea, but the model as defined above works much better. We describe both δ and s in detail below.

Clustering model The probability over clusters, $u(f)$, is just a categorical distribution. The submodel s assumes that, for each cluster f, there is a "prototype" signal $\boldsymbol{\phi}^f$ (Ristad and Yianilos, 1998). Technically, the $\boldsymbol{\phi}^f$ are parameters of the model, and will be recomputed during the M step. Then we can define:

$$s(a, b \mid f, \boldsymbol{\phi}) = \frac{\exp(-\text{DTW}(\boldsymbol{\phi}^f, \phi_a \cdots \phi_b)^2)}{\sum_{a,b=1}^{m} \exp(-\text{DTW}(\boldsymbol{\phi}^f, \phi_a \cdots \phi_b)^2)}$$

where DTW is the distance between the prototype and the segment computed using Dynamic Time Warping. Thus s assigns highest probability to spans of $\boldsymbol{\phi}$ that are most similar to the prototype $\boldsymbol{\phi}^f$.

Distortion model The submodel δ controls the reordering of the target words relative to the source frames. It is an adaptation of `fast_align` to our setting, where there is not a single source word position a_i, but a span (a_i, b_i). We want the model to prefer the middle of the word to be close to the diagonal, so we need the variable a to be somewhat to the left and b to be somewhat to the right. Therefore, we introduce an additional hyperparameter μ which is intuitively the number of frames in a word. Then we define:

$$h_a(i, j, l, m, \mu) = -\left|\frac{i}{l} - \frac{j}{m - \mu}\right| \quad \text{so that} \qquad \delta_a(a_i \mid i, l, m) = \begin{cases} p_0 & a_i = 0 \\ (1 - p_0)\frac{\exp \lambda h_a(i, a_i, l, m)}{Z_\lambda(i, l, m)} & a_i > 0 \end{cases}$$

$$h_b(i, j, l, m, \mu) = -\left|\frac{i}{l} - \frac{j - \mu}{m - \mu}\right| \quad \text{so that} \qquad \delta_b(b_i \mid i, l, m) = \begin{cases} p_0 & b_i = 0 \\ (1 - p_0)\frac{\exp \lambda h_b(i, b_i, l, m)}{Z_\lambda(i, l, m)} & b_i > 0 \end{cases}$$

$$\delta(a_i, b_i \mid i, l, m) = \delta_a(a_i \mid i, l, m)\, \delta_b(b_i \mid i, l, m)$$

where the $Z_\lambda(i, l, m)$ are set so that all distributions sum to one. Figure 3.1 shows an example visualisation of the the resulting distributions for the two variables of our model. We set μ differently for each word. For each i, we set μ_i to be proportional to the number of *characters* in e_i, such that $\sum_i \mu_i = m$.

Translation model The translation model $t(e \mid f)$ is just a categorical distribution, in principle allowing a many-to-many relation between source clusters and target words. To speed up training (with nearly no change in accuracy, in our experiments), we restrict this relation so that there are k source clusters for each target word, and a source cluster uniquely determines its target word. Thus, $t(e \mid f)$ is fixed to either zero or one, and does not need to be re-estimated. In our experiments, we set $k = 2$, allowing each target word to have up to two source-language translations/pronunciations. (If a source word has more than one target translation, they are treated as distinct clusters with distinct prototypes.)

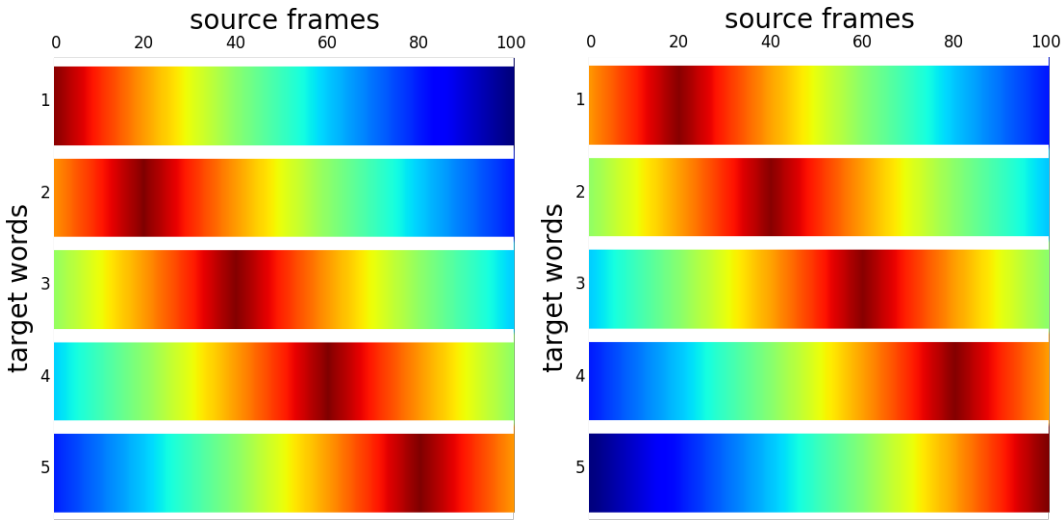


Figure 3.1. Sample distributions for the alignment variables a and b for $m = 100$, $l = 5$, $p_0 = 0$, $\lambda = 0.5$, and $\mu = 20$.

3.2.3 Training

We use the hard (Viterbi) version of the Expectation-Maximization (EM) algorithm to estimate the parameters of our model, because calculating expected counts in full EM would be prohibitively expensive, requiring summations over all possible alignments.

Recall that the hidden variables of the model are the alignments (a_i, b_i) and the source words (f_i). The parameters are the translation probabilities $t(e_i \mid f)$ and the prototypes $(\boldsymbol{\phi}^f)$.

The (hard) E step uses the current model and prototypes to find, for each target word, the best source segment to align it to and the best source word. The M step reestimates the probabilities $t(e \mid f)$ and the prototypes $\boldsymbol{\phi}^f$. We describe each of these steps in more detail below.

Initialization Initialization is especially important since we are using hard EM. To initialize the parameters, we initialize the hidden variables and then perform an M step. We associate each target word type e with $k = 2$ source clusters, and for each occurrence of

e, we randomly assign it one of the k source clusters. The alignment variables a_i, b_i are initialized to

$$a_i, b_i = \arg\max_{a,b} \delta(a, b \mid i, l, m).$$

M step The M step reestimates the probabilities $t(e \mid f)$ using relative-frequency estimation. The prototypes $\boldsymbol{\phi}^f$ are more complicated. Theoretically, the M step should recompute each $\boldsymbol{\phi}^f$ so as to maximize that part of the log-likelihood that depends on $\boldsymbol{\phi}^f$:

$$\begin{aligned} L_{\boldsymbol{\phi}^f} &= \sum_{\boldsymbol{\phi}} \sum_{i|f_i=f} \log s(a_i, b_i \mid f, \boldsymbol{\phi}) \\ &= \sum_{\boldsymbol{\phi}} \sum_{i|f_i=f} \log \frac{\exp(-\mathrm{DTW}(\phi^f, \phi_{a_i} \cdots \phi_{b_i})^2)}{Z(f, \boldsymbol{\phi})} \\ &= \sum_{\boldsymbol{\phi}} \sum_{i|f_i=f} -\mathrm{DTW}(\phi^f, \phi_{a_i} \cdots \phi_{b_i})^2 - \log Z(f, \boldsymbol{\phi}) \end{aligned}$$

where the summation over $\boldsymbol{\phi}$ is over all source signals in the training data. This is a hard problem, but note that the first term is just the sum-of-squares of the DTW distance between $\boldsymbol{\phi}^f$ and all source segments that are classified as f. This is what DBA is supposed to approximately minimize, so we simply set $\boldsymbol{\phi}^f$ using DBA, ignoring the denominator.

E step The (hard) E step uses the current model and prototypes to find, for each target word, the best source segment to align it to and the best source cluster. In order to reduce the search space for **a** and **b**, we use the unsupervised phonetic boundary detection method of (Khanagha et al., 2014). This method operates directly on the speech signal and provides us with candidate phone boundaries, on which we restrict the possible values for **a** and **b**, creating a list of candidate utterance spans.

Furthermore, we use a simple silence detection method. We pass the envelope of the signal through a low-pass filter, and then mark as "silence" time spans of 50ms or longer in which the magnitude is below a threshold of 5% relative to the maximum of the whole

signal. This method is able to detect about 80% of the total pauses, with a 90% precision in a 50ms window around the correct silence boundary. We can then remove from the candidate list the utterance spans that include silence, on the assumption that a word should not include silences. Finally, in case one of the span's boundaries happens to be within a silence span, we also move it so as to not include the silence.

Hyperparameter tuning The hyperparameters p_0, λ, and μ are not learned. We simply set p_0 to zero (disallowing unaligned target words) and set μ as described above. For λ we perform a grid search over candidate values to maximize the alignment F-score on the development set. We obtain the best scores with $\lambda = 0.5$.

3.2.4 Experiments and Results

We evaluate our method on two language pairs, Spanish-English and Griko-Italian. For Spanish-English we report results on both the CALLHOME and the Fisher dataset, comparing against two baselines.

The first is a naive baseline which assumes no reordering between the source and target language, and aligns each target word e_i to a source span whose length in frames is proportional to the length of e_i in characters. This actually performs very well on language pairs that show minimal or no reordering, and language pairs that have shared or related vocabularies.

The other baseline that we compare against is the neural network attentional model of Duong et al. (2016), which extends the attentional model of Bahdanau et al. (2015a) to be used for aligning and translating speech, and, along with several modifications, achieve good results on the phone-to-word alignment task, and almost match the baseline performance on the speech-to-word alignment task.

In both data settings, we treat the speech data as a sequence of 39-dimensional Perceptual Linear Prediction (PLP) vectors encoding the power spectrum of the speech sig-

nal (Hermansky, 1990), computed at 10ms intervals. We also normalize the features at the utterance level, shifting and scaling them to have zero mean and unit variance.

We find that our model improves upon both the baselines on all datasets. The results are summarized in Table 3.1. Our model, when compared to the baselines, improves greatly on precision, while slightly underperforming the naive baseline on recall. We note that in all cases the naive baseline is quite strong, outperforming the neural model. This is due to the minimal reordering between our tested language pairs. Our proposed model, however, builds upon the naive baseline and especially improves on precision. In certain applications, higher precision may be desirable: for example, in language documentation, it's probably better to err on the side of precision; in phrase-based translation, higher-precision alignments lead to more extracted phrases.

3.2.5 Analysis

Speaker robustness Figure 3.2 shows the alignments produced by our model for three utterances of the same sentence from the Griko-Italian dataset by three different speakers. Our model's performance is roughly consistent across these utterances. In general, the model does not seem significantly affected by speaker-specific variations, as shown in Table 3.2.

We do find, however, that the performance on male speakers is slightly higher compared to the female speakers. This might be because the female speakers' utterances are, on average, longer by about 2 words than the ones uttered by males.

Word level analysis We also compute F-scores for each Italian word type. As shown in Figure 3.3, the longer the word's utterance, the easier it is for our model to correctly align it. Longer utterances seem to carry enough information for our DTW-based measure to function properly. On the other hand, shorter utterances are harder to align. The vast majority of Griko utterances that have less than 20 frames and are less accurately aligned

TABLE 3.1

OUR ALIGNMENT MODEL ACHIEVES HIGHER PRECISION AND F-SCORE THAN BOTH THE NAIVE BASELINE AND THE NEURAL MODEL ON ALL DATASETS

		method	precision	recall	F-score
CALLHOME spa-eng	2k sents	ours	**38.8**	38.9	**38.8**
		naive	31.9	**40.8**	35.8
		neural	23.8	29.8	26.4
	17k sents	ours	**38.4**	38.8	**38.6**
		naive	31.8	**40.7**	35.7
		neural	26.1	32.9	29.1
Fisher spa-eng	143k sents	ours	**33.3**	28.7	**30.8**
		naive	24.0	**33.2**	27.8
		neural	24.7	27.9	26.2
gri-ita	300 sents	ours	**56.6**	51.2	**53.8**
		naive	42.2	**52.2**	46.7
		neural	24.6	30.0	27.0

TABLE 3.2

MODEL PERFORMANCE (F-SCORE) IS GENERALLY CONSISTENT ACROSS SPEAKERS

speaker	#utterances	average length	F-score
female 1	55	9.0	49.4
female 2	61	8.1	55.0
female 3	41	9.6	51.0
female 4	23	7.3	54.4
female 5	21	6.1	56.6
male 1	35	5.9	59.5
male 2	32	6.0	61.9
male 3	34	6.7	60.2
male 4	23	6.4	64.0

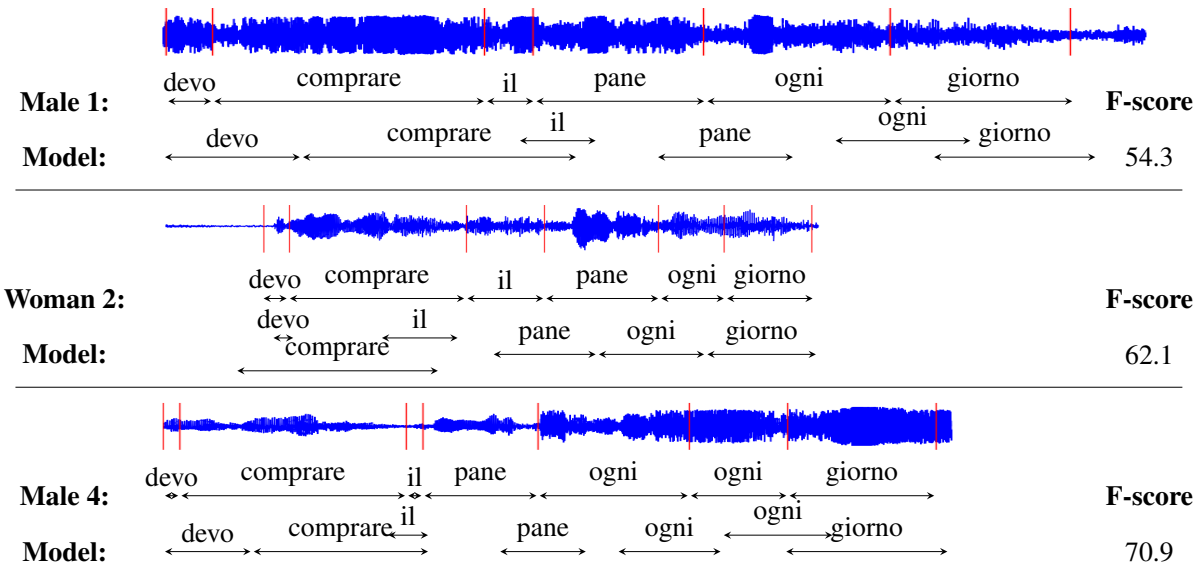


Figure 3.2: Alignments produced for the Italian sentence `devo comprare il pane ogni giorno` as uttered by three different Griko speakers.

correspond to monosyllabic determiners (`o`, `i`,`a`, `to`, `ta`) or conjunctions and prepositions (`ka`, `ce`, `en`, `na`, `an`). For such short utterances, there could be several parts of the signal that possibly match the prototype, leading the clustering component to prefer to align to wrong spans.

Furthermore, we note that rare word types tend to be correctly aligned. The average F-score for hapax legomena (on the Italian side) is 63.2, with 53% of them being aligned with an F-score higher than 70.0.

Comparison with proper model As mentioned in Section 3.2.2, our model is deficient, but it performs much better than the model that sums to one (henceforth, the "proper" model): In the Spanish-English dataset (2000 sentences sample) the proper model yields an F-score of 32.1, performing worse than the naive baseline; in the Griko-Italian dataset, it achieves an F-score of 44.3, which is better than the baselines, but still worse than our model.

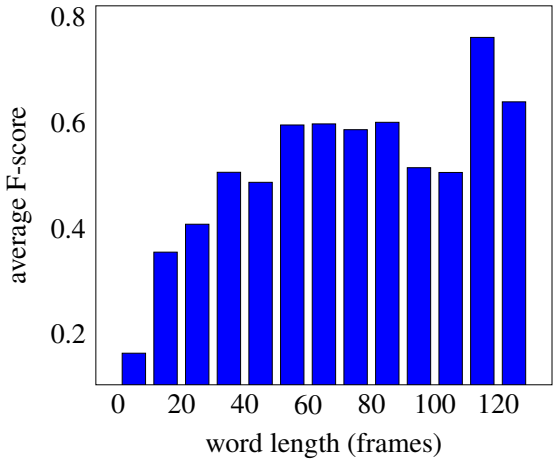


Figure 3.3. There is a positive correlation between average word-level F-score and average word utterance length (in frames).

In order to further examine why this happens, we performed three EM iterations on the Griko-Italian dataset with our model (in our experience, three iterations are usually enough for convergence), and then computed one more E step with both our model and the proper model, so as to ensure that the two models would align the dataset using the exact same prototypes and that their outputs will be comparable.

In this case, the proper model achieved an overall F-score of 44.0, whereas our model achieved an F-score of 53.6. Figures 3.4 and 3.5 show the resulting alignments for two sentences. In both of these examples, it is clear that the proper model prefers extreme spans: the selected spans are either much too short or (less frequently) much too long. This is further verified by examining the statistics of the alignments: the average span selected by the proper model has a length of about 30 ± 39 frames whereas the average span of the alignments produced by our deficient model is 37 ± 24 frames. This means that the alignments of the deficient model are much closer to the gold ones, whose average span is 42 ± 26 frames.

We think that this is analogous to the "garbage collection" problem in word alignment.

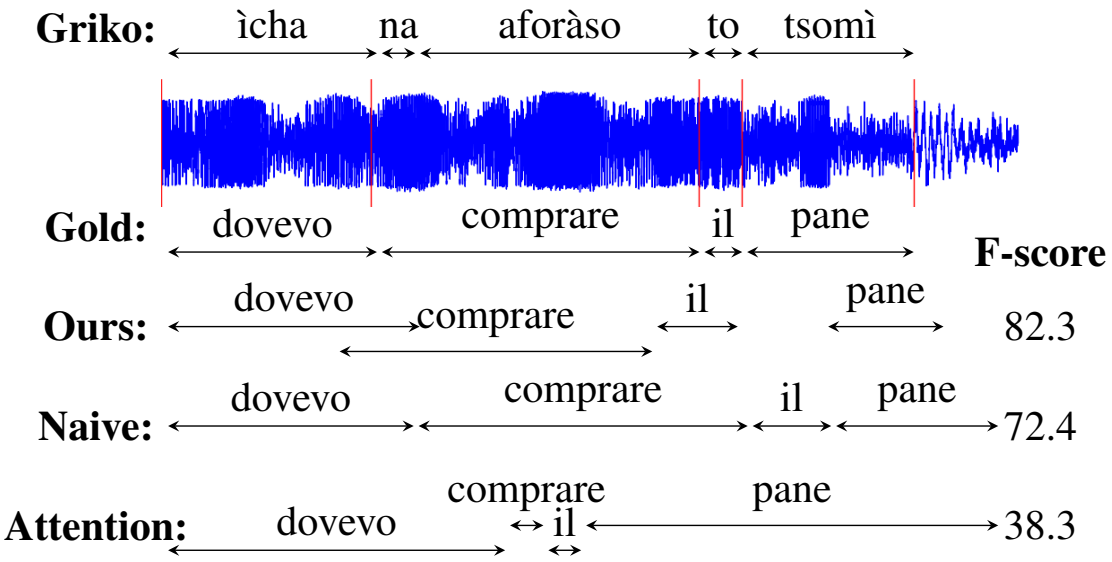


Figure 3.4: The deficient model performs very well, whereas the proper and the attentional model prefer extreme alignment spans. For example, the proper model's alignment for the words `dovevo` and `pane` are much too short.

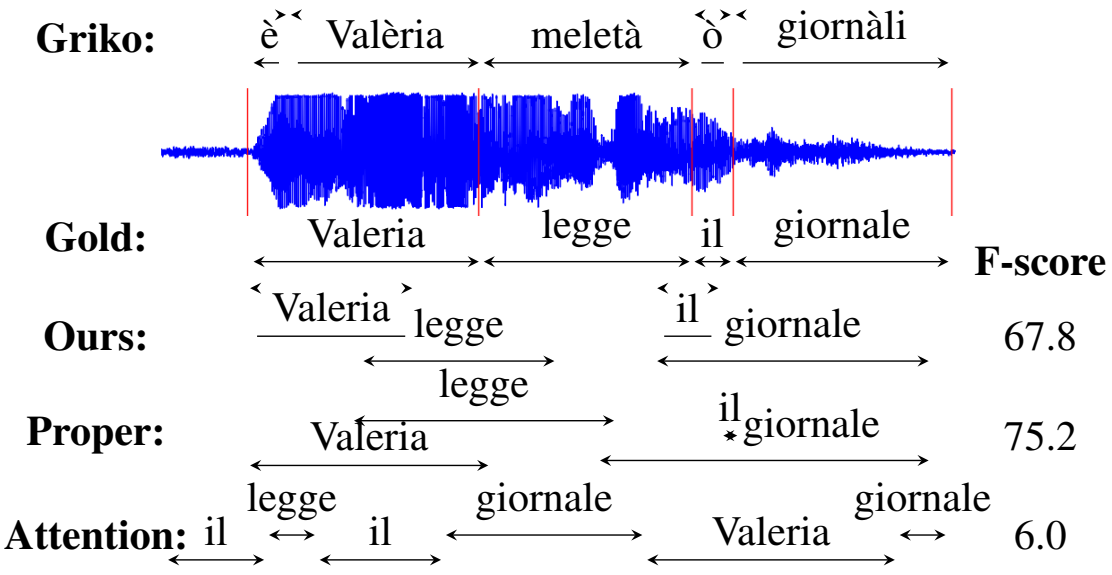


Figure 3.5: One of the rare examples where the proper model performs better than the deficient one. The hapax legomena `Valeria` and `giornali` are not properly handled by the attentional model.

In the IBM word alignment models, if a source word f occurs in only one sentence, then EM can align many target words to f and learn a very peaked distribution $t(e \mid f)$. This can happen in our model and the proper model as well, of course, since IBM Model 2 is embedded in them. But in the proper model, something similar can also happen with $s(f \mid a, b)$: EM can make the span (a, b) large or small, and evidently making the span small allows it to learn a very peaked distribution $s(f \mid a, b)$. By contrast, our model has $s(a, b \mid f)$, which seems less susceptible to this kind of effect.

3.3 A Case Study on Using Speech-to-Translation Alignments for Language Documentation

Abstract: We investigated whether providing translation and speech-to-translation alignment information can aid in producing better (mismatched) crowdsourced transcriptions from non-speakers of a language, which in turn could be valuable for training speech recognition systems. We showed that they can indeed be beneficial, through a small-scale case study on our Griko-Italian dataset, as a proof-of-concept. We also presented a simple phonetically aware string averaging technique based on DTW and DBA, that produces transcriptions of higher quality than the originally collected ones.

3.3.1 Introduction

A recent line of work (Das et al., 2016; Jyothi and Hasegawa-Johnson, 2015; Liu et al., 2016) focuses on training speech recognition systems for low-resource settings using mismatched crowdsoursed transcriptions. These are transcriptions that include some level of noise, as they are crowdsourced from workers unfamiliar with the language being spoken.

We aim to explore whether the quality of crowdsourced transcriptions could benefit from providing transcribers with speech-to-translation word-level alignments. That way, speech recognition systems trained on the higher-quality probabilistic transcriptions (of at least a sample of the collected data) could be used as part of the pipeline to document an endangered language.

3.3.2 Methodology

We randomly sampled 30 utterances from the Griko-Italian corpus and collected transcriptions through a simple online interface from 12 different participants. None of the participants spoke or had any familiarity with Griko or its directly related language, Greek. Six of the participants were native speakers of Italian, the language in which the translations are provided. Three of them did not speak Italian, but were native Spanish speakers, and the last 3 were native English speakers who also did not speak Italian but had some

level of familiarity with Spanish.

The utterances were presented to the participants in three different modes:

1. `no` mode: Only providing the translation text.
2. `auto` mode: Providing the translation text and the potentially noisy speech-to-translation alignments produced by our EM-based method (§3.2).
3. `gold` mode: Providing the translation text and the gold-standard speech-to-translation alignments.

The utterances were presented to the participants in the exact same order but under a rotation scheme that ensured that the utterances were effectively split into 3 subsets, each of which was transcribed exactly 4 times in each mode, with 2 of them by an Italian speaker, 1 time by a Spanish speaker, and 1 time by an English speaker. This enables a direct comparison of the three modes, and, hopefully, an explanation of the effect of providing the alignments. The modes under which each participant had to transcribe the utterances changed from one utterance to another, in order to minimize the potential effect of the participants' learning of the task and the language better.

The participants were asked to produce a transcription of the given speech segment, using the Latin alphabet and any pronunciation conventions they wanted. The result in almost all cases is entirely comprised of nonsense syllables. It is safe to assume, though, that the participants would use the pronunciation conventions of their native language.

Interface A simple tool for collecting transcriptions first needs to provide the user with the audio to be transcribed. The (Italian) translation of the (Griko) spoken utterance is also provided. In a real scenario, this translation would correspond to the output of a Speech Recognition system for the parallel speech, so it could potentially be somewhat noisy.For the purposes of this case study, though, we used the gold standard translations of the utterances.

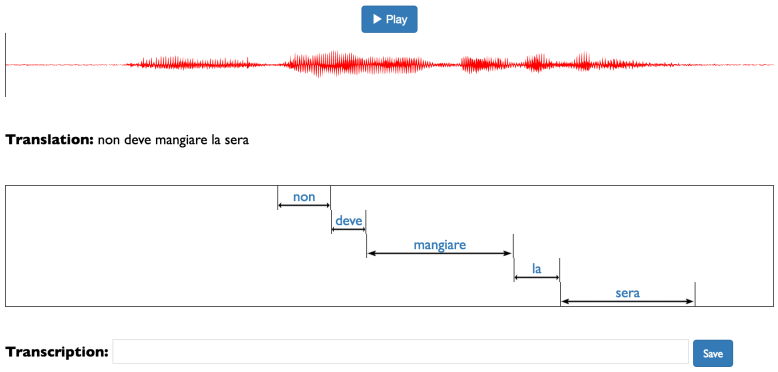


Figure 3.6: Screenshot of the interface that provides the translation `non deve mangiare la sera` *[he/she shouldn't eat at night]*, along with speech-to-translation alignment information. Clicking on a translation word would play the corresponding aligned part of the speech segment.

Our interface[1] also provides speech-to-translation alignment information as shown in Figure 3.6. Each word in the translation has been aligned to some part of the spoken utterance. Apart from listening to the whole utterance at once, the user can also click on the individual translation words and listen to the corresponding speech segment.

For the purposes of our case study, our tool collected additional information about its usage. It logged the amount of time each participant spent transcribing each utterance, as well as the amount of times that they clicked the respective buttons in order to listen to either the whole utterance or word-aligned speech segments.

Averaging the acoustic transcriptions A fairly simple way to merge several transcriptions into one is to obtain first alignments between the set of strings to be averaged by treating each substitution, insertion, deletion, or match, as an alignment. Then, we can leverage the alignments in order to create an "average" string, through an averaging scheme.

We propose a method that can be roughly described as similar to using Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) for obtaining alignments between two speech signals, and using DTW Barycenter Averaging (DBA) (Petitjean et al., 2011) for approx-

[1] Available online at `https://bitbucket.org/antonis/online-annotation-tools/src`

imating the average of a set of sequences. Instead of time series or speech utterances, however, we apply these methods on sequences of phone embeddings.

We map each IPA phone into a feature embedding, with boolean features corresponding to linguistic features.[2] Then, each acoustic transcription can be represented as a sequence of vectors, and we can use DBA in order to obtain an "average" sequence, out of a set of sequences. This "average" sequence is then mapped back to phones, by mapping each vector to the phone that has the closest phone embedding in our space.

The standard method, ROVER (Fiscus, 1997), uses an alignment module and then majority voting to produce a probabilistic final transcription. The string averaging method that we propose here is quite similar, with the exception that our alignment method and the averaging method are tied together through the iterative procedure of DBA. Another difference is that our method operates on phone embeddings and not on phones. That way, it is more phonologically informed, so that the distance between two phones that are often confused because they have similar characteristics, such as /p/ and /b/, is smaller than the distance between a pair of more distant phones such as eg. /p/ and /a/. In addition, the averaging scheme that we employ actually produces an average of the aligned phone embeddings, which in theory could result in a different output compared to simple majority voting.

3.3.3 Results

The orthography of Griko is phonetic, and therefore it is easy, using simple rules, to produce the phonetic sequences in IPA that correspond to the transcriptions. We can also use standard rules for Spanish (LDC96S35) and Italian,[3] depending on the native language of the participants, in order to produce phonetic sequences of the crowdsourced transcriptions in IPA.

[2]The features were taken from the inventories of `http://phoible.org/`

[3]Creating the rules based on (Comrie, 2009)

TABLE 3.3

BREAKDOWN OF THE QUALITY OF THE TRANSCRIPTIONS PER UTTERANCE SET

utterance set	Levenshtein distance			
	`no`	`auto`	`gold`	all modes
set 1	14.1	13.5	13.9	13.8
set 2	10.0	10.6	8.7	9.8
set 3	11.8	10.1	10.5	10.8
average	12.0	11.4	11.0	11.5

For simplicity reasons, we merge the vowel oppositions /e∼ɛ/ and /o∼ɔ/ into just /e/ and /o/ for both the Italian and Griko phonetic transcriptions, as neither of the two languages makes an orthographic distinction between the two.

For the transcriptions created by the English-speaking participants, and since most of the word-like units of the transcriptions do not exist in any English pronunciations lexicon, we use the LOGIOS Lexicon Tool (SpeechLab, 2007) that uses some simple letter-to-sound rules to produce a phonetic transcription in the ARPAbet symbol set. We map several of the English vowel oppositions to a single IPA vowel; for example, `IH` and `IY` both become /i/, while `UH` and `UW` become /u/. Phonemes `AY`, `EY`, and `OY` become /ai/, /ei/, and /oi/ respectively. This enables a direct comparison of all the transcriptions, although it might add extra noise, especially in the case of transcriptions produced by English-speaking participants.

Two examples of the resulting phonetic transcriptions as produced by the participants' transcriptions can be found in Tables 3.7 and 3.8.

We observe that the acoustic transcriptions are generally better when alignments are

TABLE 3.4

PHONE ERROR RATE (PER) OF THE PHONETIC TRANSCRIPTIONS PRODUCED BY THE ITALIAN-SPEAKING PARTICIPANTS PER UTTERANCE SET

utterance set	Phone Error Rate (PER)			
	`no`	`auto`	`gold`	all modes
set 1	23.0	25.1	23.8	24.0
set 2	25.8	26.0	23.3	25.0
set 3	32.1	26.0	24.5	28.1
avg	27.0	25.7	24.5	25.7

provided, as reported in Table 3.4 that shows the average Phone Error Rate (PER) of these phonetic sequences. The gold alignments provide more accurate information, resulting in higher quality transcriptions. However, even using the noisy alignments leads to better transcriptions in most cases.

Furthermore, using the string averaging method we combine the mismatched transcriptions into an "average" one. We can then compute the Levenshtein distance and PER between the "average" and the gold transcription in order to evaluate them. In almost all cases the "average" transcription is closer to the gold one than each of the individual transcriptions. Table 3.6 provides a more detailed analysis of the quality of the "average" transcriptions per mode and per group of participants.

We first use the transcriptions as produced by all participants, and report the errors of the averaged outputs under all modes. Again, when alignments are provided, the averaged transcriptions have lower error rates. However, the `gold` mode corresponds to an ideal scenario, which will hardly ever occur. Thus, we focus more on the combination of the `no`

TABLE 3.5

BREAKDOWN OF THE QUALITY OF THE TRANSCRIPTIONS PER PARTICIPANT GROUP

participants	PER
Italian	25.7
Spanish	28.3
English	34.3
all	28.5
best	22.8
worst	37.0

and `auto` modes, which will very likely occur in our collection efforts, as the alignments we will produce will be noisy, or we might only have translations without alignments. We also limit the input to only include the transcriptions produced by the Italian and Spanish speaking participants, as we found that the transcriptions produced by English speaking participants added more noise instead of helping. As the results in Table 3.6 show, using our averaging method we obtain better transcriptions on average, even if we limit ourselves to the more realistic scenario of not having `gold` alignments. The best result with an average PER of 23.2 is achieved using all the transcriptions produced by Italian and Spanish speaking participants. Even without gold alignments, however, the averaging method produces transcriptions with average PER of 24.0, which is a clear improvement over the average PER of the individual transcriptions (25.7).

Transcription quality As a first test, we compare the Levenshtein distances of the produced transcriptions to the gold ones. For fairness, we remove the accents from the gold

TABLE 3.6

AVERAGE LEVENSHTEIN DISTANCE AND PER OF THE "AVERAGE" TRANSCRIPTIONS OBTAINED WITH OUR STRING AVERAGING METHOD FOR DIFFERENT SUBSETS OF THE CROWDSOURCED TRANSCRIPTIONS

transcriptions used to create average		avg. distance to gold	
mode	participants' native language	Lev/tein	PER
`no`	all	8.41	27.0
`auto`	all	7.82	25.9
`gold`	all	7.58	24.3
all	Ita+Spa	**7.21**	**23.2**
`gold`	Ita+Spa	7.55	23.6
`no+auto`	Ita+Spa	7.62	24.0

Griko transcriptions, as well as any accents added by the Italian speaking participants.

The results averaged per utterance set and per mode are shown in Table 3.3. We first note that the three utterance sets are not equally hard: the first one is the hardest, with the second one being the easiest one to transcribe, as it included slightly shorter sentences. However, in most cases, as well as in the average case (last row of Table 3.3) providing the alignments improves the transcription quality. In addition, the gold standard alignments provide more accurate information that is also reflected in higher quality transcriptions.

We also evaluate the precision and recall of the word boundaries (spaces) that the transcriptions denote. We count a discovered word boundary as a correct one only if the word

TABLE 3.7

TRANSCRIPTIONS FOR THE UTTERANCE O LÀDRO ÌSOZE ÈMBI APO-TTÙ *[THE THIEF MUST HAVE ENTERED FROM HERE]* AND THEIR LEVENSHTEIN DISTANCE TO THE GOLD TRANSCRIPTION

	transcription	distance
it1	o ladro isodzeem biabiddu	5
it2	o ladro isodʒenti dabol tu	6
it3	o ladro i so ndze mia buttu	5
it4	o ladro isodzeembia po tu	2
it5	o ladroi isodʒe enbi a buttu	4
it6	o ladro idʒo dzembia a buttu	7
es1	o la vro ipsa ziem biabotu	9
es2	ola avro isonse embia butu	7
es3	o ladro isosen be abuto	9
en1	o labro ebzozaim bellato	13
en2	o laha dro iso dzenne da to	12
en3	o ladro i dzo ze en habito	11
average	o ladro isodʒe mbia buttu	3
correct	o ladro isodʒe embi apo ttu	

boundary in the transcription is matched with a boundary marker in the gold transcription, when we compute the Levenshtein distance.

Under no mode (without alignments), the transcribers achieve 58% recall and 70% precision on correct word boundaries. However, when provided with alignments, they achieve 66% recall and 77% precision; in fact, when provided with gold alignments (un-

TABLE 3.8

TRANSCRIPTIONS FOR THE UTTERANCE PÀO CERKÈONTA ÈNA FÙRNO KA KÀNNI RÙSTIKU *[I'M LOOKING FOR A BAKERY THAT MAKES RUSTIC (BREAD)]* AND THEIR LEVENSHTEIN DISTANCE TO THE GOLD TRANSCRIPTION

participant	acoustic transcription	distance
it1	bau tʃerkianta ena furno e tranni e rustiku	9
it2	pau tʃerkianta ena furna kanni e rustiku	7
it3	pau tʃerkianta na furno kakanni rustiko	5
it4	po ʃerkieunta na furna ka kanni rustiku	6
it5	pau tʃerkeunta en furno ganni rustiku	6
it6	pa u tʃerkionta en na furno kahanni rustiko	5
es1	pogurʃe kiunta en a furna e kakani e rustiku	12
es2	pao ʃerkeonta ena furna ka kani rustigo	5
es3	bao tʃerke on ta e na furno e kagani e rustiko	6
en1	paoje kallonta e un forno e grane e rustiko	15
en2	pao tʃerkeota eno furno e kakarni e rustiko	5
en3	pouʃa kianta e a forno e tagani e rustiko	14
average	pao tʃerkionta ena furno kaanni e rustiku	3
correct	pao tʃerkeonta ena furno ka kanni rustiku	

der gold mode) recall increases to 70% and precision to 81%. Therefore, the speech-to-translation alignments seem to provide information that helped the transcribers to better identify word boundaries, which is arguably hard to achieve from just continuous speech.

Phonetic transcription quality We observe the same pattern when evaluating using the average PER of these phonetic sequences, as reported in Table 3.4: the acoustic transcriptions are generally better when alignments are provided. Also, the gold alignments provide more accurate information, resulting in higher quality transcriptions. However, even using the noisy alignments leads to better transcriptions in most cases.

It is worth noting that out of the 30 utterances, only 4 included words that are shared between Italian and Griko (`ancora` *[yet]*, `ladro` *[thief]*, `giornale` *[newspaper]*, and `subito` *[immediately]*) and only 2 of them included common proper names (`Valeria` and `Anna`). The effect of having those common words, therefore, is minimal.

Non-Italian speaking participants The scenario where the crowdsourcers do not even speak the language of the translations is possibly too extreme. It still could be applicable, though, in the case where the language of the translations is not endangered but still low-resource (Tok Pisin, for example) and it's hard to find annotators that speak the language. In any case, we show that if the participants speak a language related to the translations (and with a similar phonetic inventory, like Spanish in our case) they can still produce decent transcriptions.

Table 3.5 shows the average on the performance of the different groups of participants. As expected, the Italian-speaking participants produced higher quality transcriptions, but the Spanish-speaking participants did not perform much worse. Also in the case of non-Italian speaking participants, we found that providing speech-to-translation alignments (under `auto` and `gold` modes) improves the quality of the transcriptions, as we observed a similar trend as the ones shown in Tables 3.3 and 3.4.

The noise in the non-Italian speaker annotations, and especially the ones produced by English speakers, can be explained in two ways. One, it could be caused by annotation scheme employed by the English speakers, which must be more complicated and noisy, as English does not have a concrete letter-to-sound system. Or two, it could be explained by

the fact that English is much more typologically distant from Griko, meaning, possibly, that some of the sounds in Griko just weren't accessible to English speakers. The latter effect could indeed be real, as it has been shown that a language's phonotactics can affect what sounds a speaker is actually able to perceive (Dupoux et al., 2008; Peperkamp et al., 1999). The perceptual "illusions" created by one's language can be quite difficult to overcome.

3.4 Spoken Term Discovery for Language Documentation using Translations

Abstract: We presented a method for partially labeling unlabelled speech with translation labels in a scenario where only a small amount of data is translated. We modified an unsupervised speech-to-translation alignment model and obtained prototype speech segments that match the translation words, which are in turn used to discover terms in the unlabelled data. We evaluated our method on a Spanish-English speech translation corpus and on two corpora of endangered languages, Arapaho and Ainu, demonstrating its appropriateness and applicability in an actual very-low-resource scenario.

3.4.1 Introduction

Vast amounts of speech data collected for language documentation and research remain untranscribed and unsearchable, but often a small amount of speech may have text translations available. We focused on this scenario, and explored whether we can partially label additional speech with translation keywords. We used our EM-based alignment method (§3.2, henceforth `s2t`) to obtain labeled "prototypical" speech segments, which we use for term discovery. Hopefully, those translation keywords could render the unlabeled linguistic archives more searchable.

Background The only previous system we know of to address the same very-low-resource scenario and provide translation terms for unlabeled audio is that of Bansal et al. (2017) (henceforth `UTD-align`), who used an unsupervised term discovery system (Jansen et al., 2010) to cluster recurring audio segments into pseudowords. The pseudowords occurring in the parallel section of the corpus were then aligned to the translation text using IBM Model 1, and used to translate instances occurring in the test (audio-only) section.

3.4.2 Method

The main difference between our method and `UTD-align` is that `UTD-align` clusters the audio prior to aligning with the translations, whereas we start by performing joint align-

ment and clustering using an improved version of the `s2t` method. The resulting aligned clusters are represented by one or more prototype speech segments. We extended `s2t` to identify new instances of those prototypes in the unlabeled speech, using a modified version of `ZRTools`, the same UTD toolkit used by `UTD-align`.[4] (Jansen et al., 2010)

We first modified `s2t` so that, before the M-step, each cluster's segments are grouped into sub-clusters using connected components clustering with a similarity threshold d, following Park and Glass (2008). That way, the number of sub-clusters and prototypes for each translation word is determined automatically based on the acoustic similarity of the segments.

Our preliminary analysis showed that shorter alignments tend to introduce significantly more noise than longer ones. Therefore, in the final M-step of `s2t`, we also discard all segments shorter than a length threshold t before computing the prototypes. We used the default values for the rest of the `s2t` parameters, and obtained speech-to-translation alignments, as a first step.

Another pragmatic choice we made based on the performance of our method was to remove the stopwords from the translations, following Bansal et al. (2017). The rationale is that translation stopwords would not be particularly useful for labelling speech in our envisioned use cases.

In the second stage, we use the approximate DTW-based pattern matching method of `ZRTools` to search for the obtained prototypes in the test data. We require that each discovered term matches at least $k\%$ of a prototype's length and that its DTW similarity score is higher than a threshold s. By varying s we can control the number of discovered terms, trading off precision and recall. Also, we do not allow overlapping matches; in the case of an overlap, we output the match with the higher score.

[4] https://github.com/arenjansen/ZRTools

TABLE 3.9

RESULTS OF OUR KEYWORD-SPOTTING METHOD AND BASELINE WORK ON THE CALLHOME DATASET

Method	Precision	Recall	F-score	Coverage
`UTD-align`	5.1	2.1	3.0	27%
ours	4.2	3.5	3.8	59%
ours (oracle)	5.3	4.9	5.1	65%

3.4.3 Experiments and Results

We evaluated our spoken term discovery method on CALLHOME and on the Arapaho and Ainu datasets. We first evaluated the effect of our modifications to the `s2t` method, by calculating alignment F-score on links between speech frames and translation words. The intermediate sub-clustering step between the E- and M-steps results in a more informed selection of the number of sub-clusters that increases the alignment F-score by 1.5%. Also, removing translation stopwords further leads to higher alignment precision by +4%. Alignment recall is lower since it's computed over the alignments of both content and stopwords. Although both improvements are small, the higher alignment precision leads to better prototypes.

Out of the eight Arapaho narratives, we select the longest (18 minutes of audio, 233 English word types) for training, using the other seven (32 minutes total) for evaluation. The Ainu collection provides ten narratives, so we use the first two for training (24 minutes of audio, 494 English word types) and the rest (133 minutes total) as test data.

Treating each narrative as a bag of words, the precision and recall results at the token level are shown in Tables 3.10 and 3.11. The last columns of these Tables correspond to

TABLE 3.10

KEYWORD SPOTTING RESULTS ON ARAPAHO TEST NARRATIVES

Arapaho narrative	Terms found	Prec (%)	Rec (%)	Oracle Recall
1	29	31.0	4.7	32.3
2	65	21.5	8.0	44.3
3	91	7.7	6.4	54.5
4	158	13.9	8.4	53.4
6	1	100.0	0.7	41.4
7	104	7.7	7.1	44.6
8	10	30.0	4.5	65.2
average-ours	65	14.0	6.0	
`UTD-align`	2	26.7	0.4	

the highest possible recall that we could get if we discovered all the training terms that also appear in the test set. Precision-recall curves can be seen in Figure 3.7.

On both corpora, `UTD-align` identifies hardly any translation terms, with recall scores below 1% and average F-scores of 0.8% and 0.2% for Arapaho and Ainu, respectively. Our method, instead, outputs several terms per narrative without the need to readjust preprocessing decisions, with F-scores of 8.4% (Arapaho) and 7.2% (Ainu). Two exceptions are Arapaho narratives #6 and #8, which, unlike our training data, are narrated by a woman. Although there is clearly room for improvement in terms of recall, as shown by the last columns of Table2 3.10 and 3.11, we are generally able to identify meaningful terms.

TABLE 3.11

KEYWORD SPOTTING RESULTS ON AINU TEST NARRATIVES

Ainu narrative	Terms found	Prec (%)	Rec (%)	Oracle Recall
3	80	50.0	3.8	63.0
4	73	49.3	4.5	67.1
5	199	49.7	5.1	61.8
6	174	22.4	9.0	65.0
7	123	19.5	8.9	56.1
8	122	57.4	3.9	67.8
9	59	62.7	1.5	63.0
10	149	46.3	6.6	69.7
average-ours	122	42.3	4.2	
`UTD-align`	4	24.2	0.1	

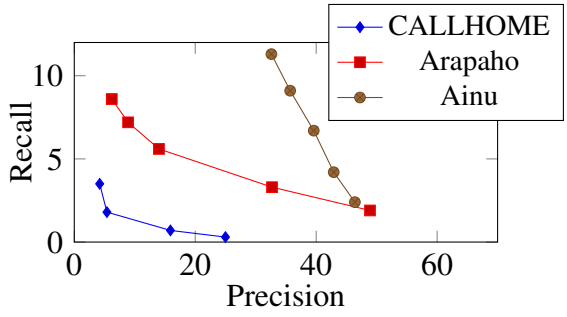


Figure 3.7. Average precision and recall curve for our discovered matches in CALLHOME and the Arapaho and Ainu test narratives (varying the output threshold s between 0.90 and 0.94).

CHAPTER 4

NEURAL SPEECH TRANSCRIPTION AND TRANSLATION

This chapter describes our contributions that use translations in order to produce higher-quality transcriptions in an endangered language. Our extremely low-resource setting requires that we explore methods that take full advantage of the available signals.

In Section §4.1 we focus on the scenario where translations are available both at training and inference time, and show that indeed access to translations can lead to better performance, especially in noisy audio scenarios. Then, in §4.2 we explore scenarios where translations are not available at inference time, but can be used at training time as privileged information. We extend the Learning Under Privileged Information paradigm and achieve improvements comparable to the §4.1 models, which however, use translations at test time. Finally, Section §4.3 presents a multitask setting where we train a neural model to produce both transcriptions and translations. Training jointly for the two tasks, and encouraging the model to obey the notion of *transititivity*, we attain further improvements in performance.

4.1 Leveraging Translations for Speech Transcription in Low-Resource Settings

Abstract: In our data collection framework, we end up with a significant amount of parallel speech in a low-resource and a high-resource language. We explore whether we can leverage these translations, in order to improve the quality of the endangered language transcriptions. We find that sharing the parameters of the attention mechanisms improves the performance of the multi-source models, in most cases outperforming the single-source baselines.

4.1.1 Introduction

We focus on the language documentation scenario where we already have translations for speech utterances (e.g. through a parallel data collection pipeline like Aikuma). The goal is to produce transcriptions for these utterances. Therefore, we explore methods that learn from a small number of transcribed speech utterances along with their translations.

We use the neural attentional model (Bahdanau et al., 2015a) and experiment with extensions that take both speech utterances and their translations as input sources. We assume that the translations are in a high-resource language that can be automatically transcribed; therefore, in our experiments, the translation input is text instead of speech. We also explore different parameter-sharing methods across the attention mechanisms.

We experiment on three diverse low-resource language pairs. One is Ainu, a severely endangered language, with translations in English. We also experiment on a recently collected speech corpus of Mboshi (Godard et al., 2017), with translations in French. Lastly, we evaluate our models on Spanish-English, using the CALLHOME dataset.

Our proposed multi-source model that employs a shared attention mechanism outperforms the baselines in almost all cases. In Mboshi, we find that our model reduces character error rates (CER) by 1.2 points. In Spanish, we observe a reduction of 4.6 points in CER over the strongest baseline, and more than 14.4 points over a speech-only baseline. In Ainu, although our multi-source model doesn't reduce the overall CER, we show that it actually is beneficial in the cases where the single-source speech transcription model has greatest

difficulty.

4.1.2 Model

Unlike the traditional attentional model (see Figure 4.1a), in a *multi-source* model we have two encoders (Figure 4.1b); one that transforms the input sequence of speech frames $\mathbf{f}_1, \ldots, \mathbf{f}_N$ into a sequence of input states $\mathbf{h}^1_1 \ldots \mathbf{h}^1_N$, and one that transforms an input sequence of translation words $\mathbf{x}_1, \ldots, \mathbf{x}_M$ into another sequence of input states $\mathbf{h}^2_1 \ldots \mathbf{h}^2_M$:

$$\mathbf{h}^1_n = \text{enc1}(\mathbf{h}^1_{n-1}, \mathbf{f}_n)$$
$$\mathbf{h}^2_m = \text{enc2}(\mathbf{h}^2_{m-1}, \mathbf{x}_m)$$

An attention mechanism transforms the two sequences of input states into a sequence of *context vectors* via two matrices of *attention weights*:

$$\mathbf{c}_k = \left[\sum_n \alpha^1_{kn} \mathbf{h}^1_n \quad \sum_m \alpha^2_{km} \mathbf{h}^2_m \right].$$

Finally, the decoder computes a sequence of *output states* from which a probability distribution over output words can be computed.

$$\mathbf{s}_k = \text{dec}(\mathbf{s}_{k-1}, \mathbf{c}_k, \mathbf{y}_{k-1})$$
$$P(\mathbf{y}_m) = \text{softmax}(\mathbf{s}_m).$$

The attention mechanisms produce the attention weights with the following computations, as in (Luong et al., 2015), with $\mathbf{v}^1, \mathbf{v}^2, \mathbf{W}^s_{\alpha^1}, \mathbf{W}^s_{\alpha^2}, \mathbf{W}^h_{\alpha^1}$, and $\mathbf{W}^h_{\alpha^2}$ being parameters to

be learnt:

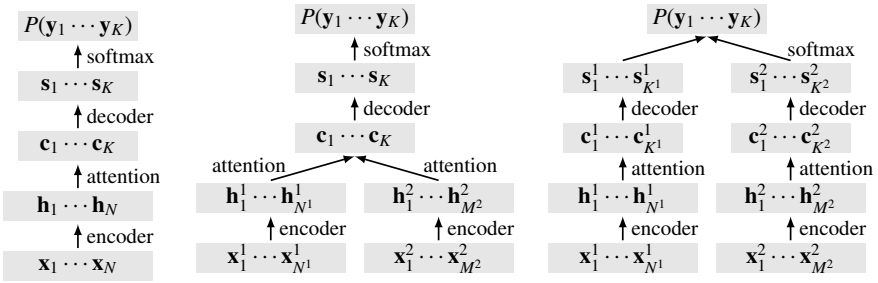
$$\alpha^1_{kn} = \text{softmax}(\mathbf{v}^1 \tanh(\left[\mathbf{W}^s_{\alpha^1}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^1}\mathbf{h}^1_n\right]))$$
$$\alpha^2_{km} = \text{softmax}(\mathbf{v}^2 \tanh(\left[\mathbf{W}^s_{\alpha^2}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^2}\mathbf{h}^2_m\right])).$$

Since both attention mechanisms, though, provide context to the same decoder, we can tie the computation of the weights so that the two mechanisms share the $\mathbf{v}$ and $\mathbf{W}^s_{\mathbf{a}}$ parameters. We refer to them as *tied* attention mechanisms:

$$\alpha^1_{kn} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^1}\mathbf{h}^1_n\right]))$$
$$\alpha^2_{km} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^2}\mathbf{h}^2_m\right])).$$

If the two encoders share the same output size for their $\mathbf{h}^1$ and $\mathbf{h}^2$ vectors, then the two attentions could further share the $\mathbf{W}^h_{\alpha}$ parameters, effectively merging into one, *shared* attention mechanism.

A baseline that has to be compared with our work is *ensembling*. Traditionally, ensembles refer to models that have been trained on similar data for the similar task, with their predictions only combined at inference time. In our case, we explore an ensemble of a transcription and a translation model. In the *simple ensemble* case, the two models are trained separately. Recently, *coupled ensembles* were shown to outperform simple ensembles (Dutt et al., 2017). In the *coupled ensemble* setting (see Figure 4.1c), the two models are trained jointly, albeit they don't share any parameters. The two decoder outputs are averaged right before the softmax layer, in order to produce a single output probability distribution. It was shown (Dutt et al., 2017) that this approach works better than combining the two predictions after the softmax layer:



(a) single-source (b) multi-source (c) coupled ensemble

Figure 4.1: Source-side variations on the standard attentional model. In the standard *single-source* model, the decoder attends to a single encoder's states. In our *multi-source* setup, we have two input sequences encoded by two different encoders, and attention mechanisms provide two context to the decoder. Note that for clarity's sake there are dependencies not shown.

$$\mathbf{s}_k^1 = \text{dec1}(\mathbf{s}_{k-1}^1, \mathbf{c}_k^1, \mathbf{y}_{k-1})$$

$$\mathbf{s}_k^2 = \text{dec2}(\mathbf{s}_{k-1}^2, \mathbf{c}_k^2, \mathbf{y}_{k-1})$$

$$P(\mathbf{y}_k) = \text{softmax}(\frac{\mathbf{s}_k^1 + \mathbf{s}_k^2}{2}).$$

4.1.3 Experiments and Results

We evaluate our models on three datasets: Ainu-English, Mboshi-French, and the CALLHOME Spanish-English dataset. The results on Ainu are calculated over the concatenated outputs of 10 cross-validation folds, in each of which each narrative becomes the test set. The results on Mboshi are the outputs of the best model (selected from dev performance) of 10 restarts.

Table 4.1 summarizes the outcome of some initial experiments. We find that the best models (with the lower Character Error Rates and higher BLEU) in our endangered languages are the ones that take advantage of both the spoken speech and the translations, and combine them with a parameter sharing mechanism for attention. In the CALLHOME

TABLE 4.1

BOTH CHARACTER ERROR RATES (CER) AND WORD-LEVEL BLEU OF OUR BEST MULTI-SOURCE MODELS ON MBOSHI AND SPANISH OUTPERFORM THE SINGLE-SOURCE BASELINES

Source	Target CER			Target BLEU	
	Ainu	Mboshi	Spanish	Ainu	Spanish
speech	40.7	29.8	52.0	**28.92**	9.41
translation	74.9	68.2	44.6	5.89	14.73
coupled ensemble	**40.6**	36.8	42.2	26.99	16.94
speech+translation	46.0	37.5	41.6	24.03	17.59
+tied	41.4	32.6	**37.6**	26.95	**20.82**
+shared	**40.6**	**28.6**	38.7	**28.57**	19.47

dataset, both the *tied* and the *shared* attention mechanisms significantly outperform the baselines, with the *tied* attention mechanism producing slightly better results. It is also worth noting that the coupled ensemble model also outperforms the single-source baselines. However, they only perform en par with our proposed multi-source models with parameter sharing mechanisms for attention in the case of Ainu.

4.1.4 Analysis

The performance of each fold of cross-validation for Ainu is shown in Figure 4.2. For each narrative, it compares the speech-only baseline system with our best multi-source system. The overall performance of the speech-only single-source model and our best model is similar with a CER of 40.7 and 40.6 respectively. A possible reason is that all the Ainu stories are narrated by the same speaker, making it a generally easier task for a speech

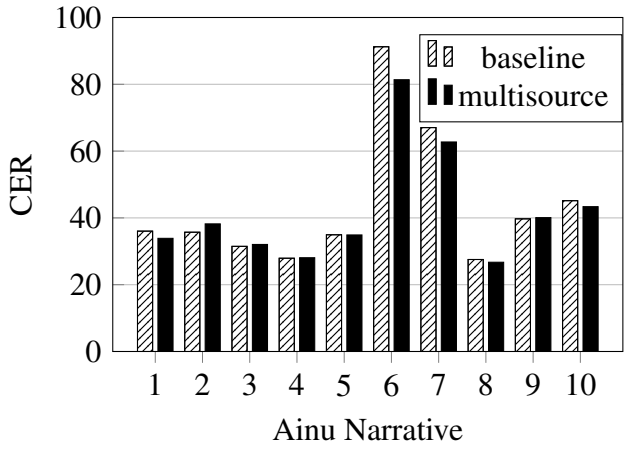
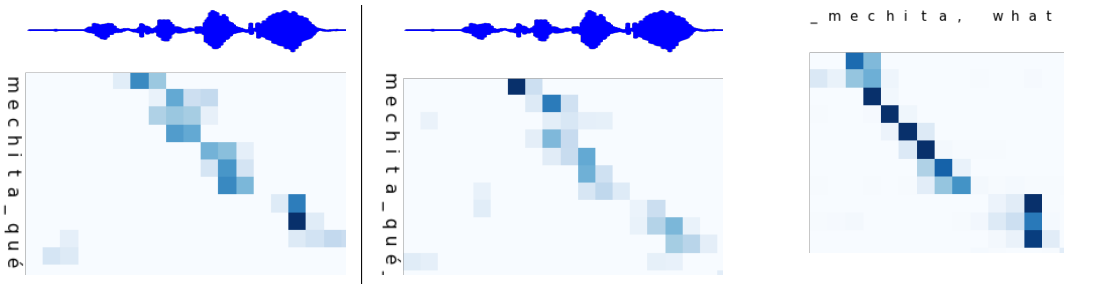


Figure 4.2. Character Error Rates of the best baseline system and our best multisource system for each Ainu narrative. The gains of using the translations are apparent in the cases that are harder for a speech-only system: narratives 6 and 7 are more sung than narrated, rendering them harder to transcribe.

recognition system. But we also see that in the cases where speech transcription is harder, translation information does help. Namely, narratives 6 and 7 are *sung*, making them harder to transcribe with a speech-only system trained on spoken data, as indicated by the higher error rates: 91.2 and 67.0, respectively. The multi-source models achieve noticeable improvements of 9.9 and 4.3 points on these narratives.

We further quantify the effect of the different sharing mechanisms for the attentions. Using word-level forced alignments on the CALLHOME dataset (Duong et al., 2016) we can evaluate the accuracy of the attention. Treating the forced alignments as reference, we compute the percentage of the weights of the attention over the speech source that fall within the boundaries of the forced alignment spans. Note that the forced alignments naturally include noise, so they should be treated as a "silver standard." However, they can still provide indications that could reveal the effect of parameter sharing.

We computed the average sum of this *attention accuracy* by forced decoding on the CALLHOME development set. We find that the average sum for the speech single-source model is almost 71%, a value similar to the average sums of the attention accuracy of



(a) speech-source (b) multi-source: speech (c) multi-source: translation

Figure 4.3: Attentions on a speech sample from the dev set, that includes a proper name ("Mechita") unseen during training. The multi-source model (using a *shared* attention mechanism) receives informative context from the translation so as to produce the output.

the coupled ensemble and the multi-source model that employs no sharing mechanism. Instead, the attention accuracy of the model with the shared mechanism is almost 75%. The model with tied attentions, which achieves the best results on CALLHOME, has an attention accuracy of 76%.

Figure 4.3 presents the attention weights over a sample taken from the development set, produced by forced decoding. The segment includes an out-of-vocabulary word, the name *Mechita*, never seen during training. The attention weights over the speech source with the single-source model (3a) are not too different from the weights of the multi-source model with tied attentions (3b). However, the multi-source model in this case takes advantage of the translation and receives most of its context from the text source (3c), as the attention weights over the characters of the name are quite high (albeit, off-by-one, as often is the case in neural attention-based translation).

4.2 Translations as Privileged Information for Low-Resource Speech Transcription

Abstract: We explore the recently proposed Learning Under Privileged Information for deep neural models (DLUPI) framework (Lambert et al., 2018). We adapt it to the speech transcription task and enhance it with attention mechanism in order to receive fine-grained privileged information from translations. We show that in low-resource settings we can achieve performance comparable to the best multi-source models, despite not having access to translations at inference time. At the same time the DLUPI models surpass single-source baselines that do not use translations at all.

4.2.1 Introduction

Learning Under Privileged Information (LUPI) is a novel machine learning paradigm, which attempts to imitate the role of a teacher that provides intuitive comments or comparisons, rather than just right or wrong answers. Originally introduced by Vapnik and Vashist (2009), it addresses an important shortcoming of typical supervised learning: in addition to the correct answer, the teacher also supplies the student with an "explanation".

More formally, a standard machine learning setup the model is trained using tuples $\{x, y\}$ of input x and desired outputs y. Under the LUPI paradigm, the model is trained using triplets $\{x, x^*, y\}$, where x^* denotes some sort of privileged information that the teacher has. Note that inference is still performed only on the standard input x, without access to the privileged information (as the "student" operates without the assistance of the "teacher").

Originally applied on Support Vector Machines, the LUPI paradigm has been extended to several models; importantly, it is theoretically proven that this algorithm accelerates the rate at which the upper bound of the error drops, from $O(\sqrt{\frac{1}{n}})$ to $O(\frac{1}{n})$, effectively leading to a steeper learning curve (Vapnik and Vashist, 2009).

Since then, LUPI has been applied to several problems in computer vision (Motiian et al., 2016; Sharmanska et al., 2014), ranking (Sharmanska et al., 2013), or clustering problems (Feyereisl et al., 2014). Hernández-Lobato et al. (2014) extended the LUPI framework to Gaussian Processes, while Lopez-Paz et al. (2015) showed that LUPI and

knowledge distillation are closely related.

Lambert et al. (2018) recently extended the LUPI framework to deep neural models (henceforth DLUPI), by using the privileged information to inform the variance of heteroscedastic dropout. We explain this method in the next section, applying it on the task of speech transcription and extending it to use more fine-grained privileged information.

4.2.2 Method

For the speech transcription task, the input x is a sequence of audio feature frames $\mathbf{f}$, and the desired output y is a sequence of the transcription characters or words. We will use the translations of the speech utterances as the privileged information $\mathbf{x}^*$. We describe our encoder-decoder attentional model using our previously introduced notation, with the encoder producing an intermediate representation of the input:

$$\mathbf{h}_n = \text{enc}(\mathbf{h}_{n-1}, \mathbf{f}_n)$$

Following Lambert et al. (2018), we apply heteroscedastic dropout (that is, dropout with varying variance)[1] on this representation such that:

$$\mathbf{h}'_n = \mathbf{h}_n \odot \mathcal{N}(1, \mathbf{h}^*(x^*)).$$

Effectively, the privileged information x^* is only used to estimate the variance $h^*(x^*)$ of the (Gaussian) heteroscedastic dropout of the representation of the input. In order to compute the variances, we employ an LSTM encoder over the translation. We use the final output state as the representation of the sentence, which is the passed through a fully connected layer.

These additional weights for the translation encoder and the fully connected layer are

[1] Standard Gaussian dropout (or additive Gaussian noise) with fixed variance is *homoscedastic*.

also learned. The loss includes the standard cross-entropy loss, plus regularization over the logarithm of the computed variances of the heteroscedastic dropout:[2]

$$\mathcal{L}(\theta) = \Sigma \log p(y|x) - \beta \parallel \log \mathbf{h}^*(x^*) \parallel .$$

Furthermore, we can extend the model to receive more fine-grained privileged information at each time step. Instead of using the same representation of the translation x^* at each time step, we can employ an attention scheme.

The translation encoder encodes the translation and produces an intermediate representation h^*:

$$\mathbf{h}^*(x^*_m) = \text{enc}(\mathbf{h}_{m-1}, \mathbf{x}^*_m).$$

and then an attention model transforms the input states into a sequence of *context vectors* via a matrix of *attention weights* for each of the n steps of encoded input x:

$$\mathbf{c}^*_n = \sum_n \alpha_{nm} \mathbf{h}^*_m.$$

Then, the heteroscedastic dropout at time step n is computed as $\mathcal{N}(1, \text{MLP}(\mathbf{c}^*_n))$ and applied on the intermediate representations produced by the transcription encoder.

In either case, the decoder attends with another set of attention weights over the intermediate representations $\mathbf{h}'$, receiving a context $\mathbf{c}$ and computing a sequence of *output states* from which a probability distribution over output words can be computed.

$$\mathbf{s}_k = \text{dec}(\mathbf{s}_{k-1}, \mathbf{c}_k, \mathbf{y}_{k-1})$$

$$P(\mathbf{y}_m) = \text{softmax}(\mathbf{s}_m).$$

A schematic representation of our proposed model is displayed in Figure 4.4.

[2] See (Lambert et al., 2018) for the complete derivation.

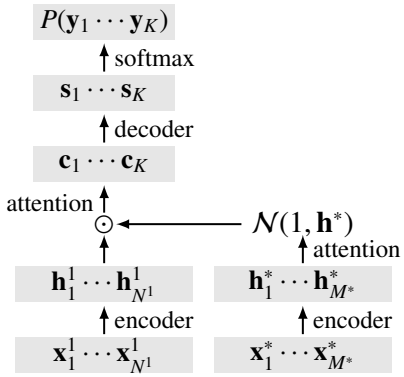


Figure 4.4: Schematic representation of our DLUPI model that includes an attention mechanism for obtaining the variance of the heteroscedastic dropout.

4.2.3 Experiments and Discussion

We test our approach on the Mboshi dataset, and compare the performance of our DLUPI models with the performance of our single-source and multi-source transcription models.

Our results, shown in Table 4.2, are in line with the theoretical and empirical results of Lambert et al. (2018). In our low-resource setting, providing additional "explanations" in the form of translations is beneficial to training (28.8 CER), as we observe an improvement of 1% Character Error Rate over the single-source baseline (29.8CER) which does not use translations at all. Moreover, the performance of the DLUPI model is comparable to the performance of our previously discussed multi-source model (28.6 CER), despite the fact that our DLUPI model does not have access to the translations at inference time.

This confirms that the LUPI paradigm indeed has huge potential for low-resource applications. With a few additional annotations that could be provided as privileged information, the model's learning curve is steeper and likely leads to better performance, even if the additional annotations are not available for the test data.

TABLE 4.2

THE PERFORMANCE OF THE DLUPI MODEL IS BETTER THAN THE SINGLE-SOURCE MODEL AND COMPARABLE TO THE MULTI-SOURCE MODEL

Model	Mboshi CER	Ainu CER
single source	29.8	40.7
multi-source	**28.6**	40.6
multi-task	30.2	**40.1**
DLUPI	28.8	40.6

4.3 Tied Multitask Models for Speech Transcription and Translation

Abstract: We explore multitask models for neural translation of speech, augmenting them in order to reflect two intuitive notions. First, we introduce a model where the second task decoder receives information from the decoder of the first task, since higher-level intermediate representations should provide useful information. Second, we apply regularization that encourages *transitivity* and *invertibility*. We show that the application of these notions on jointly trained models improves performance on the tasks of low-resource speech transcription and translation.

4.3.1 Introduction

Speech can be interpreted either by transcription in the original language or translation to another language. Since the size of the data is extremely small, multitask models that jointly train a model for both tasks can take advantage of both signals. Our contribution lies in improving the sequence-to-sequence multitask learning paradigm, by drawing on two intuitive notions: that higher-level representations are more useful than lower-level representations, and that translation should be both transitive and invertible.

Higher-level intermediate representations, such as transcriptions, should in principle carry information useful for an end task like speech translation. A typical multitask setup (Weiss et al., 2017) shares information at the level of encoded frames, but intuitively, a human translating speech must work from a higher level of representation, at least at the level of phonemes if not syntax or semantics. Thus, we present a novel architecture for *tied* multitask learning with sequence-to-sequence models, in which the decoder of the second task receives information not only from the encoder, but also from the decoder of the first task.

In addition, *transitivity* and *invertibility* are two properties that should hold when mapping between levels of representation or across languages. We demonstrate how these two notions can be implemented through regularization of the attention matrices, and how they lead to further improved performance.

In the speech transcription and translation tasks, our proposed model leads to improved performance against all baselines as well as previous multitask architectures. We observe improvements of up to 5% character error rate in the transcription task, and up to 2.8% character-level BLEU in the translation task.

4.3.2 Background

Multitask learning (Caruana, 1998) has found extensive use across several machine learning and NLP fields. For example, Luong et al. (2016) and Eriguchi et al. (2017) jointly learn to parse and translate; Kim et al. (2017) combine CTC- and attention-based models using multitask models for speech transcription; Dong et al. (2015) use multitask learning for multiple language translation. Toshniwal et al. (2017) apply multitask learning to neural speech recognition in a less traditional fashion: the lower-level outputs of the speech encoder are used for fine-grained auxiliary tasks such as predicting HMM states or phonemes, while the final output of the encoder is passed to a character-level decoder.

Our work is most similar to the work of Weiss et al. (2017). They used sequence-to-

sequence models to transcribe Spanish speech and translate it in English, by jointly training the two tasks in a multitask scenario where the decoders share the encoder. In contrast to our work, they use a large corpus for training the model on roughly 163 hours of data, using the Spanish Fisher and CALLHOME conversational speech corpora. The parameter number of their model is significantly larger than ours, as they use 8 encoder layers, and 4 layers for each decoder. This allows their model to adequately learn from such a large amount of data and deal well with speaker variation. However, training such a large model on endangered language datasets would be infeasible.

Our model also bears similarities to the architecture of the model proposed by Tu et al. (2017). They report significant gains in Chinese-English translation by adding an additional *reconstruction* decoder that attends on the last states of the *translation* decoder, mainly inspired by auto-encoders.

4.3.3 Model

Our models are based on a sequence-to-sequence model with attention (Bahdanau et al., 2015a). In general, this type of model is composed of three parts: a recurrent encoder, the attention, and a recurrent decoder (see Figure 4.5a).[3]

The encoder transforms an input sequence of words or feature frames $\mathbf{x}_1, \ldots, \mathbf{x}_N$ into a sequence of *input states* $\mathbf{h}_1, \ldots, \mathbf{h}_N$:

$$\mathbf{h}_n = \text{enc}(\mathbf{h}_{n-1}, \mathbf{x}_n).$$

The attention transforms the input states into a sequence of *context vectors* via a matrix of *attention weights*:

$$\mathbf{c}_m = \sum_n \alpha_{mn} \mathbf{h}_n.$$

[3]For simplicity, we have assumed only a single layer for both the encoder and decoder. It is possible to use multiple stacked RNNs; typically, the output of the encoder and decoder ($\mathbf{c}_m$ and $P(\mathbf{y}_m)$, respectively) would be computed from the top layer only.

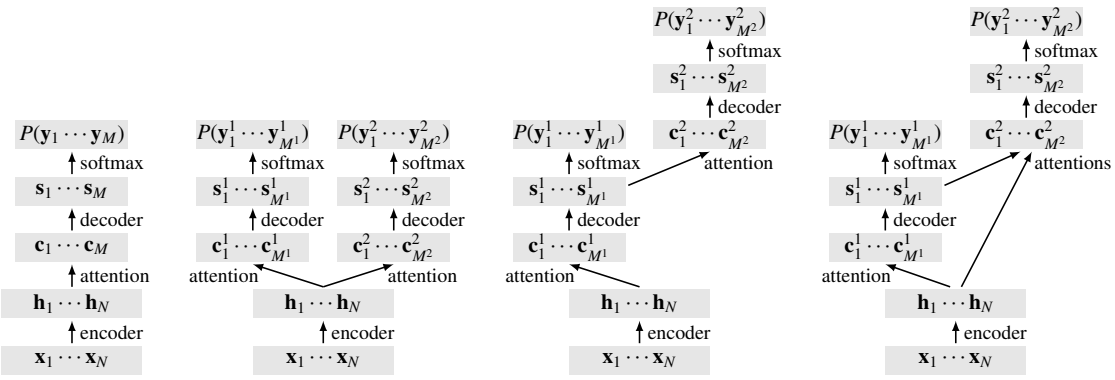


Figure 4.5: Target-side variations on the standard attentional model. In the standard *single-task* model, the decoder attends to the encoder's states. In a typical *multitask* setup, two decoders attend to the encoder's states. In the *cascade* (Tu et al., 2017), the second decoder attends to the first decoder's states. In our proposed *triangle* model, the second decoder attends to both the encoder's states and the first decoder's states. Note that for clarity's sake there are dependencies not shown.

Finally, the decoder computes a sequence of *output states* from which a probability distribution over output words can be computed.

$$\mathbf{s}_m = \text{dec}(\mathbf{s}_{m-1}, \mathbf{c}_m, \mathbf{y}_{m-1})$$

$$P(\mathbf{y}_m) = \text{softmax}(\mathbf{s}_m).$$

In a standard encoder-decoder *multitask* model (Figure 4.5b) (Dong et al., 2015; Weiss et al., 2017), we jointly model two output sequences using a shared encoder, but separate attentions and decoders:

$$\mathbf{c}^1_m = \sum_n \alpha^1_{mn}\mathbf{h}_n \qquad \mathbf{c}^2_m = \sum_n \alpha^2_{mn}\mathbf{h}_n$$
$$\mathbf{s}^1_m = \text{dec}^1(\mathbf{s}^1_{m-1}, \mathbf{c}^1_m, \mathbf{y}^1_{m-1}) \qquad \mathbf{s}^2_m = \text{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1})$$
$$P(\mathbf{y}^1_m) = \text{softmax}(\mathbf{s}^1_m) \qquad P(\mathbf{y}^2_m) = \text{softmax}(\mathbf{s}^2_m).$$

We can also arrange the decoders in a *cascade* (Figure 4.5c), in which the second decoder attends only to the output states of the first decoder:

$$\mathbf{c}^2_m = \sum_{m'} \alpha^{12}_{mm'}\mathbf{s}^1_{m'}$$
$$\mathbf{s}^2_m = \text{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1})$$
$$P(\mathbf{y}^2_m) = \text{softmax}(\mathbf{s}^2_m).$$

Tu et al. (2017) use exactly this architecture to train on bitext by setting the second output sequence to be equal to the input sequence ($\mathbf{y}^2_i = \mathbf{x}_i$).

In our proposed *triangle* model (Figure 4.5d), the first decoder is as above, but the second decoder has two attentions, one for the input states of the encoder and one for the output states of the first decoder:

$$\mathbf{c}^2_m = \left[\textstyle\sum_{m'} \alpha^{12}_{mm'}\mathbf{s}^1_{m'} \quad \sum_n \alpha^2_{mn}\mathbf{h}_n\right]$$
$$\mathbf{s}^2_m = \text{dec}^2(\mathbf{s}^2_{m-1}, \mathbf{c}^2_m, \mathbf{y}^2_{m-1})$$
$$P(\mathbf{y}^2_m) = \text{softmax}(\mathbf{s}^2_m).$$

4.3.4 Learning and Inference

For compactness, we will write $\mathbf{X}$ for the matrix whose rows are the $\mathbf{x}_n$, and similarly $\mathbf{H}$, $\mathbf{C}$, and so on. We also write $\mathbf{A}$ for the matrix of attention weights: $[\mathbf{A}]_{ij} = \alpha_{ij}$. Let θ be the parameters of our model, which we train on sentence triples $(\mathbf{X}, \mathbf{Y}^1, \mathbf{Y}^2)$.

Maximum likelihood estimation Define the score of a sentence triple to be a log-linear interpolation of the two decoders' probabilities:

$$\text{score}(\mathbf{Y}^1, \mathbf{Y}^2 \mid \mathbf{X}; \theta) = \lambda \log P(\mathbf{Y}^1 \mid \mathbf{X}; \theta) + (1 - \lambda) \log P(\mathbf{Y}^2 \mid \mathbf{X}, \mathbf{S}^1; \theta)$$

where λ is a parameter that controls the importance of each sub-task. In all our experiments, we set λ to 0.5. We then train the model to maximize the score over all sentence triples in the training data:

$$\mathcal{L}(\theta) = \sum \text{score}(\mathbf{Y}^1, \mathbf{Y}^2 \mid \mathbf{X}; \theta).$$

Regularization We can optionally add a regularization term to the objective function, in order to encourage our attention mechanisms to conform to two intuitive principles of machine translation: *transitivity* and *invertibility*.

***Transitivity* attention regularizer** To a first approximation, the translation relation should be transitive (Levinboim and Chiang, 2015; Wang et al., 2006): If source word $\mathbf{x}_i$ aligns to target word $\mathbf{y}^1_j$ and $\mathbf{y}^1_j$ aligns to target word $\mathbf{y}^2_k$, then $\mathbf{x}_i$ should also probably align to $\mathbf{y}^2_k$. To encourage the model to preserve this relationship, we add the following *transitivity* regularizer to the loss function of the *triangle* models with a small weight $\lambda_{\text{trans}} = 0.2$:

$$\mathcal{L}_{\text{trans}} = \text{score}(\mathbf{Y}^1, \mathbf{Y}^2) - \lambda_{\text{trans}} \left\| \mathbf{A}^{12}\mathbf{A}^1 - \mathbf{A}^2 \right\|_2^2.$$

***Invertibility* attention regularizer** The translation relation also ought to be roughly invertible (Levinboim et al., 2015): if, in the *reconstruction* version of the *cascade* model, source word $\mathbf{x}_i$ aligns to target word $\mathbf{y}^1_j$, then it stands to reason that $\mathbf{y}_j$ is likely to align to $\mathbf{x}_i$. So, whereas Tu et al. (2017) let the attentions of the translator and the reconstructor be unrelated, we try adding the following *invertibility* regularizer to encourage the attentions to each be the inverse of the other, again with a weight $\lambda_{\text{inv}} = 0.2$:

$$\mathcal{L}_{\text{inv}} = \text{score}(\mathbf{Y}^1, \mathbf{Y}^2) - \lambda_{\text{inv}} \left\| \mathbf{A}^1 \mathbf{A}^{12} - \mathbf{I} \right\|_2^2.$$

Decoding Since we have two decoders, we now need to employ a two-phase beam search, following Tu et al. (2017):

1. The *first decoder* produces, through standard beam search, a set of triples each consisting of a candidate transcription $\hat{\mathbf{Y}}^1$, a score $P(\hat{\mathbf{Y}}^1)$, and a hidden state sequence $\hat{\mathbf{S}}$.
2. For each transcription candidate from the *first decoder*, the *second decoder* now produces through beam search a set of candidate translations $\hat{\mathbf{Y}}^2$, each with a score $P(\hat{\mathbf{Y}}^2)$.
3. We then output the combination that yields the highest total score$(\mathbf{Y}^1, \mathbf{Y}^2)$.

Implementation All our models are implemented in DyNet (Neubig et al., 2017).[4] We use a dropout of 0.2, and train using Adam with initial learning rate of 0.0002 for a maximum of 500 epochs. For testing, we select the model with the best performance on dev. At inference time, we use a beam size of 4 for each decoder (due to GPU memory constraints), and the beam scores include length normalization (Wu et al., 2016) with a weight of 0.8, which Nguyen and Chiang (2017) found to work well for low-resource NMT.

[4] Available online at `https://bitbucket.org/antonis/dynet-multitask-models`

TABLE 4.3

THE MULTITASK MODELS OUTPERFORM THE BASELINE SINGLE-TASK MODEL AND THE PIVOT APPROACH (AUTO/TEXT) ON ALL LANGUAGE PAIRS TESTED AND THE *TRIANGLE* MODEL ALSO OUTPERFORMS THE SIMPLE MULTITASK MODELS ON BOTH TASKS IN ALMOST ALL CASES

	Model		Search		Mboshi	French	Ainu	English	Spanish	English
	ASR	MT	ASR	MT	CER	BLEU	CER	BLEU	CER	BLEU
1	auto	text	1-best	1-best	42.3	21.4	44.0	16.4	70.2	24.2
2	gold	text	—	1-best	0.0	31.2	0.0	19.3	0.0	51.3
3	single-task		1-best		—	20.8	—	12.0	—	21.6
4	multitask		4-best	1-best	36.9	21.0	40.1	18.3	**57.4**	26.0
5	triangle		4-best	1-best	32.5	22.0	39.9	19.2	58.9	**28.6**
6	tr.+$\mathcal{L}_{\text{trans}}$		4-best	1-best	33.1	**23.4**	43.3	**20.2**	59.3	**28.6**
7	triangle		1-best	1-best	**31.9**	17.4	**38.9**	**19.8**	58.4	**28.8**
8	tr.+$\mathcal{L}_{\text{trans}}$		1-best	1-best	32.3	19.3	43.0	**20.3**	59.1	**28.5**

4.3.5 Experiments and Results

In Table 4.3, we present results on three small datasets that demonstrate the efficacy of our models. We compare our proposed models against three baselines and one "skyline." The first baseline is a traditional pivot approach (line 1), where the ASR output, a sequence of characters, is the input to a character-based NMT system (trained on gold transcriptions). The "skyline" model (line 2) is the same NMT system, but tested on gold transcriptions instead of ASR output. The second baseline is translation directly from source speech to target text (line 3). The last baseline is the standard *multitask* model (line 4), which is

similar to the model of Weiss et al. (2017).

On all three datasets, the *triangle* model (lines 5, 6) outperforms all baselines, including the standard *multitask* model. On Ainu-English, we even obtain translations that are comparable to the "skyline" model, which is tested on gold Ainu transcriptions.

Comparing the performance of all models across the three datasets, there are two notable trends that verify common intuitions regarding the speech transcription and translation tasks. First, an increase in the number of speakers hurts the performance of the speech transcription tasks. The character error rates for Ainu are smaller than the CER in Mboshi, which in turn are smaller than the CER in CALLHOME. Second, the character-level BLEU scores increase as the amount of training data increases, with our smallest dataset (Ainu) having the lowest BLEU scores, and the largest dataset (CALLHOME) having the highest BLEU scores. This is expected, as more training data means that the translation decoder learns a more informed character-level language model for the target language. (Note that Weiss et al. (2017) report much higher BLEU scores on CALLHOME: our model underperforms theirs by almost 5 *word-level* BLEU points. However, their model has significantly more parameters and is trained on 10 times more data than ours. Such an amount of data would never be available in our endangered languages scenario.)

To evaluate the effect of using the combined score from both decoders at decoding time, we evaluated the *triangle* models using only the 1-best output from the speech model (lines 7, 8). One would expect that this would favor speech at the expense of translation. In transcription accuracy, we indeed observed improvements across the board. In translation accuracy, we observed a surprisingly large drop on Mboshi-French, but little effect on the other language pairs – in fact, BLEU scores tended to go up slightly, but not significantly.

CHAPTER 5

CROSS-LINGUAL MORPHOSYNTACTIC ANALYSIS ON AN ENDANGERED LANGUAGE

The previous sections dealt with the basic set of annotations that render a corpus interpretable: alignments, transcriptions, and translations. After collecting these, linguistic research requires additional levels of annotation that highlight specific phenomena. We suggest that translation information can be leveraged to computationally assist this analysis. This chapter presents work that confirms this suggestion in the setting of Griko, an endangered language spoken in South Italy.

The next section §5.1 described how we collaborated with linguists in order to collect a new parallel resousrce in Griko and Italian. We built Part-of-Speech taggers to produce annotations on the data, which were then updated in an active learning scenario as the linguists corrected our automatic annotations.

5.1 POS-tagging on an Endangered Language: a parallel Griko-Italian resource

Abstract: Most work on part-of-speech (POS) tagging is focused on high resource languages, or examines low-resource and active learning settings through simulated studies. We evaluate POS tagging techniques on an actual endangered language, Griko. We present a resource that contains 114 narratives in Griko, along with sentence-level translations in Italian, and provides gold annotations for the test set. Based on a previously collected small corpus, we investigate several traditional methods, as well as methods that take advantage of monolingual data or project cross-lingual POS tags. We show that the combination of a semi-supervised method with cross-lingual transfer is more appropriate for this extremely challenging setting, with the best tagger achieving an accuracy of 72.9%. With an applied active learning scheme, which we use to collect sentence-level annotations over the test set, we achieve improvements of more than 21 percentage points.

5.1.1 The Griko Language

Griko is a Greek dialect spoken in southern Italy, in the Grecìa Salentina area southeast of Lecce.[1] There is another endangered Italo-Greek variety in southern Italy spoken in the region of Calabria, known as Grecanico or Greco. Both languages, jointly referred to as *Italiot Greek*, were included as seriously endangered in the UNESCO *Red Book of Endangered Languages* in 1999.

Griko is only partially intelligible with modern Greek, and unlike other Greek dialects, it uses the Latin alphabet. Less than 20,000 people (mostly people over 60 years old) are believed to be native speakers (Douri and De Santis, 2015; Horrocks, 2009), a number which is quite likely an overestimation (Chatzikyriakidis, 2010).

5.1.2 Background

Naturally, textual resources of an endangered language are very hard to find, let alone in the form of parallel text in another language. Any available data is usually the result of

[1] A discussion on the possible origins of Griko can be found in the paper by Manolessou (2005).

documentation efforts by linguists, but rarely are all collected resources properly annotated, as we discussed earlier (§3.4).

For example, resources on Griko are very scarce. Other than the small corpus (Lekakou et al., 2013) of 330 spoken utterances, annotated with transcriptions, morphosyntactic tags, and glossing in Italian, no other resources exist online, at least not in a form suitable for traditional or computational linguistics research. The digital footprint of the language only includes a few websites. One of the websites[2] presents, among others, some narratives in Griko, also translated in Italian.

Part-of-Speech Tagging is a very well studied problem; probabilistic models like Hidde Markov Models and Conditional Random Fields were initially proposed (Lafferty et al., 2001), with neural network approaches taking over in the last years (Huang et al., 2015; Mikolov et al., 2010). Rarely are such methods applied on low-resource languages, due mostly to the lack of annotated data. To our knowledge, no other previous work has been tested on an actual endangered language.

The lack of high quality annotated data lead to approaches that attempt to use minimal such resources. Garrette and Baldridge (2013) used about 200 annotated sentences along with monolingual corpora improving the accuracy of an HMM-based model.

The use of parallel data for projecting POS tag information across languages was introduced by Yarowsky and Ngai (2001), and further improved at a large scale by Das and Petrov (2011) who used graph-based label propagation to expand the coverage of labelled tokens. Täckström et al. (2013) used high-quality alignments to construct type and token level dictionaries. Zhang et al. (2016) used only a few word translations in order to train cross-lingual word embeddings, using them in an unsupervised setting. Fang and Cohn (2017), on the other hand, used parallel dictionaries of 20k entries along with 20 annotated sentences. They use the cross-lingual embeddings when training a tagger in a high resource language and using it to tag monolingual corpora in the low-resource language, which are

[2] `https://www.ciuricepedi.it`

TABLE 5.1

STATISTICS ON OUR COLLECTED GRIKO-ITALIAN RESOURCE

	Stories	Sentences	Griko		Italian	
			Types	Tokens	Types	Tokens
train	104	9.2k	13.5k	197.6k	10.6k	169.7k
test	10	885	2.4k	14.0k	2.3k	13.1k
all	114	10.1k	14.1k	211.6k	11.0k	182.7k

in turn used as distant supervision for the transferred neural model.

5.1.3 Resource

Resources in Griko are very scarce. The German scholar Gerhard Rohlfs pioneered research on Griko and composed the first grammar of the language (Rohlfs, 1977), also heavily influencing the subsequent grammar created by Karanastasis (1997). Although the language has been further studied, almost no corpora are available for linguistics research.

The only Griko corpus available online[3] (Lekakou et al., 2013) consists of about 20 minutes of speech in Griko, along with text translations into Italian. The corpus (henceforth `UoI` corpus, as it is hosted at the University of Ioannina, Greece) consists of 330 mostly elicited utterances by nine native speakers, annotated with transcriptions, morphosyntactic tags, and glossing in Italian.

The most noted Griko scholar is Vito Domenico Palumbo (1854–1928) who made the first serious attempts to create a literary Griko for the dialect of Calimera (the most populous of the nine remaining communities where Griko is still spoken), based on modified

[3] `http://griko.project.uoi.gr`

Italian orthography. Salvatore Tommasi and Salvatore Sicuro then edited and published Palumbo's manuscripts (Palumbo, 1998, 1999), a part of which we now make available for computational and linguistic research.

After scraping from their website[4] 114 narratives that Palumbo had collected, along with their Italian translations, we removed all HTML markup and normalized the orthography: we substituted all curly quotes and apostrophes with simple ones, and substituted the vowels with circumflex (â, ô, û) that were used in a few contractions with the more common accented vowel–apostrophe combination (à', ò' ù'). Using the Moses tools (Koehn et al., 2007) with the Italian settings, we lowercased and tokenized our parallel dataset. For completeness purposes, we also make available the untokenized and proper-case versions of the corpus. The statistics of the resource are shown in Table 5.1.

We chose the first 10 narratives to be our test set, as they correspond to about 10% of all sentences. The rest of the narratives are treated as a monolingual or parallel resource to be leveraged. The test set was, in addition, hand-annotated by linguists: they corrected any tokenization errors that were introduced by the automatic process (for example, regarding the use of the apostrophe) and produced POS tags for every test sentence.

For every narrative, one of the linguists was presented with the produced output of the tagger, and proceeded to correct it. In order to ensure the quality of the annotations, a second linguist was then presented with the result of the work of the first linguist and tasked with correcting it, until all disagreements were resolved. Although it significantly slowed down the annotation process, we hope that this scheme ensured the quality of our annotations.

5.1.4 Differences from Previous Griko Resources

Orthography Griko has never had a consistent orthography. The transcriptions in the `UoI` corpus are based on orthographic conventions found in the few textual resources such

[4] `https://www.ciuricepedi.it`

TABLE 5.2

LIST OF TAGS AND THEIR FREQUENCY IN THE ANNOTATED TEST PART OF THE CORPUS

tag	freq	tag	freq	tag	freq
V *(verb)*	24.4	Prt *(particle)*	2.2	Adv+Adv	0.4
PUNCT *(punctuation)*	18.3	P+D	1.8	X *(other)*	0.3
Pr *(pronoun)*	12.5	P *(adposition)*	1.6	V+Pr	0.3
N *(noun)*	11.6	Adj *(adjective)*	1.2	P+P	0.1
C *(complementizer)*	11.4	Num *(numeral)*	0.7	Pr+Pr	0.1
D *(determiner)*	7.2	N+Pr	0.6	C+Pr	0.1
Adv *(adverb)*	5.0	V+C	0.4		
Adv+P, Adv+Pr, Adv+Prt, Adj+Pr, Prt+N, Prt+Pt, D+N, Adv+Adv+Prt, Adv+Adv+Pr					< 0.1

as the local magazine *Spìtta*, that closely follow conventions adopted in Italian, aiming to be familiar to the speakers of the language. This non-standardization of the orthography leads to variations in the transcription of the same words.

In addition, we find that the word segmentation in our collected narratives follows more the concept of a phonological word. As a result, words that are segmented in the `UoI` corpus, in our narratives are often fused in a single token. The most common case that also appears in both Italian and Greek, is the contraction of prepositions and subsequent articles, such as the Italian *alla* or the Greek *στη (sti)* 'to the.Fem'. Other examples of word fusion that is not permitted in either Italian or Greek but appear in our narratives are nouns and possessive pronouns, or adverbs with other adverbs or prepositions. A direct result of this phenomenon is that annotating such tokens with single POS tags does not capture all

of the necessary information.

Therefore, we chose to annotate such words with multiple POS-tags, effectively making our tag dictionary the superset of the universal tagset. The final tags that appear in practice in our corpus, and their respective frequencies, are listed in Table 5.2. Examples of fused words and their glosses and associated tags are shown in Table 5.3.

Phonosyntactic Gemination One important difference is that the `UoI` corpus explicitly annotates the phenomenon of *raddoppiamento fonosintattico* (phonosyntactic gemination, or doubling of the initial consonant of the word in certain contexts) with a hyphen that separates the two words. The transcriptions that we collected do not mark for this phenomenon. The two words are often fused into a single token, and the doubling is not always present. For example, both following types appear in our corpus: *aderfòmmu* and *aderfòmu* 'my brother'.

Furthermore, the `UoI` corpus also uses apostrophes to mark word boundaries within which the *raddoppiamento fonosintattico* takes place. The use of apostrophes in our collected narratives is more loose. They are used both to mark elision/apocope, stress, as well as what it seems to be instances of raddoppiamento fonosintattico. This poses further issues that are discussed in the next paragraph.

Code Switching There are three languages present in the region of Salento: the regional variety of Italian, the Italo-Romance dialect of Salentino, and Griko.[5] In modern day all members of the Griko community are bilingual or trilingual. The generations before the Second World War are considered to have been predominantly monolingual, and our narratives were collected at that time, around the beginning of the 20th century. However, elements of Salentino do appear in the narratives, either as passing words, or as full sentences, mostly in dialogue turns. Note that resources on Salentino are also extremely scarce

[5] See (Golovko and Panov, 2013) for a broader overview of the linguistic diversity in the Salento area.

TABLE 5.3

EXAMPLES OF FUSED TYPES THAT RECEIVE MULTIPLE TAGS IN OUR ANNOTATION

word:	*stì*	*mànassu*	*cikau*	*ènna*	*vàleti*
morphemes:	*s[e]-tì*	*màna-su*	*ci-kau*	*è-na*	*vàle-ti*
POS tag:	P+D	N+Pr	Adv+Adv	V+C	V+Pr
gloss:	to-the.Fem.SG	mother-your.SG	there-down	have-COMP	put-her
translation:	‘to the’	‘your mother’	‘down there’	‘will’	‘put her’

if not non-existent.

In order to deal with such examples, we decided to distinguish two scenarios. Tag switching or intra-sentential switching instances were fully annotated. So, any Salentino words or phrases that appear *within* a Griko sentence, are used for training and evaluation. However, in the few cases where we encounter full sentences in Salentino, we opt to not use them for training or evaluation. Such sentences are marked with distinctive tags in the released corpus. Note that the `UoI` corpus does not include any non-Griko words or phrases. An extensive study of the code-switching phenomena that occur in our corpus is left for future work.

The following is an example of usage of a Salentino phrase (italicized) within a Griko sentence, taken from story 4. Note that there exists a Griko word for ‘olive oil’, namely *alài* or *alàdi*, as well words for ‘good’, namely *kalòn* or *brao*. However, the Salentino phrase *oju finu* ‘fine oil’ is chosen:

leo	ti	vastò	*oju*	*finu*
say-1SG	COMP	hold-1SG	oil	fine
V	C	V	N	Adj

'I say that I have good olive oil'

Tokenization The `UoI` corpus has been carefully crafted to make sure that word boundaries are clearly denoted by spaces or hyphens. This unfortunately is not the case in our collected narratives. The "loose" use of apostrophes complicates the work of the tokenizer. We chose to tokenize all apostrophes as a single token, except for the cases of known elisions that were present in previous corpora, such as the case of the conjunction *c' (ce)* 'and'. In addition, in the manually annotated test set, the linguists corrected any clear tokenization issues regarding the apostrophe.

Stress Marking In the `UoI` corpus, all words with two or more syllables have a diacritic mark to indicate the location of stress. However, the resources that we collected are not consistent in the use of such a diacritic. Its use is, besides, not standardized and not well studied. Although in most cases such a diacritic is used, there are several instances of polysyllabic words that have no stress marks.

Metadata We further provide as much information as possible for each narrative, in the form of metadata. This includes the original url of the narrative, the title of the narrative in Griko and its translation in Italian. Whenever they were reported (more than 95% of the narratives) we include the location where the narrative was collected, and we anticipate that further analysis could possibly reveal any regional variations. The vast majority of the stories were naturally collected in Calimera, the largest village and the center of the Griko community, but the resource also includes 10 stories collected in Martano, as well as stories collected in Corigliano and Martignano, two smaller villages. We also include information about the date that a story was collected, as well as the narrator of the story. There are

a total of 37 different narrators, while the 10 stories from Martano were retrieved from anonymous manuscripts. There are also 11 stories where the narrator is not known. Two thirds of the stories were narrated by women, while 15% of the narrators were male. The oldest manuscript dates back to 1883, while the most recent story was collected in 1998. We hope that this additional information will further allow us to investigate morphosyntactic phenomena in relation to their temporal or location context, but this is left as future work.

5.1.5 Part-of-Speech Tagging

First, we construct a mapping of the tags of the `UoI` corpus to the Universal Part-of-Speech tagset (Petrov et al., 2012). This mapping is available as part of the complementary material of our resource.

Starting with the tagged `UoI` corpus, we can use several methods to train a tagger, which we use as baselines. We use the Stanford Log-linear POS-tagger (Toutanova et al., 2003) (henceforth `stanford`), trained and tested with the default settings. We also test a simple feature-based CRF tagger (henceforth `crf`), using the implementation of the `nltk` toolkit (Bird and Loper, 2004). We extended the implementation to also use prefix and suffix features of up to 4 characters, along with bigram and trigram features.[6] We will refer to this method as `crf-mod`.

Finally, we also investigate the use of a simple neural model. It uses a single bi-LSTM layer to encode the input sentence, and it outputs tags after a fully connected layer applied on the output of the recurrent encoder, as was described in Lample et al. (2016). The model is implemented in DyNet (Neubig et al., 2017), with input embedding and hidden sizes of 128, and output (tag) embedding size of 32. It is trained with the Adam optimizer with an initial learning rate of 0.0002 and for a maximum of 50 epochs. We select the best model based on the performance on a small dev set of 40 sentences that we sampled randomly from the training set.

[6] Our extensions will be submitted to the `nltk` codebase.

TABLE 5.4

THE BEST PERFORMING MODEL IS THE ONE THAT COMBINES SEMI-SUPERVISED LEARNING WITH CROSS-LINGUAL PROJECTED TAGS (G&B+CLP), WITH ALL MODELS EXCEPT FOR CRF-MOD BENEFITING FROM TRANSFER LEARNING THROUGH ALIGNMENTS (+CLP)

Model	Data		
	no transduction		*transduction*
	UoI	+clp	+clp-all
stanford	62.90	67.10	67.11
crf	57.79	59.12	59.26
crf-mod	67.52	62.89	66.50
neural	45.27	53.24	58.50
	UoI+mono	+clp	+clp-all
G&B	71.67	**72.92**	72.07

The tagging performance of all methods is shown in the first column of Table 5.4. We find that the crf-mod model is the best baseline model. With such few data to train on, both the crf and the neural model do not perform well. The bi- and tri-gram features that the crf-mod model uses are very sparse, while the neural model has to deal with a very large number of unknown words, as discussed below in the Analysis subsection.

In line with previous work, we find that semi-supervised training achieves better results in such low-resource settings. We exploit all the narratives that we collected by treating them as an additional monolingual corpus, used in the framework proposed by Gar-

rette and Baldridge (2013). This approach (henceforth G&B) significantly improves upon all baselines, achieving an accuracy of 71.67% in the test set, an improvement of more than 4 percentage points.

Cross-lingual projected tags So far, our results have not used the Italian translations of our resource. We can follow a procedure similar to the one of Täckström et al. (2013), and extract word alignments from the Griko-Italian parallel data of the training set. We use a pre-trained Italian tagger[7] in order to tag the Italian side, and we map those tags to the universal tagset. We can then project the tags of the Italian tokens to the aligned Griko ones.[8] For the cross-lingual projected tags, we found that in practice type-level predictions work better, and thus we only report results with such models. The tags of the Italian side of our resource, the Griko-Italian alignments, and the cross-lingual POS projections on Griko types are available through the complementary material of our resource.

Augmenting the training set with the type-level projected tags (clp in Table 5.4), we achieve improvements for all models, except crf-mod. The crf-mod method uses sparser features and is more prone to errors due to the noise of the projections. The best performance is achieved when we combine the projected tags, as type-level supervision, with the G&B method that leverages monolingual data. Their combination achieves the best overall performance, with an accuracy of 72.9%, a significant improvement over all other methods. As far as we know, this is the first time that cross-lingual projected tags are combined with the method of Garrette and Baldridge (2013).

Transduction An additional approach that needs to be studied is the transductive approach. Since we have translations both for the training and the test sets, we can extract word alignments and project POS tags also for the test set. The results of the transductive

[7] http://elearning.unistrapg.it/TreeTaggerWeb/TreeTagger.html

[8] The type-level projections are also provided with the Supplementary Material.

approach using cross-lingual projected tags from all the data that we have are shown in the third column of Table 5.4 (under `clp-all`).

We find that most methods benefit from the transductive approach, with the `stanford` and `crf` methods exhibiting minimal improvements, while the `neural` method improves significantly by about 5 percentage points as now there are even less out-of-vocabulary words in the input. The `crf-mod` method improves over the `UoI+clp` version, but still does not surpass the `UoI` only version. The only method that does not benefit from the transduction setting is the `G&B` method, where the performance drops.

An additional transductive step that can be taken with the `G&B` method is to also add the test set as part of the monolingual data that it uses. However, including the test set in the monolingual data also resulted in a drop in performance. Using all monolingual data along with the train-only cross-lingual types (`clp`) leads to accuracy around 69.9% (a drop of 3 points from the best model), while using all monolingual data with `clp-all` leads to a drop of another 1.4 points, to an accuracy of only 68.5%, which however is still better than all other taggers. These accuracy drops are probably justifiable, since `G&B` was not developed under a transductive assumption.

Analysis It is worth noting that our choice of using combined tags for fused/contracted words means that our training sets, under all settings, do not contain all tags that we encounter in the test set. The tagset of the `UoI` corpus only had 14 tags (the 12 universal ones plus `P+D` and `C+Pr`), indicative of its small size. As more narratives were annotated, the size of the necessary tagset increased to the final 29. However, the additional tags that we had to use are rather rare and do not severely affect the performance of our models. The tags that are present in the `UoI` corpus in fact account for 96.7% of all target tags in the test set, a value that could be considered as a skyline for all methods.

The explanation of our models' performance lies in vocabulary coverage. The `UoI` corpus only includes 46.6% of the test set tokens (8.9% of the test set types). The augmented

training set with type-level projections increases those numbers to 48.7% of test tokens and 14.8% of test types. Even though we restrict ourselves to high quality alignments,[9] we are able to project tags to 3870 types (3911 in the transductive scenario), an amount higher than the amount of tags that a trained linguist can produce within four hours of annotation (Garrette et al., 2013).

The G&B method deals with the vocabulary coverage issue by introducing a tag dictionary expansion as a first step. They use a label propagation algorithm —Modified Adsorption (Talukdar and Crammer, 2009)— in order to spread labels between related items. In our framework, the cross-lingual projected tags provide labels for a subset of the types, in a way similar that an annotator would, partially alleviating the difficulty of the method's first step. This leads to less noise in the created tag dictionary, leading to increased accuracy. Note that, out of the cross-lingual projected tags that correspond to types that appear in the test set (about 10% of the test set types, in the *no transduction* setting), more than 65% were correctly projected.

5.1.6 Active Learning

We further explored the use of active learning while tagging our test set. Our active learning scheme is as follows: We first sorted the test set narratives according to length, and starting only with the UoI corpus, we trained all taggers, producing annotations for the first story of the test set. After the corrections on the annotation of each narrative were completed, it was added as gold training data and the taggers were re-trained. For each subsequent story, the linguists were provided with the output of the tagger that achieved the highest accuracy in the previous iteration.

The main reason why we decided to follow this narrative-level active learning scheme instead of collecting type-level annotations is that a noisy corpus is not very helpful for

[9] An alignment is used if either its probability is 1, or its probability is higher than 0.9 *and* the frequency of both tokens is higher than 5. Relaxing those conditions leads to worse performance due to noise.

TABLE 5.5

TAGGING ACCURACY FOR EACH TEST NARRATIVE WITH AND WITHOUT ACTIVE LEARNING, OBTAINING SIGNIFICANT IMPROVEMENTS (Δ COLUMN) BY ADDING EACH ANNOTATED NARRATIVE TO THE TRAINING SET BEFORE RETRAINING AND TAGGING THE NEXT NARRATIVE

Iter.	Narrative	Best accuracy		Δ	Accuracy	Best method
		no AL	with AL		on story 9	
1	story-1	77.89	—	0.0	78.13	G&B+clp
2	story-8	72.76	78.48	5.72	82.12	
3	story-7	75.07	85.17	10.10	83.57	
4	story-10	70.88	79.98	9.10	85.08	
5	story-5	72.26	82.34	10.08	88.21	crf-mod+clp
6	story-4	74.03	86.30	12.27	90.32	
7	story-3	72.48	89.67	17.19	92.13	crf-mod
8	story-6	74.67	91.80	17.13	93.64	
9	story-2	70.78	92.67	21.89	94.17	
10	story-9	72.97	94.17	21.20	—	

linguistics research; at least some part of the resource should have to be checked for quality and accuracy by hand. In addition, the translations of the narratives can provide such information, as we already showed in the previous section. As we expand the coverage of our POS annotations over the whole corpus, we will explore other methods for selecting the types or sentences to be annotated through an active learning scheme.

The results, per narrative, with and without active learning, in the order that they were

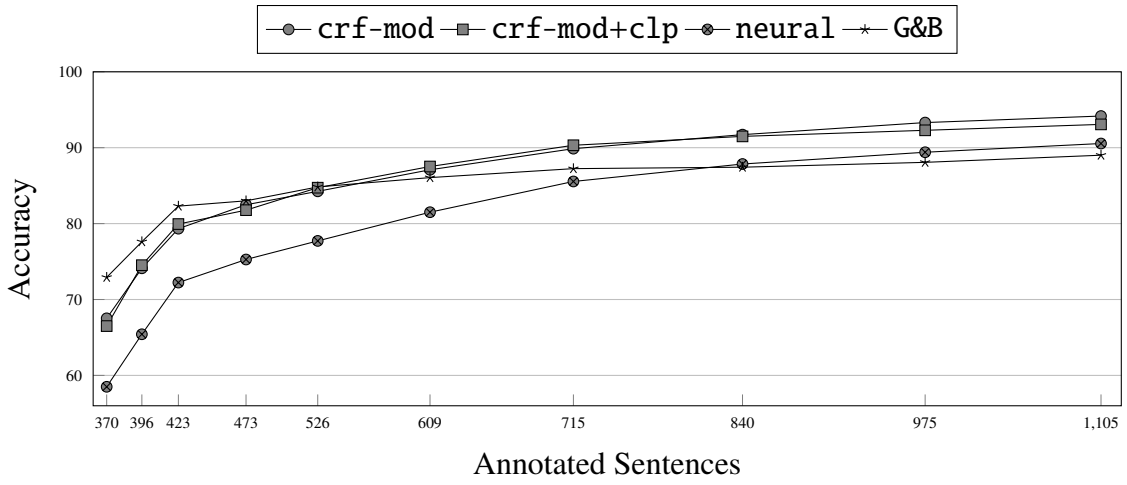


Figure 5.1: Accuracy on the (remaining) test set as we add annotated narratives to the training set. All methods benefit from the active learning approach, with G&B displaying better performance due to its use of monolingual data in the first iterations, but the crf-mod approach achieving the best results in the last iterations, eventually not even needing the cross-lingual type-level projections (+clp).

annotated by our linguists (from the shortest narrative to the longest) are outlined in Table 5.5. The results for each narrative in the active learning scenario ("with AL" column) report the best performing model that is trained on the concatenation of the UoI corpus and all the stories that were annotated in previous iterations. It is clear that the performance of the taggers improved continuously, as we added more training data. This is further outlined by each iteration's tagging accuracy on story 9, the last and longest narrative of the test set. Of course, when a narrative is added in the training set, it is then excluded from the test set, and the performance is reported on the rest of the narratives.

All methods display notable improvement as we added the annotated narratives to the training set. The performance trends are outlined in Figure 5.1. Firstly, it is notable that as the training set increases, the advantage of the model of Garrette and Baldridge (2013) that leverages monolingual data diminishes, compared to our simpler crf-mod tagger, both with or without cross-lingual projected tags. Before the first iteration, the accuracy gap is 6.4 percentage points in favor of G&B. However, after adding around 4-5 narratives so that there are around 500 training sentences, our crf-mod+clp method surpasses the G&B

method and keeps improving. This is also outlined by the dashed line in Table 5.5. As we add more training instances, the accuracy of the G&B method plateaus around 85% and does not improve further.

Furthermore, after a couple more iterations, when more than 800 annotated sentences are available for training, the crf-mod method without cross-lingual projected tags achieves higher accuracy than all others. We identify this point as the one where simple token-level supervision is efficient enough to outperform semi-supervised or transfer-learning approaches.

Finally, we observe that the accuracy of the neural bi-LSTM approach that only uses the tagged corpus without further use of monolingual data, improves significantly as the training set increases. With only 370 training sentences, the gap between the neural and the best method is more than 14 percentage points. With 1,100 training sentences, the accuracy gap diminishes to only 2 percentage points.

5.1.7 Cross-Validation

Our end goal is to annotate the whole corpus with POS tags, as well as richer annotations. Towards that direction, our gold annotated test data could be used to train a higher quality POS tagger, which we will use to annotate the rest of the corpus. In Section §5.1.6, we found that including all but one annotated narratives for training, and testing on the last one (story-9) we were able to obtain an accuracy of more than 94.17%. In order to get a better estimation of how well a tagger trained on our gold data would work, we perform a cross-validation experiment, using crf-mod, our best performing model.

For each cross-validation instance, one of the annotated narratives becomes the test set, and the rest will be included in the training set. This allows us to obtain an average performance over 10 instances. The average accuracy of the crf-mod model is about 91.9%, with a standard deviation of about 2 percentage points (minimum is 88.5% on story 5, and maximum is 94.9% on story 2).

The main obstacle to annotating the rest of the corpus with higher quality is out-of-vocabulary words. The combined vocabulary of the `UoI` corpus and our 10 annotated narratives covers 16% of the vocabulary of the 104 unannotated sentences (but 85% of the total tokens). As part of our future work, we plan to incorporate word-level active learning in our annotation/correction scheme, similar to the approaches proposed by Fang and Cohn (2017).

5.1.8 Conclusion

We presented a parallel corpus of 114 narratives on an endangered language, Griko, with translations in Italian. For now, a test set of 10 narratives is hand-annotated with Part-of-Speech tags, but in the future we will enrich the resource with annotations on the rest of the corpus, as well as with richer syntactic and morphological annotations. We also plan on contributing our corpus to the Universal Dependencies treebanks (Nivre et al., 2016) as Griko is absent from the supported languages.

We extensively evaluated several POS tagging approaches, and found that the method of Garrette and Baldridge (2013) can be combined with cross-lingual type-level projected tags, outperforming all other methods, with an accuracy of 72.9%, when less than 500 sentences are available. As data was added in the training set in an active learning scenario, a simple feature-based CRF approach outperforms all other models, with accuracy improvements of over 21 percentage points and over 94% accuracy on the last narrative. In fact, when more than 800 sentences are available for training, cross-lingual tag projections hurt performance.

The collected annotations from our test set could form the basis for training a high-accuracy POS tagger for Griko, so that we can expand the POS annotations to the rest of our corpus with only a small amount of noise. We aim to explore this direction in our future work, along with other active learning methods that require less human intervention. In addition, we plan to further enrich the annotations of our corpus with morphological tags

similar to the UoI corpus, that will provide even more insight in Griko and its usage. When the full annotations of the corpus are completed, we plan to use statistical methods to study specific phenomena regarding the grammar and syntax of Griko.

Finally, and most importantly, we hope that the release of this corpus will spark further interest for computational approaches applied on endangered languages documentation and on under-represented languages in general.

CHAPTER 6

CONCLUSION

The pace at which endangered languages disappear is extremely fast: about one language every week. In contrast, the process of documenting a language is very slow, requiring trained linguists to devote a lot of time and effort in data collection and, especially, annotation.

In this document, we proposed to assist language documentation efforts by evaluating previous machine learning techniques and by developing new ones, that address the various sub-tasks involved in a documentation pipeline. All of these techniques leverage translations in a high-resource language, that provide a signal useful for those tasks.

We first presented novel contributions that address the problem of aligning speech to its translation (Anastasopoulos et al., 2016; Duong et al., 2016). We also showed that such alignments can be useful for other tasks of language documentation, both earlier in the pipeline, when collecting transcriptions (Anastasopoulos and Chiang, 2017), as well as further down in the pipeline, when annotating already collected data with translation keywords (Anastasopoulos et al., 2017).

Then, we focused on the tasks of transcribing and translating speech of an endangered or low-resource language. We presented novel multi-source and tied multitask neural approaches, that are useful both when translations are available at test time (Anastasopoulos and Chiang, 2018a) and when translations are another target to be produced (Anastasopoulos and Chiang, 2018b). We also showed that translations can be used as privileged information during training, leading to better performance when transcribing speech, even without access to translations at test time.

Finally, we collected a parallel textual resource of 114 narratives in an endangered language, Griko, with translations in Italian (Anastasopoulos et al., 2018). In collaboration with linguists, we produced Part-of-Speech annotations for a test set, correcting the automatic annotations produced by a tagger trained in an active learning schema. We showed that cross-lingual information can be combined with semi-supervised approaches to produce more accurate POS-tagging in an endangered language, while the active learning scenario can further significantly improve the performance.

The combination of our contributions will support the overall goal: to aid the process of documenting an endangered language with computational methods that take advantage of translations.

6.1 Future Directions

Here, we list three directions for future work in computational linguistics, which have the potential to positively affect language documentation.

Tools for Linguists: Throughout this dissertation we showcased new machine learning methods that could be applied in each step of the documentation process. A natural next step is actually building a user-friendly interface that will allow linguists to utilize these methods. Such an interface should make the integration seamless, possibly only presenting the user with suggestions for boundaries, transcriptions, translation, or other annotation. Ideally, an active learning scheme should also be employed, so that the back-end models get updated and improve, as the linguist produces more and more annotations. Another option is equipping the tool with back-end multilingual models, which could produce candidate annotations even for unseen languages, an especially common case in language documentation.

As many researchers have pointed out, e.g. Thieberger (2017), such a tool would be undeniably useful to the field and documentary linguistics community, as it has the potential

to significantly accelerate the pace of language documentation. It is important to note that prototypes of such tools have already been developed (Bettinson and Bird, 2017; Neubig et al., 2018).

Low-Resource Speech Transcription and Translation: Our work has pushed the boundaries of speech transcription and translation when applied on extremely low-resource settings. However, these problems are far from solved. Following our contributions (Anastasopoulos and Chiang, 2018a,b; Anastasopoulos et al., 2016), the field has further advanced. Bansal et al. (2018b) shows that pretraining models similar to ours on a high-resource language leads to better performance. Di Gangi et al. (2018) showed that fine-tuning on clean data has the same positive effect, while Jia et al. (2018) demonstrated that monolingual data can be leveraged, by generating synthetic speech with a text-to-speech system. The interest of the community, as well as the need for further progress, is further reinforced by the organization of shared tasks like the IWSLT shared task on speech translation (Jan et al., 2018). Nevertheless, a lot of work is still needed in order to obtain performance comparable to high-resource settings

Automatic Glossing and Morphological Analysis: The latter stages of documentation consist of creating word- and morpheme-level glosses with annotations. In fact, automatic glossing is an under-studied problem, although it potentially is an easier task than translation: there is an one-to-one correspondence between every source word and its target. Most of advances on automatic glossing focuses on glossing for specific phenomena, using a minimal amount of previously glossed data (e.g. the work of Xia and Lewis (2009); Zamaraeva et al. (2017)). A non-trivial hurdle in tackling automatic glossing is the lack of big collections of interlinear gloss text in many languages, but the community has started to address it with project such as the Online Database of Interlinear Text (ODIN) (Lewis, 2006).

Meanwhile, automated study of morphology is another emerging field that has huge potential for application within the language documentation process. The UniMorph project (Kirov et al., 2018) and the CoNLL shared tasks (Cotterell et al., 2018) provide a continuously updated collection of data, even in low-resource languages, which could be leveraged to train multilingual models.

BIBLIOGRAPHY

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 *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. Near or far, wide range zero-shot cross-lingual dependency parsing. arXiv:1811.00570.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *Proc. ICML*.

Antonios Anastasopoulos, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. 2017. Spoken term discovery for language documentation using translations. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 53–58, Copenhagen, Denmark. Association for Computational Linguistics.

Antonios Anastasopoulos and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Honolulu. Association for Computational Linguistics.

Antonios Anastasopoulos and David Chiang. 2018a. Leveraging translations for speech transcription in low-resource settings. In *Proc. INTERSPEECH*.

Antonios Anastasopoulos and David Chiang. 2018b. Tied multitask learning for neural speech translation. In *Proc. NAACL HLT*, volume 1, pages 82–91.

Antonios Anastasopoulos, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *Proc. EMNLP*.

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource. In *Proc. COLING*. To appear.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015a. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2015b. End-to-end attention-based large vocabulary speech recognition. *CoRR*, abs/1508.04395.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018a. Low-resource speech-to-text translation. arXiv:1803.09164.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018b. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. arXiv:1809.01431.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proc. EACL*.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. arXiv:1802.04200.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *Proc. NIPS Workshop on End-to-end Learning for Speech and Audio Processing*.

Donald J. Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proc. KDD*, pages 359–370.

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 156–164.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February*.

Bruce Birch, Sebastian Drude, Daan Broeder, Peter Withers, and Peter Wittenburg. 2013. Crowdsourcing and apps in the field of linguistics: Potentials and challenges of the coming technology.

Steven Bird. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the Role of Digital Libraries in a Time of Global Change, and 12th International Conference on Asia-Pacific Digital Libraries*, ICADL'10, pages 5–14, Berlin, Heidelberg. Springer-Verlag.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proc. Posters*, pages 125–134.

Steven Bird, David Chiang, Friedel Frowein, Andrea L. Berez, Mark Eby, Florian Hanke, Ryan Shelby, Ashish Vaswani, and Ada Wan. 2013. The International Workshop on Language Preservation: An experiment in text collection and language technology. *Language Documentation and Conservation*.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014a. Collecting bilingual audio in remote indigenous communities. In *Proc. COLING*.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014b. Aikuma: A mobile app for collaborative language documentation. In *Proc. of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proc. ACL*, page 31. Association for Computational Linguistics.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proc. SLTU (Spoken Language Technologies for Under-Resourced Languages)*, volume 81.

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5.

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. 2018. A small griko-italian speech translation corpus. In *Proc. SLTU*.

Marcely Zanon Boito, Alexandre Bérard, Aline Villavicencio, and Laurent Besacier. 2017. Unwritten languages demand attention too! word discovery with encoder-decoder models. In *Pric. ASRU*.

Peter Brown, John Cocke, S Della Pietra, V Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Roossin. 1988. A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 71–76. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Lukáš Burget, Sanjeev Khudanpur, Najim Dehak, Jan Trmal, Reinhold Haeb-Umbach, Graham Neubig, Shinji Watanabe, Daichi Mochihashi, Takahiro Shinozaki, Ming Sun, et al. 2016. Building speech recognition system from untranscribed data report from jhu workshop.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, Kaori Ueki, and John Van Way. 2013. New knowledge: findings from the catalogue of endangered languages (elcat). In *3rd International Conference on Language Documentation and Conservation*.

Rich Caruana. 1998. Multitask learning. In *Learning to learn*, pages 95–133. Springer.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. ICASSP*, pages 4960–4964. IEEE.

Stergios Chatzikyriakidis. 2010. *Clitics in four dialects of Modern Greek: A dynamic account*. Ph.D. thesis, University of London.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent NN: first results. In *Deep Learning and Representation Learning Workshop*.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. 2016. Hierarchical multiscale recurrent neural networks. arXiv:1609.01704.

Bernard Comrie. 2009. *The world's major languages*. Routledge.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. arXiv:1810.07125.

David Crystal. 2000. *Language death*. Cambridge University Press.

Siddharth Dalmia, Xinjian Li, Florian Metze, and Alan W Black. 2018a. Domain robust feature extraction for rapid low resource asr development. arXiv:1807.10984.

Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black. 2018b. Sequence-based multi-lingual low resource speech recognition. arXiv:1802.07420.

Amit Das, Preethi Jyothi, and Mark Hasegawa-Johnson. 2016. Automatic speech recognition using probabilistic transcriptions in swahili, amharic, and dinka. *Proc. Interspeech*, pages 3524–3528.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. ACL*, pages 600–609. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Roberto Dessì, Roldano Cattoni, Matteo Negri, and Marco Turchi. 2018. Fine-tuning on clean data for end-to-end speech translation: Fbk@ iwslt 2018. arXiv:1810.07652.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proc. ACL-IJCNLP*.

Angeliki Douri and Dario De Santis. 2015. Griko and modern Greek in Grecìa Salentina: an overview. *L'Idomeneo*, 2015(19):187–198.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. arXiv:1712.04313.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proc. NAACL HLT*.

Emmanuel Dupoux, Núria Sebastián-Gallés, Eduardo Navarrete, and Sharon Peperkamp. 2008. Persistent stress 'deafness': The case of french learners of spanish. *Cognition*, 106(2):682–706.

Anuvabh Dutt, Denis Pellerin, and Georges Quénot. 2017. Coupled ensembles of neural networks. arXiv:1709.06053.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. NAACL HLT*.

Janek Ebbers, Lukas Drude Jahn Heymann, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. 2017. Hidden markov model variational autoencoder for acoustic unit discovery. In *Proceedings of INTERSPEECH*, volume 2017.

Micha Elsner and Cory Shain. 2017. Speech segmentation with a neural encoder model of working memory. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1080.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proc. ACL*.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910.

Nicholas Evans. 2011. *Dying words: Endangered languages and what they have to tell us*, volume 22. John Wiley & Sons.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proc. ACL*, pages 587–593. Association for Computational Linguistics.

Jan Feyereisl, Suha Kwak, Jeany Son, and Bohyung Han. 2014. Object localization based on structural svm using privileged information. In *Advances in Neural Information Processing Systems*, pages 208–216.

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

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochv'il, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 200–204.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proc. NAACL-HLT*, pages 138–147. Association for Computational Linguistics.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 583–592.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. Lig-aikuma: a mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*.

Thomas Glarner, Benedikt Boenninghoff, Oliver Walter, and Reinhold Haeb-Umbach. 2017. Leveraging text data for word segmentation for underresourced languages. *System*, 9(10):11.

P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, et al. 2017. A very low resource language speech corpus for computational language documentation experiments. arXiv:1710.03501.

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Annie Rialland, et al. 2018. Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Ekaterina Golovko and Vladimir Panov. 2013. Salentino dialect, Griko and regional Italian: Linguistic diversity of Salento. *Working Papers of the Linguistics Circle of the University of Victoria*, 23(1):51.

Alex Graves, Santiago Fernández, and Faustino Gomez. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

Ken Hale. 1992. Language endangerment and the human value of linguistic diversity. *Language*, 68(1):35–42.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv:1412.5567.

David Harwath and James Glass. 2017. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–517. Association for Computational Linguistics.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. arXiv:1804.01452.

John Hatton. 2013. Saymore: Language documentation productivity.

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *J. Acoustical Society of America*, 87(4):1738–1752.

Daniel Hernández-Lobato, Viktoriia Sharmanska, Kristian Kersting, Christoph H Lampert, and Novi Quadrianto. 2014. Mind the nuisance: Gaussian process classification using privileged noise. In *Advances in Neural Information Processing Systems*, pages 837–845.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

Geoffrey Horrocks. 2009. *Greek: A History of the Language and its Speakers*. Wiley-Blackwell.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv:1508.01991.

Hirofumi Inaguma, Xuan Zhang, Zhiqi Wang, Adithya Renduchintala, Shinji Watanabe, and Kevin Duh. 2018. The jhu/kyotou speech translation system for iwslt 2018. In *International Workshop on Spoken Language Translation*.

Niehues Jan, Roldano Cattoni, Stüker Sebastian, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–6.

Aren Jansen, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. In *Proc. INTERSPEECH*.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, et al. 2013. A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8111–8115. IEEE.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2018. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. arXiv:1811.02050.

Robbie Jimerson, Kruthika Simha, Raymond Ptucha, and Emily Prudhommeaux. 2018. Improving ASR Output for Endangered Language Documentation. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 182–186.

Robert 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)*.

Preethi Jyothi and Mark Hasegawa-Johnson. 2015. Transcribing continuous speech using mismatched crowdsourcing. In *Proc. Interspeech*.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2015. Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE Trans. Audio, Speech, and Language Processing*.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. arXiv:1804.06024.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2018. Structured-based curriculum learning for end-to-end english-japanese speech translation. arXiv:1802.06003.

Anastasios Karanastasis. 1997. *Grammatiki ton ellinikon idiomaton tis Kato Italias [Grammar of the Greek dialects of south Italy]*. Akadimia Athinon.

Kazuya Kawakami, Chris Dyer, and Phil Blunsom. 2018. Unsupervised word discovery with segmental neural language models. arXiv:1811.09353.

Santosh Kesiraju, Raghavendra Pappagari, Lucas Ondel, Lukáš Burget, Najim Dehak, Sanjeev Khudanpur, Jan Černockỳ, and Suryakanth V Gangashetty. 2017. Topic identification of spoken documents using unsupervised acoustic unit discovery. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5745–5749. IEEE.

Vahid Khanagha, Khalid Daoudi, Oriol Pont, and Hussein Yahia. 2014. Phonetic segmentation of speech signal using local singularity analysis. *Digital Signal Processing*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *Proc. ICASSP*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. arXiv:1810.11101.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL: Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Michael Krauss. 2007. Classification and terminology for degrees of language endangerment. *Language diversity endangered*, 181:1.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*.

John Lambert, Ozan Sener, and Silvio Savarese. 2018. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. NAACL-HLT*, pages 260–270.

Nala Huiying Lee and John Van Way. 2016. Assessing levels of endangerment in the catalogue of endangered languages (elcat) using the language endangerment index (lei). *Language in Society*, 45(2):271–292.

Marika Lekakou, Valeria Baldiserra, and Antonis Anastasopoulos. 2013. Documentation and analysis of an endangered language: aspects of the grammar of Griko.

Tomer Levinboim and David Chiang. 2015. Multi-task word alignment triangulation for low-resource languages. In *Proc. NAACL HLT*.

Tomer Levinboim, Ashish Vaswani, and David Chiang. 2015. Model invertibility regularization: Sequence alignment with or without parallel data. In *Proc. NAACL HLT*.

M Paul Lewis, Gary F Simons, Charles D Fennig, et al. 2009. *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX.

William D Lewis. 2006. Odin: A model for adapting and enriching legacy infrastructure. In *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, pages 137–137. IEEE.

Xinjian Li, Siddharth Dalmia, David R Mortensen, Florian Metze, and Alan W Black. 2018. Zero-shot learning for speech recognition with universal phonetic model.

M. Liberman. 2006. The problems of scale in language documentation. plenary talk at tlsx texas linguistics society 10: Computational linguistics for less-studied languages.

Chunxi Liu, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-Johnson, and Sanjeev Khudanpur. 2016. Adapting asr for under-resourced languages using mismatched transcriptions. In *Proc. ICASSP*, pages 5840–5844.

Chunxi Liu, Jan Trmal, Matthew Wiesner, Craig Harman, and Sanjeev Khudanpur. 2017. Topic identification for speech without asr. arXiv:1703.07476.

Dan Liu, Junhua Liu, Wu Guo, Shifu Xiong, Zhiqiang Ma, Rui Song, Chongliang Wu, and Quan Liu. 2018. The ustc-nel speech translation system at iwslt 2018. arXiv:1812.02455.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2015. Unifying distillation and privileged information. arXiv:1511.03643.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proc. ICLR*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Andrew Maas, Ziang Xie, Dan Jurafsky, and Andrew Ng. 2015. Lexicon-free conversational speech recognition with neural networks. In *Proc. NAACL HLT*, pages 345–354.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Ioanna Manolessou. 2005. The greek dialects of southern Italy: an overview. *KAMPOS: Cambridge Papers in Modern Greek*, 13:103–125.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *INTERSPEECH*, pages 3177–3180.

Evgeny Matusov, Patrick Wilken, Parnia Bahar, Julian Schamper, Pavel Golik, Albert Zeyer, Joan Albert Silvestre-Cerda, Adria Martınez-Villaronga, Hendrik Pesch, and Jan-Thorsten Peter. 2018. Neural speech translation at apptek. In *International Workshop on Spoken Language Translation*.

Fergus McInnes and Sharon Goldwater. 2011. Unsupervised extraction of recurring words from infant-directed speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

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 and Conservation*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3*, pages 1–10. Association for Computational Linguistics.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. Interspeech*.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore. Association for Computational Linguistics.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2016. Information bottleneck learning using privileged information for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1496–1505.

Armando Muscariello, Guillaume Gravier, and Frédéric Bimbot. 2009. Audio keyword extraction by unsupervised word discovery. In *Proc. INTERSPEECH*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. DyNet: The dynamic neural network toolkit. arXiv:1701.03980.

Graham Neubig, Patrick Littell, Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin, and Yuyan Zhang. 2018. Towards a general-purpose linguistic annotation backend. arXiv:1812.05272.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proc. ACL*.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proc. NAACL HLT*, pages 632–641.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. ICASSP*, volume 1.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource related languages for neural machine translation. In *Proc. IJCNLP*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. LREC*.

Lucas Ondel, Lukáš Burget, and Jan Černockỳ. 2016. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86.

Vito Domenico Palumbo. 1998. *Io' mia fora' - Fiabe e Racconti della Grecìa Salentina [Once upon a time - Fairy Tales and Stories from Grecìa Salentina]*. Calimera (LE): Ghetonìa. A cura di S. Tommasi.

Vito Domenico Palumbo. 1999. *'Itela na su pò - Canti popolari della Grecìa Salentina [I wanted to tell you - Folk songs of Grecìa Salentina]*. Calimera (LE): Ghetonìa. A cura di S. Sicuro.

Alex S. Park and James R. Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Trans. Audio, Speech, and Language Processing*, 16(1):186–197.

Sharon Peperkamp, Emmanuel Dupoux, and Núria Sebastián-Gallés. 1999. Perception of stress by french, spanish, and bilingual subjects. In *Eurospeech*. Citeseer.

François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. LREC*.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620. Association for Computational Linguistics.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan VuliÄ‡. 2018. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus. In *Proc. IWSLT*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Michal Ptaszynski and Yoshio Momouchi. 2012. Part-of-speech tagger for Ainu language based on higher order Hidden Markov Model. *Expert Systems with Applications*, 39(14):11576–11582.

Lawrence R Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs.

Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Gerhard Rohlfs. 1977. *Grammatica storica dei dialetti italogreci (Calabria, Salento) dt. Original [1949–1954] [Historical Grammar of the Italiot Greek dialects (Calabria, Salento)]*. CH Beck.

Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, et al. 2006. The human speechome project. In *Symbol Grounding and Beyond*, pages 192–196. Springer.

Deb K Roy and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.

Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, et al. 2018a. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the" speaking rosetta" jsalt 2017 workshop. arXiv:1802.05092.

Odette Scharenborg, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson. 2017. Building an asr system for a low-research language through the adaptation of a high-resource language asr system: Preliminary results.

Odette Scharenborg, Patrick Ebel, Mark Hasegawa-Johnson, and Najim Dehak. 2018b. Building an ASR System for Mboshi Using A Cross-Language Definition of Acoustic Units Approach. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 167–171.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. ACL*.

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. 2013. Learning to rank using privileged information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 825–832.

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. 2014. Learning to transfer privileged information. arXiv:1410.0389.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association of Computational Linguistics*, 1:255–266.

Man-hung Siu, Herbert Gish, Steve Lowe, and Arthur Chan. 2011. Unsupervised audio patterns discovery using hmm-based self-organized units. In *Twelfth Annual Conference of the International Speech Communication Association*.

SpeechLab. 2007. Speech at cmu, logios lexicon tool. *http://www.speech.cs.cmu.edu/tools/lextool.html*.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 11–14.

Felix Stahlberg, Tim Schlippe, Sue Vogel, and Tanja Schultz. 2012. Word segmentation through cross-lingual word-to-phoneme alignment. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

Umut Sulubacak, Jörg Tiedemann, Aku Rouhe, Stig-Arne Grönroos, and Mikko Kurimo. 2018. The memad submission to the iwslt 2018 speech translation task. arXiv:1810.10320.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer.

Louis Ten Bosch and Bert Cranen. 2007. A computational model for unsupervised word discovery. In *Proc. INTERSPEECH*, pages 1481–1484.

Alexis Michaud Thi-Ngoc-Diep Do and Eric Castelli. 2014. Towards the automatic processing of yongning na (sino-tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *Proc. SLTU*, pages 153–160.

Nick Thieberger. 2017. Ld&c possibilities for the next decade.

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu. 2017. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. In *Proc. Interspeech*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL-HLT*, pages 173–180. Association for Computational Linguistics.

Tasaku Tsunoda. 2017. *Language endangerment and language revitalization: An introduction*. De Gruyter Mouton.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proc. AAAI*.

UNESCO. 2010. *Atlas of the World's Languages in Danger*. United Nations Educational, Scientific and Cultural Organization. 3rd edition.

Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 165–168. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proc. COLING/ACL*, pages 874–881.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. In *Proc. INTERSPEECH*.

Matthew Wiesner, Chunxi Liu, Lucas Ondel, Craig Harman, Vimal Manohar, Jan Trmal, Zhongqiang Huang, Sanjeev Khudanpur, and Najim Dehak. 2018. The jhu speech lorehlt 2017 system: Cross-language transfer for situation-frame detection. arXiv:1802.08731.

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.

Anthony C Woodbury. 2003. Defining documentary linguistics. *Language documentation and description*, 1(1):35–51.

Philip C Woodland, Julian J Odell, Valtcho Valtchev, and Steve J Young. 1994. Large vocabulary continuous speech recognition using htk. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, pages II–125. Ieee.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

Fei Xia and William D Lewis. 2009. Applying nlp technologies to the collection and enrichment of language data on the web to aid linguistic research. In *Proceedings of the*

EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, pages 51–59. Association for Computational Linguistics.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 523–530. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora. In *Proc. NAACL*.

Olga Zamaraeva, František Kratochvíl, Emily M Bender, Fei Xia, and Kristen Howell. 2017. Computational support for finding word classes: A case study of abui. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 130–140.

Daniel Zeman, Jan Haji, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

Thomas Zenkel, Matthias Sperber, Jan Niehues, Markus Müller, Ngoc-Quan Pham, Sebastian Stüker, and Alex Waibel. 2018. Open source toolkit for speech to text translation. *The Prague Bulletin of Mathematical Linguistics*, 111(1):125–135.

Yaodong Zhang and James R Glass. 2010. Towards multi-speaker unsupervised speech pattern discovery. In *Proc. ICASSP*, pages 4366–4369.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proc. NAACL-HLT*, pages 1307–1317. Association for Computational Linguistics.