A small Griko-Italian speech translation corpus
Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, Laurent Besacier

To cite this version:
Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, Laurent Besacier. A small Griko-Italian speech translation corpus. 6th international workshop on spoken language technologies for under-resourced languages(SLTU’18), Aug 2018, New Delhi, India. hal-01962528

HAL Id: hal-01962528
https://hal.archives-ouvertes.fr/hal-01962528
Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A small Griko-Italian speech translation corpus

Marcely Zanon Boito¹, Antonios Anastasopoulos², Marika Lekakou³, Aline Villavicencio⁴,⁵, Laurent Besacier¹

¹Laboratoire d’Informatique de Grenoble, Univ. Grenoble Alpes (UGA), France
²Department of Computer Science and Engineering, Univ. of Notre Dame, USA
³Department of Philology, Univ. of Ioannina, Greece
⁴Institute of Informatics, UFRGS, Brazil
⁵CSEE, University of Essex, UK

contact: marcely.zanon-boito@univ-grenoble-alpes.fr and aanastas@nd.edu

Abstract

This paper presents an extension to a very low-resource parallel corpus collected in an endangered language, Griko, making it useful for computational research. The corpus consists of 330 utterances (about 20 minutes of speech) which have been transcribed and translated in Italian, with annotations for word-level speech-to-transcription and speech-to-translation alignments. The corpus also includes morphosyntactic tags and word-level glosses. Applying an automatic unit discovery method, pseudo-phones were also generated. We detail how the corpus was collected, cleaned and processed, and we illustrate its use on zero-resource tasks by presenting some baseline results for the task of speech-to-translation alignment and unsupervised word discovery. The dataset is available online, aiming to encourage replicability and diversity in computational language documentation experiments.

1. Introduction

For many low-resource and endangered languages, speech data is easier to obtain than textual data. Oral tradition has historically been the main medium for passing cultural knowledge from one generation to the next, and at least 43% of the world’s languages are still unwritten [1]. Traditionally, documentary records of endangered languages are created by highly trained linguists in the field. However, modern technology has the potential to enable creation of much larger-scale (but lower-quality) resources. Recently, proposed frameworks [2, 3] propose collection of bilingual audio, rendering the resource interpretable through translations.

New technologies have been developed to facilitate collection of spoken translations [4] along with speech in an endangered language, and there already exist recent examples of parallel speech collection efforts focused on endangered languages [5, 6, 7]. The translation is usually in a high-resource language that functions as a lingua franca of the area, as it is common for members of an endangered-language community to be bilingual.

Tackling the issue of the possible vanishing of more than 50% of the current spoken languages by the year 2100 [8], the Computational Language Documentation (CLD) field assembles two different communities: linguistics and informatics, proposing challenges [9, 10, 11] and frameworks from speech signal [12, 13, 14]. However, as the interest on CLD approaches grows, it becomes clear the urgent need of more publicly available low-resource corpora to provide replicable evaluation of the proposed methods. We are aware of only a few endangered languages whose corpora are publicly available [15,16].

Our work is part of this effort to share resources, and with this paper we present a corpus on a truly endangered dialect from south Italy, Griko. The corpus has several levels of information (speech, machine extracted pseudo-phones, transcriptions, translations and sentence alignment), and we believe it can be an interesting resource for evaluating documentation techniques on (very) low-resource settings.

In addition, we provide baseline results for two tasks: speech-to-translation alignment and unsupervised word discovery. We encourage the community to challenge these results by using their own techniques. For word discovery, we also provide the gold standard for evaluation following the Track 2 of the Zero Resource Challenge (ZRC) 2017 [11]. These metrics were extensively described in [17,11], and are another important community effort for increasing reproducibility.

This paper is organized as follows: after a quick related work (section 2), the Griko language is presented (section 3). Data processing methodology (section 4) and dataset are then presented. Our baseline systems and results for two tasks (sections 5 and 6) are finally described.

2. Related Work

Unsupervised Word Discovery (UWD) systems operate on unsegmented speech utterances and their goal is to output timestamps delimiting stretches of speech, associated with class labels, corresponding to real words in the language. This task is already considered in the Zero Resource Speech Challenge¹ in a fully unsupervised setting: systems must learn to segment from a collection of raw speech signals only. Here, we investigate a slightly more favorable case where speech utterances are multilingually grounded (using cross-lingual supervision, where a written translation is available for each utterance). In CLD scenarios, this task helps to attenuate the heavy charge on field linguists: the output vocabulary can be used as a first clue of the lexicon present in the language of interest. As a monolingual setup, UWD was previously investigated from text input [13] and from speech [19, 20, 15, 21].

The speech translation problem has been traditionally approached by feeding the output of a speech recognition system into a Machine Translation (MT) system. Speech recognition uncertainty was integrated with MT by using speech output latents as input to translation models [22, 23]. A sequence-to-sequence model for speech translation without transcriptions has been introduced [24], but was only evaluated on align-

¹http://zerospeech.com/2017
ment. Synthesized speech data were translated in \textsuperscript{[25]} using a model similar to the Listen Attend and Spell model \textsuperscript{[29]}, while a larger-scale study \textsuperscript{[27]} used an end-to-end system for translating audio books between French and English. Sequence-to-sequence models to both transcribe Spanish speech and translate it in English have also been proposed \textsuperscript{[25]}, by jointly training the two tasks in a multitask scenario with two decoders sharing the speech encoder. This model was further extended \textsuperscript{[29]} with the translation decoder receiving information both from the speech encoder and the transcription decoder.

For endangered languages (extremely low-resource settings) the lack of training data leads to the problem being framed as a sparse translation problem. This semi-supervised task lies between speech translation and keyword spotting, with cross-lingual supervision being used for word segmentation \textsuperscript{[30, 31, 32, 33]}. Bilingual setups for word segmentation were discussed by \textsuperscript{[43]}, \textsuperscript{[31, 36, 37]}, but applied to speech transcripts (true phones). Among the most relevant to our approach are the works of \textsuperscript{[24]} on speech-to-translation alignment using attentional Neural Machine Translation (NMT) and of \textsuperscript{[31, 32]} for language documentation. However, the former does not address word segmentation and is not applied to a language documentation scenario, while the latter methods do not provide a full coverage of the speech corpus analyzed. A neural approach for word segmentation in documentation scenarios using the soft attention matrices (which we also use for our baseline experiments) was investigated in \textsuperscript{[37]}.  

3. The Griko Language  

Griko is a Greek dialect spoken in southern Italy, in the Greca Salentina area southeast of Lecce. 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 \textit{Italiot Greek}, were included as seriously endangered in the UNESCO \textit{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. In addition, it is rare among the Greek dialects, due to its retention of terms from 9 different speakers (5 male, 4 female) from the 4 villages (Calimera, Sternatia, Martano, Corigliano) where native speakers could still be found. The digitally collected audio files (16-bit PCM, 44.1kHz, stereo) were manually segmented into utterances, transcribed, glossed in Italian, and annotated with extensive morphosyntactic tags by a trained linguist.

4. Data Processing  

The original \textit{UoI} corpus was collected during a field trip in Puglia, Italy by two linguists, with a particular focus on the use of imperative and verbal morphosyntax. The corpus contains utterances from 9 different speakers (5 male, 4 female) from the 4 villages (Calimera, Sternatia, Martano, Corigliano) where native speakers could still be found. The digitally collected audio files (16-bit PCM, 44.1kHz, stereo) were manually segmented into utterances, transcribed, glossed in Italian, and annotated with extensive morphosyntactic tags by a trained linguist.

4.1. Annotation Extensions  

In order to render the \textit{UoI} corpus useful for speech-related computational research on Griko, we extend the corpus with the following annotations:

1. Free-form Italian translations for every utterance, created by a bilingual speaker,
2. gold-standard word-level alignment information for every utterance, including annotated silences,
3. gold-standard speech-to-translation alignments,
4. pseudo-phones representation, obtained by using the acoustic unit discovery (AUD) method presented in \textsuperscript{[48]},
5. \textit{ZRC} gold standard for standard evaluation, described in the next section.

Figure \textsuperscript{[1]} shows an example of sentence pairs from our collection, and Table \textsuperscript{[1]} presents some statistics on these aligned transcriptions and translations. We observe that both sides of the parallel corpus are considerably similar with respect to the metrics presented here (sentence structure and vocabulary). This is reasonable: the two languages belong to the same family and have been in contact for centuries.

4.2. A reference compatible with the \textit{ZRC} metrics  

In addition to the word-level annotations, we built and make available a reference (in the format of the \textit{ZRC} challenge) in order to allow evaluation of different word discovery approaches using this corpus. We had a manual alignment between speech and words, but no possibility to obtain an accurate automatic alignment between speech and phones (or graphemes) due to the very small amount of data available (not possible to train an acoustic model using a Kaldi pipeline on 330 signals, for instance).

Thus, we used the word-level alignment information between speech and transcription, and the silence annotation available in our corpus, to approximate a speech-to-grapheme alignment. For each word present in the corpus, we retrieve its time window and segment this time window into smaller ones, giving to each existing grapheme an equal portion of its word time window. We manually corrected some of the silence and word annotations to ensure that we had no overlap between silence and words time windows. This approximation was necessary to make the \textit{ZRC} metrics work.

The final reference can be considered correct for evaluation of word discovery tasks (which do not take into account subword annotation), but should be consider with caution for evaluation of subword discovery tasks. Finally, we created two \textit{ZRC} versions, one removing the silence tokens, used for grapheme evaluation, and a second one with all the information, used for pseudo-phones evaluation.

\footnote{A discussion on the possible origins of Griko can be found in \textsuperscript{[48].}\url{http://griko.project.uoi.gr}}
of unsupervised word discovery. We use three different base-

Table 1: Statistics of the 330 sentences in our parallel Griko-Italian corpus.

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportional</td>
<td>42.2</td>
<td>52.2</td>
<td>46.7</td>
</tr>
<tr>
<td>neural</td>
<td>24.6</td>
<td>30.0</td>
<td>27.0</td>
</tr>
<tr>
<td>DTW-EM</td>
<td>56.6</td>
<td>51.2</td>
<td>53.8</td>
</tr>
</tbody>
</table>

Table 2: On speech-to-translation alignment, the unsupervised model outperforms the neural attentional model and the naive baseline in terms of Precision and F-score.

5. Speech-to-Translation Alignment

The task of speech-to-translation alignment is the problem of identifying portions in an audio segment that should be aligned to words in (text) translation, without access to transcriptions [24]. Our speech-to-translation alignment annotations allow us to evaluate such methods on our corpus. Evaluation is performed by computing standard precision, recall, and F-score.

Providing a baseline for future work, Table 3 reiterates previous results on speech-to-translation alignment. We present results with three methods: a naive proportional baseline (proportional), a neural alignment model [23] (neural), and an unsupervised model (DTW-EM) [49]. The naive baseline assumes no reordering and simply segments the audio to as many segments as the translation words, each with a length proportional to the word’s length in characters. The neural alignment model trains a speech-to-translation end-to-end sequence-to-sequence system with attention on all the data, and then the soft attention matrices are converted to hard alignments between audio segments and translation words. DTW-EM is an unsupervised model that extends the IBM Model 2 alignment model [50] to work on speech segments, combining it with a Dynamic Time-Warping-based clustering approach [51].

Since the two languages have several similar characteristics, the naive proportional baseline is already very competitive; its recall is better than both other evaluated methods. The unsupervised model, however, achieves much higher Precision and F-score than the rest. Unsurprisingly, the neural approach performs significantly worse in this setting: 330 sentences are clearly not enough to train a robust word-level model.

6. Unsupervised Word Discovery

In this section we illustrate the use of our corpus for the task of unsupervised word discovery. We use three different baselines, one monolingual and two bilingual, and two different representation levels, graphemes (from text) and pseudo-phones (automatically extracted from speech). Evaluation is performed using the Boundary metric from the Zero Resource Challenge 2017 (Track 2) [11]. We compute recall, precision and F-score. Below, we describe the three baselines evaluated in this work.

- **Dpseg (monolingual)**: dpseg is the non-parametric bayesian model introduced in [24]. On this setup, words are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet-Process. Estimation is performed through Gibbs sampling. This approach is known as being very robust on low-resource scenarios. The hyper-parameters used here are the same from [53].

- **Proportional Segmentation (bilingual)**: this baseline uses the word boundaries in the translation to segment the input proportionally. We can expect considerable good results for proportional segmentation when applied on language pairs similar on sentence structure and average token length, and therefore, we expect good results for this baseline when applied to the Griko-Italian corpus (see Table 1).

- **Neural Segmentation (bilingual)**: the method applied in this paper was presented in [37]. It post-processes a NMT system’s soft-alignment probability matrices to generate hard segmentation. Due to the length discrepancy between the symbols (graphemes and pseudophones) and the translations, our post-processing included alignment smoothing. This procedure, proposed by [23], consists of adding temperature $T$ to the softmax function used by the attention mechanism. Resulting soft-alignments matrices are further smoothed by averaging each probability by its right and left neighborhood. However, in this work we use $T = 1$ for all setups, and only the alignment matrices smoothing (averaging with the right and left neighbors) is used here. Also, for stability reasons, we report the averaged scores over 5 different trained models.

- **Merged Neural Segmentation (bilingual)**: the same methodology from the previous baseline, with the difference of averaging the soft-alignment probability matrices before post-processing, instead of averaging only the scores. We use the same 5 runs from the previous setup to generate an averaged (merged) segmentation.

Table 3 presents the achieved results. Even on this very low-resource scenario, dpseg has a remarkable performance for

---

Available at http://homepages.inf.ed.ac.uk/sgwater/resources.html
Table 3: Boundary scores for the task of unsupervised word segmentation. Results for neural segmentation are the average over 5 runs. Best results for each metric are presented in bold.

<table>
<thead>
<tr>
<th></th>
<th>dpseg</th>
<th>proportional</th>
<th>neural</th>
<th>merged neural</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>grapheme</td>
<td>68.50</td>
<td>75.10</td>
<td>71.60</td>
<td>44.70</td>
</tr>
<tr>
<td>pseudo-phones</td>
<td>23.30</td>
<td>36.90</td>
<td>28.50</td>
<td>28.50</td>
</tr>
</tbody>
</table>

Table 4: A comparison between the generated segmentation by the four baselines. For the neural baseline, results are the arithmetic mean between the statistics for the 5 runs.

<table>
<thead>
<tr>
<th></th>
<th>grapheme</th>
<th>pseudo-phones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># tokens</td>
<td>Vocabulary Size</td>
</tr>
<tr>
<td>proportional</td>
<td>2.370</td>
<td>1.715</td>
</tr>
<tr>
<td>dpseg</td>
<td>2.629</td>
<td>567</td>
</tr>
<tr>
<td>neural (average)</td>
<td>2.972</td>
<td>1.462</td>
</tr>
<tr>
<td>merged neural</td>
<td>2.573</td>
<td>1.476</td>
</tr>
</tbody>
</table>

the task of word segmentation working with graphemes. It retrieved 75.10% of the correct boundaries (recall). The second best method from the baselines for grapheme segmentation was the merged version of the neural segmentation. The remaining two baselines (proportional and neural) had close performance, achieving retrieval between 44 and 52%.

For the pseudo-phones segmentation, all methods had a considerable drop in performance, specially dpseg. They all achieved similar F-scores, with the merged neural baseline being slightly more effective. Table 3 presents some numbers for the generated segmentation of all methods presented in this section. We observe that, for pseudo-phones, dpseg seems to over-segment the input (average tokens per sentence), while the neural baselines segmented the input considerably less.

Lastly, pseudo-phones were obtained through an unsupervised unit discovery system, which inevitably adds noise to the representation. This noise is then propagated to the word discovery system. We believe the achieved results for pseudo-phones illustrate the difficulty of the task of word discovery on extreme low-resource setups.

7. Conclusion

In this paper we presented an extension of a very small parallel corpus on an endangered language called Griko. We make this corpus, with all its different levels of representation, freely available to the community as an effort in the direction of research replicability for low-resource approaches.

We illustrated the potential of this parallel corpus by performing the tasks of speech-to-text alignment and unsupervised word discovery. We encourage the community to challenge the baselines presented here.

Future work includes comparing the tasks results from this extreme case of language documentation with other low-resource corpora, such as the one presented in [15].

8. Acknowledgements

This work was partly funded by French ANR and German DFG under grant ANR-14-CE35-0002 (BULB project). Antonis Anastasopoulos was generously supported by NSF Award 1464553. Marika Lekakou was supported by the John S. Labatsis Public Benefit Foundation under the project ‘Documentation and analysis of an endangered language: aspects of the grammar of Griko’.

9. References


