



A small Griko-Italian speech translation corpus

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

¹Laboratoire d'Informatique de Grenoble, Univ. Grenoble Alpes, **France**

²Department of Computer Science and Engineering, Univ. of Notre Dame, **USA**

³Department of Philosophy, Univ of Ioannina, **Greece**

⁴Institute of Informatics, UFRGS, **Brazil** and CSEE, Univ. of Essex, **UK**

GOAL: An under-resourced language corpus for Computational Language Documentation

- **REAL endangered language**
- **Very low-resource**
- **Parallel corpus (Griko-Italian)**
- **Several levels of representation**

GOAL: An under-resourced language corpus **for Computational Language Documentation**

- **Highly interdisciplinary research field**
- **Leverage of Computational Models and Machine Learning**
- **Focus on endangered/unwritten languages (speech!)**

OUTLINE

01. THE GRIKO CORPUS

- a. Presentation
- b. Data collection
- c. Post-processing

02. CASE STUDY

- a. Speech-to-translation Alignment
- b. Unsupervised Word Discovery

THE GRIKO CORPUS

The Griko Dialect

- A **Greek** dialect spoken on south **Italy** (Grecia Salentina, southeast of Lecce)
- Included as **seriously endangered** in the UNESCO Red Book of Endangered Languages in 1999
- **Less than 20,000 native speakers** (overestimation)*



The Corpus

- Collected by 2 linguists during a field trip to Puglia, Italia
- Particular focus on the use of infinitive and verbal morphosyntax
- **9 different speakers** (5 male, 4 female) from 4 villages in Grecia Salentina



The Corpus

- 330 manually segmented sentences
- The only Griko speech dataset available online¹
(roughly 20 minutes of speech in Griko)



¹<http://griko.project.uoi.gr/> (University of Ioannina)

Annotation Extensions

1. Italian translations for every utterance

Griko: Jatì ìche pòlemìsonta òli tin addomàda.

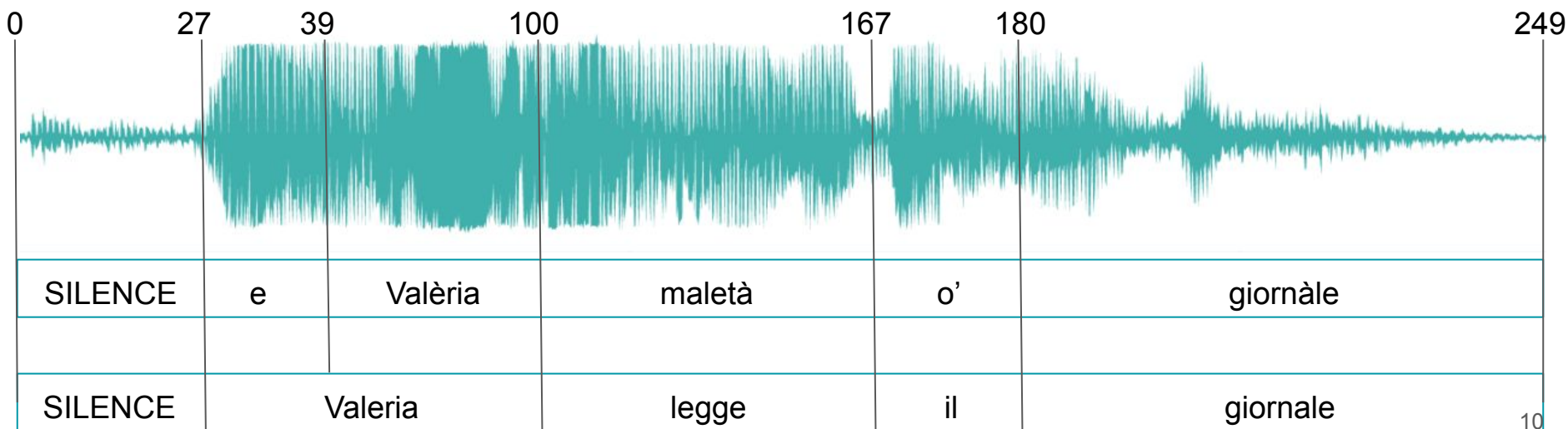
Italian: Preche aveva lavorato tutta la settimana.

Example: Sentence 144

Annotation Extensions

2. Gold-standard word-level alignment (including silences)
3. Gold-standard speech-to-translation alignments

Example: Sentence 1



Annotation Extensions

4. Since no Griko acoustic model is available, we use unsupervised Acoustic Unit Discovery (AUD) to extract pseudo-phone units (see [1]).

Griko: Ìsose èmbi àtti finèstra

Pseudo-phones: a46 a31 a8 a31 a15 a13 a24 a31
a18 a31 a20 a11 a8 a31 a8 a31

Example: Sentence 102

Annotation Extensions

5. Zero Resource Challenge 2017 Track 2¹ reference for Spoken Term (word) Discovery

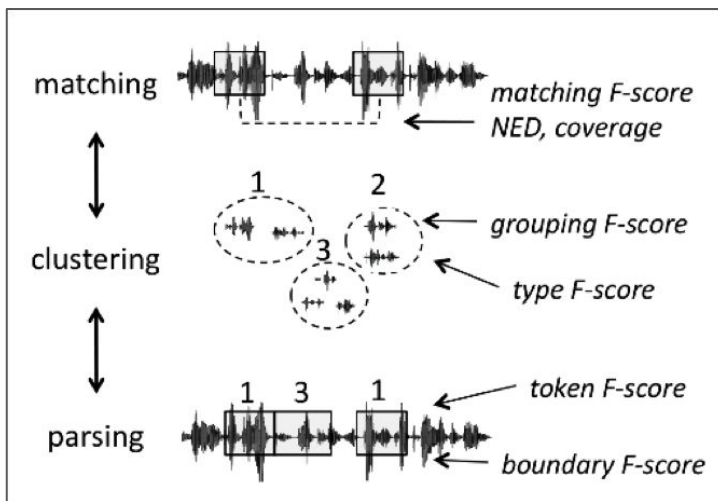


Image from The Zero Resource Speech Challenge 2017 web page¹.

¹http://sapience.dec.ens.fr/bootphon/2017/page_3.html

Corpus Statistics Summary

- 330 sentences, 20 minutes of speech
- 48 different pseudo-phones (average 24 per sentence)

	# tokens	# types	Avg tokens length	Avg tokens per sentence	Shortest token	Largest token
Griko	2,374	691	5.68	7.19	1	16
Italian	2,384	456	5.76	7.22	1	13

Table: Statistics for the transcriptions and their aligned translations.

CASE STUDY

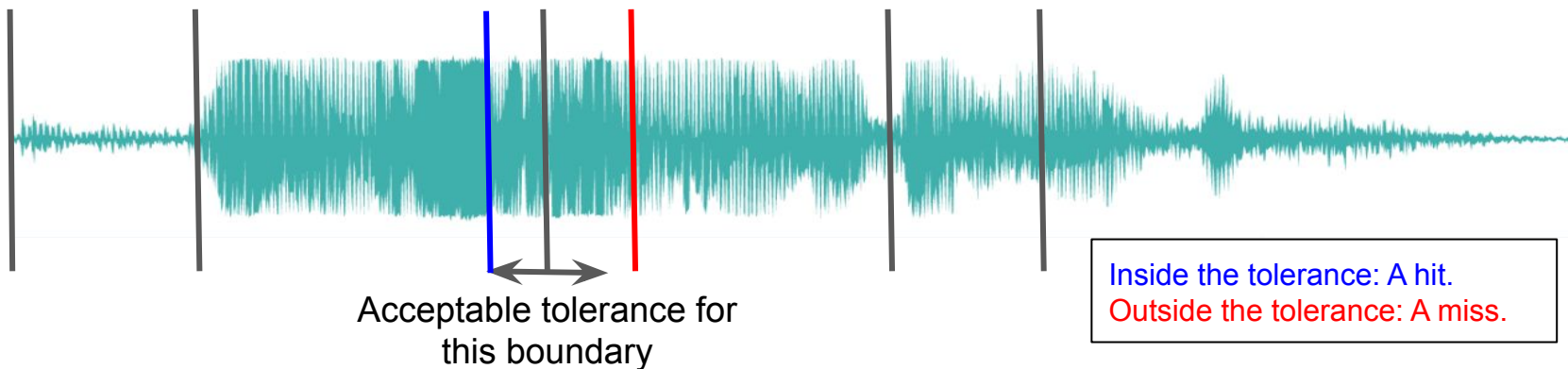
Speech-to-Translation Alignment

- **Task:** to align portions of the audio with its translation, without access to the transcription.
- **Evaluation:** Links between speech frames and translation words (Precision, Recall and F-score).

Method	Precision	Recall	F-score
Proportional	42.2	52.2	46.7
Neural	24.6	30.0	27.0
DTW-EM	56.6	51.2	53.8

Unsupervised Word Discovery

- **Task:** to cluster unsegmented graphemes/pseudo-phones into word-like units.
- **Evaluation:** ZRC Track 2, Boundary scores (Precision, Recall and F-score).



Unsupervised Word Discovery

	GRAPHEMES (TOPLINE)			PSEUDO-PHONES (SPEECH)		
Method	Precision	Recall	F-score	Precision	Recall	F-score
dpseg ¹	68.5	75.1	71.6	23.3	36.9	28.5
proportional	44.7	44.8	44.7	28.5	29.9	29.2
neural ²	42.6	51.8	46.7	32.0	27.6	29.5
merged neural	50.2	54.0	52.1	34.3	26.7	30.0

Table: For graphemes segmentation, dpseg beats all other baselines, while for pseudo-phones the merged neural approach achieves better Precision and F-score.

*Available at <https://homepages.inf.ed.ac.uk/sqwater/>, Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 317–325. Association for Computational Linguistics, 2009.

**Unwritten languages demand attention too! Word Discovery using Encoder-Decoder Models

M. Zanon Boito, A. Berard, A. Villavicencio, L. Besacier, ASRU 2017

CONCLUSION

Conclusion

We presented a corpus extension for computational documentation research. The result:

- Several levels of representation available
- Reference for ZRC spoken **word** discovery experiments

Conclusion

→ **Freely Available on Github**

[antonisa/griko-italian-parallel-corpus](https://github.com/antonisa/griko-italian-parallel-corpus)





Thank you!

Questions?

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



[antonisa/griko-italian-parallel-corpus](https://antonisa.github.io/griko-italian-parallel-corpus)