Intelligent Linguistic Annotation Interface

Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin, Yuyan Zhang
Language Technologies Institute
Carnegie Mellon University
{chianyuc, jeanl1, ziruil, yuhsianl, yuyanz1}@andrew.cmu.edu

Graham Neubig
Language Technologies Institute
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

Based on [1].

1 Data: Griko–Italian

1.1 Data acuqisition

The human process of gathering and producing data is roughly as follows:

1. Griko speech → Griko transcription
2. Griko transcription → Italian glossing
3. Produced POS and morphosyntactic tags for each Griko token, and everything follows

Logically, the data can be understood by the following relations:

1. Raw wav files (Griko speech)
2. wav2gr (written Griko token) ⇐ extracted from raw wav files
3. silences ⇐ extracted from raw wav files[1]
4. Transcriptions = concatenating wav2gr
5. itgloss (Italian gloss, an Italian corresponding word is given to each Griko token) ⇐ obtained from wav2gr
6. Translations ⇐ performed from raw wav files (not necessary)
7. wav2it ⇐ some Italian words in translations are aligned with corresponding segments of raw wav files

For the transcription task, the most important data are the first 4 items.

1.2 Basic statistics

Total 331 data in this dataset, from number 1 to 332, with number 5 missing.

[1] But silence does not exactly fill the spaces between the intervals listed in wav2gr.

1.3 Raw wav file format

Python document for wave library: https://docs.python.org/3/library/wave.html

Useful ref: https://stackoverflow.com/questions/18625085/how-to-plot-a-wav-file

1.4 wav2gr format

Griko token, start time, end time. Time in the unit of 10 ms.

1.5 Silences format

Start time, end time. Time in the unit of 10 ms.

1.6 Transcriptions format

The string of Griko transcription.

1.7 Example: data number 6, clean, only one speaker

Raw wav file: [raw/6.wav]

Griko tokens: [wav2gr/6.words]
ti 57 70
kànni 70 93
e 93 102
Anna 102 156
o 185 198
sàmba 198 240
pornò 255 306

Silences: [silences/6.txt]
0 54
155 185
239 255
307 349

Transcription: [transcriptions/6.gr]
ti kànni e Anna o sàmba pornò

1.8 Example: data number 10, has another background speaker

Raw wav file: [raw/10.wav]

Griko tokens: [wav2gr/10.words]
allòra 81 120
fèti 120 149
p 149 156
òrkete 156 225
èrkome 249 300
mapàle 293 351
ettù 351 390
ce 410 428
tròo 428 453
poddù 453 497
pasticciòttu 497 584

Silences: [silences/10.txt]
0 80
224 249
391 410
589 599

Transcription: [transcriptions/10.gr]
allòra fèti p òrkete èrkome mapàle ettù ce tròo poddù pasticciòttu

1.9 Example: data number 30, two equally strong speakers

The female speaker is transcribed.

Raw wav file: [raw/30.wav]

Griko tokens: [wav2gr/30.words]
ste 33 49
kammèni 49 108
sto' 108 128
giardìno 128 197
sto' 197 219
cìpo 219 274

Silences: [silences/30.txt]
0 32
274 299

Transcription: [transcriptions/30.gr]
ste kammèni sto' giardìno sto' cìpo

References

[1] Oliver Adams, Trevor Cohn, Graham Neubig, and Alexis Michaud. Phonemic transcription of low-resource tonal languages. In *Australasian Language Technology Association Workshop 2017*, December 2017.