Forced Alignement Task – Training Set Description

The training set for the EVALITA forced alignement task contains 8064 training units. Each participant should extract a portion of this data set and use it as development set. Each unit contains data regarding exactly one dialogic turn and comprises three files

- A WAV file containing the utterance (clean speech, close mic, high quality).
- A Wrd file containing the transcription of the utterance at word level
- A Phn file containing the transcription of the utterance at phone level

WAV files are encoded as PCM 16Khz mono.

Information regarding speech data contained in each file can be extrapolated from the file name. File names are composed as follows:

- Corpus type: **DG** (Only dialogic data are contained in this set)
- Task type: **MT** (Map Task)
  TD (Difference Test)
- Test ID: **A/B 0#** (A03, B05...)
- Dialect: Please refer to Table 1 for details regarding dialect coding
- Speaker ID: **_p1**, **_p2** for each test ID
- Role: **G/F** (Giver/Follower. Map task dialogues only)
- Turn number: **####** (#23, #254...)

Dialect	Label
Bari	B
Cagliari	C
Bergamo	D
Parma	E
Firenze	F
Genova	G
Catanzaro	H
Lecce	L
Milano	M
Napoli	N
Perugia	O
Palermo	P
Roma	R
Torino	T
Venezia	V

Table 1: Dialect codes

More details regarding each speaker (Gender, age, birthplace...) can be found in the included dialogue description files (TXT) along with the full dialogue transcription.
Transcription files are encoded as ASCII files containing an N*3 table where N is the number of segments (words or phones) in the file. Each row of the table reports the starting sample of the corresponding segment, its end sample and its label.

In Table 2 the set of SAMPA symbols used in Phn files is reported. Due to the difficulty of finding a marker between two vowels, the annotation rule was not to split the occurrence of this situation.

a	k	m	ja
e	g	n	je
E	ts	J	jo
i	dz	r	ju
o	tS	l	oj
O	dZ	L	wa
u	f	__	we
p	v	aj	wi
b	s	aw	wo
t	z	ej	# (garbage)
d	S	ew	

Table 2: Phn files symbols set

Symbols other than words used in the Wrd files are reported in Table 3

#	Garbage
<sp>	Short pause
<lp>	Long pause
<P>	Medium/long pause with discourse interruption
<ehm>, <eeh>...	Filled pauses
word<vv> <vv>word	Filled pauses with vowel lengthening (allora<aa>, <ee>eccolo)
<cc>word word<cc>	Filled pauses with consonant lengthening (<ss>senti, non<nn>...)
wo_rd	Internal interruptions (mon_tato)

Table 3: Wrd files special symbols set