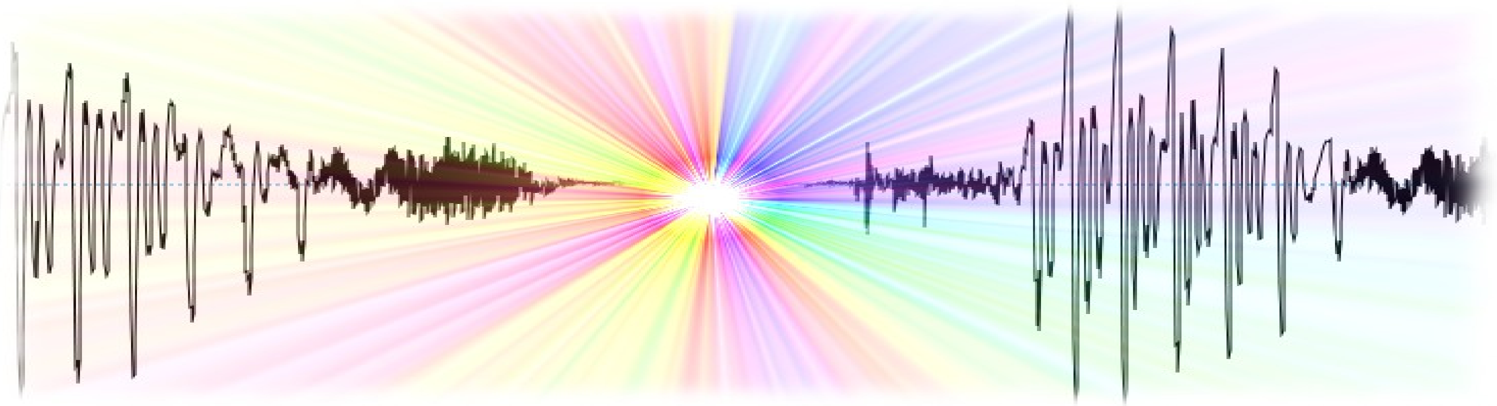
SPPAS: a tool for the phonetic segmentation of speech

Brigitte Bigi



Keywords:

Phonetizaton Automatic Speech
Syllabification Segmentation
Alignment Prosody

What SPPAS can do today?

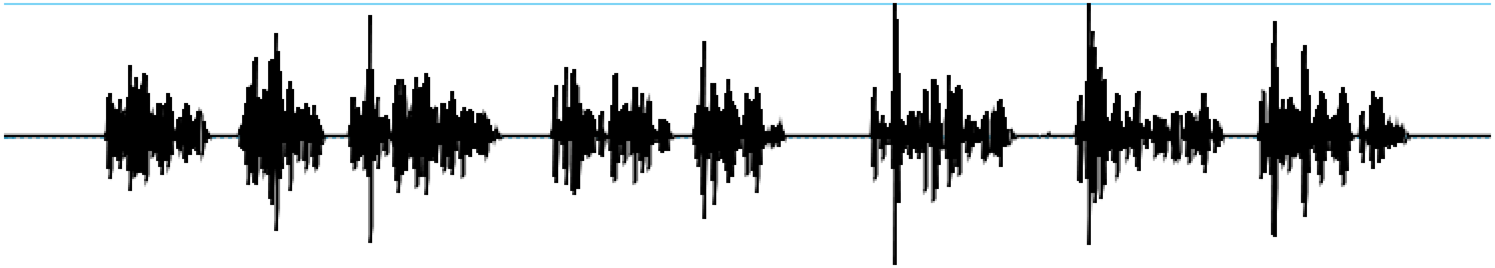
- Automatic annotations:
  - **Momel/INTSINT**: Modelisation of Mélodie
  - **IPUs segmentation**: utterance level segmentation
  - **Phonetization**: grapheme to phoneme conversion
  - **Alignment**: phonetic segmentation
  - **Syllabification**: group phonemes into syllables
- Goodies:
  - Get files information
  - Play sound (mono wav)
  - Manual transcription based on IPUs
  - Filter tiers

Key-points

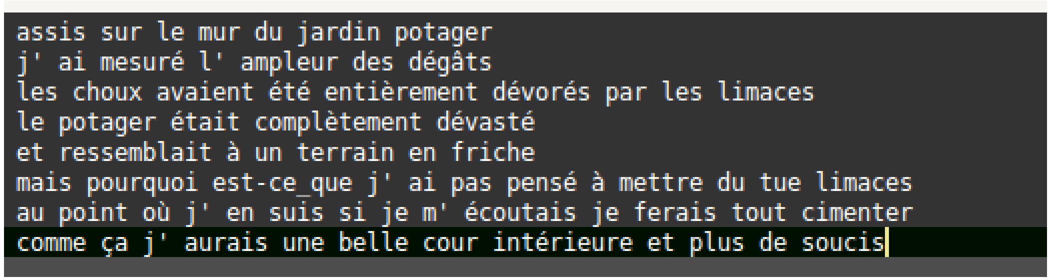
- A tool dedicated to computer scientists **and** linguists
- Language-independent algorithms
  - Resources for French, English, Italian and Chinese and there is an easy way to add other languages
- GNU Public License

SPPAS inputs

- Speech signal: wav file



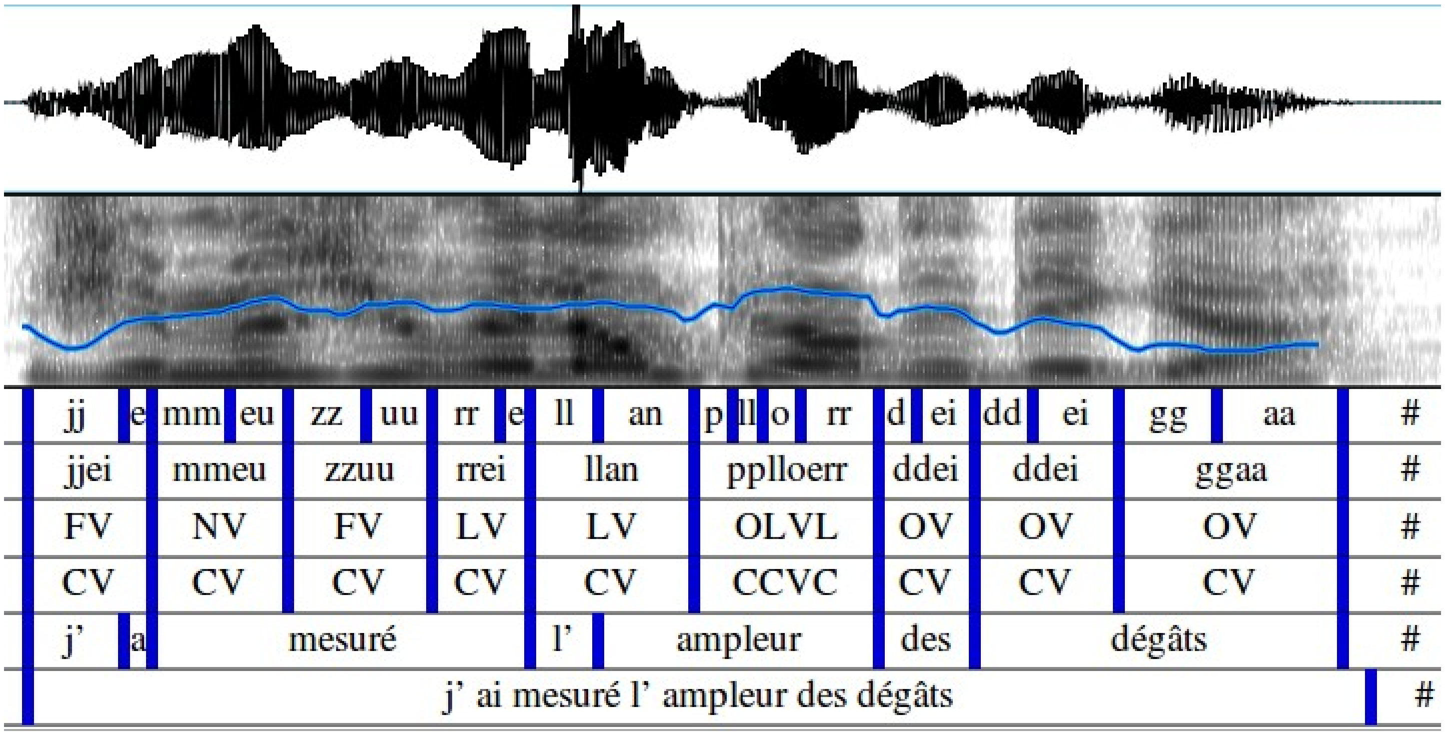
- Transcription: txt or TextGrid



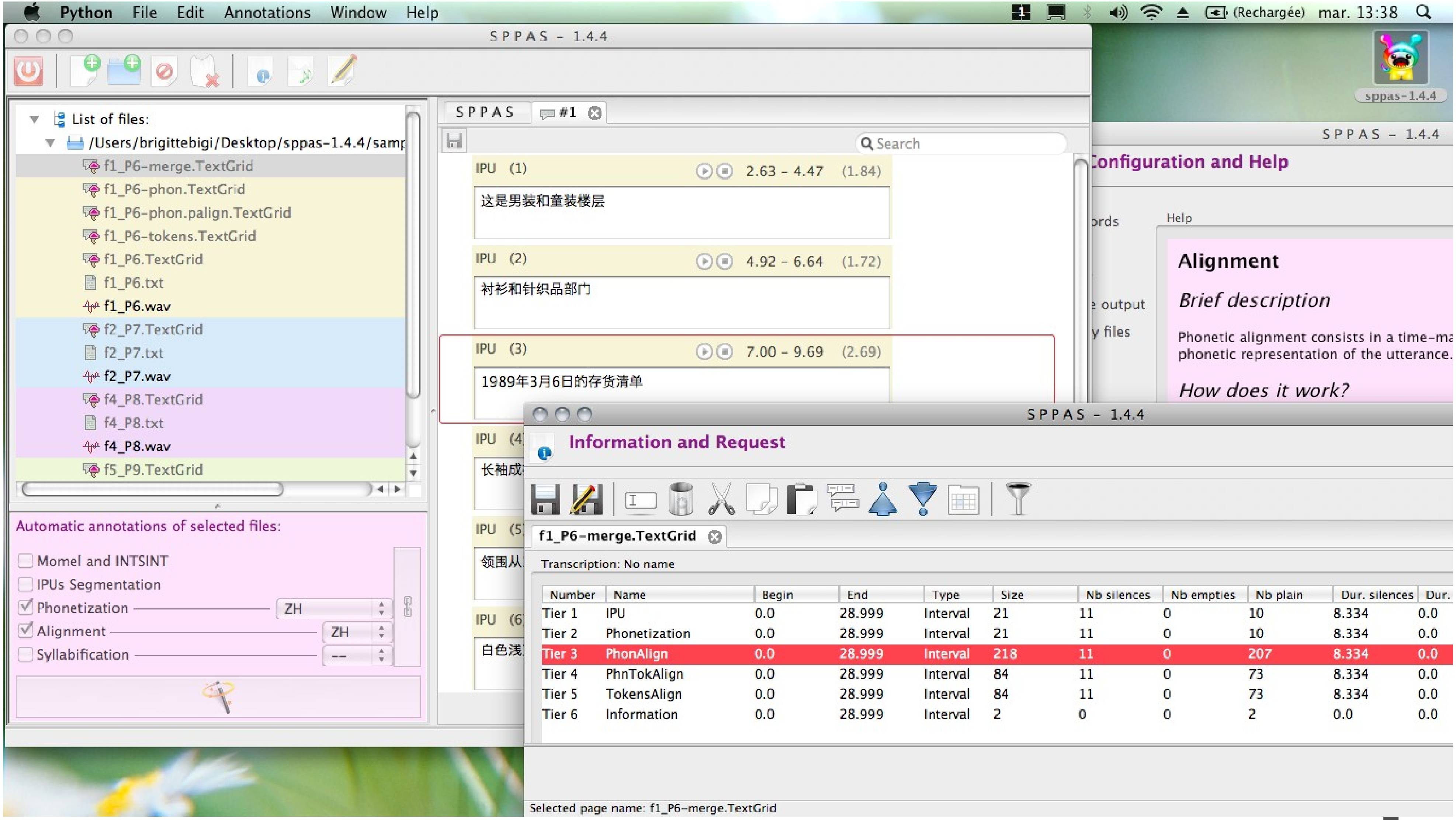
```
assis sur le mur du jardin potager
j' ai mesuré l' ampleur des dégâts
les choux avaient été entièrement dévorés par les limaces
le potager était complètement dévasté
et ressemblait à un terrain en friche
mais pourquoi est-ce_que j' ai pas pensé à mettre du tue limaces
au point où j' en suis si je m' écoutais je ferais tout cimenter
comme ça j' aurais une belle cour intérieure et plus de soucis
```

SPPAS outputs

- A set of TextGrid files



Screenshot



List of files:
/Users/brigittebigi/Desktop/sppas-1.4.4/samp

- f1_P6-merge.TextGrid
- f1_P6-phon.TextGrid
- f1_P6-phon.palign.TextGrid
- f1_P6-tokens.TextGrid
- f1_P6.TextGrid
- f1_P6.txt
- f1_P6.wav
- f2_P7.TextGrid
- f2_P7.txt
- f2_P7.wav
- f4_P8.TextGrid
- f4_P8.txt
- f4_P8.wav
- f5_P9.TextGrid

Automatic annotations of selected files:

- □ Momel and INTSINT
- □ IPUs Segmentation
- ☑ Phonetization — ZH
- ☑ Alignment — ZH
- □ Syllabification — --

IPU (1) 2.63 - 4.47 (1.84)
这是男装和童装楼层

IPU (2) 4.92 - 6.64 (1.72)
衬衫和针织品部门

IPU (3) 7.00 - 9.69 (2.69)
1989年3月6日的存货清单

IPU (4
长袖成

IPU (5
领围从

IPU (6
白色浅

Alignment

Brief description

Phonetic alignment consists in a time-ma
phonetic representation of the utterance.

How does it work?

SPPAS - 1.4.4

Information and Request

f1_P6-merge.TextGrid

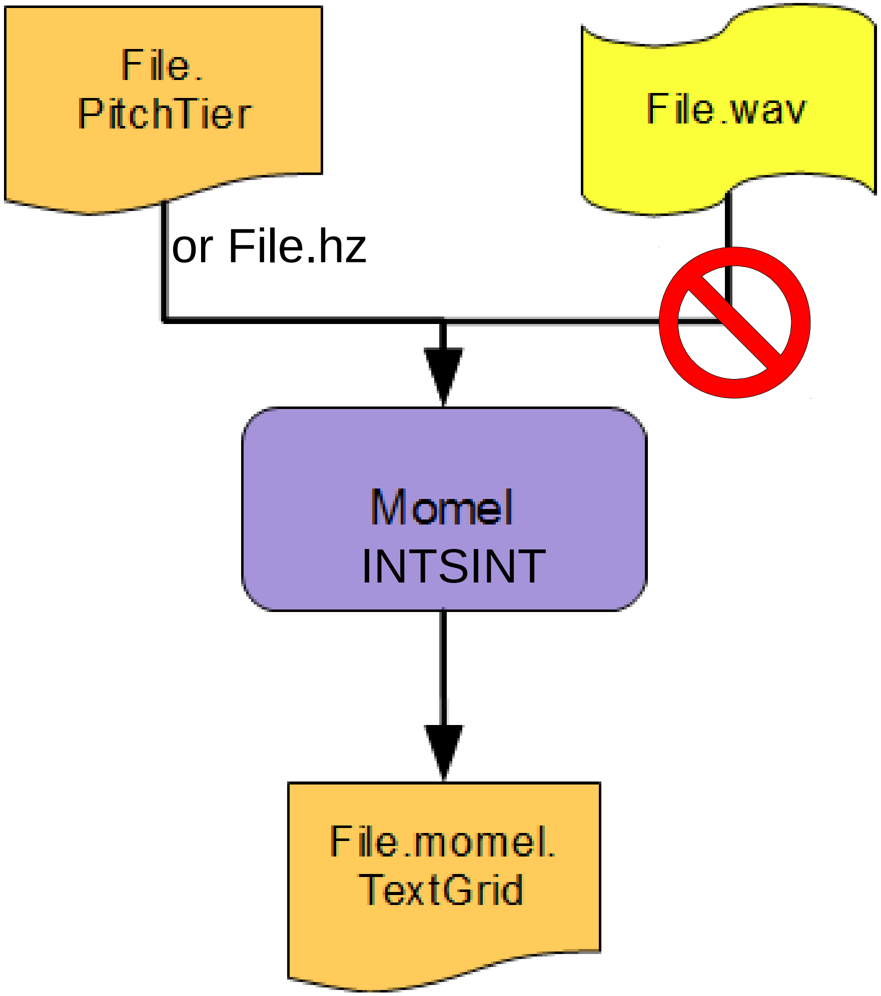
Transcription: No name

Number	Name	Begin	End	Type	Size	Nb silences	Nb empties	Nb plain	Dur. silences	Dur.
Tier 1	IPU	0.0	28.999	Interval	21	11	0	10	8.334	0.0
Tier 2	Phonetization	0.0	28.999	Interval	21	11	0	10	8.334	0.0
Tier 3	PhonAlign	0.0	28.999	Interval	218	11	0	207	8.334	0.0
Tier 4	PhnTokAlign	0.0	28.999	Interval	84	11	0	73	8.334	0.0
Tier 5	TokensAlign	0.0	28.999	Interval	84	11	0	73	8.334	0.0
Tier 6	Information	0.0	28.999	Interval	2	0	0	2	0.0	0.0

Selected page name: f1_P6-merge.TextGrid

Momel and INTSINT

- SPPAS implements Momel
- But... in the today's version: is missing!



File. PitchTier

or File.hz

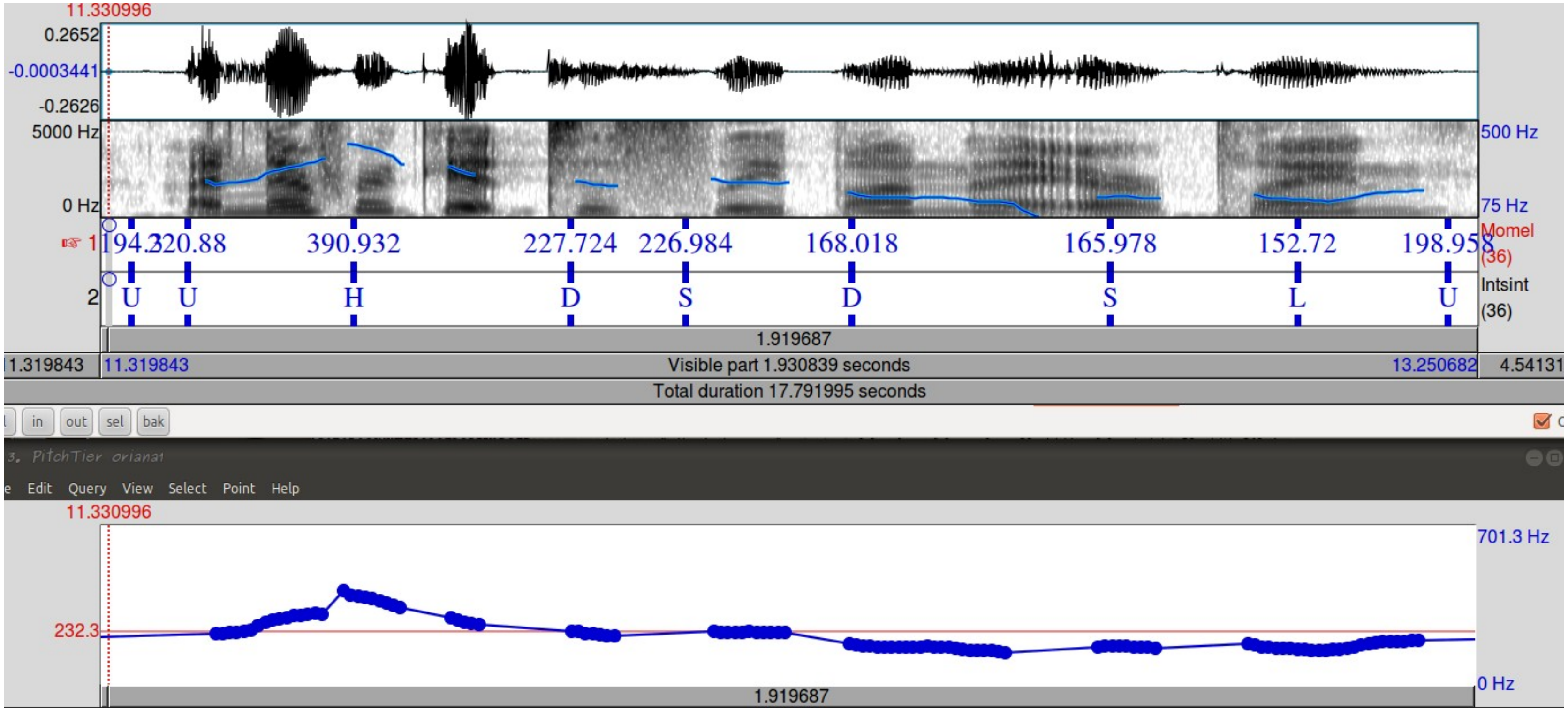
File.wav

Momel INTSINT

File.momel. TextGrid

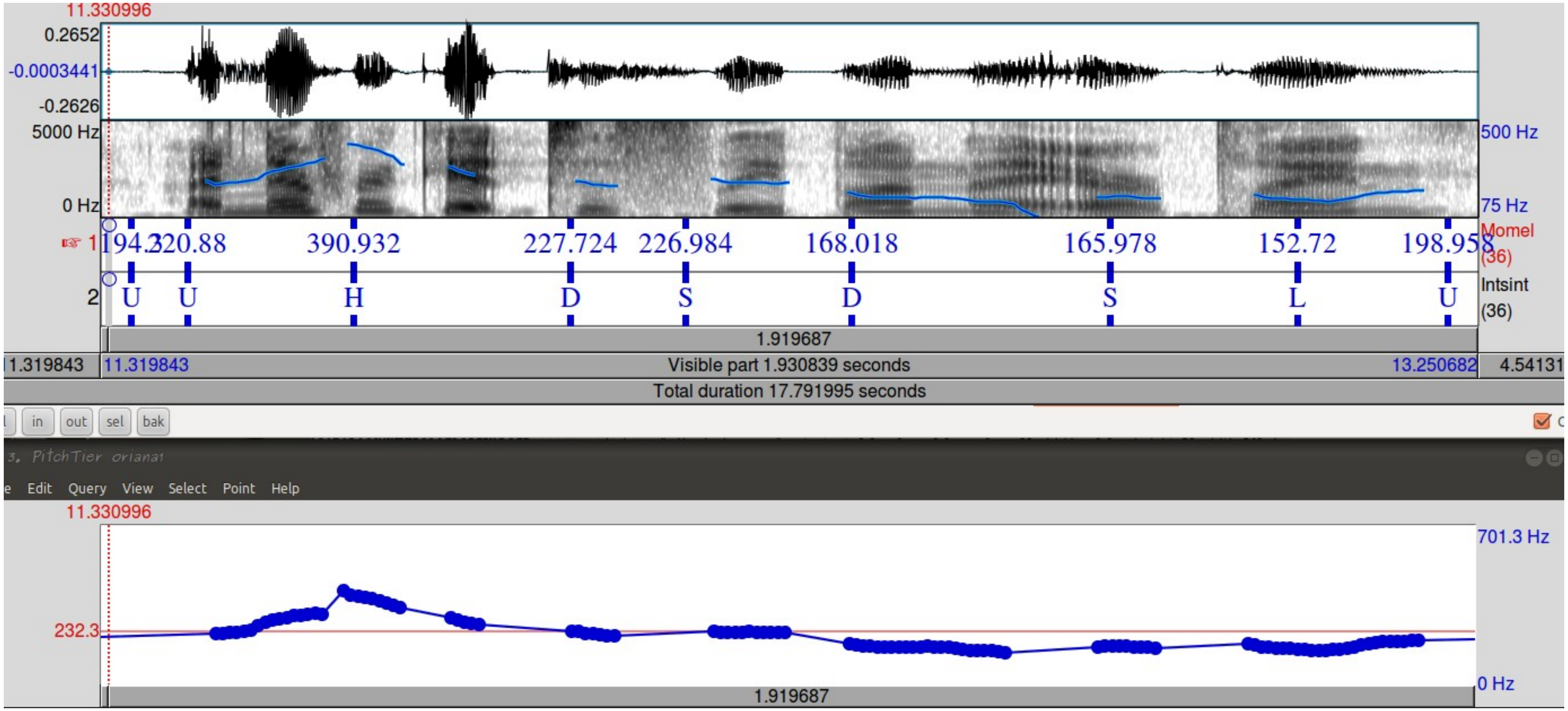
Momel/INTSINT: example

- Output: a TextGrid file with 2 tiers
  - Momel targets (pitch values)
  - INTSINT annotation of these targets



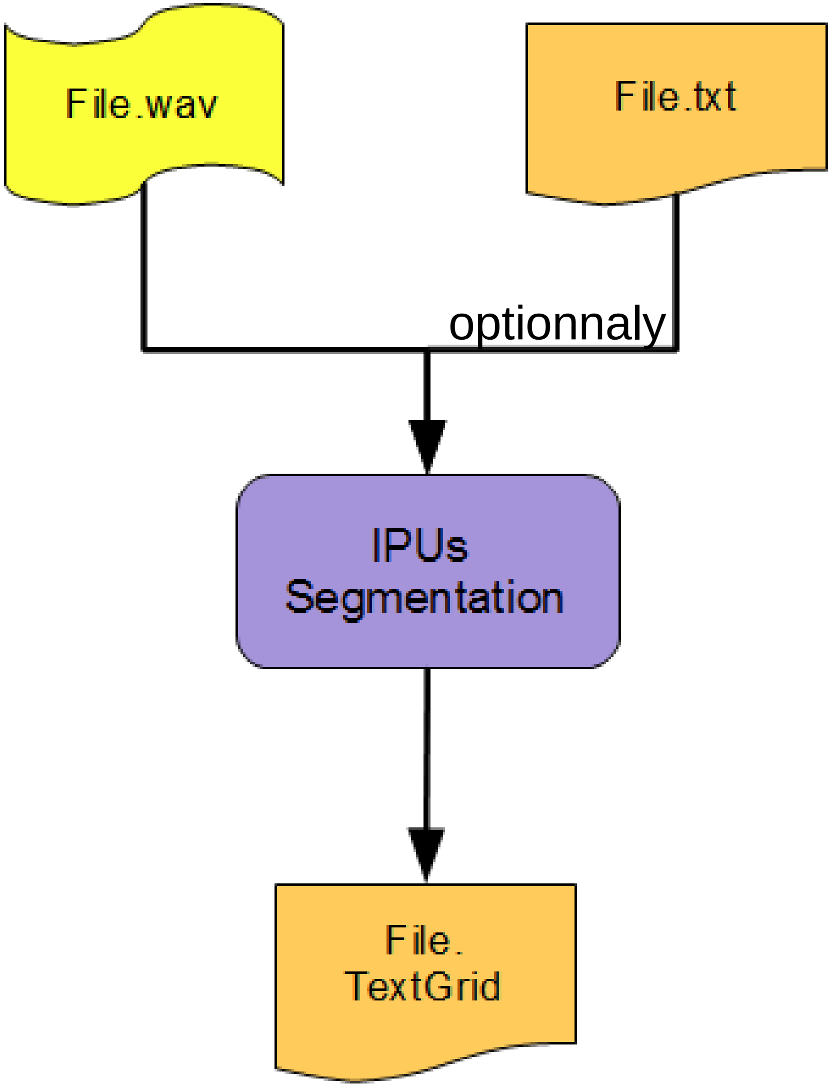
Momel/INTSINT: example

- Output: a TextGrid file with 2 tiers
  - Momel targets (pitch values)
  - INTSINT annotation of these targets



IPUs segmentation

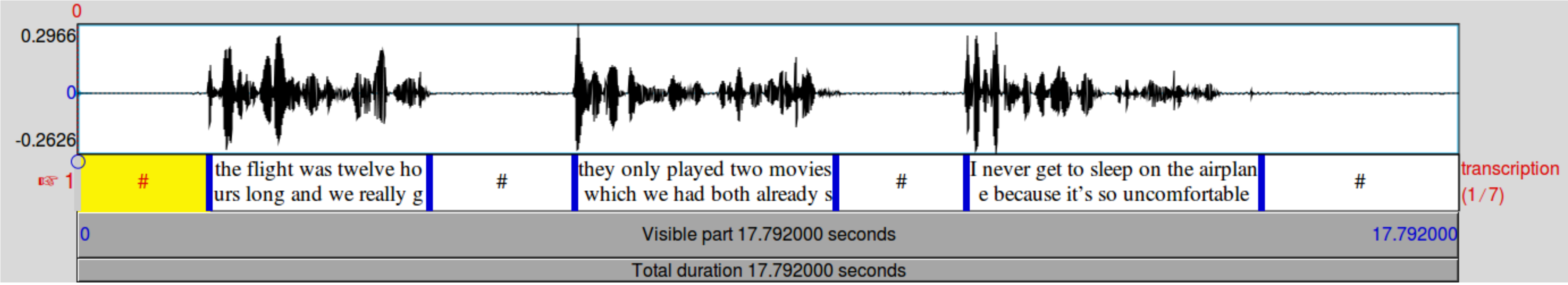
- Inter-Pausal Units segmentation
- The algorithm computes a heuristics based on the detection of silences, by using:
  - volume
  - min silence duration
  - min speech duration



IPUs segmentation: example

Transcription: silences are indicated by newlines or '#'

```
the flight was twelve hours long and we really got bored
they only played two movies which we had both already seen
I never get to sleep on the airplane because it's so uncomfortable
```



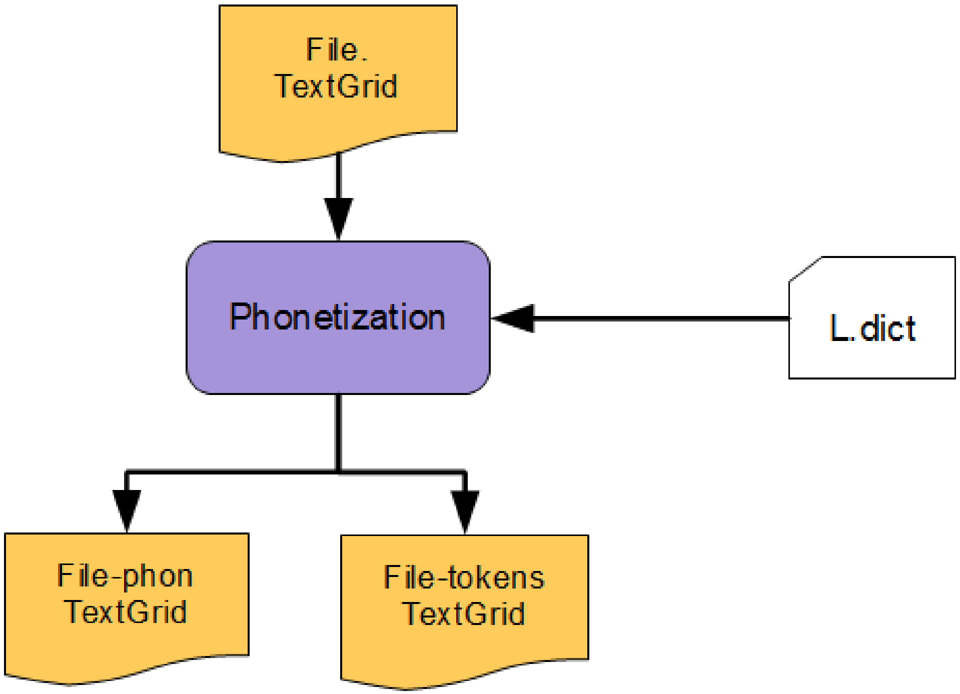
#	the flight was twelve ho urs long and we really g	#	they only played two movies which we had both already s	#	I never get to sleep on the airplan e because it's so uncomfortable	#

Visible part 17.792000 seconds

Total duration 17.792000 seconds

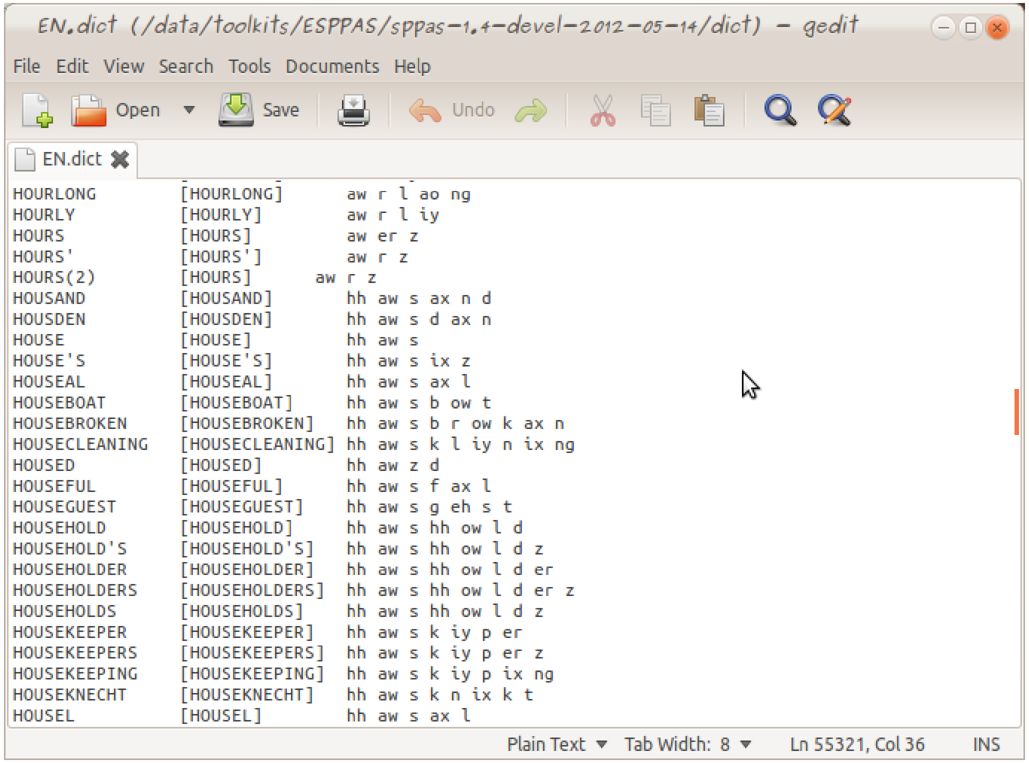
Phonetization

- Process of representing sounds with phonetic signs
- The phonetization is the equivalent of a sequence of dictionary look-ups.
- Phonetic variants:
  - no rules are applied, all possibilities are stored



Phonetization: example

- Resources:
  - a dictionary (HTK-ASCII format)



EN.dict (/data/toolkits/ESPPAS/sppas-1.4-devel-2012-05-14/dict) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
EN.dict
Plain Text
Tab Width: 8
Ln 55321, Col 36
INS

```
HOURLONG        [HOURLONG]      aw r l ao ng
HOURLY          [HOURLY]        aw r l iy
HOURS           [HOURS]         aw er z
HOURS'          [HOURS']        aw r z
HOURS(2)        [HOURS]     aw r z
HOUSAND         [HOUSAND]       hh aw s ax n d
HOUSDEN         [HOUSDEN]       hh aw s d ax n
HOUSE           [HOUSE]         hh aw s
HOUSE'S         [HOUSE'S]       hh aw s ix z
HOUSEAL         [HOUSEAL]       hh aw s ax l
HOUSEBOAT       [HOUSEBOAT]     hh aw s b ow t
HOUSEBROKEN     [HOUSEBROKEN]   hh aw s b r ow k ax n
HOUSECLEANING   [HOUSECLEANING] hh aw s k l iy n ix ng
HOUSED          [HOUSED]        hh aw z d
HOUSEFUL        [HOUSEFUL]      hh aw s f ax l
HOUSEGUEST      [HOUSEGUEST]    hh aw s g eh s t
HOUSEHOLD       [HOUSEHOLD]     hh aw s hh ow l d
HOUSEHOLD'S     [HOUSEHOLD'S]   hh aw s hh ow l d z
HOUSEHOLDER     [HOUSEHOLDER]   hh aw s hh ow l d er
HOUSEHOLDERS    [HOUSEHOLDERS]  hh aw s hh ow l d er z
HOUSEHOLDS      [HOUSEHOLDS]    hh aw s hh ow l d z
HOUSEKEEPER     [HOUSEKEEPER]   hh aw s k iy p er
HOUSEKEEPERS    [HOUSEKEEPERS]  hh aw s k iy p er z
HOUSEKEEPING    [HOUSEKEEPING]  hh aw s k iy p ix ng
HOUSEKNECHT     [HOUSEKNECHT]   hh aw s k n ix k t
HOUSEL          [HOUSEL]        hh aw s ax l
```

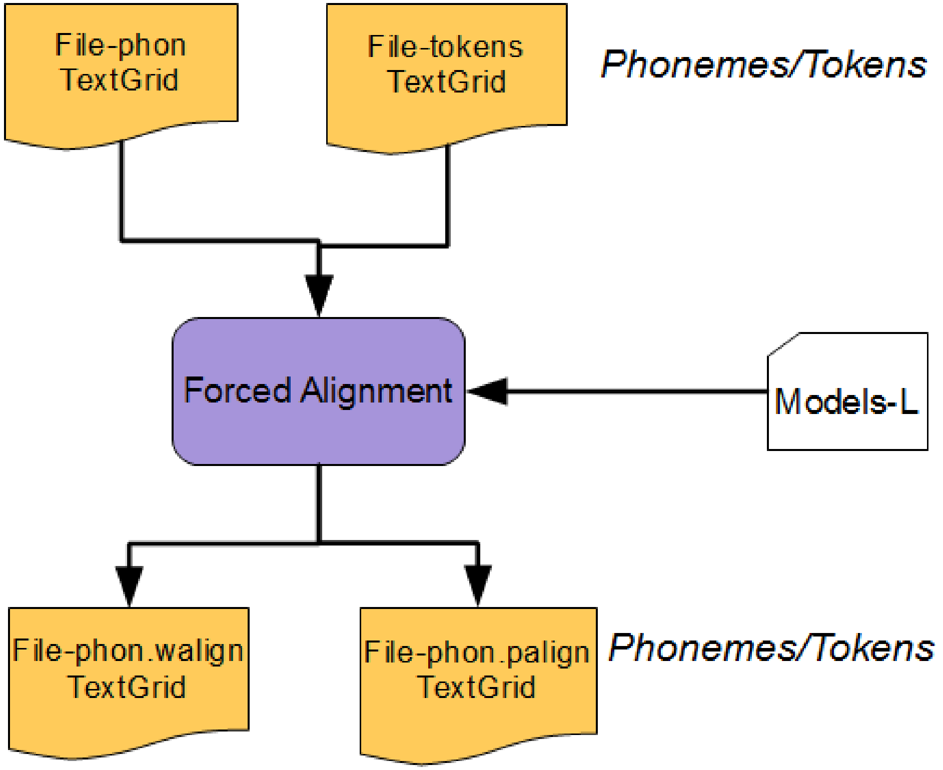
```
the flight was twelve hours long and we really got bored
```

is phonetized as follow:

```
dh.ax|dh.ah|dh.iy    f.l.ay.t    w.aa.z|w.ah.z|w.ax.z|w.ao.z    t.w.eh.l.v
aw.er.z|aw.r.z   l.ao.ng   ae.n.d|ax.n.d   w.iy   r.ih.l.iy|r.iy.l.iy   g.aa.t
b.ao.r.d
```

Alignment

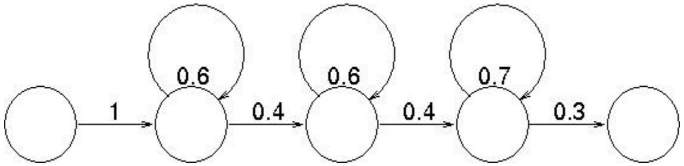
- A time-matching between a given speech utterance along with a phonetic representation of the utterance
- Forced-alignment in SPPAS is based on the **Julius** Speech Recognition Engine

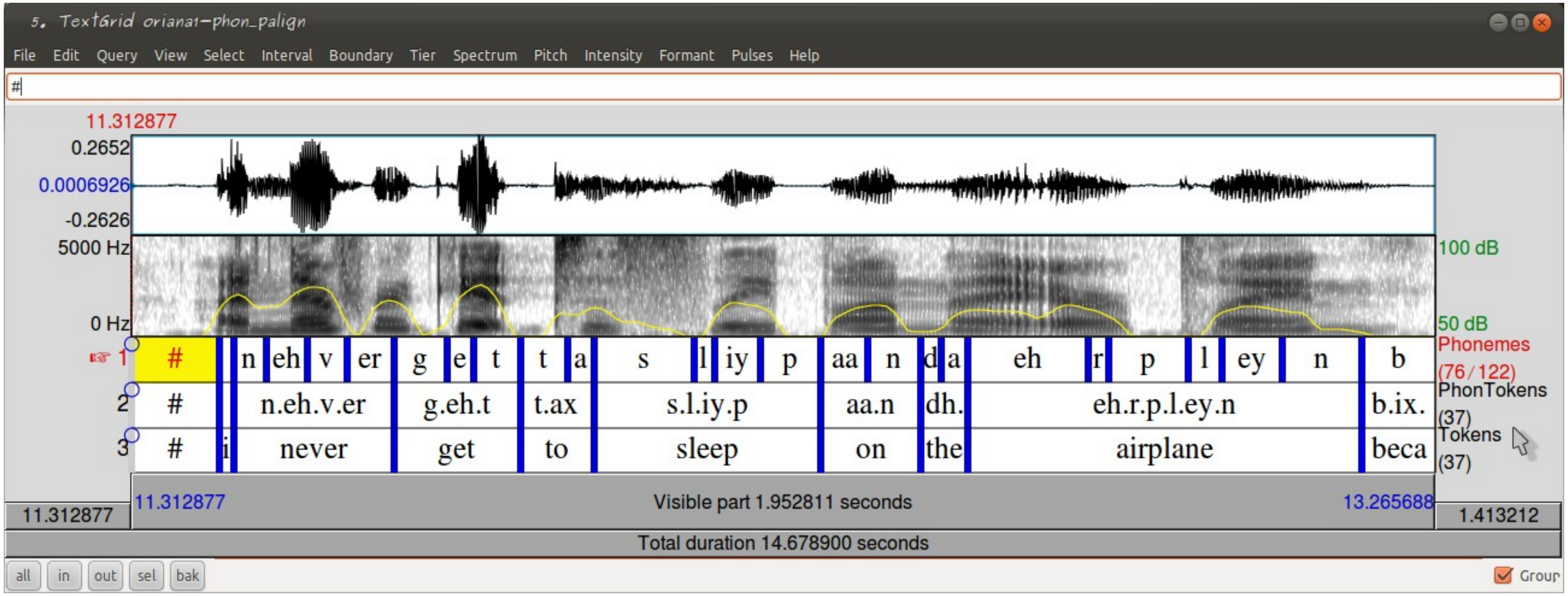


segmentation.

Alignment: example

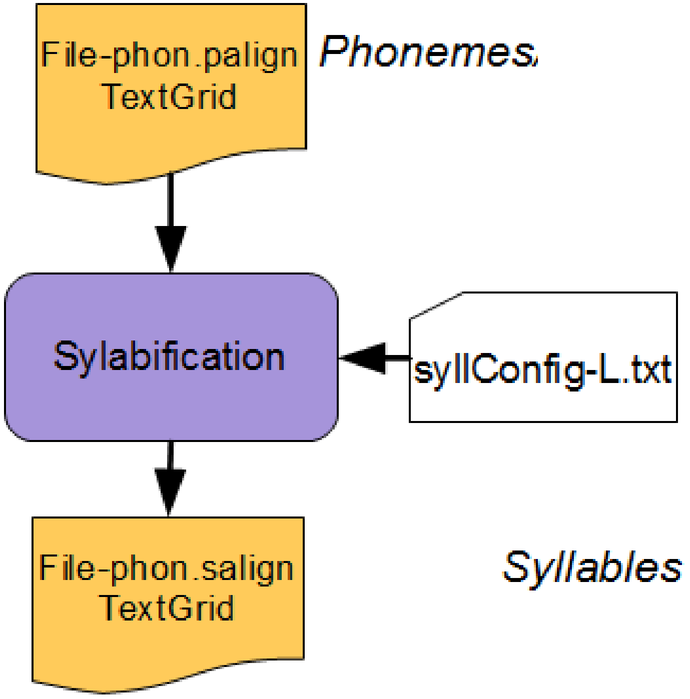
- Resources:
  - A finite state grammar that describes sentence patterns to be recognized (created by SPPAS);
  - An acoustic model.




5. TextGrid oriana1-phon_palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
#
11.312877
0.2652
0.0006926
-0.2626
5000 Hz
0 Hz
100 dB
50 dB
1 # n eh v er g e t t a s l iy p aa n d a eh r p l ey n b
2 # n.eh.v.er g.eh.t t.ax s.l.iy.p aa.n dh. eh.r.p.l.ey.n b.ix.
3 # i never get to sleep on the airplane beca
Phonemes (76/122)
PhonTokens (37)
Tokens (37)
11.312877
11.312877
Visible part 1.952811 seconds
13.265688
1.413212
Total duration 14.678900 seconds
all
in
out
sel
bak
Group

Syllabification

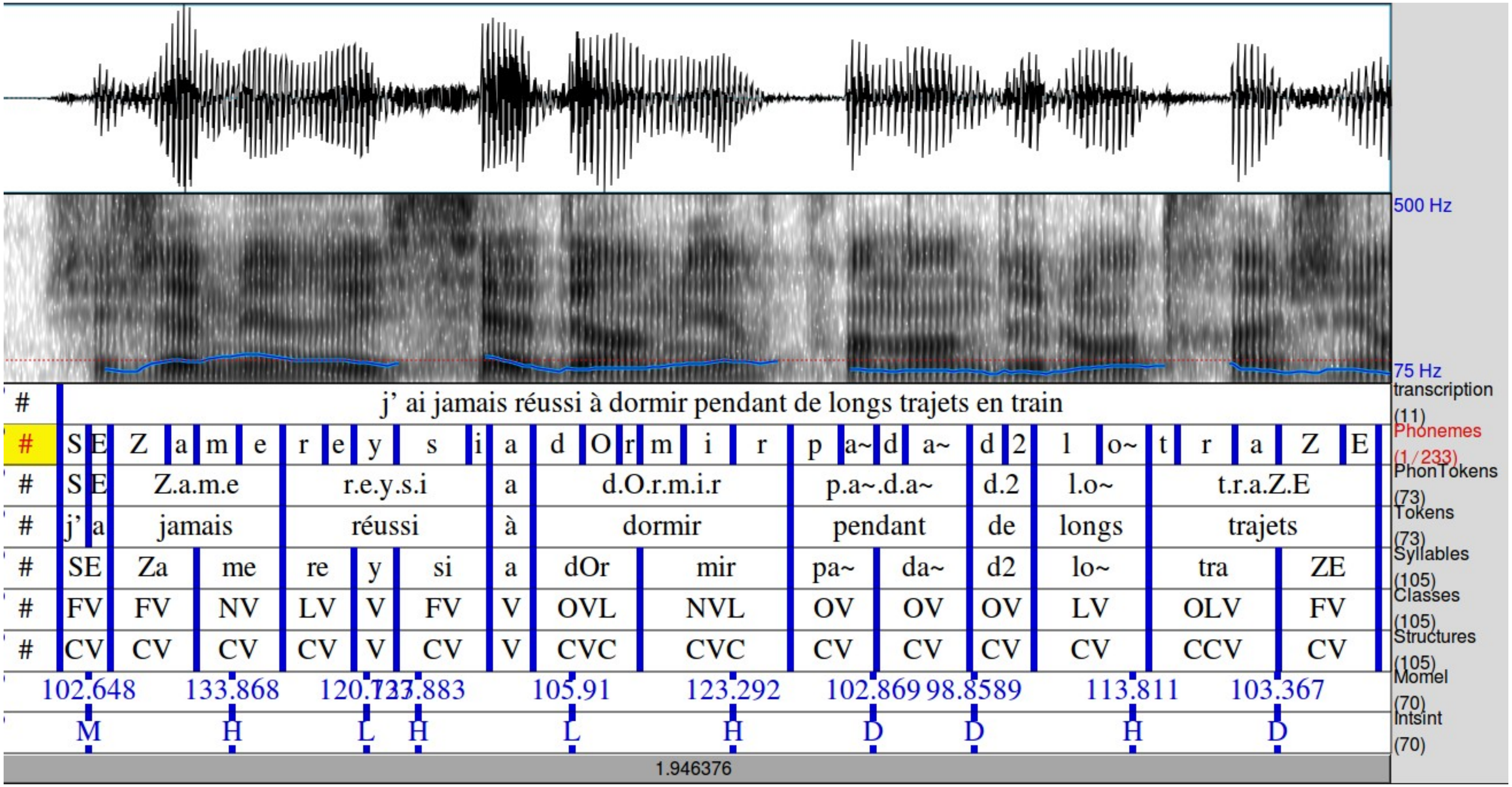
- Development of a Rule-Based System for automatic syllabification of phonemes' strings
- The syllabification is based on 2 principles:
  - a syllable contains a vowel, and only one;
  - a pause is a syllable boundary.



V C C V

Syllabification: example

- Resources (FR and IT):
  - a configuration file with the phoneme set, the classes and all rules



j' ai jamais réussi à dormir pendant de longs trajets en train

S E Z a m e r e y s i a d O r m i r p a~ d a~ d 2 l o~ t r a Z E

S E Z.a.m.e r.e.y.s.i a d.O.r.m.i.r p.a~.d.a~ d.2 l.o~ t.r.a.Z.E

j' a jamais réussi à dormir pendant de longs trajets

SE Za me re y si a dOr mir pa~ da~ d2 lo~ tra ZE

FV FV NV LV V FV V OVL NVL OV OV OV LV OLV FV

CV CV CV CV V CV V CVC CVC CV CV CV CV CCV CV

102.648 133.868 120.723 123.883 105.91 123.292 102.869 98.8589 113.811 103.367

M H L H L H D D H D

1.946376

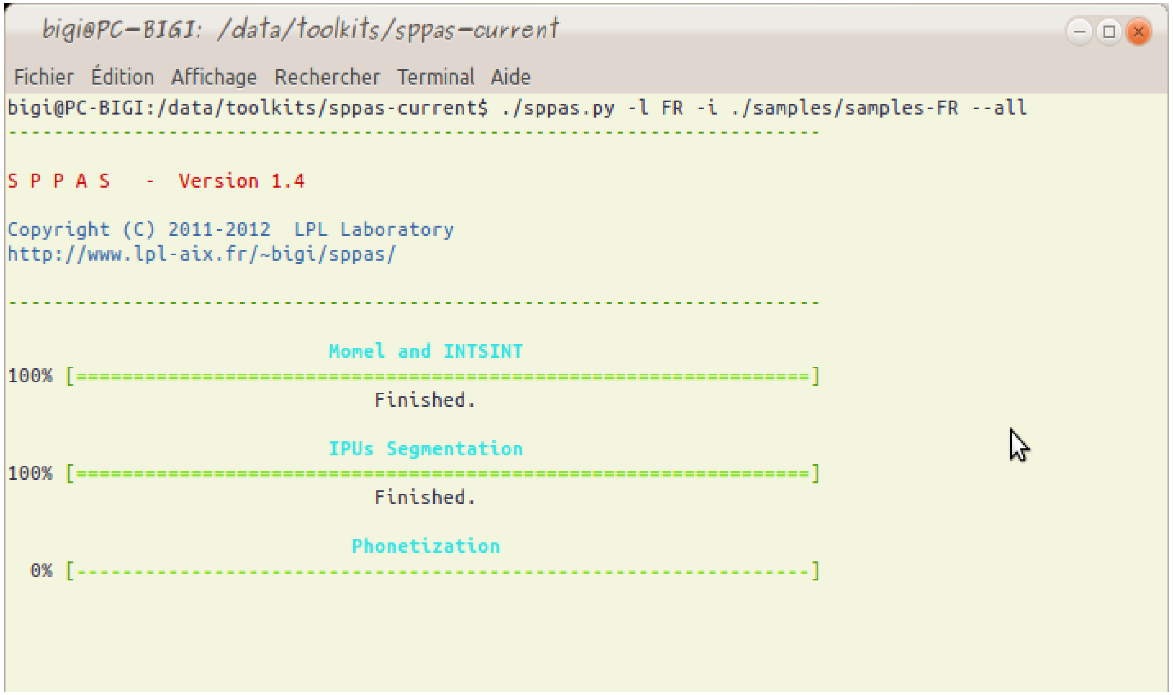
Resources summary

	FR	IT	ZH	EN
Dictionary : Number of entries	350k words and 300k variants	390k words and 5k variants	88k words (350 syllables)	121k words and 10k variants
Acoustic model: Data to train	Triphones - 7h30 CID +30min read	Triphones - 3h30 map-task	Monophones - 90min read	Triphones See voxforge.org

SLDR forge   Evalita 2011   Eurom1   CMU dictionary

Development

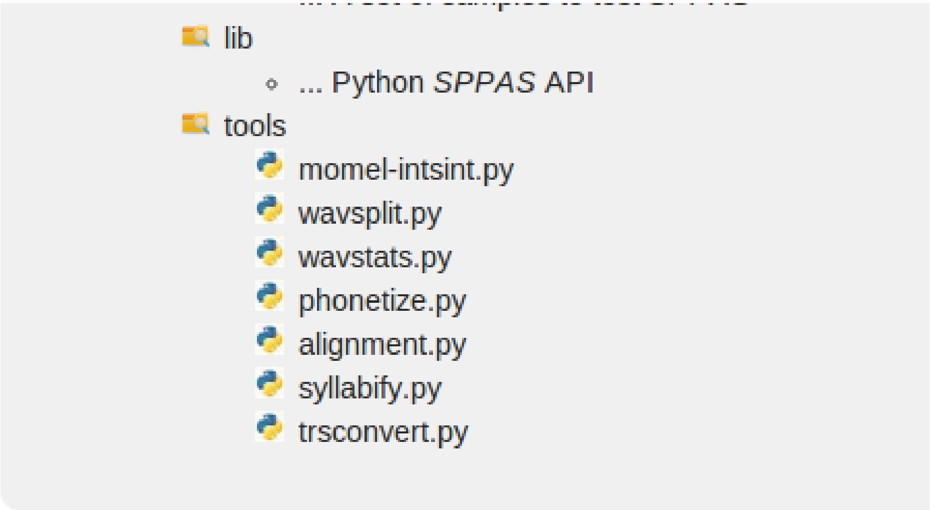
- Based on Python and wxPython (v2.7)
- 21000 lines (25% are comments)
- sppas.py: GUI or Inline usage


bigi@PC-BIGI: /data/toolkits/sppas-current
Fichier Édition Affichage Rechercher Terminal Aide

```
bigi@PC-BIGI:/data/toolkits/sppas-current$ ./sppas.py -l FR -i ./samples/samples-FR --all
-----------------------------------------------------------------------

S P P A S   -  Version 1.4

Copyright (C) 2011-2012  LPL Laboratory
http://www.lpl-aix.fr/~bigi/sppas/

-----------------------------------------------------------------------


                            Momel and INTSINT
100% [================================================================]
                                Finished.

                            IPUs Segmentation
100% [================================================================]
                                Finished.

                              Phonetization
  0% [----------------------------------------------------------------]
```

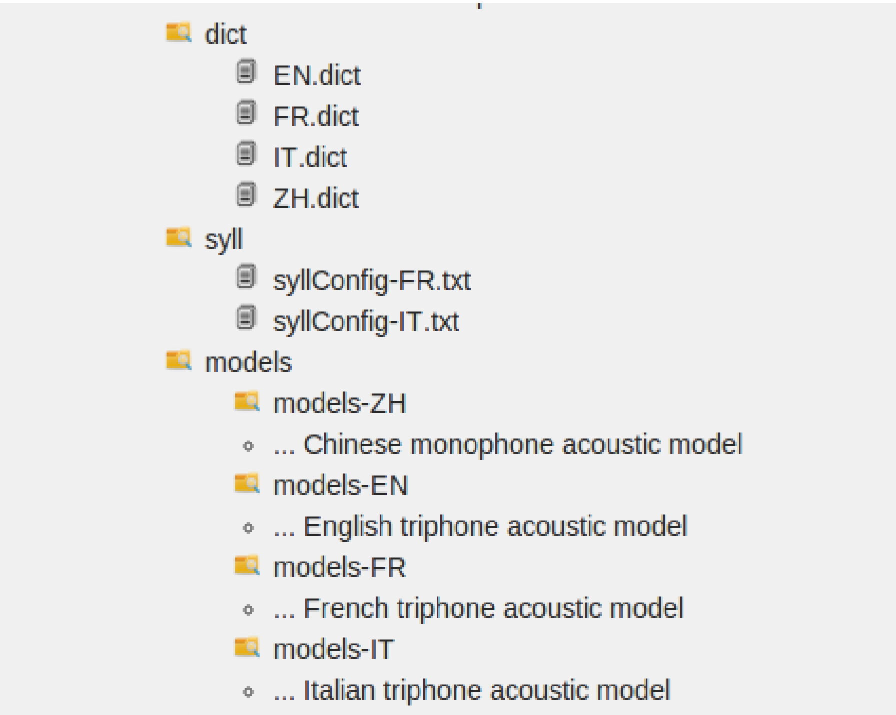
Architecture

- One directory with the API
  - One package per annotation
  - One package to deal with “Tiers”
- A set of inline tools



- lib
  - ... Python *SPPAS* API
- tools
  - momel-intsint.py
  - wavsplit.py
  - wavstats.py
  - phonetize.py
  - alignment.py
  - syllabify.py
  - trsconvert.py

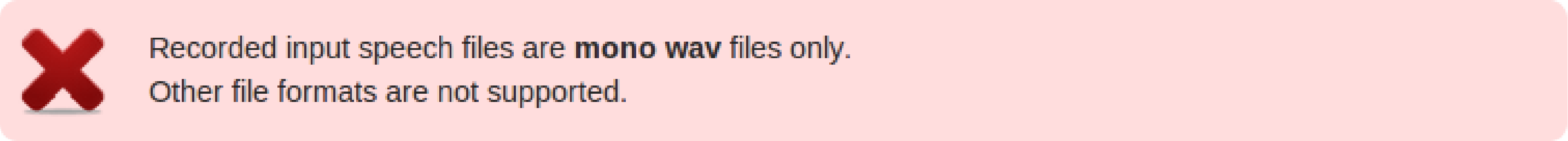
- 3 directories for resources



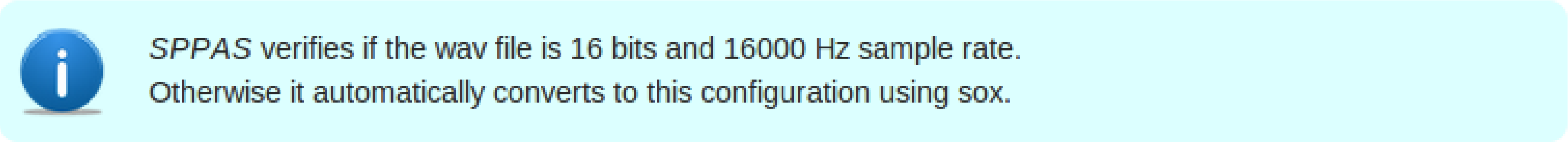
- dict
  - EN.dict
  - FR.dict
  - IT.dict
  - ZH.dict
- syll
  - syllConfig-FR.txt
  - syllConfig-IT.txt
- models
  - models-ZH
    - ... Chinese monophone acoustic model
  - models-EN
    - ... English triphone acoustic model
  - models-FR
    - ... French triphone acoustic model
  - models-IT
    - ... Italian triphone acoustic model

A few words about technical stuff...

- The transcription encoding must correspond to that of SPPAS dictionary:
  - UTF-8 for French, Chinese or Italian,
  - us-ascii for English.
- The transcription and the audio files must have the same name (except for the extension)
- Windows: No spaces or accentuated chars



Recorded input speech files are **mono wav** files only.
Other file formats are not supported.



SPPAS verifies if the wav file is 16 bits and 16000 Hz sample rate.
Otherwise it automatically converts to this configuration using sox.

About

- URL: http://www.lpl-aix.fr/~bigi/sppas/
- SPPAS is still in progress...
    - Suggestions are welcome
    - New resources are welcome
        - Help in this development is also welcome!
- SPPAS can achieve a set of automatic phonetic annotations of speech; results are depending on...
    - The input wav quality
    - The transcription quality...

Orthographic Transcription: which Enrichment is required for Phonetization?

(Brigitte Bigi, Pauline Péri, Roxane Bertrand)

- Hypothesis:
  - The better transcription is:
    - the better phonetization...
    - thus, the better alignment,
    - thus, the better syllabification!
- But... what is a « better » transcription

Transcription:	I	never	get	to	sleep	on	the	airplane
Phonetization:	ay	n.eh.v.e.r	g.eh.t g.ih.t	t.uw t.ix t.ax	s.l.iy.p	aa.n ao.n	dh.ax dh.ah dh.iy	eh.r.p.l.ey.n

Context of this study

- OTIM: Tools for Multimodal Information Processing
  - Http://www.lpl-aix.fr/~otim/
- Aims to develop an annotation scheme and tools for face to face interaction.
- Corpus of Conversational Data

- Transcription of the speech signal is the first annotation.
- How to reflect the orality of speech?

Three different transcriptions

- This study focused on 3 different transcription enrichments

1. TOS: standard orthographic written text
2. TOE1: TOS + the following specific speech phenomena: short pauses, various noises, laughter, filled pauses, truncated words, repeats.
3. TOE2: TOE1 + elisions, particular pronunciations and unusual liaisons

- Evaluations compare phonetizations obtained from automatic systems to a reference phonetized manually

Test corpus: MARC-Fr

- The corpus was transcribed using the three transcriptions.
- In parallel, it was manually phonetized by an expert.
- Freely available: http://www.sldr.fr
- Made of parts of three different French corpora:
    - CID - Corpus of Interactional Data
    - AixOx - read speech
    - Grenelle – political debate
- About 7 minutes altogether

Test corpus description

	CID	AixOx	Grenelle
Duration	143s	137s	134s
Nb speakers	12	4	1
Nb Phonemes	1876	1744	1781
Nb Tokens	1269	1059	550
Silent Pauses	10	23	28
Hesitations	21	0	5
Noise, breath...	0	8	0
Laughts	4	0	0
Truncations	6	2	1
Elisions	60	21	43
Special pron.	58	37	23

TOE1

TOE2

Automatic Phonetization

- There are two general ways to construct a phonetization process. We experimented:
  - SPPAS: dictionary based solutions which consist in storing a maximum of phonological knowledge in a lexicon. Phonetic variants are choose by the aligner.
    - Dictionary: 350k words, 300k variants
    - Acoustic model trained from 8h of speech
  - LIA_Phon: rule based systems, with rules based on inference approaches or proposed by expert linguists.
    - Without phonetic variants: a POS-tagger is used to disambiguate pronunciations.
    - Standard liaisons

LIA_Phon + TOE?

- LIA_Phon was conceived to take as input a standard orthographic transcription. The pronunciation is supposed to correspond to a standard French.
- We proposed a tree-based approach to use LIA_Phon with an enriched transcription as input

Results

- Evaluations were carried out with Sclite:
    - accuracy is calculated as a function of phonemes, by estimating the sum (Err) of the following errors: Substitution (sub), Deletion (del), Insertion (ins)

• 3 transcription enrichments	TOS, TOE1, TOE2
X 3 corpus types	CID, AixOx, Grenelle
X 3 systems	SPPAS, LIA_Phon, Tree-
X 4 values per result	err, sub, del, ins

= too many results for this presentation!

Results

	LIA_Phon
	Err %
CID	
TOS	**17.3**
TOE1	**14.4**
TOE2	**6.5**
AixOx	
TOS	**9.5**
TOE1	**6.5**
TOE2	**5.6**
Grenelle	
TOS	**8.0**
TOE1	**6.3**
TOE2	**4.0**

Results

	LIA_Phon	Tree-based + LIA_Phon
	Err %	Err %
CID		
TOS	17.3	
TOE1	14.4	
TOE2	6.5	5.6
AixOx		
TOS	9.5	
TOE1	6.5	
TOE2	5.6	5.2
Grenelle		
TOS	8.0	
TOE1	6.3	
TOE2	4.0	3.7

Other results...

French only system

Room for Improvements: Dict/Model

Language independent algorithms

	LIA_Phon:				SPPAS:
	Sub	Del	Ins	Err	Err
CID					
TOS	2.8	4.5	10.0	17.3	
TOE1	2.7	1.4	10.3	14.4	12.5
TOE2	1.8	1.3	3.4	6.5	
AixOx					
TOS	1.4	5.0	3.0	9.5	
TOE1	1.4	2.3	2.9	6.5	8.2
TOE2	1.3	1.8	2.5	5.6	
Grenelle					
TOS	1.1	2.8	4.1	8.0	
TOE1	1.0	1.2	4.1	6.3	7.2
TOE2	1.3	1.0	1.7	4.0	

Conclusion

- We showed how transcription can impact on the performances of automatic phonetization
- Evaluations were carried out on 3 different types of speech
- We proposed a solution to improve the rule-based system which obtained a phonetization of about 95.2% correct:
    - from 3.7% to 5.6% error rates depending on the corpus
- *Orthographic transcription .... which *manual* enrichment is required for *automatic* phonetization?*
    - Although if the transcription enrichment is more time consuming, it constitutes therefore an effective alternative to phonetize all corpus types

References

B. Bigi, C. Meunier, I. Nesterenko, R. Bertrand. *Automatic detection of syllable boundaries in spontaneous speech.* Language Resource and Evaluation Conference (LREC), pp 3285-3292, La Valetta, Malte, 2010.

B. Bigi. *The SPPAS participation to Evalita 2011.* Working Notes of EVALITA 2011, Rome, Italy, ISSN: 2240-5186, January 2012.

S. Herment, A. Loukina, A. Tortel, D. Hirst, B. Bigi. *A multi-layered learners corpus: automatic annotation.* 4th International Conference on Corpus Linguistics, Jaèn, Spain, March 2012.

B. Bigi, D. Hirst. *SPeech Phonetization Alignment and Syllabification: a tool for the automatic analysis of speech prosody.* Speech Prosody, Shanghai, China, May 2012, Accepted.

B. Bigi, P. Péri, R. Bertrand. *Orthographic Transcription: which Enrichment is required for phonetization?* Language Resource and Evaluation Conference, Istanbul, Turkey, May 2012. Accepted.

B. Bigi. *SPPAS: a tool for SPeech Phonetization Alignment and Syllabification.* Language Resource and Evaluation Conference (LREC), Istanbul, Turkey, May 2012. Accepted.

B. Bigi. *Forced Alignment on Spontaneous Speech for Italian: the SPPAS tool.* B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 312--321. Springer, Heidelberg.

http://www.lpl-aix.fr/~bigi/sppas/