OTIM project

Primary data :

Transcription, Phonetization,Alignment

Part 1

Robert Espesser

23 mai 2011, Aix-enProvence

Å Corpus involved : CID

Å Enriched Orthographic Transcription (TOE)

Å Phoneme alignment

Å Evaluation of the alignment

Å Descriptive data about phonetic (and non phonetic) phenomena (elision, overlap ....)

Conclusion

CID: Corpus of Interactional Data

(Bertrand & al, 2008)

Å 8 dialogs, ~ 1 hour /dialog

Å 1 channel /speaker (head-mounted microphone)

Å recorded in a sound booth

Å speakers from southeastern France or long-term residents

Pre-segmentation of the speech signal

Å Inter Pausal Unit segmentation (silent pause >= 200 ms)

~ 13000 IPUs

median: 1390 ms quartiles: 600, 2770 ms

Å Manual transcription (Praat)

Enriched Orthographic Transcription (TOE)

Transcription Orthographique Enrichie (TOE): why ?

Å Available speech tools designed for standard (read) French

Å Results on uncontrolled speech are likely to be unreliable

Å Extent of the difference between the 2 styles is unknown.

↳ transcription of a maximum of information

Å get data on the oral phenomena (frequency, patterns)

Å Improve the performance of the speech tools involved
to get an acceptable phonem alignment.

TOE main conventions

Derived from the works of GARS (Blanche-Benveniste, 1987)

Å Laugh	il est @ parti loin	
Å Laughing speech	il est @@ parti loin @@	
Å Elision	p(e)tit	/pti/
Å Truncated word	s- c’est non	/s se no~/
Å unexpected liaison	les =z= haricots	/lezaRiko/
Å Non-standard realization		
ï assimilation	[je sais pas, Sepa]	/Sepa/
ï realization of final schwa (southern French)		
	le [verre, veR2]	l2 veR2
[…]		

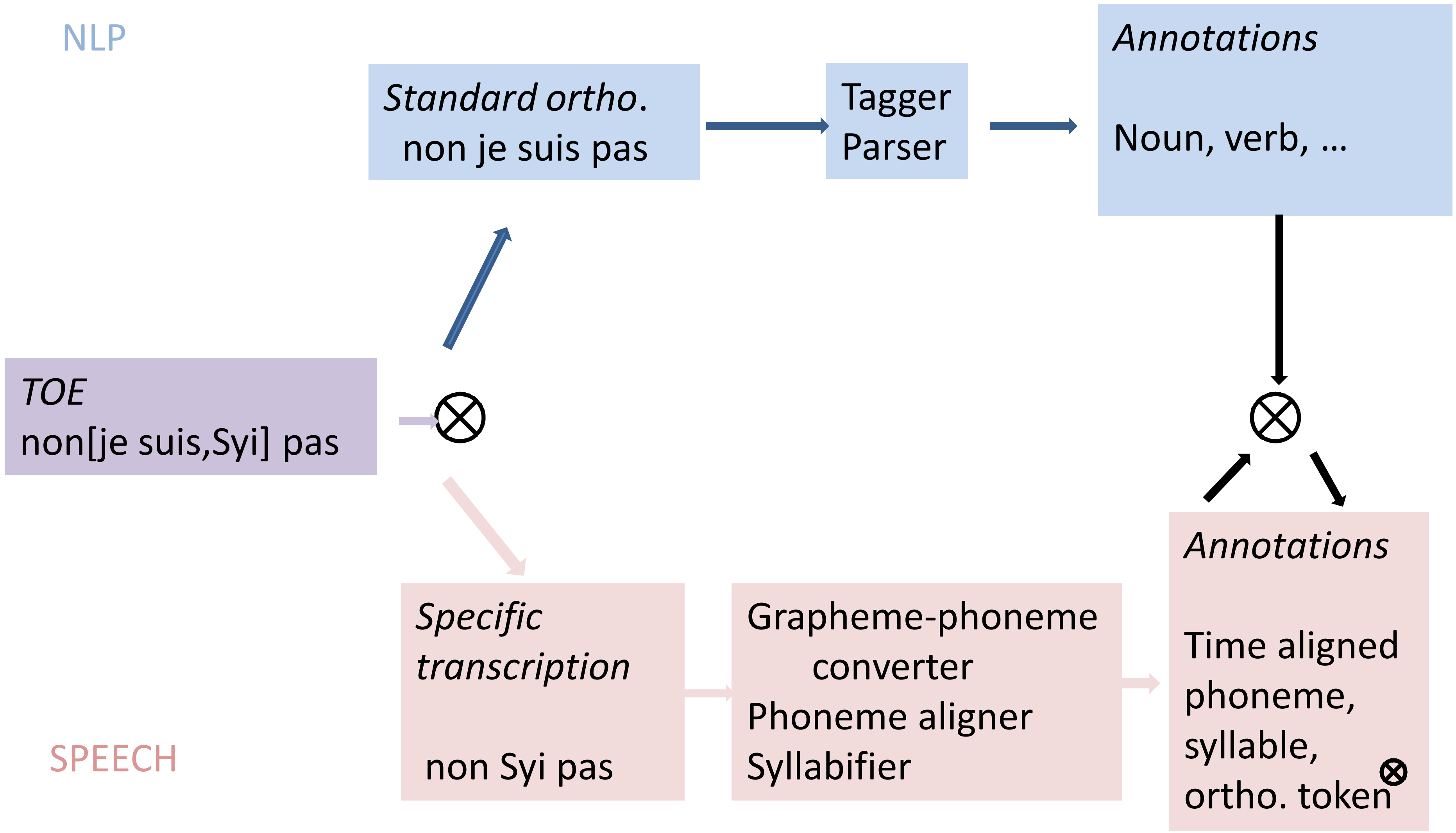
TOE main instructions

Annotators were instructed to:

- Favor elision notation, e.g. p(e)tit NOT [petit, pti]
- In case of doubt: orthographic transcription

- Not to use spectrogram ….
- Avoid attending to fine-grained detail (if possible…)

Structure of automatic processing



NLP

Standard ortho.
non je suis pas

Tagger
Parser

Annotations

Noun, verb, …

TOE
non[je suis,Syi] pas

⊗

⊗

Specific transcription

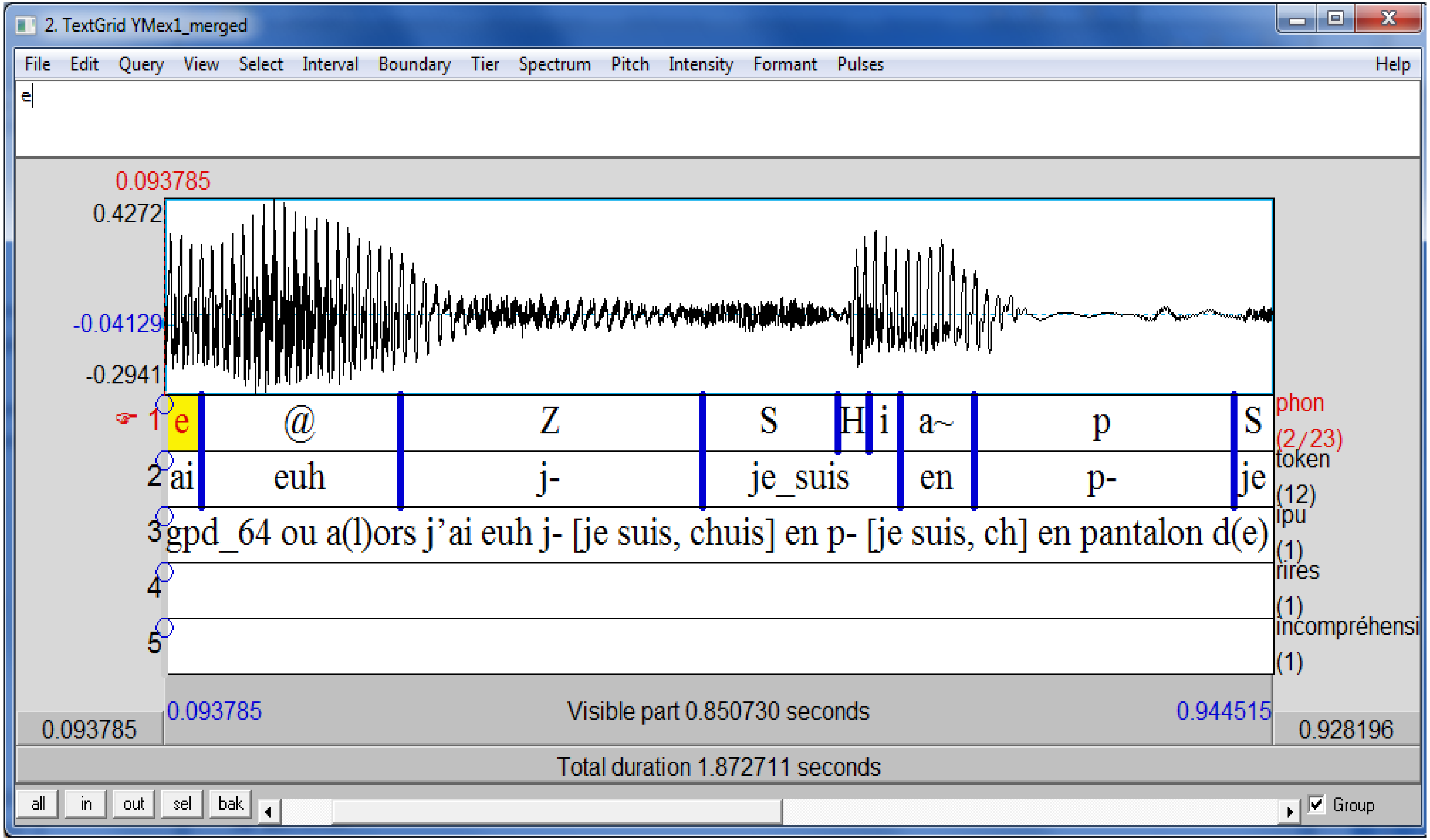
non Syi pas

Grapheme-phoneme converter
Phoneme aligner
Syllabifier

Annotations

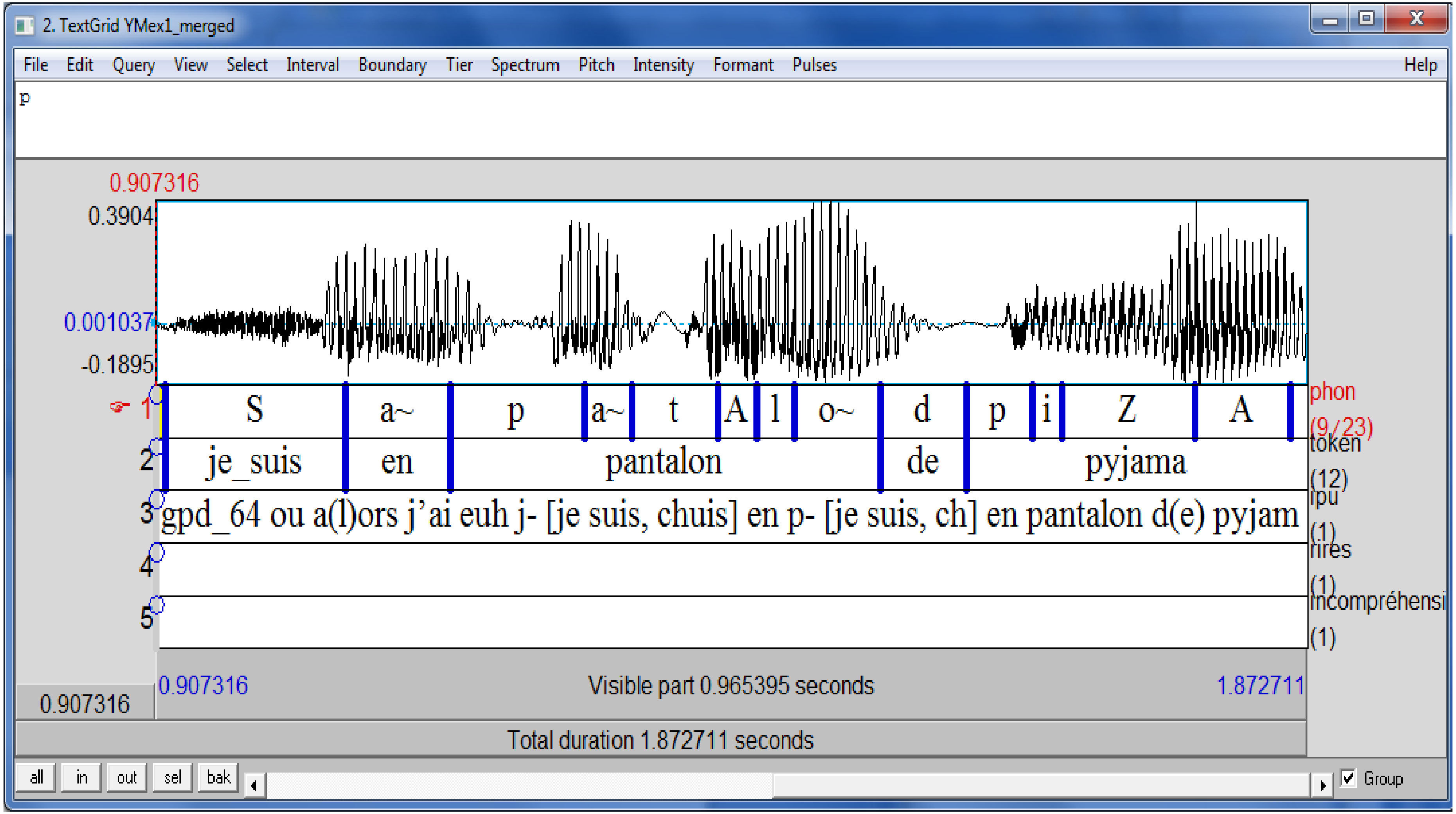
Time aligned phoneme, syllable, ortho. token ⊗

SPEECH



Å Ex1a Time aligned phonemes , orthographic tokens

YM_ex1.wav


2. TextGrid YMex1_merged
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
p
0.907316
0.3904
0.001037
-0.1895
0.907316
Visible part 0.965395 seconds
1.872711
0.907316
Total duration 1.872711 seconds
all in out sel bak
☑ Group

1 S | a~ | p | a~ | t | A | l | o~ | d | p | i | Z | A — phon (9/23)

2 je_suis | en | pantalon | de | pyjama — token (12)

3 gpd_64 ou a(l)ors j'ai euh j- [je suis, chuis] en p- [je suis, ch] en pantalon d(e) pyjam — ipu (1)

4 — rires (1)

5 — incompréhensi (1)

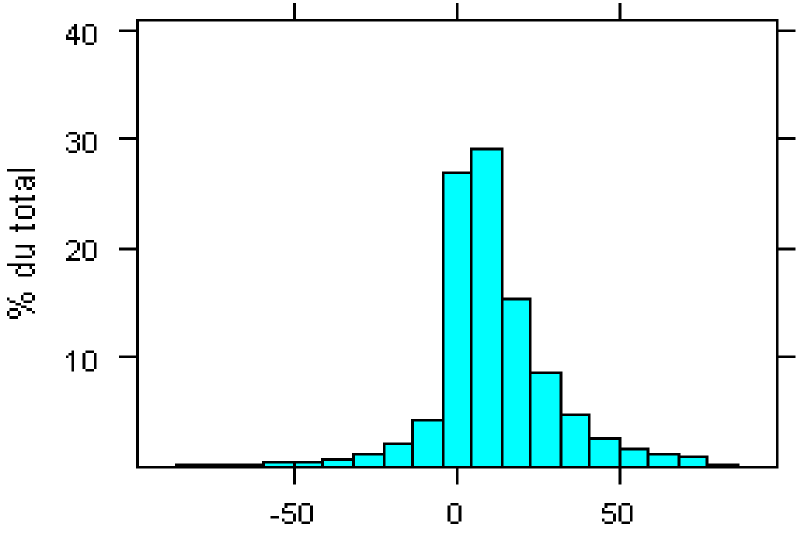
Å Ex1b

YM_ex1.wav

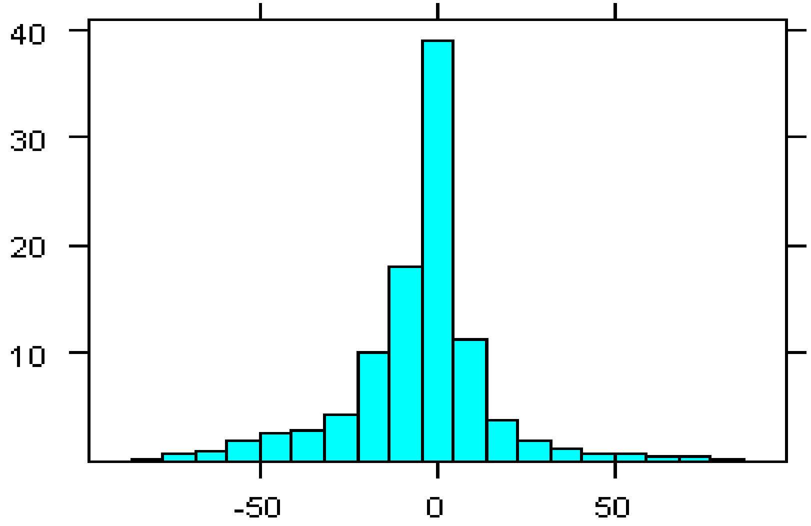
Alignment evaluation

Å 2 speakers (1 male, 1 female)

Å ~13000 vowels corrected



v.begin gap auto – manual (ms)



v.end gap auto – manual (ms)

Vowel duration underestimated: 14 ms (median)

(auto – manual)	v. begin (ms)	v.end (ms)	midpoint(ms)
Median	9	0	3
| auto – manual| 3rd Quart.	20	23	16

Alignment evaluation

Å 7 macro-classes of oral vowels:

A(A,a) e(E,e) o(O,o) @(2,9,@) i y u

Å 4378 ”automatic” vowels [30,300] ms

5367 ”manual” vowels

Å 3 formants estimated at the midpoint (ESPS, standard parameters)

Å F1, F2, F3: Manual vs Auto segmentation

ï insignificant differences or < 0.2 Bark

ï Formant value variability very similar

Difference limen discriminating formants = 0.28 Bark

(D. Kewley-Port, Y. Zheng 1999)

Truncated words

Å 1730 items

Å 455 patterns

Å The 18 most frequent patterns (> 1%) = 50% of the items

Å /i/ /i/ /i/ /va/ /parle/:

1) i- i- i(l) va parler

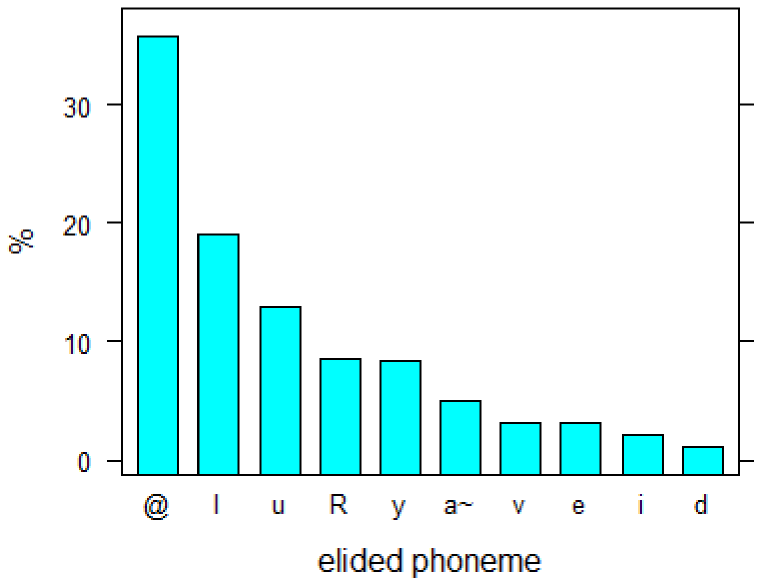
2) i(l) i(l) i(l) va parler

Elision

~ 11000 elided phonemes
(3.6 % of 302,000 phonemes)

187 patterns

The 10 patterns with frequency > 1%
= 88% of the elided phonems



Non-standard phonetic realizations

Å 2810 items , 1300 patterns

[je , S] :	7.7 %	~ 17 %
[je sais, Se] :	6 %	
[je suis, SHi] :	2.9 %	
[je suis, Sy] :	0.9 %	

% items	#occurrence	
37	1	(half = 520 items = final schwas)
5	2	
1.6	3	

~50 % [ ] could be automatically processed

(Final schwas + 4 most freq. patterns)

LAUGHS

Å 2111 laughing sequences

Å 367 speech laughing sequences

Å 844 single laughing sequence (IPU without speech)

~ 16% of the 13000 alignable IPUs contain (at least) one laughing sequence

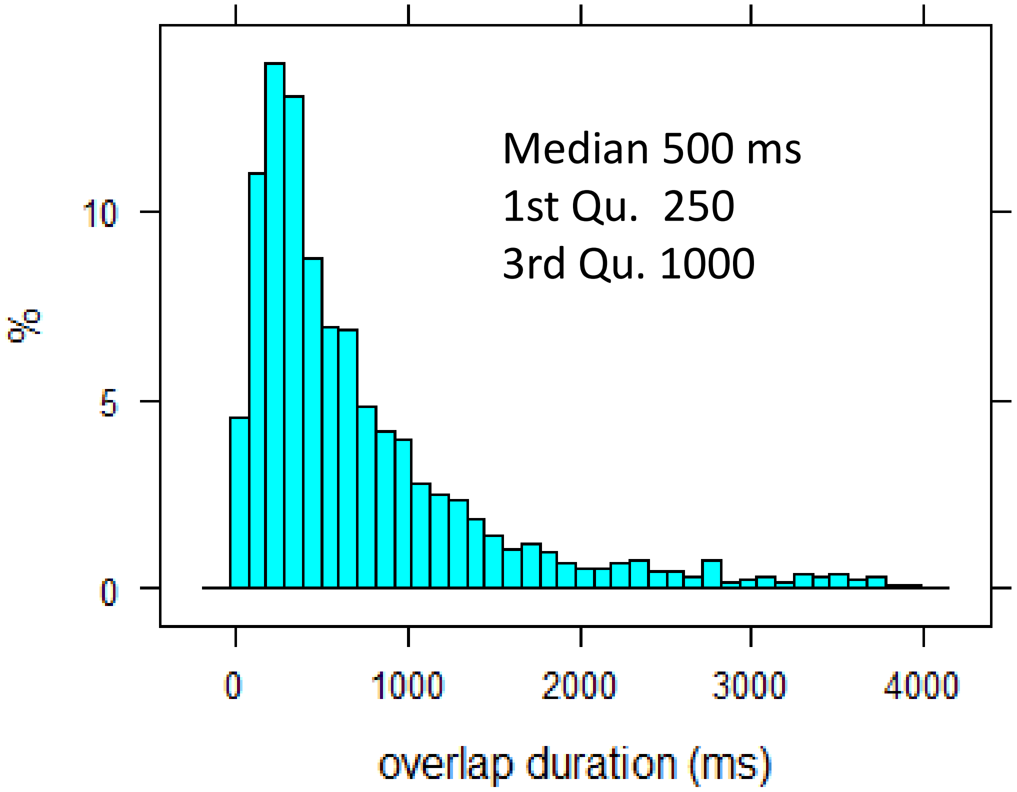
overlaps

4753 overlaps (ipu overlapping)

12.6% <= 150 ms
(min value for overlapping ?)

6% <= 80 ms

63 % of the ~13000 IPUs
are involved in an
overlap (>150ms)



Conclusion

1) Enriched orthographic transcription
+ simple pre- and post-processing
+ standard speech processing tools

➔ **Some phonetic analyses (at vowel- or syllable-level) are possible on a ″large ″ corpus of very uncontrolled conversational speech(*)**

2) **TOE may be simplified** :
reducing human work transcription , depending more on the abilities of the automatic aligner .
e.g., for standard elisions & liaisons, final schwas (?)

3) **Enhancement of the grapheme-to-phoneme process**

4) **Enhancement of the alignment tool (new acoustic models..)**

(*) Meunier C. & Espesser R. Vowel reduction in conversational speech in French: The role of lexical factors. Journal of Phonetics (2011) (in press, already published online)