The Pauses & Lexical Stress Processing Pipeline (PLSPP)

Source code: https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp

Sylvain Coulange

Univ. Grenoble Alpes, Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), Grenoble, France
Univ. Grenoble Alpes, CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), Grenoble, France
Doshisha University, Spoken Language Processing Laboratory (SLPL), 610-0394 Kyoto, Japan

sylvain.coulange@univ-grenoble-alpes.fr

This research is supported by JSPS Kakenhi n°23H00648

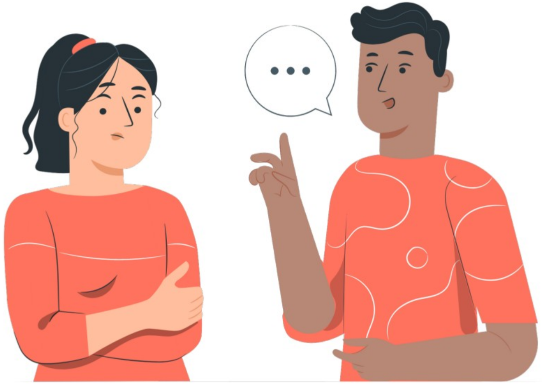
financé par IDEX Université Grenoble Alpes

Assessing L2 pronunciation: From nativelikeness to intelligibility

Native speaker as a target → Be (easily) understood

“Intelligibility”

“Comprehensibility”



Isaacs, T., Trofimovich, P., and Foote, J. A. (2018) Developing a user-oriented L2 comprehensibility scale for english-medium universities. Language Testing 35(2), 193–216.
Jenkins, J., Baker, W., & Dewey, M. (Eds.). (2017) The Routledge Handbook of English as a Lingua Franca (1st ed.). Routledge.
Frost, D., O’Donnell, J. (2018) Evaluating the essentials, the place of prosody in oral production. In J. Volín (ed.). Pronunciation of EFL.
Council of Europe (2020) Common European framework of reference for languages. Strasbourg, France.
Walker, R., Low, E., & Setter, J. (2021) English pronunciation for a global world. Oxford: Oxford University Press

Assessing L2 pronunciation: From nativelikeness to intelligibility

Rhythm

Beats

Speech flow

Parameters related to L2 English comprehensibility:

- Hesitation markers position (pauses, false starts, repetitions…)
- Lexical stress (presence, position, quality)
- Speech rate (not too fast, not too slow)
- Pitch variation (make the speech sound lively and engaging)
- Phonemes quality (depending on their functional load)

Isaacs, T., Trofimovich, P., and Foote, J. A. (2018) Developing a user-oriented L2 comprehensibility scale for english-medium universities. Language Testing 35(2), 193–216.
Jenkins, J., Baker, W., & Dewey, M. (Eds.). (2017) The Routledge Handbook of English as a Lingua Franca (1st ed.). Routledge.
Frost, D., O'Donnell, J. (2018) Evaluating the essentials, the place of prosody in oral production. In J. Volín (ed.). Pronunciation of EFL.
Council of Europe (2020) Common European framework of reference for languages. Strasbourg, France.
Walker, R., Low, E., & Setter, J. (2021) English pronunciation for a global world. Oxford: Oxford University Press

Assessing L2 pronunciation: From nativelikeness to intelligibility

Rhythm

Beats

Speech flow

Parameters related to L2 English comprehensibility:

- Hesitation markers position (pauses, false starts, repetitions…)
- Lexical stress (presence, position, quality)

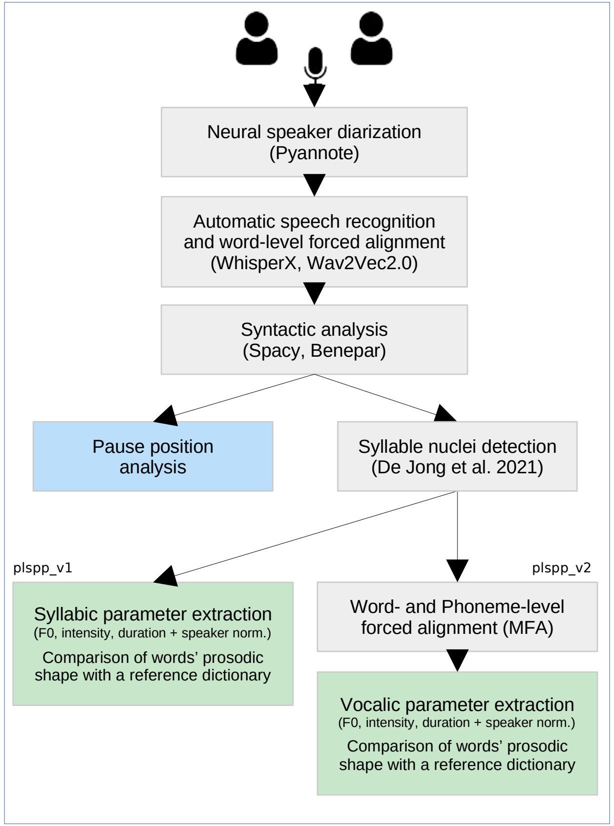
PhD

Université Grenoble Alpes (France) - 3rd year

Doshisha University (Japan)

Semi-automatic diagnosis of spontaneous English as a foreign language: the role of rhythm in speaker comprehensibility

The Pauses & Lexical Stress Processing Pipeline (PLSPP)



Neural speaker diarization (Pyannote)

Automatic speech recognition and word-level forced alignment (WhisperX, Wav2Vec2.0)

Syntactic analysis (Spacy, Benepar)

Pause position analysis

Syllable nuclei detection (De Jong et al. 2021)

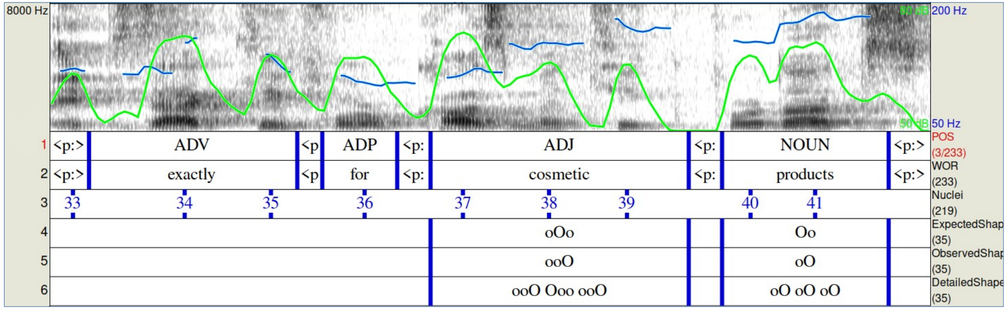
plspp_v1

Syllabic parameter extraction
(F0, intensity, duration + speaker norm.)
Comparison of words' prosodic shape with a reference dictionary

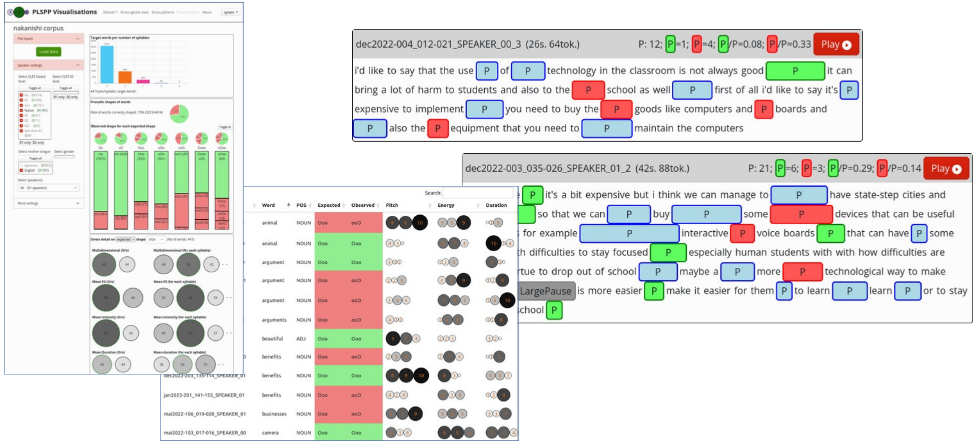
plspp_v2

Word- and Phoneme-level forced alignment (MFA)

Vocalic parameter extraction
(F0, intensity, duration + speaker norm.)
Comparison of words' prosodic shape with a reference dictionary



1	<p:>	ADV	<p	ADP	<p:	ADJ	<p:	NOUN	<p:>	POS (3/233)
2	<p:>	exactly	<p	for	<p:	cosmetic	<p:	products	<p:>	WOR (233)
3	33	34 35		36	37	38 39		40 41		Nuclei (219)
4						oOo		Oo		ExpectedShap (35)
5						ooO		oO		ObservedShap (35)
6						ooO Ooo ooO		oO oO oO		DetailedShape (35)



PLSPP Visualisations
nakanishi corpus
File inputs
Load data
Speaker settings
Target words per number of syllables
Prosodic shapes of words
Observed shape for each expected shape
(Nb of words: 447)
Multidimensional (for each syllable)
Mean F0 (for each syllable)
Mean intensity (for each syllable)
Mean duration (for each syllable)

Search:

	Word	POS	Expected	Observed	Pitch	Energy	Duration
	animal	NOUN	Ooo	ooO	8 8 10	5 5 9	1 5
	animal	NOUN	Ooo	Ooo	4 3	7 4	10 1 4
	argument	NOUN	Ooo	Ooo	3 1 1	1 3 5	7 3
	argument	NOUN	Ooo	ooO	2 2 5	4 9	1 2 8
	argument	NOUN	Ooo	ooO	3 5 4	7 5	1 5 10
	arguments	NOUN	Ooo	ooO	4 1 1	1	1 1 8
	beautiful	ADJ	Ooo	Ooo	9 7 4	2 2 3	3 2
	benefits	NOUN	Ooo	ooO	2 5	2 7 4	1 7
dec2022-203_135-114_SPEAKER_01	benefits	NOUN	Ooo	Ooo	8 9 10	8 3	5 5 3
jan2023-201_141-155_SPEAKER_01	benefits	NOUN	Ooo	ooO	4 2 4	3 5	1 3 8
mai2022-106_019-020_SPEAKER_01	businesses	NOUN	Ooo	ooO	7 9	5 5	3 3 7
mai2022-103_017-016_SPEAKER_00	camera	NOUN	Ooo	Ooo	3 4	8 8	7 7 4

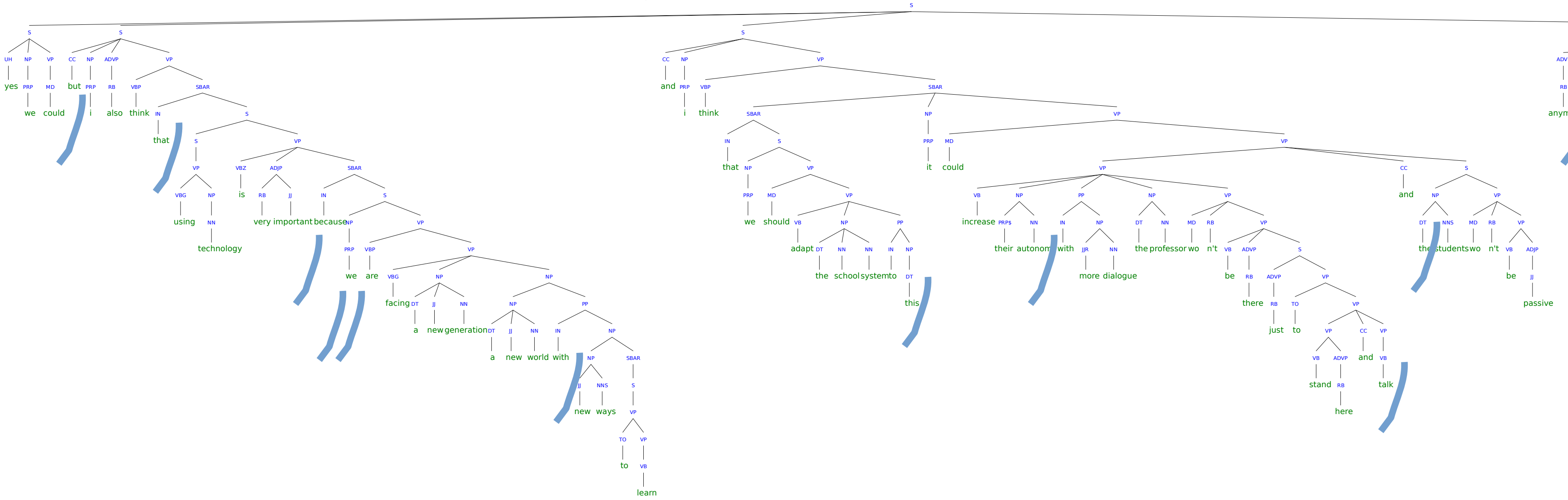
dec2022-004_012-021_SPEAKER_00_3 (26s. 64tok.) P: 12; P=1; P=4; P/P=0.08; P/P=0.33 Play

i'd like to say that the use P of P technology in the classroom is not always good P it can bring a lot of harm to students and also to the P school as well P first of all i'd like to say it's P expensive to implement P you need to buy the P goods like computers and P boards and P also the P equipment that you need to P maintain the computers

dec2022-003_035-026_SPEAKER_01_2 (42s. 88tok.) P: 21; P=6; P=3; P/P=0.29; P/P=0.14 Play

P it's a bit expensive but i think we can manage to P have state-step cities and so that we can P buy P some P devices that can be useful for example P interactive P voice boards P that can have P some th difficulties to stay focused P especially human students with with how difficulties are rtue to drop out of school P maybe a P more P technological way to make LargePause is more easier P make it easier for them P to learn P learn P or to stay school P

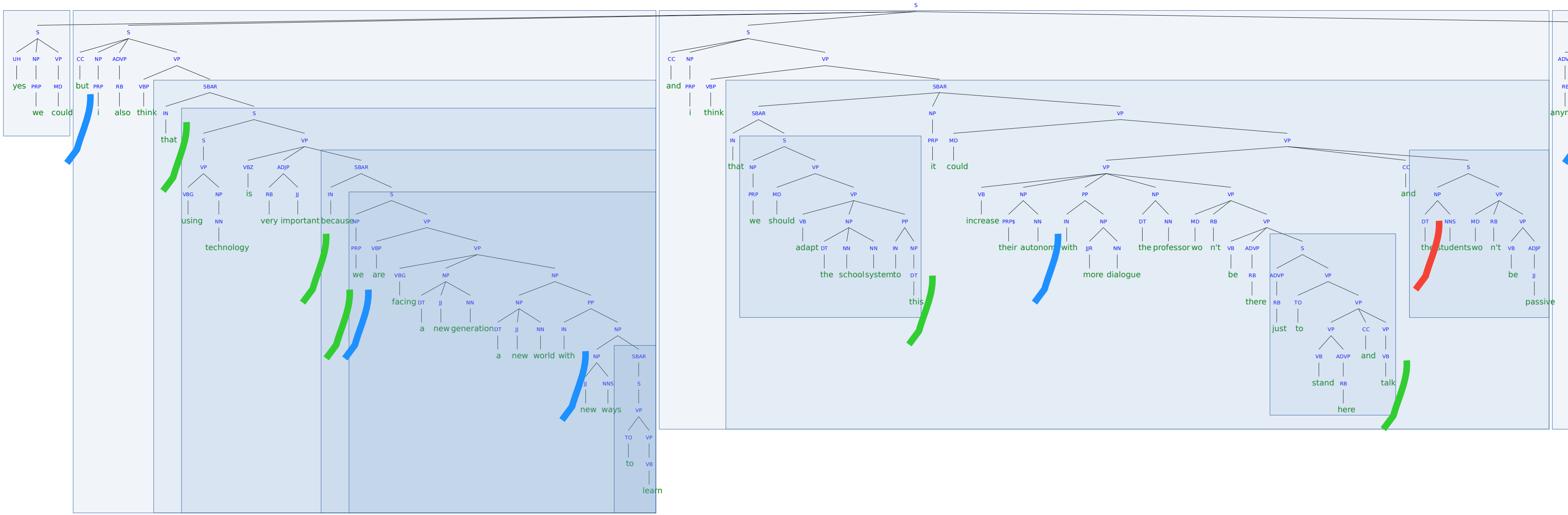
Pauses processing


S
yes we could
but i also think that using technology is very important because we are facing a new generation a new world with new ways to learn
and i think that we should adapt the school system to this it could increase their autonom with more dialogue the professor wo n't be there just to stand here and talk and the students wo n't be passive
anym

Customisable fixed duration threshold (here 180ms-2s)

file: dec2022-003_039-040_SPEAKER_01_5
Speaker total speech duration: 6’33’’

Pauses processing


yes we could
but i also think that using technology is very important because we are facing a new generation a new world with new ways to learn
and i think that we should adapt the school system to this it could increase their autonom with more dialogue the professor wo n't be there just to stand here and talk and the students wo n't be passive anyr

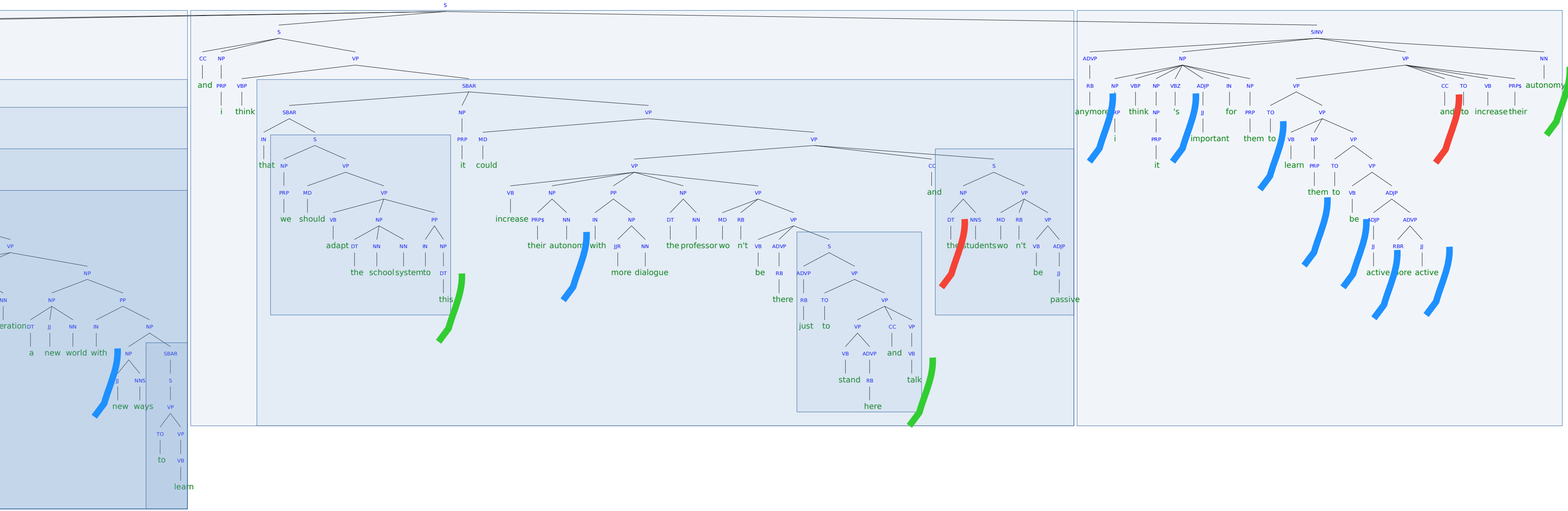
3 categories:

- Pauses between clauses
- Pauses between phrases
- Pauses within phrases

Customisable fixed duration threshold (here 180ms-2s)

file: dec2022-003_039-040_SPEAKER_01_5
Speaker total speech duration: 6’33’’

Pauses processing



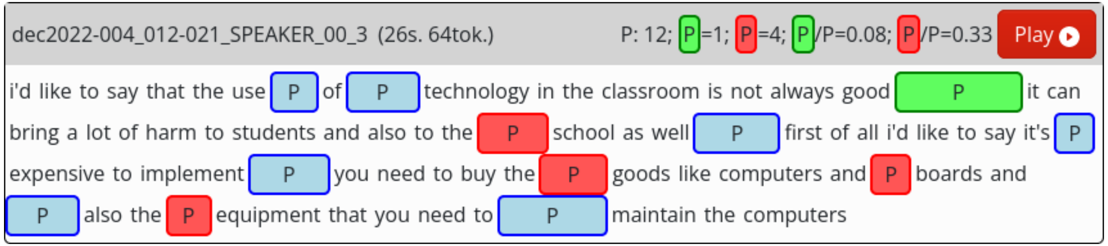
3 categories:

- Pauses between clauses
- Pauses between phrases
- Pauses within phrases

Customisable fixed duration threshold (here 180ms-2s)

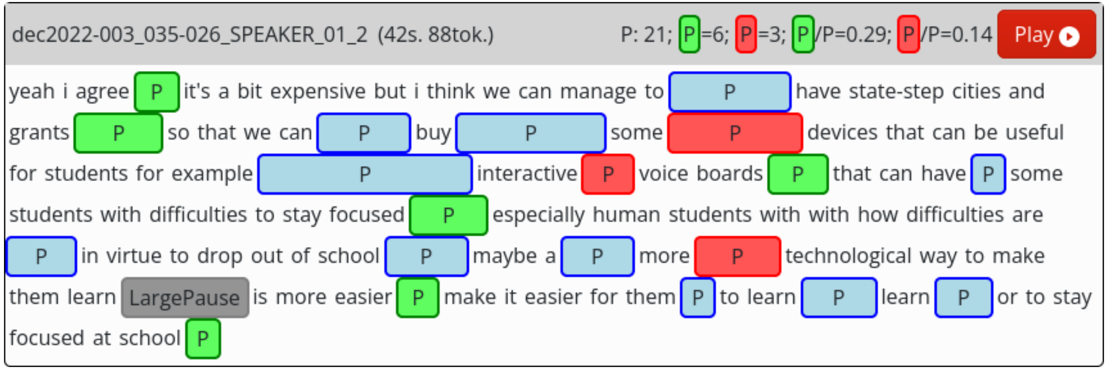
file: dec2022-003_039-040_SPEAKER_01_5
Speaker total speech duration: 6’33’’

Pauses processing



dec2022-004_012-021_SPEAKER_00_3 (26s. 64tok.) P: 12; P=1; P=4; P/P=0.08; P/P=0.33 Play

i'd like to say that the use P of P technology in the classroom is not always good P it can bring a lot of harm to students and also to the P school as well P first of all i'd like to say it's P expensive to implement P you need to buy the P goods like computers and P boards and P also the P equipment that you need to P maintain the computers

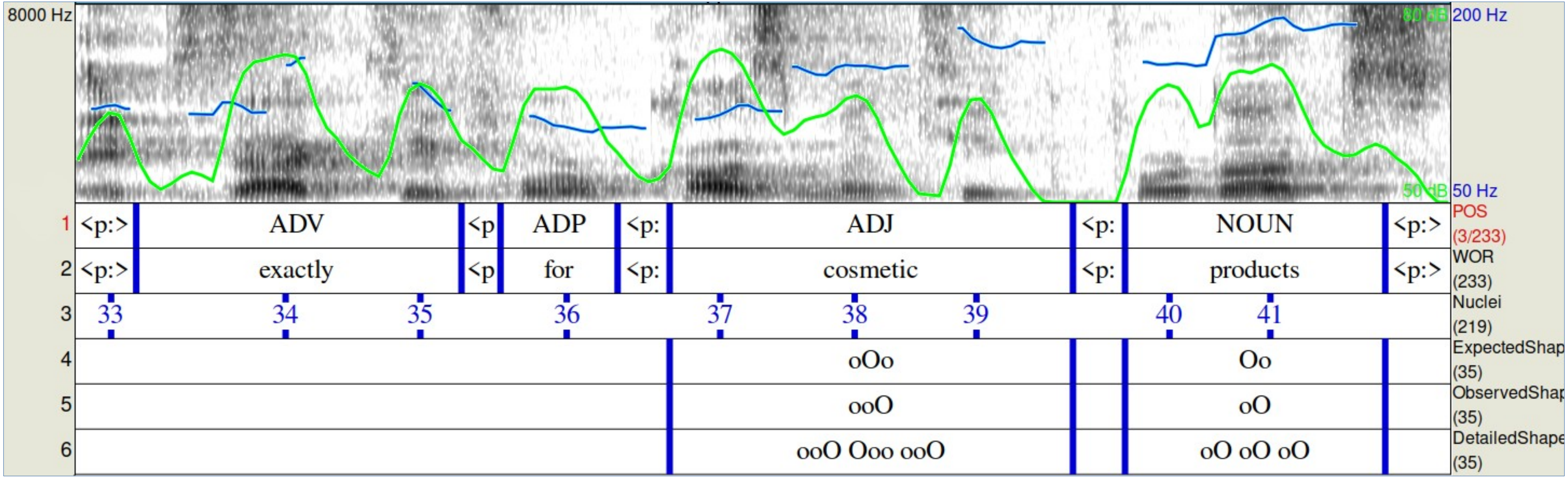


dec2022-003_035-026_SPEAKER_01_2 (42s. 88tok.) P: 21; P=6; P=3; P/P=0.29; P/P=0.14 Play

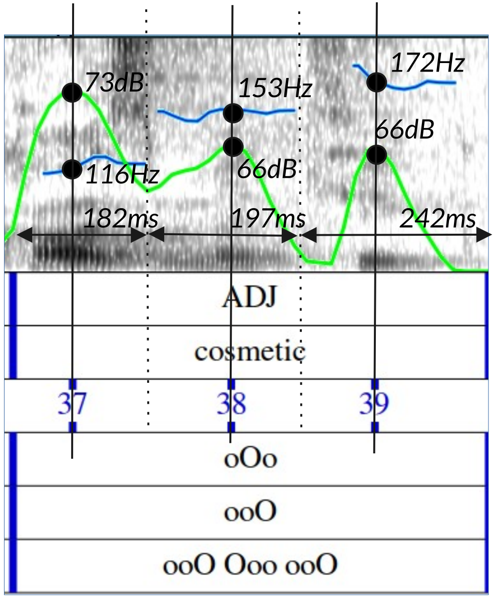
yeah i agree P it's a bit expensive but i think we can manage to P have state-step cities and grants P so that we can P buy P some P devices that can be useful for students for example P interactive P voice boards P that can have P some students with difficulties to stay focused P especially human students with with how difficulties are P in virtue to drop out of school P maybe a P more P technological way to make them learn LargePause is more easier P make it easier for them P to learn P learn P or to stay focused at school P

Online example

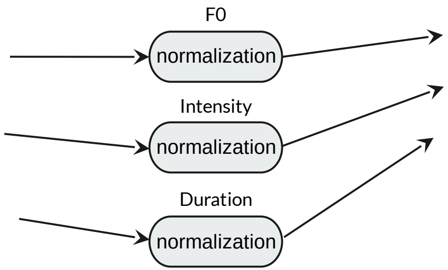
Stress processing (PLSPP v1)

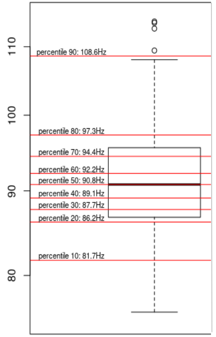

8000 Hz
80 dB
200 Hz
50 dB
50 Hz
1 <p:> ADV <p ADP <p: ADJ <p: NOUN <p:> POS (3/233)
2 <p:> exactly <p for <p: cosmetic <p: products <p:> WOR (233)
3 33 34 35 36 37 38 39 40 41 Nuclei (219)
4 oOo Oo ExpectedShap (35)
5 ooO oO ObservedShap (35)
6 ooO Ooo ooO oO oO oO DetailedShape (35)

Stress processing (PLSPP v1)



(F0, Intensity, Duration)





Percentiles
of the speaker's
distribution

	37	38	39
F0:	(19),	(80),	(90)
dB:	(95),	(28),	(28)
dur:	(55),	(61),	(78)

⇩

Observed pattern

(56), (56), (65)

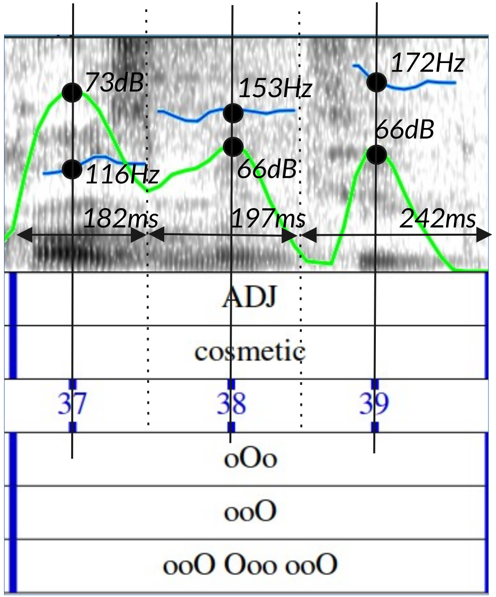
○ ○ O

⇕

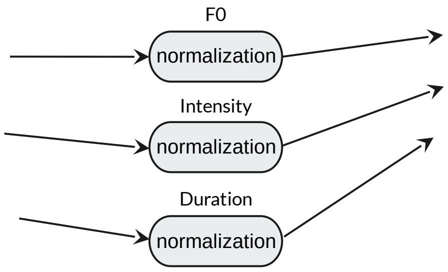
Reference pattern

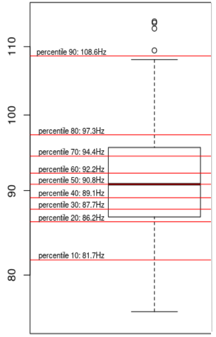
○ O ○

Stress processing (PLSPP v1)



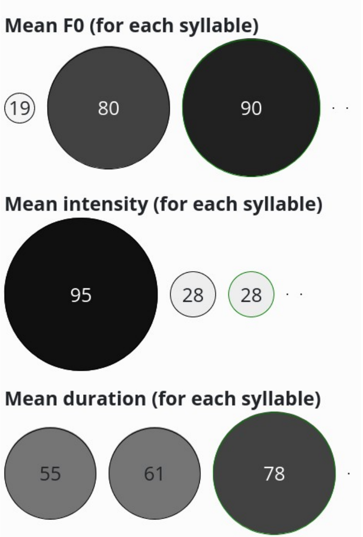
(F0, Intensity, Duration)



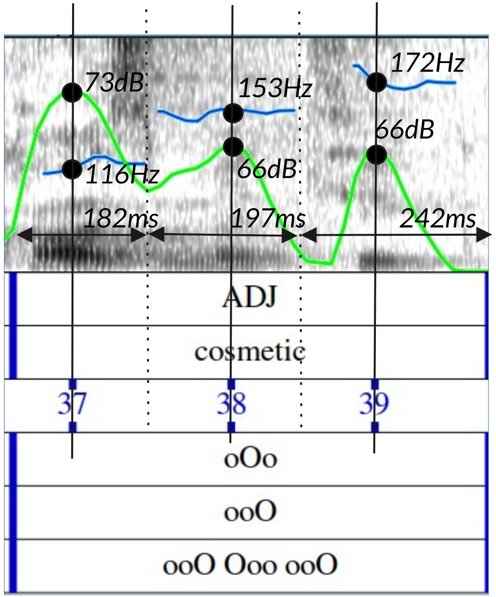


Percentiles
of the speaker's
distribution

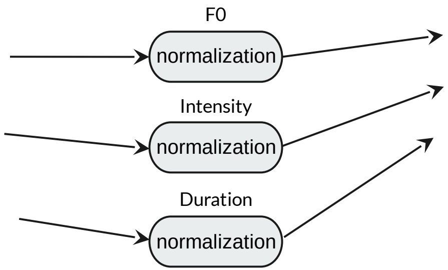
	37	38	39
F0:	(19),	(80),	(90)
dB:	(95),	(28),	(28)
dur:	(55),	(61),	(78)

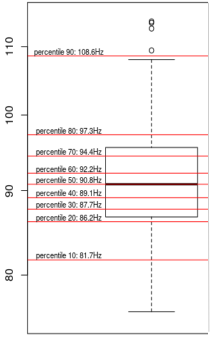


Stress processing (PLSPP v1)

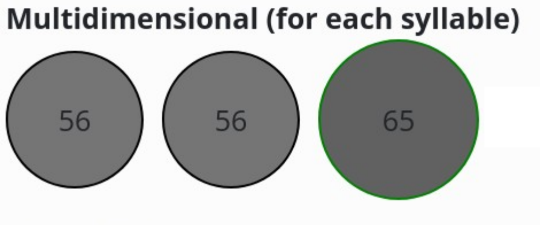


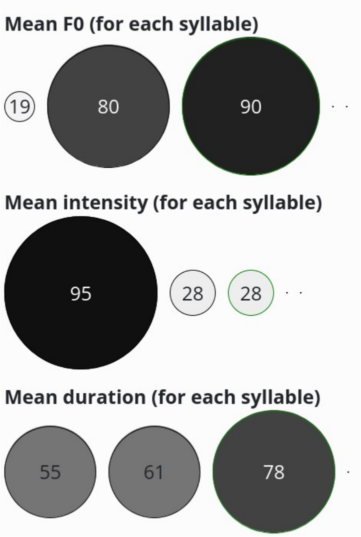
(F0, Intensity, Duration)



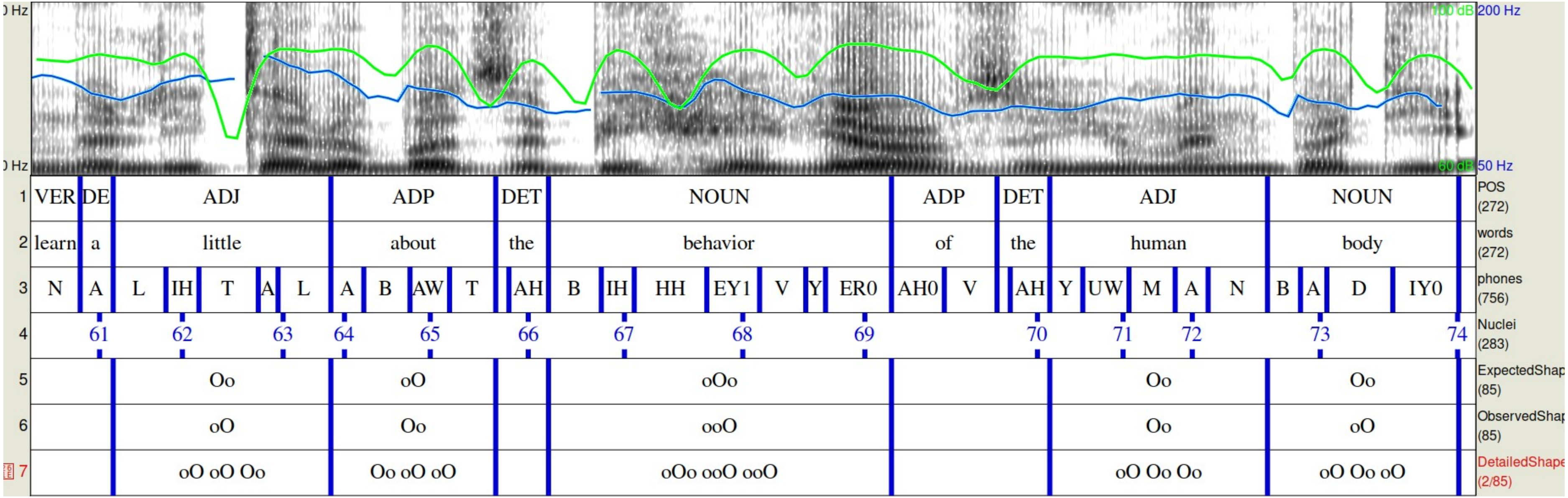


Percentiles
of the speaker's
distribution





Stress processing (PLSPP v2)



1 VER DE ADJ ADP DET NOUN ADP DET ADJ NOUN POS (272)

2 learn a little about the behavior of the human body words (272)

3 N A L IH T A L A B AW T AH B IH HH EY1 V Y ER0 AH0 V AH Y UW M A N B A D IY0 phones (756)

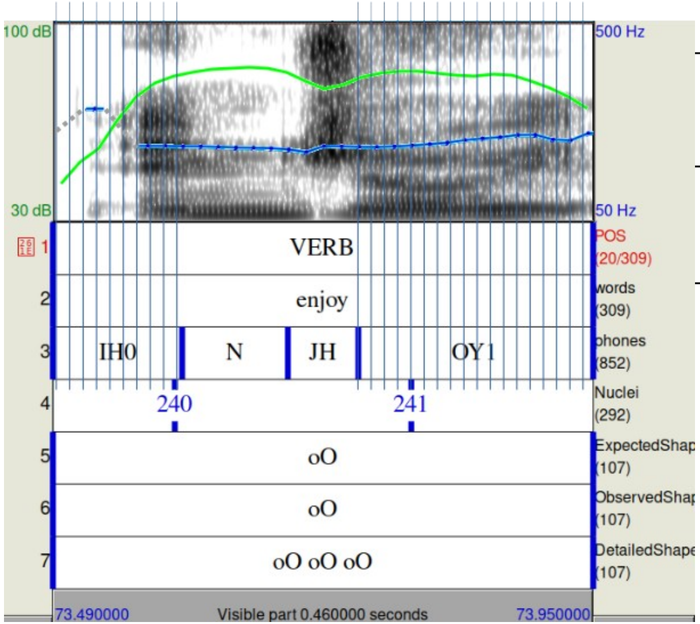
4 61 62 63 64 65 66 67 68 69 70 71 72 73 74 Nuclei (283)

5 Oo oO oOo Oo Oo ExpectedShap (85)

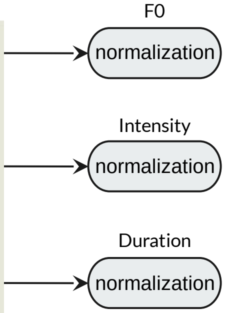
6 oO Oo ooO Oo oO ObservedShap (85)

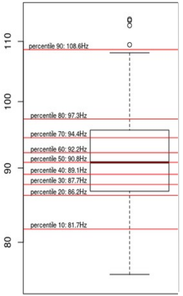
7 oO oO Oo Oo oO oO oOo ooO ooO oO Oo Oo oO Oo oO DetailedShape (2/85)

Stress processing (PLSPP v2)



(F0, Intensity, *Duration)*





Percentiles
of the speaker's
distribution

time_step = 10ms
(customizable)

F0

- mean(F0s)
- (Min, max, sd)

pitch linear interpolation

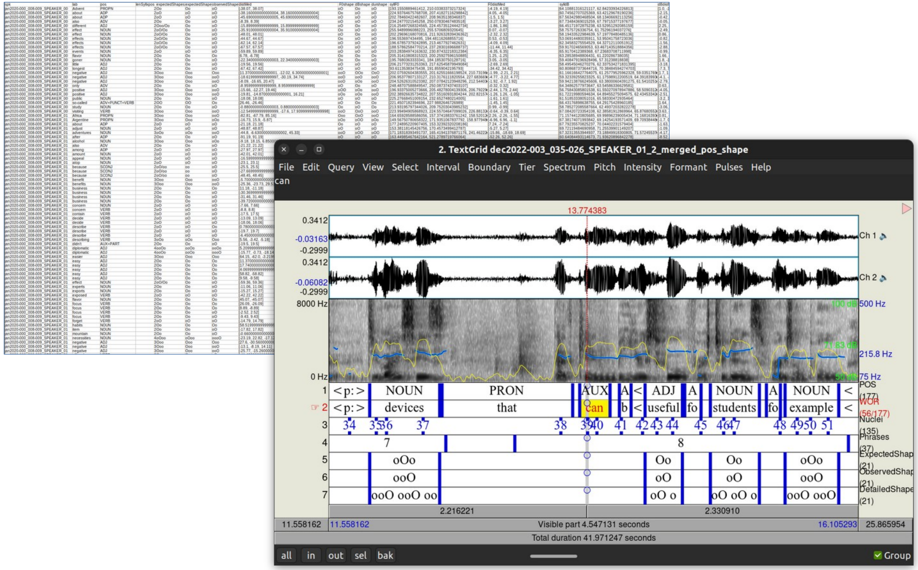
Intensity

- max(dBs)

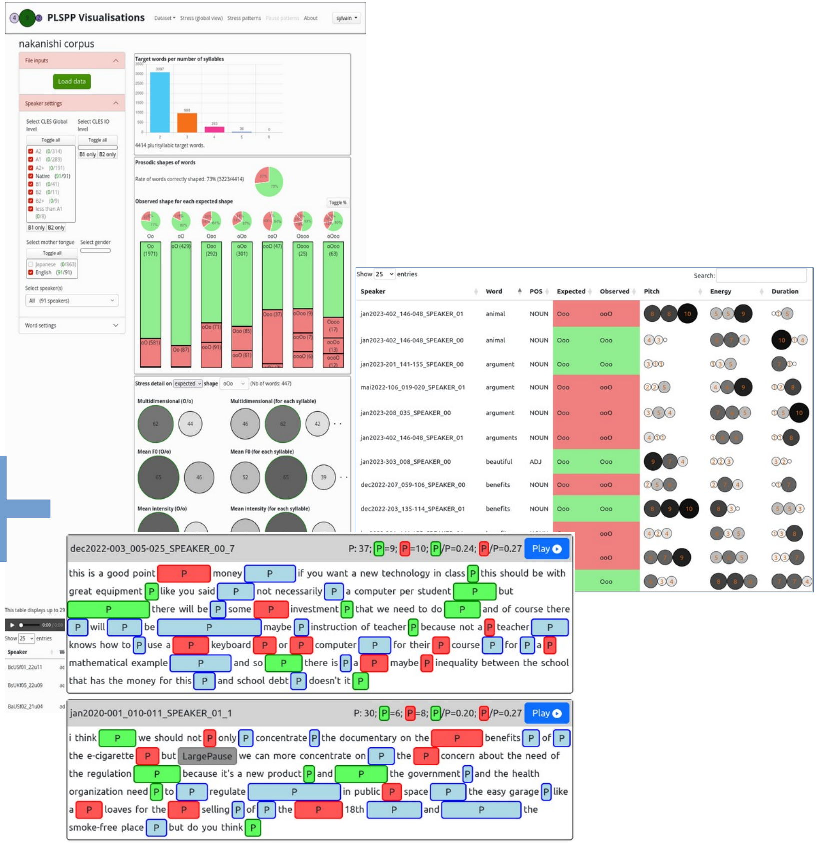
Duration

- Length of vowel interval

Visualizations



+



Speaker	Word	POS	Expected	Observed	Pitch	Energy	Duration
jan2023-402_146-048_SPEAKER_01	animal	NOUN	Ooo	ooO			
jan2023-402_146-048_SPEAKER_00	animal	NOUN	Ooo	Ooo			
jan2023-201_141-155_SPEAKER_00	argument	NOUN	Ooo	Ooo			
mai2022-106_019-020_SPEAKER_01	argument	NOUN	Ooo	ooO			
jan2023-208_035_SPEAKER_00	argument	NOUN	Ooo	ooO			
jan2023-402_146-048_SPEAKER_01	arguments	NOUN	Ooo	ooO			
jan2023-303_008_SPEAKER_00	beautiful	ADJ	Ooo	Ooo			
dec2022-207_059-106_SPEAKER_00	benefits	NOUN	Ooo	ooO			
dec2022-203_135-114_SPEAKER_01	benefits	NOUN	Ooo	Ooo			

dec2022-003_005-025_SPEAKER_00_7 P: 37; P=9; P=10; P/P=0.24; P/P=0.27 Play

this is a good point P money P if you want a new technology in class P this should be with great equipment P like you said P not necessarily P a computer per student P but P there will be P some P investment P that we need to do P and of course there P will P be P maybe P instruction of teacher P because not a P teacher P knows how to P use a P keyboard P or P computer P for their P course P for P a P mathematical example P and so P there is P a P maybe P inequality between the school that has the money for this P and school debt P doesn't it P

jan2020-001_010-011_SPEAKER_01_1 P: 30; P=6; P=8; P/P=0.20; P/P=0.27 Play

i think P we should not P only P concentrate P the documentary on the P benefits P of P the e-cigarette P but LargePause we can more concentrate on P the P concern about the need of the regulation P because it's a new product P and P the government P and the health organization need P to P regulate P in public P space P the easy garage P like a P loaves for the P selling P of P the P 18th P and P the smoke-free place P but do you think P

https://plspp.univ-grenoble-alpes.fr/

</> https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plsppviz

Studies using PLSPP

PLSPP v1

PLSPP v2

CLES Spontaneous speech

Université Grenoble Alpes

PhD

Multispeaker spontaenous speech
University students (B1~B2)
L1: **French**

CLES

- Coulange S, Kato T, Rossato S, Masperi M. (2024). Enhancing Language Learners' Comprehensibility through Automated Analysis of Pause Positions and Syllable Prominence. Languages 9(3):78
- Coulange, S., Kato, T., Rossato, R., Masperi, M. (2023). Automatic Measurement of Lexical Stress in Spontaneous L2 English Speech of French Learners. Phonetic Society of Japan, Sep 2023, Sapporo, Japan. pp. 126-131
- Coulange, S., Kato, T. (2023). Pause position analysis in spontaneous speech for L2 English fluency assessment. Acoustic Society of Japan, Sep. 2023, Nagoya, Japan. pp. 991-994

Corpus:

- Coulange, S., Fries, M.-H., Masperi, M., Rossato, R. (2024). A corpus of spontaneous L2 English speech for real-situation speaking assessment. LREC-COLING 2024, 20-25 May, Torino, Italy.

CLES-jp Spontaneous speech

同志社大学 Doshisha University
WASEDA University

Multispeaker spontaenous speech
University students (A2~C1)
L1: **Japanese**, **English**

Corpus:

- Coulange, S., Konishi, T., Kato, T., Sugahara, M., Rossato, R., Masperi, M. (2024). A corpus of spontaneous dialogues in L2 English by French and Japanese L1 speakers for automated assessment of fluency. 6th International Symposium on Learner Corpus Studies in Asia and the World (LCSAW6), Feb. 2024, Kobe, Japan.

Fluency evaluation

Université Grenoble Alpes

Read-aloud
University students (B1-B2)
L1: **French**

PIC

Prosodie, Intelligibilité, Communication (PIC) (Frost, D.)

- Paper coming soon :)

Stress awareness vs. stress production

同志社大学 Doshisha University

Carrier phrases
L1: **Japanese**, **Korean**, **English**

- Sugahara, M., Coulange, S., Kato, T. (2024). English Lexical Stress in Awareness and Production: Native and Non-native Speakers. The 19th Conference on Laboratory Phonology, June 27-29, Seoul.
- Sugahara, M., Coulange, S., Kato, T. (2023). Stress awareness vs. stress production: Comparison of primary stress assignment to English words between Japanese and Korean university students. 347th regular meeting of the Phonetic Society of Japan, Nov 25, online.

Fluency evaluation

神戸学院大学 KOBE GAKUIN UNIVERSITY

Read-aloud
University students (A1-B2)
L1: **Japanese**, **English**

PLSPP v3...

- Nakanishi, M., Coulange, S. (2024). Measuring speech rhythm through automated analysis of syllabic prominences. Prosodic features of language learners' fluency (Speech Prosody WS), July 1, Leiden.

Automatic vs. native speakers' evaluation of lexical stress

同志社大学 Doshisha University

Text recitation
Elementary school children (A2-B1)
L1: **Japanese**

- Kimura, T., Coulange, S., Kato, T. (2024). Automatic estimation and native speakers' evaluation of lexical stress positions in English recitation speech produced by Japanese elementary school children. Spring Meeting of the Acoustic Society of Japan, Mar 6-8, Tokyo.

Current PhD experiment: Corpus

Corpus:

- ✔ L2 English spontaneous speech from 176 French learners recorded during CLES certification speaking session.
- ✔ Situation: 2 or 3 candidates discussing a polemical topic (role play) during 10min.

- ➢ Total 11 hours of continuous speech (per speaker: mean 3’44’’, min 32’’, max 6’51)
- ➢ Speaking B1 level: 34%, B2 level: 66%
- ➢ Speech duration: B1≈B2, Nb tokens: B1<B2, Nb pauses: B1<B2, Silence proportion: B1≈B2

Hypothesis:

- **Pauses:**
  - More random pauses with B1
  - More structurant pauses with B2
- **Stress:**
  - Stress position accuracy B2>B1
  - Lower contrast stressed/unstressed
  - Stress shift to last syllable

CLES official website: https://www.certification-cles.fr/english/
Raw data is available for research: https://hdl.handle.net/11403/cles-spontaneous-english
See Coulange, S., Fries, M.-H., Masperi, M., Rossato, R. (2024). A corpus of spontaneous L2 English speech for real-situation speaking assessment. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May, Torino, Italy.

Current PhD experiment: Corpus

Corpus:

✔ L2 English spontaneous speech from 176 French learners recorded during CLES certification speaking session.

✔ Situation: 2 or 3 candidates discussing a polemical topic (role play) during 10min.

➢ Total 11 hours of continuous speech (per speaker: mean 3’44’’, min 32’’, max 6’51)

➢ Speaking B1 level: 34%, B2 level: 66%

➢ Speech duration: B1≈B2, Nb tokens: B1<B2, Nb pauses: B1<B2, Silence proportion: B1≈B2

Hypothesis:

- **Pauses:**
  - More random pauses with B1
  - More structurant pauses with B2
- **Stress:**
  - Stress position accuracy B2>B1
  - Lower contrast stressed/unstressed
  - Stress shift to last syllable

intra-phrase

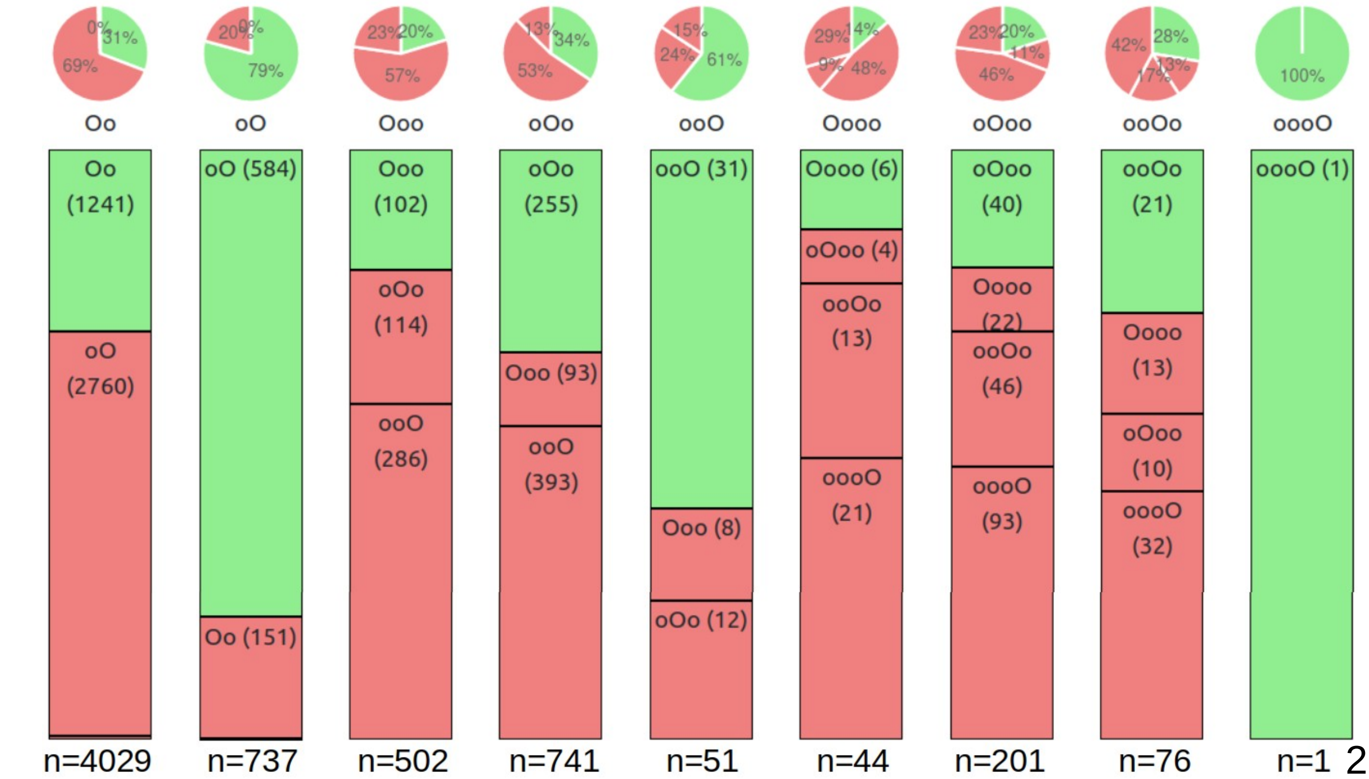
inter-clause

CLES official website: https://www.certification-cles.fr/english/
Raw data is available for research: https://hdl.handle.net/11403/cles-spontaneous-english
See Coulange, S., Fries, M.-H., Masperi, M., Rossato, R. (2024). A corpus of spontaneous L2 English speech for real-situation speaking assessment. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May, Torino, Italy.

Stress position

Expected pattern →

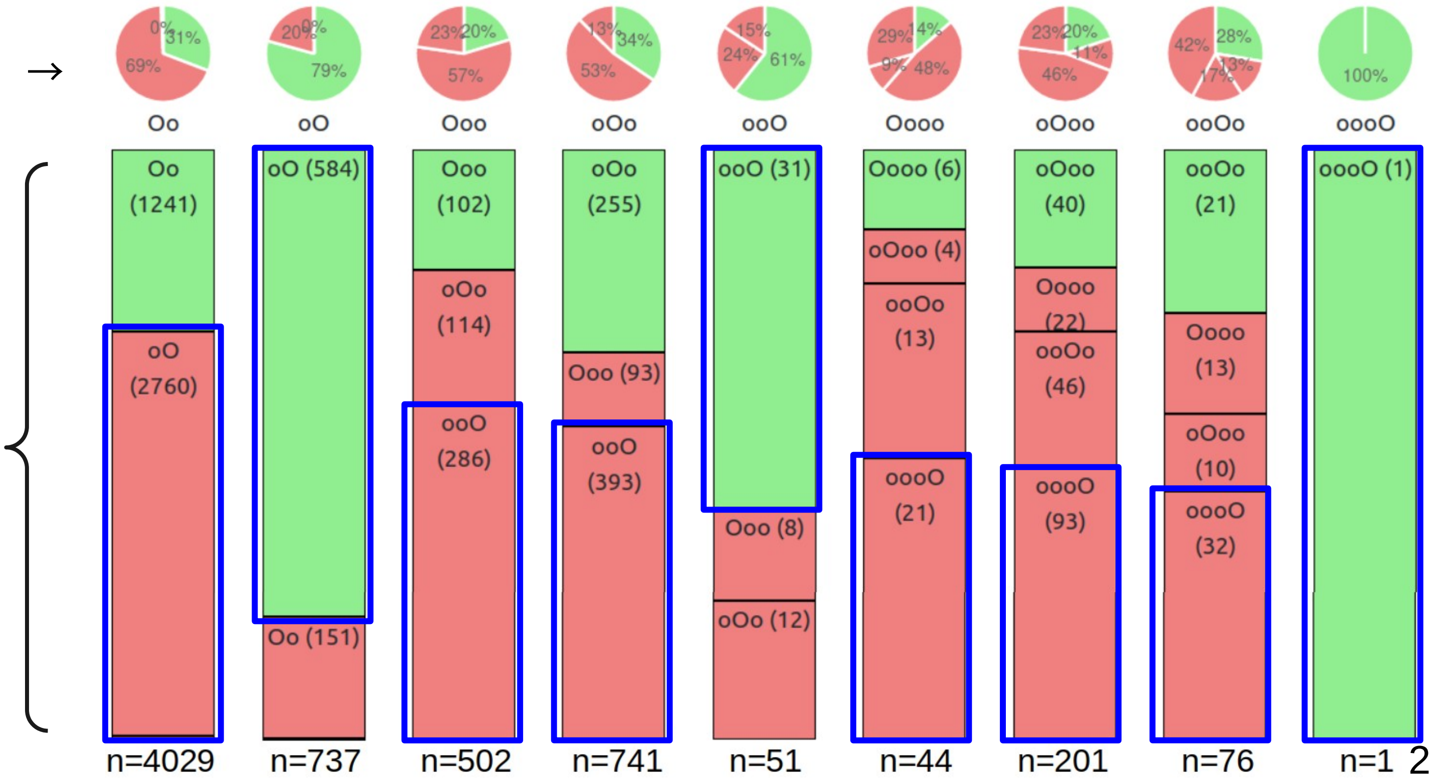
Observed patterns



Stress position

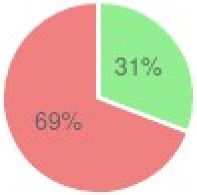
Expected pattern →

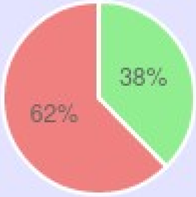
Observed patterns



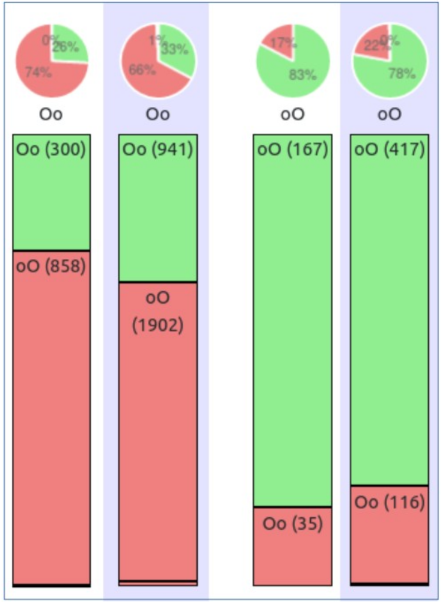
Stress position

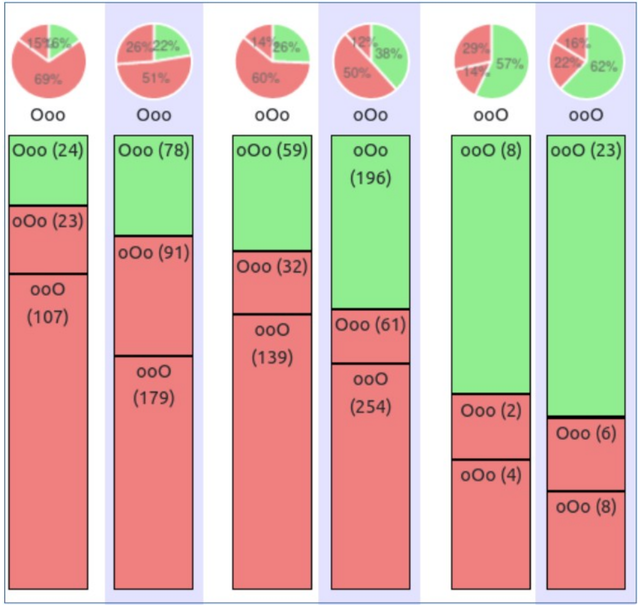
B1 speakers
spk=59
words=1873

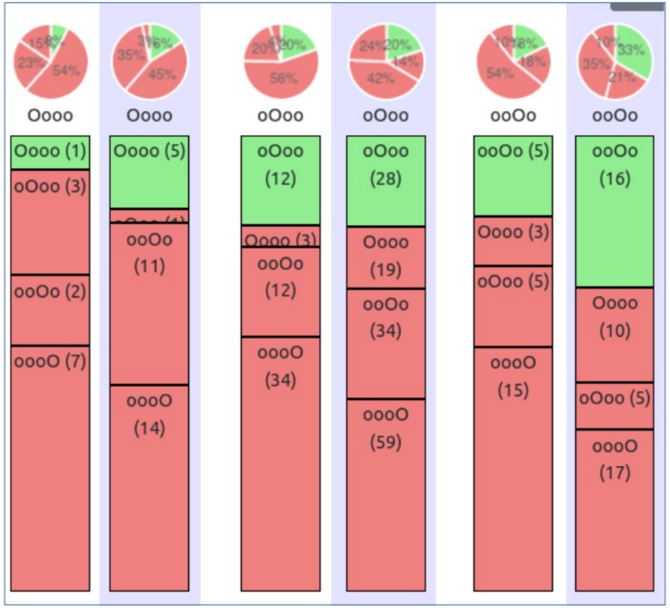




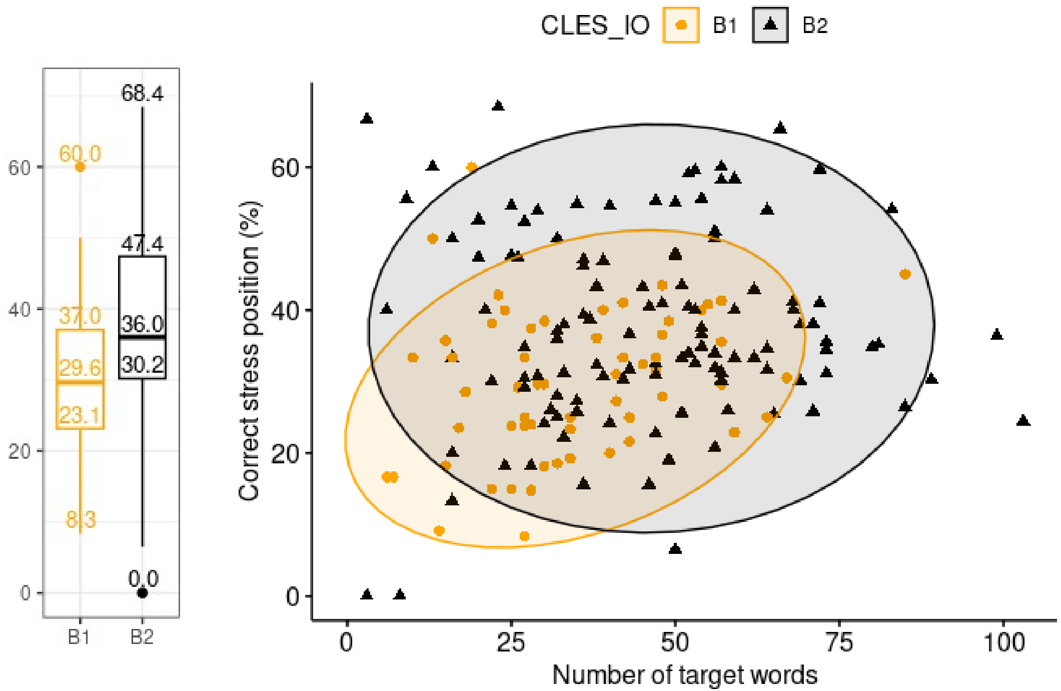
B2 speakers
spk=117
words=4551







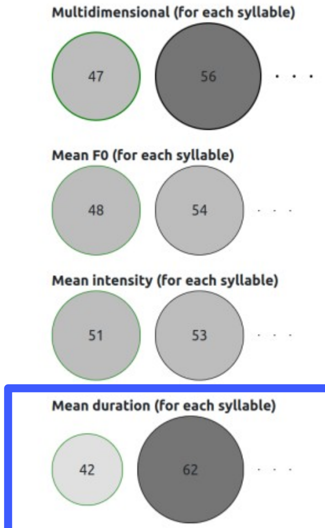
Current PhD experiment: Stress position analysis


Proportion of target words with correct stress position per speaker (n=176)

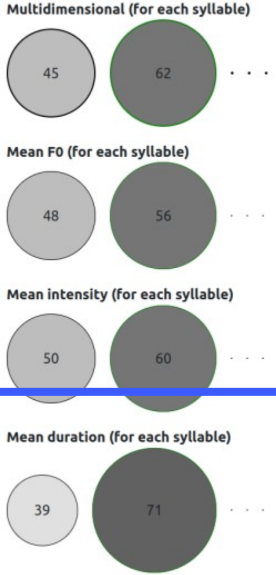
- *Mean stress position accuracy: 35.4 %*
- *Stress accuracy per speaker: 0 % ～ 68.4 %*
- *Stress accuracy per CEFR level: B1 = 29.6 % B2 = 36 % (p<.001)*

Stress quality: dimension

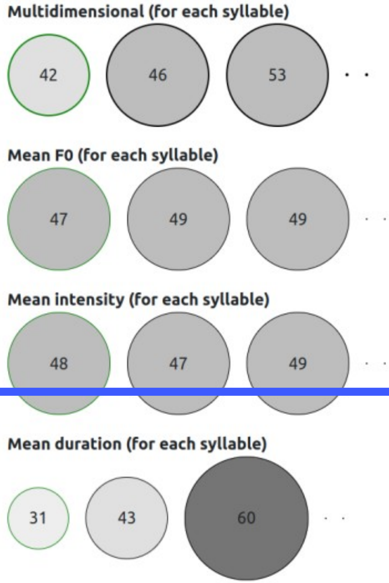
Expected: Oo
(n=4029)



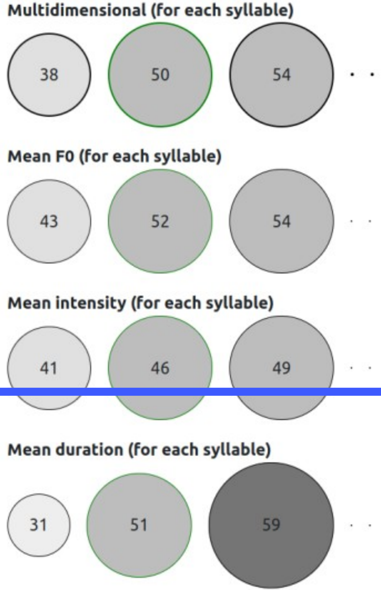
Expected: oO
(n=737)



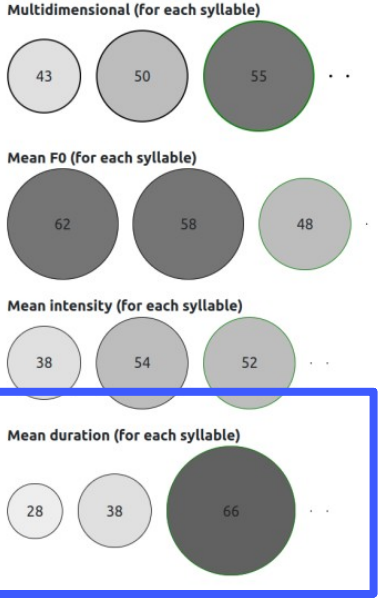
Expected: Ooo
(n=502)



Expected: oOo
(n=741)



Expected: ooO
(n=51)



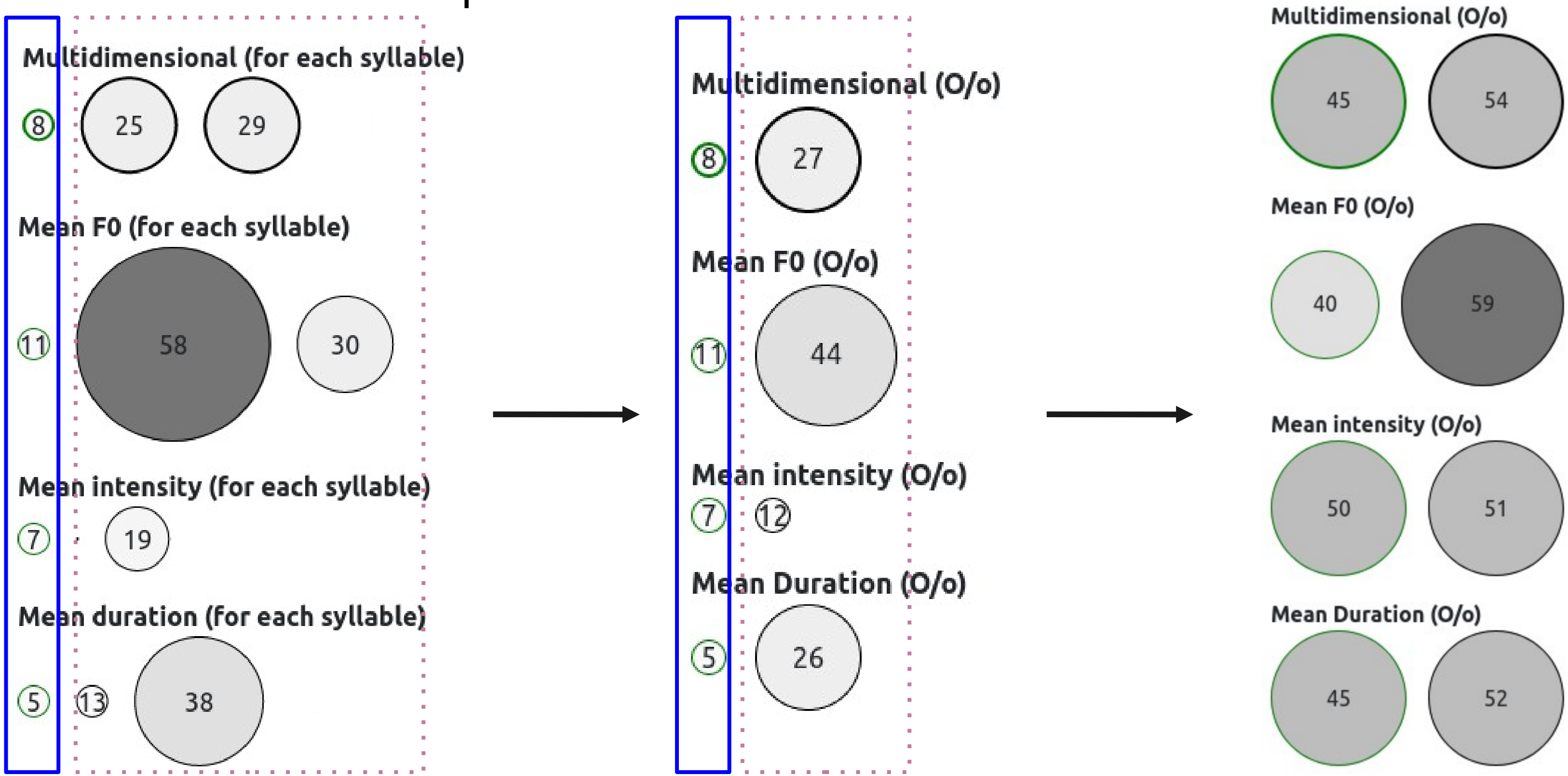
全ての話者（176人）

Stress quality: dimension

Speaker jan2020-001_020-022_SPEAKER_00

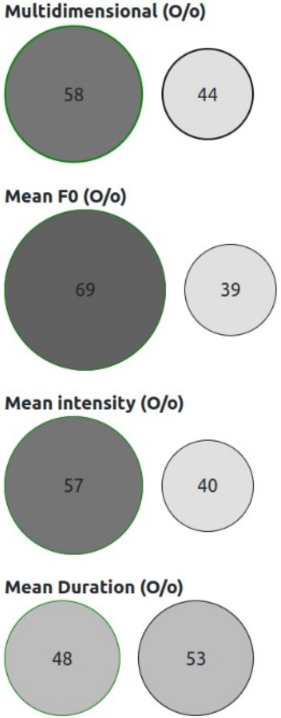
- 42 target words
- Stress position accuracy: 19%
- Mean prosodic contrast: -9 points

Expected Ooo

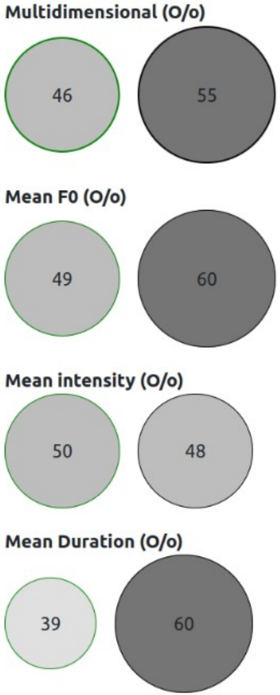


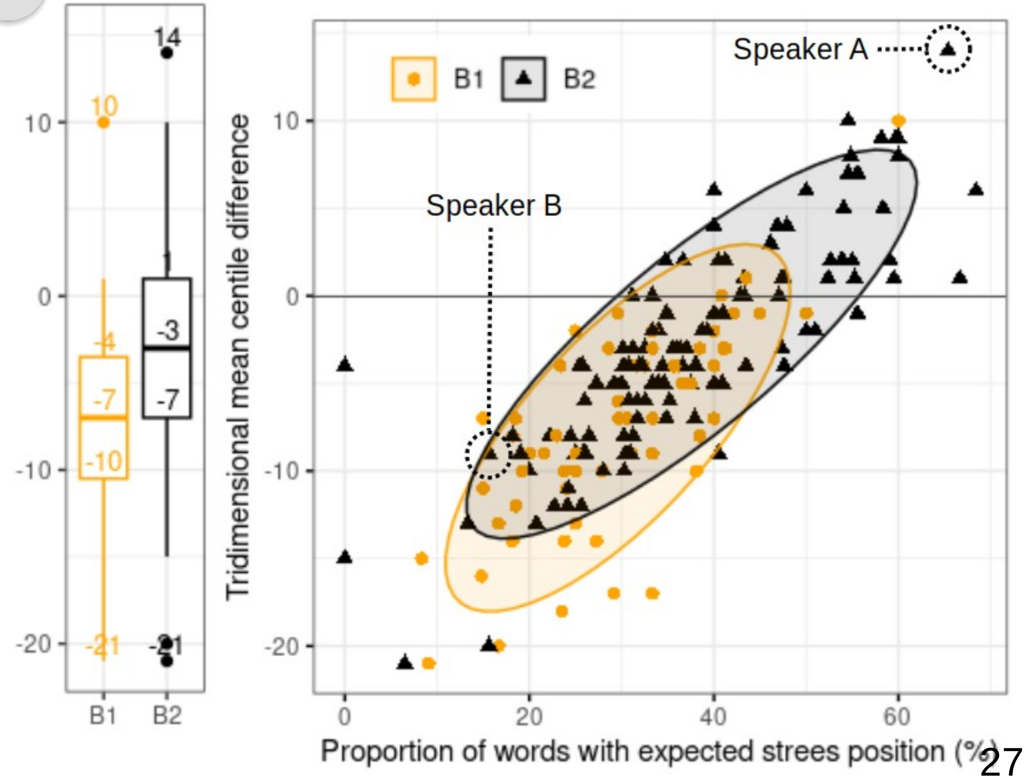
Current PhD experiment: Stress quality analysis

Speaker A
(stress position score: 65%)



Speaker B
(stress position score: 16%)

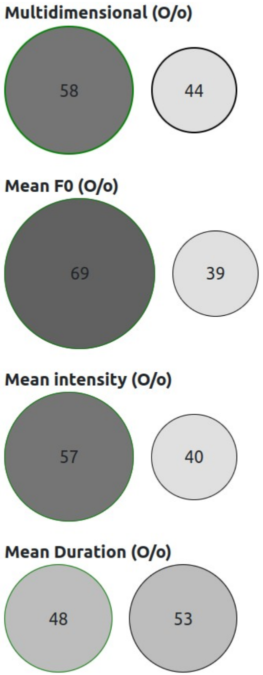




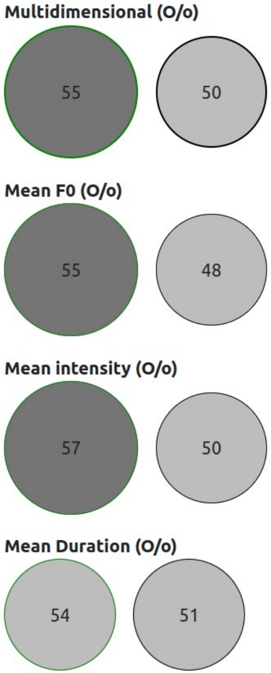
Current PhD experiment: Stress quality analysis

Stress position accuracy:

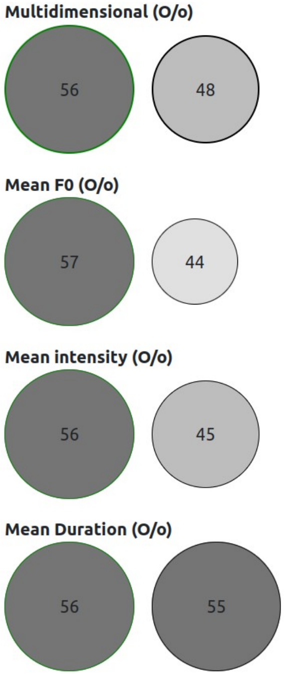
65%



58%

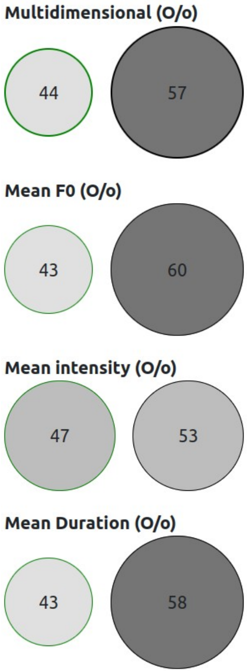


60%

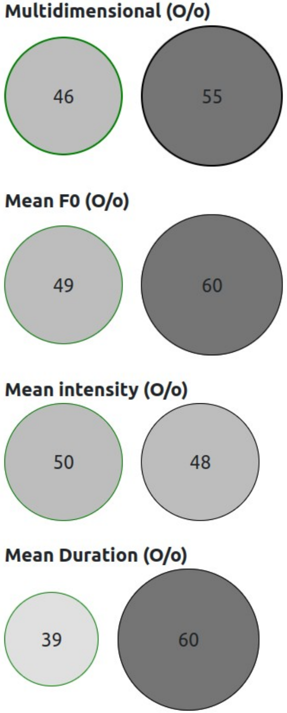


Stress position accuracy:

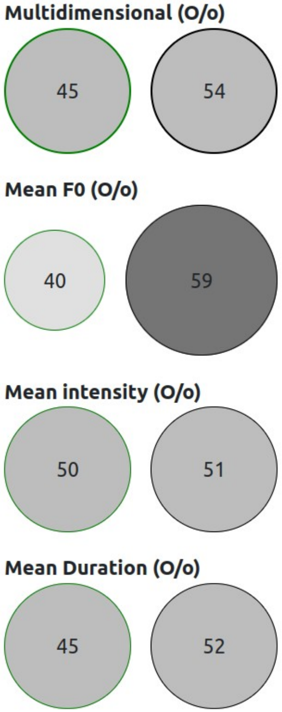
21%



16%



19%



Current PhD experiment: Main observations

- First prototype of the Pauses and Lexical Stress Processing Pipeline
- Analysis of B1 and B2 speaking level French-L1 university students
  11 hours of speech 6350 target words 21 831 pauses

➢ Pause position:

- ○ Great variation of number of pauses within phrases among speakers, less with pauses between clauses
- ○ B2 speakers make less pauses within phrases than B1 speakers (p<0.01)
- ○ Difference between B1 and B2 is small
- ○ High intra-speaker variability

➢ Lexical stress position:

- ○ *Mean stress position accuracy: 35.4 %*
- ○ *Stress accuracy per speaker: 0 % ~ 68.4 %*
- ○ *Stress accuracy per CEFR level:*
  B1 = 29.6 % B2 = 36 % (p<0.001)
- ○ *Frequent stress shift to the last syllable*

➢ Lexical stress quality:

- ○ *Low accuracy speakers: lengthening of the last syllable*
  tendency to make it higher
  No change in intensity
- ○ *High accuracy speakers: the expected syllable is higher in F0 and intensity*
  No change in duration

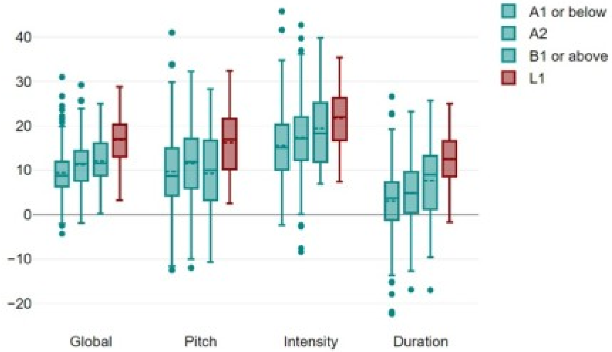
Nakanishi & Coulange (2024)

神戸学院大学 KOBE GAKUIN UNIVERSITY

- 34 hours read-aloud speech
- 877 Japanese-L1 samples (42 speakers, <A1-B2)
- 91 Native English samples (7 professional narrators)
- PLSPP extension to monosyllabic words

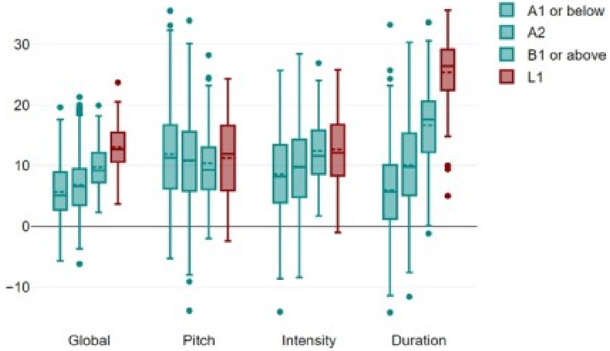
> analysis of contrast between **content and function words**

Figure 1. Syllabic Contrast Scores within Monosyllabic Words by CEFR Level.



Global scores between groups ($p < .001$)
A1 <*** A2 *n.s.* B1 < *** L1

Figure 2. Lexical Contrast Scores between Content and Function Words by CEFR Level.



Global scores between groups ($p < .001$)
A1 <*** A2 <*** B1 < *** L1

Pipeline Evaluation & Limitations:

- As the pipeline combines several modules, errors can occur at different levels, often leading to incorrect annotations.

▲Syllable detection and word alignment often mismatches, leading to a limited nb. of target words (only **41%** of polysyllabic words in the study below were **target words**).

▲Manual evaluation of random 100 target words showed that **17%** were miss-recognized or miss-aligned, potentially leading to wrong judgments that can be problematic in a real assessment context.

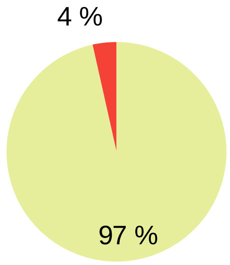
▲**Intrinsinc vowel length** and **word ending lengthening** need to be considered in order to improve stress estimation.

▲Some cases of **vowel devoicing** also impacted F0 measures (tackled with linear interpolation for now)

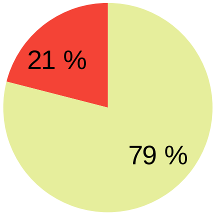
Word alignment precision

Number of target words with totally wrong alignment,
among the first 200 plain target words in the visualization interface:

plspp : 7 target words

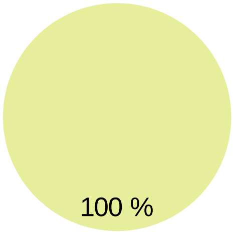


plspp_mfa : 42 target words



Corpus PIC (Frost, D.)
(280 speakers Read speech
~1min20s/spk)

Plspp: 0 words



plspp_mfa: 7 words