

# ***Annotations minimales multi-niveaux d'un corpus de parole spontanée d'apprenants japonais de FLE et traitement automatique : perspectives didactiques***

*Work In Progress...*

Sylvain Detey (U. Waseda, Japon), Maxime Le Coz (Archean Technologies, France), Lionel Fontan (Archean Technologies, France), Corentin Barcat (TUFS, Japon), Yuji Kawaguchi (TUFS, Japon), Hisae Akihiro (TUFS, Japon), Kaori Sugiyama (Seinan Gakuin U., Japon) & Nori Kondo (NUFS, Japon).

IPFC2018 – Paris MSH – 26-27 novembre 2018



# Plan

- 1) Le corpus: objectifs et enjeux
- 2) Des annotations minimales à une description automatisée
- 3) Perspectives d'analyses
- 4) Perspectives didactiques

# 1) Le corpus: objectifs et enjeux

**CLIJAF:** *Corpus longitudinal interphonologique d'apprenants japonais de français* (Detey, 2011-2019)

1) Grant-in-Aid for Scientific Research (B) n°23320121 (2011-2015)

2) Grant-in-Aid for Scientific Research (B) n°15H03227 (2015-2019)

Japanese Society for  
the Promotion of  
Science (JSPS)

1) 2011-2015 : *Etude longitudinale sur corpus des traits interphonologiques des apprenants japonais de français.*

2) 2015-2019 : *Analyse multi-niveaux sur corpus du français parlé par des apprenants japonais de niveau pré-avancé.*

Collaborateurs : Y. Kawaguchi (TUFS), M. Kondo (Waseda), H. Akihiro (TUFS), K. Sugiyama (Seinan Gakuin), K. Kawashima (Fukuoka)

# 1) Le corpus: objectifs et enjeux

## Volet 1:

- longitudinal (4 sessions sur 2 ans)
- apprenants débutants (A1.1-B1)
- focus sur la prononciation
- perception & production
- parole non-spontanée (partie 1 du protocole IPFC)

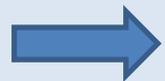
# 1) Le corpus: objectifs et enjeux

## Volet 2:

- parole spontanée (partie 2 du protocole IPFC)
- apprenants de niveau intermédiaire (- B2+)
- extension aux autres niveaux: lexique, syntaxe...

## Question:

Quelles divergences (« erreurs ») persistantes à l'oral ?



Interface prononciation/lexique/grammaire

cf. difficultés de transcription & codage: nature de la « divergence »

## Finalité: didactique

aider les apprenants à résoudre les divergences persistantes

# 1) Le corpus: objectifs et enjeux

## Le corpus:

Corpus (n=108 <sup>i</sup> )	Niveau	Participants	Tâche
CLIJAF 1 (2011-2015)	A1.1-B1 (n=48)	12 – TUFS (4 sessions sur 2 ans, après 4, 7, 12 et 19 mois à Tokyo).	Répétition IPFC, Lecture PFC, Lecture IPFC, Lecture Texte PFC
CLIJAF 2 (2015-2019)	B1-C1 (n=60)	16 (1 <sup>st</sup> 2016) + 4 (2 <sup>nd</sup> 2017) - Waseda 11 (1 <sup>st</sup> 2016) + 7 (2 <sup>nd</sup> 2017) - TUFS 8 (1 <sup>st</sup> 2015) + 6 (2 <sup>nd</sup> 2016) - Seinan 8 – Fukuoka	Lecture Texte PFC, Conversations avec natif et non-natif.  Certains étudiants enregistrés 2 fois en longitudinal après 1 an.

## Analyse phonético-phonologique:

Exploitation de CLIJAF 1 & 2 (mots, texte, conversations)

## Pour cette présentation:

Focus sur CLIJAF 2 – parole spontanée (conversations)

# 1) Le corpus: objectifs et enjeux

## CLIJAF 2: parole spontanée

- Locuteurs: 39 (31F & 8H)
- Conversations guidées: 46
- Conversations libres: 67
- Durée : environ 26h
- Format: son + transcription orthographique alignés  
(*Transcriber*)

## *Sous-corpus Waseda + TUFS dans la BDD*

- Nbre mots: 167 172
- Durée: 18h33
- Nbre annotations: 56 137

## 2) Des annotations minimales à une description automatisée

### Analyse multiniveaux d'un corpus oral:

énorme chantier... méthodologie, outils, RH... coûteux...+ les défis du traitement d'une L2 ! (e.g. FLLOC, Myles & Mitchell, <http://www.flloc.soton.ac.uk/>)

### Approche didactique:

- annotations minimales: repérage des divergences de surface
- point de vue de l'enseignant de FLE: "divergences" et "modèles"

### Méthodologie:

- Transcription orthographique
- Conventions ad hoc (GARS, IPFC...)
- Transcriber (simple, gratuit, testé)

## 2) Des annotations minimales à une description automatisée

### Annotations manuelles minimales des divergences:

- 1) [v]\_ ou [c] ou [v/c]\_\_ modification vocalique ou consonantique simple ou multiple (sans précision)
- 2) [e]\_...\_[xxx]: formes inacceptables transcrites telles quelles puis forme jugée souhaitable par le transcripateur:  
[e]\_à\_[en] Allemagne, je [e]\_vas\_[vais]
- 3) [ac]\_[ ]: formes inacceptables auto-corrigées (dernier énoncé) :  
je suis allé [ac]\_à Allemagne euh je suis allé en Allemagne\_[ ]

# Exemple de séquence

File Edit Signal Segmentation Options Help

report

[no speaker]

- AH1 - ok moi je veux parler [e]\_sur\_[de] euh # respecter s- l'heure
- YI1 - mh
- AH1 - parce qu'à Québec # à Montréal il y avait beaucoup de gens q- [e]\_[v:]\_de\_[du] Japon de France
- YI1 - <mh mh>
- AH1 - <[e]\_du\_[de] Montréal> # du Québec # et euh # les gens du Jap- les Japonais et les Québécois # respectaient l'heure parc- <donc euh>
- YI1 - <ah c'est> vrai ?
- AH1 - ouais ouais
- YI1 - <oh>
- AH1 - donc euh l- les Québécois c'était comme
- YI1 - <mh>
- AH1 - <les Japonais> par exemple euh
- YI1 - <[c:]\_mh>
- AH1 - <[v]\_quand on dit euh> # (X) on va [e]\_\_[se] voir euh à lib- # [v:]\_à [e]\_\_[la] biblio
- YI1 - <mh mh>
- AH1 - <bibliothèque> à douze heures
- YI1 - <ouais>



jpto2ah1yi1\_1l



report

[no speaker]

AH1 - ok moi je veux... ... l'heure	YI1 - mh	AH1 - parce qu'à Québec #... ... France	AH1 - <[e]_du_[de] Montréal> # du Québec... ... <donc euh>	YI1 - <ah c'est> vrai ?	AH1 - ouais ouais	YI1 - <oh>	AH1 - donc euh l- les Québécois c'était comme	YI1 - <mh>	AH1 - <les Japonais> par exemple euh	YI1 - <[c:]_mh>	AH1 - <[v]_quand on dit euh> # (X) on va [e]__[se] voir euh à lib- # [v:]_à [e]__[la] biblio	YI1 - <mh mh>	AH1 - <bibliothèque> à douze heures	YI1 - <ouais>
--	----------	--	---	-------------------------	-------------------	------------	---	------------	--------------------------------------	-----------------	--	---------------	-------------------------------------	---------------

0 5 10 15 20 25 30

## **2) Des annotations minimales à une description automatisée**

Traitement automatisé pour une description à visée didactique:

Partenariat avec Archean Labs (L. Fontan & M. Le Coz)

- Un serveur de stockage et de traitement
- Une description quantitative du corpus
- Un concordancier texte-son (didactique & recherche)
- Une interface d'évaluation pédagogique

## 2) Des annotations minimales à une description automatisée: Fonctionnalités Recherche

The screenshot shows a web browser window with the URL 192.168.33.82:8001. The page title is 'Concordancier'. The main content area is titled 'Rechercher' and shows a search for 'université'. The search results are displayed in a list format, with the word 'université' highlighted in red in each snippet. The interface includes a sidebar with 'Recherche', 'Vue globale', and 'Apprenants'. The search options are set to 'dans toutes les productions', 'dans les productions erronées', 'Phonétiques', and 'Autres'. The sorting is set to 'Contexte droit'.

Concordancier x

← → ↻ Non sécurisé | 192.168.33.82:8001

**CLIJAF**

Recherche  
Vue globale  
Apprenants

**Rechercher**

université

dans toutes les productions  dans les productions erronées  dans les corrections

Phonétiques  Autres

**Trier par**

Contexte gauche  Contexte droit

...ai # réétudié le français à l'université

f euh internet de ce f cette université

français que j'ai étudié à l'université

...donc à l'université

...'ai # je je l'ai commencé à l'université

...oui avant que je entrer # Ø l'université #

...euh depuis entrée le l de l'université #

an pour étudier euh dans une université # et euh malheureusement comm

...uand j'étais au débutant # de université # je suis allé # au Canada # ...

...ah # oui # à Ø université # les les z # les étudiants v...

...ongue pause au Japon èh # Ø l'université # on # on doit euh ét on doit...

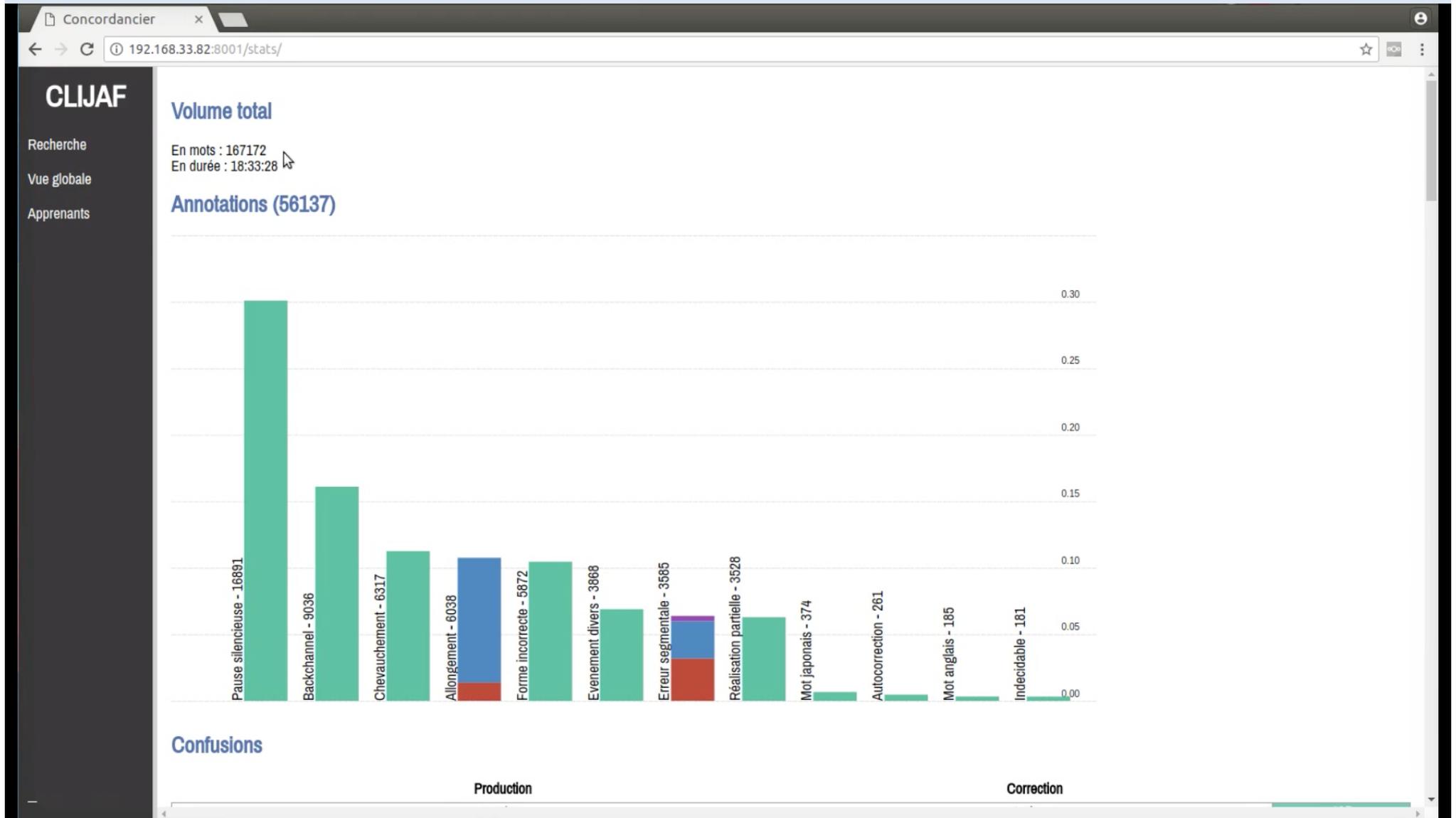
...oui à l'université ?

...'a- j'appris le français à au université c'était le conversation de # ...

...t je peux le l'étudier au à l'université comme une première langue étr...

...le travail ne dépend pas de l'université de # inspiration # mh chigau ...

## 2) Des annotations minimales à une description automatisée Statistiques Globales



## 2) Des annotations minimales à une description automatisée Profils d'apprenants

The screenshot shows a web browser window with the URL `192.168.33.82:8001/learners/`. The page is titled "CLIJAF" and has a sidebar with "Recherche", "Vue globale", and "Apprenants". The main content is divided into two sections: "Waseda" and "TUFs". Each section displays a grid of learner profiles, each represented by a box containing a code (e.g., AH1, FK1), a year, and a gender (e.g., 1995 F).

**Waseda**

AH1 1995 F	FK1 1993 F	JO1 1994 M	JT1 1993 M	KI1 1994 F	KT1 1994 M	MO1 1992 F	RS1 1993 F	SO1 1996 F	SY1 1995 F	TS1 1995 M	YI1 1994 F	YK1 1993 F
YN1 1994 F	HK1 1993 F	RI1 1993 M	SD1 None None									

**TUFs**

CS1 1988 F	HK1 1996 F	KH1 1991 F	KK1 1996 F	MK1 1993 F	MT1 1995 F	SH1 1993 F	SM1 1994 M	AS1 1991 M	RI1 1963 F	WK1 None None
---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	------------------

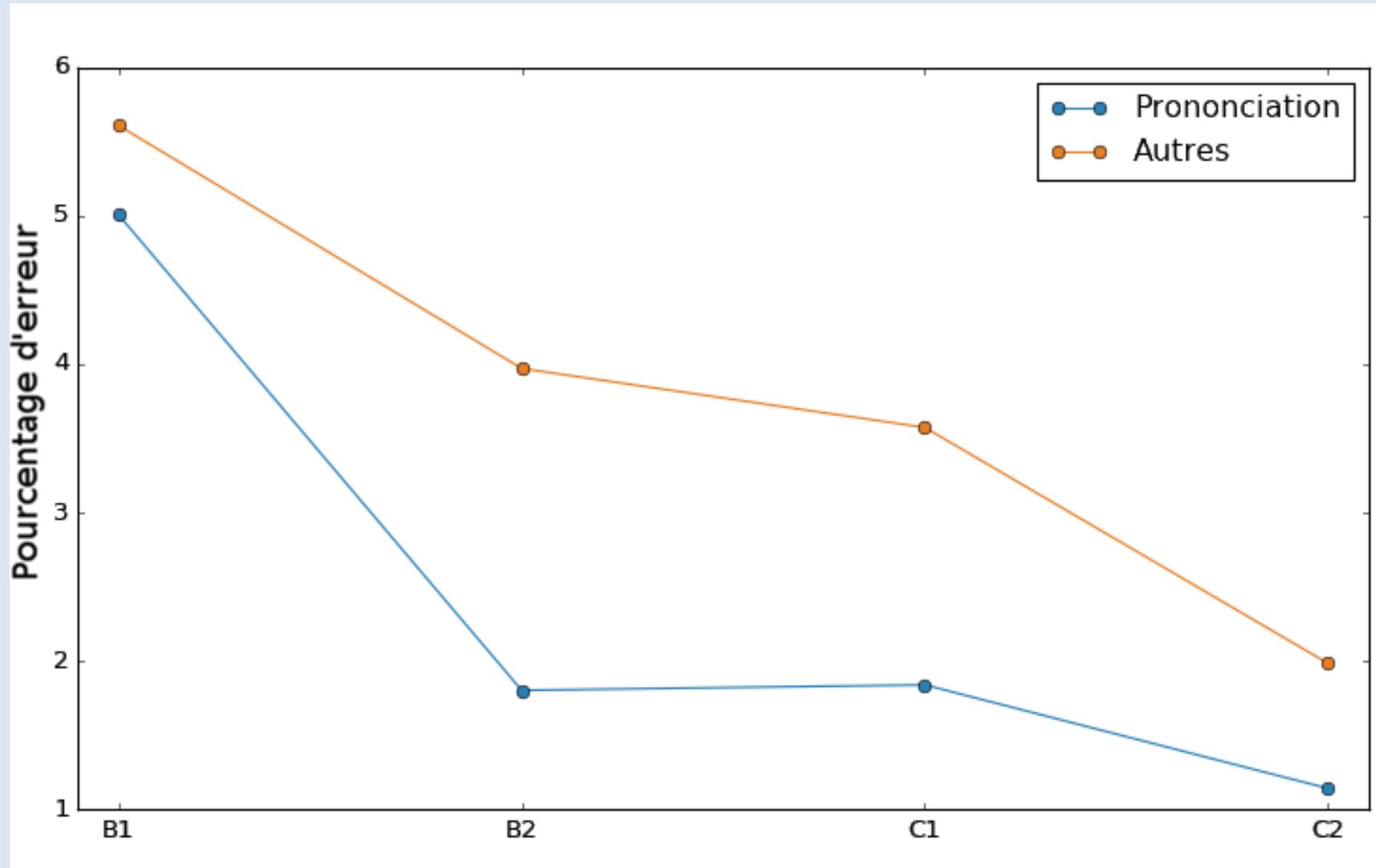
### 3) Perspectives d'analyse

#### Caractéristiques :

- Par apprenant: nbre & type d'erreurs
- Inter-apprenants: erreurs récurrentes, par niveaux, par type de conversation
- Longitudinales: 2 sessions

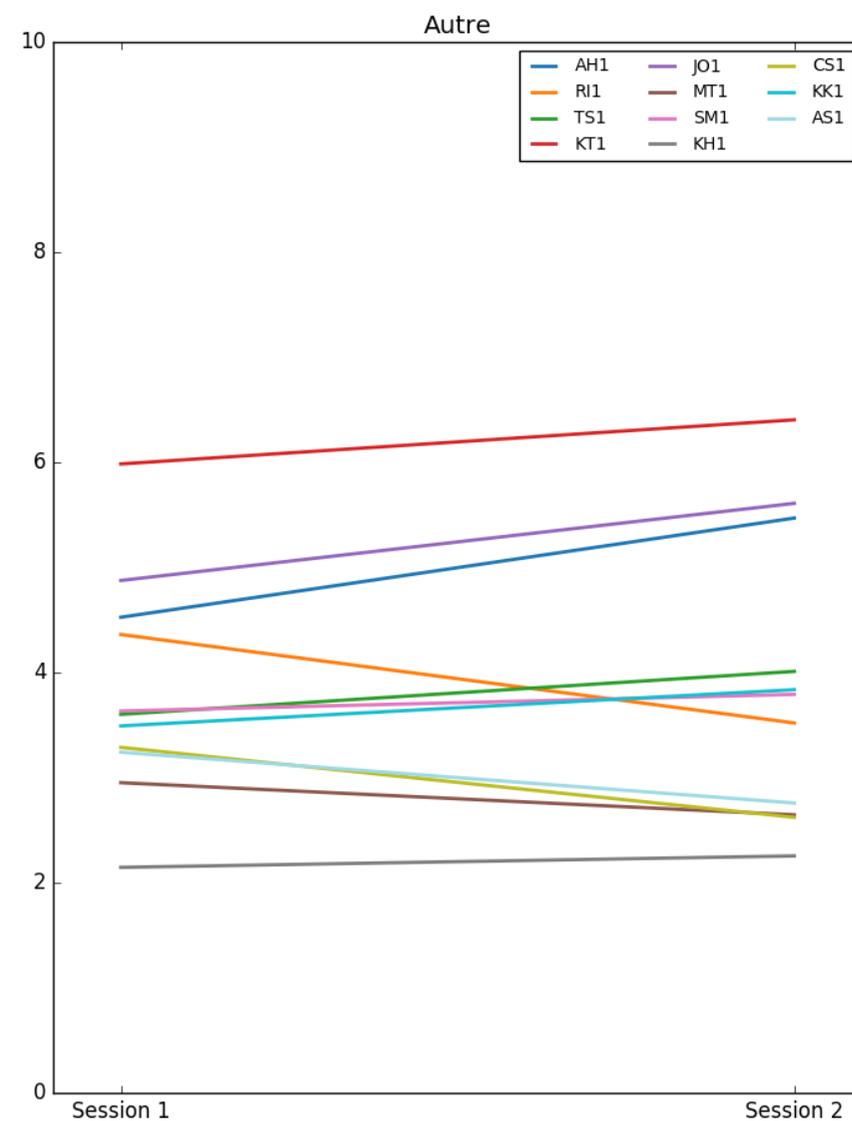
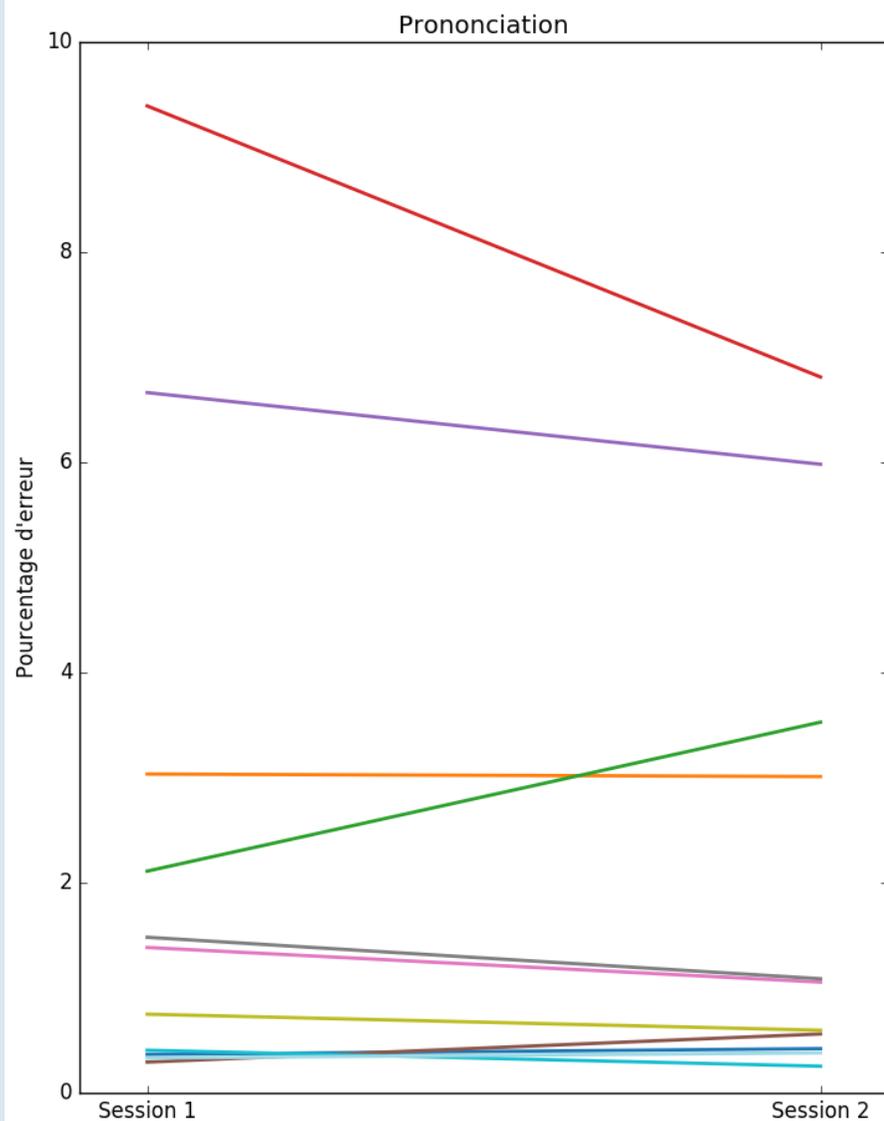
### 3) Perspectives d'analyse

Pourcentage d'erreurs en fonction du niveau de l'apprenant



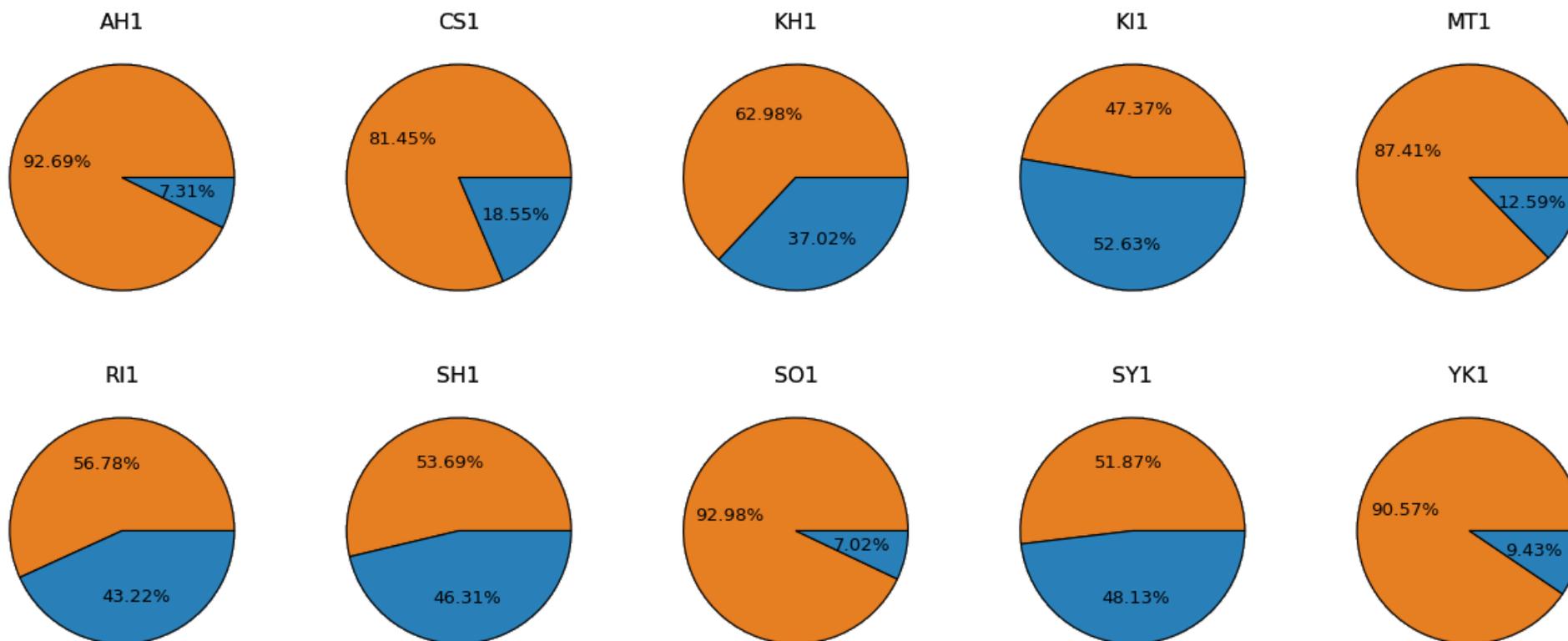
# 3) Perspectives d'analyse

## Evolution longitudinale du pourcentage d'erreurs



# 3) Perspectives d'analyse

Répartition des erreurs entre prononciation et autre pour 10 apprenants B2



### 3) Perspectives d'analyse

Au niveau phonético-phonologique:

apprentissage semi-supervisé avec Thomas Pellegrini (IRIT, Toulouse), ANR Jeune Chercheur

LUDAU (Lightly-supervised and Unsupervised

Discovery of Audio Units using Deep Learning) :

→ position et nature de la divergence segmentale

# 3) Perspectives d'analyse

## Approche:

- Recherche à partir de la cible vers des divergences (correction)
- Recherche d'un item divergent et récupération du contexte
- Recherche d'une catégorie PDD (en cours)

## Intérêt:

- 1) Adéquation lexicogrammaticale vs. Adéquation phonétique :  
récupération du signal sonore en contexte →  
nature de la divergence: « du »/ « du »
- 2) Double entrée : production vs cible possible

# 4) Perspectives didactiques

Des « divergences »: génération d'activités didactiques:

- Phonético-phonologiques
- Lexicales
- Morpho-syntaxiques
- Discursives

Connexions avec CAPT-L2 et Lexpro

- Profil phonologique des apprenants
- Profil lexical des apprenants

En projet:

- Profil grammatical des apprenants
- Dimension discursive et sociolinguistique

→ Personnalisation du contenu didactique selon le profil

# Conclusion

Apport par rapport à un concordancier standard:

- Évaluation de l'oral
- Rapport entre formes produites et formes attendues (double entrée de recherche)

Défis:

- 1) Transcription orthographique: manuelle ?
- 2) Annotation minimale: manuelle ?
- 3) Multiplicité des modèles: degré d'acceptabilité ?  
Nécessité de codage multiple

Objectif ultime: un système de correction automatique de la parole L2

→ Besoin de système(s) de référence:

- phonologie de référence (CAPT-L2)
- lexique de référence (Lexpro)
- grammaire de référence (cf. correcteurs grammaticaux – mais de l'oral !)

# Remerciements

- Japanese Society for the Promotion of Science
- Archean Technologies
- Les étudiants du corpus
- Laboratoire Praxiling UMR 5267 U. Montpellier 3

# ***Annotations minimales multi-niveaux d'un corpus de parole spontanée d'apprenants japonais de FLE et traitement automatique : perspectives didactiques***

*Work In Progress...*

Sylvain Detey (U. Waseda, Japon), Maxime Le Coz (Archean Technologies, France), Lionel Fontan (Archean Technologies, France), Corentin Barcat (TUFS, Japon), Yuji Kawaguchi (TUFS, Japon), Hisae Akihiro (TUFS, Japon), Kaori Sugiyama (Seinan Gakuin U., Japon) & Nori Kondo (NUFS, Japon).

IPFC2018 – Paris MSH – 26-27 novembre 2018

