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)	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 (1st 2016) + 4 (2nd 2017) - Waseda 11 (1st 2016) + 7 (2nd 2017) - TUFS 8 (1st 2015) + 6 (2nd 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

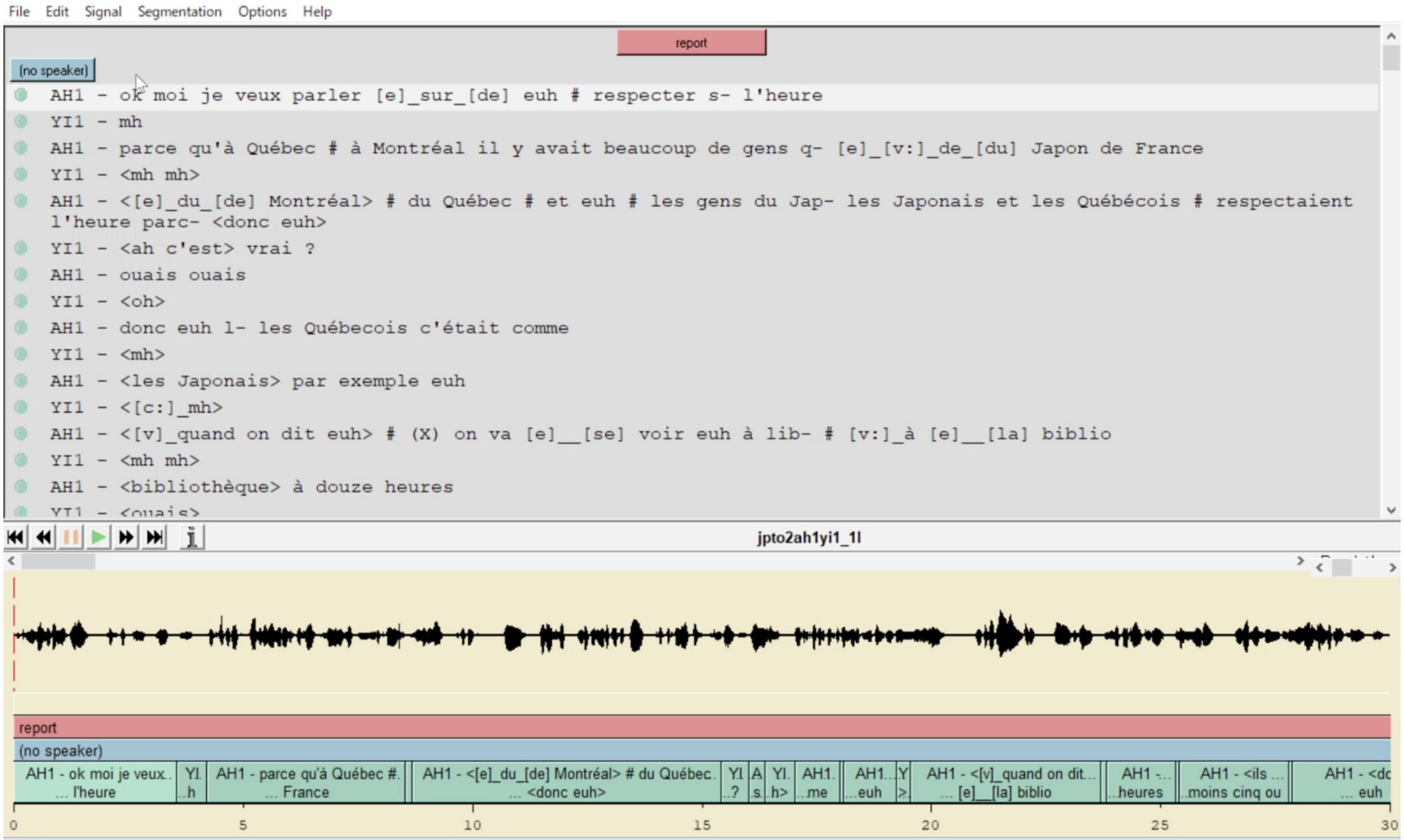
Annotations manuelles minimales des divergences:

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 transcripteur:
[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



2) Des annotations minimales à une description automatisée

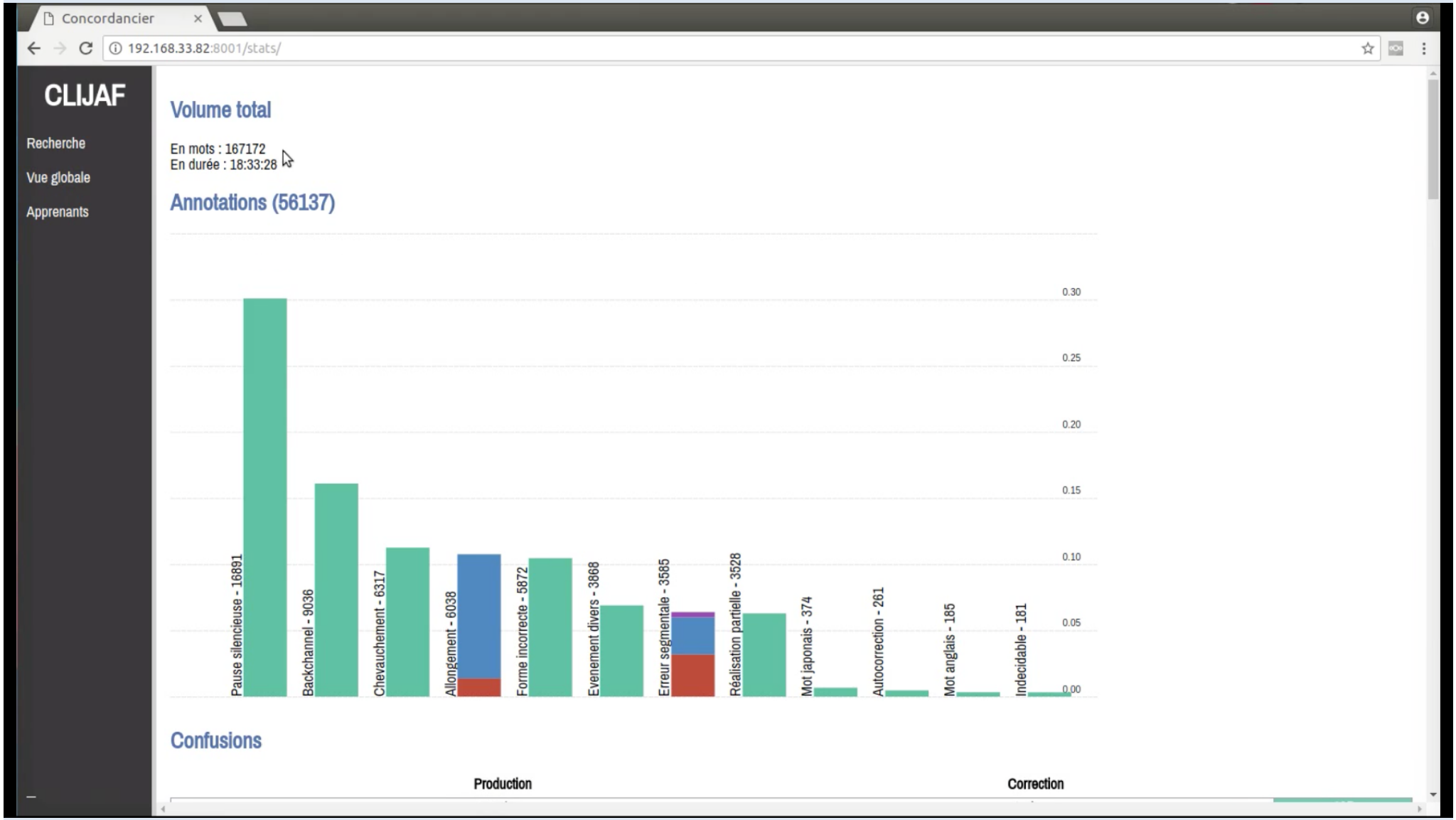
Traitement automatisé pour une description à visée didactique:

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

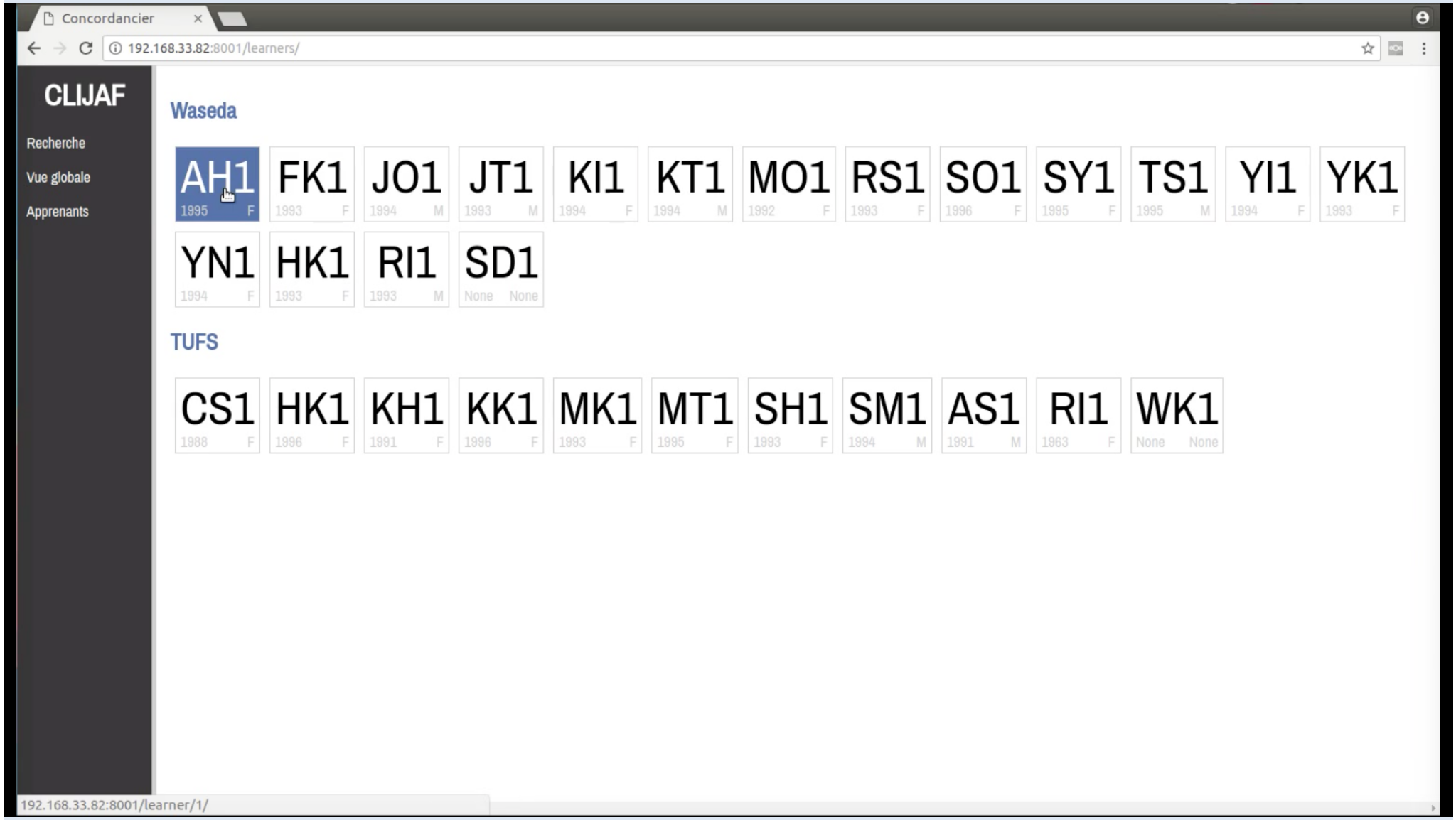
- Un serveur de stockage et de traitement
- Une description quantitative du corpus
- Un concordancier texte-son (didactique & recherche)
- Une interface d'évaluation pédagogique

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

2) Des annotations minimales à une description automatisée
Statistiques Globales



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



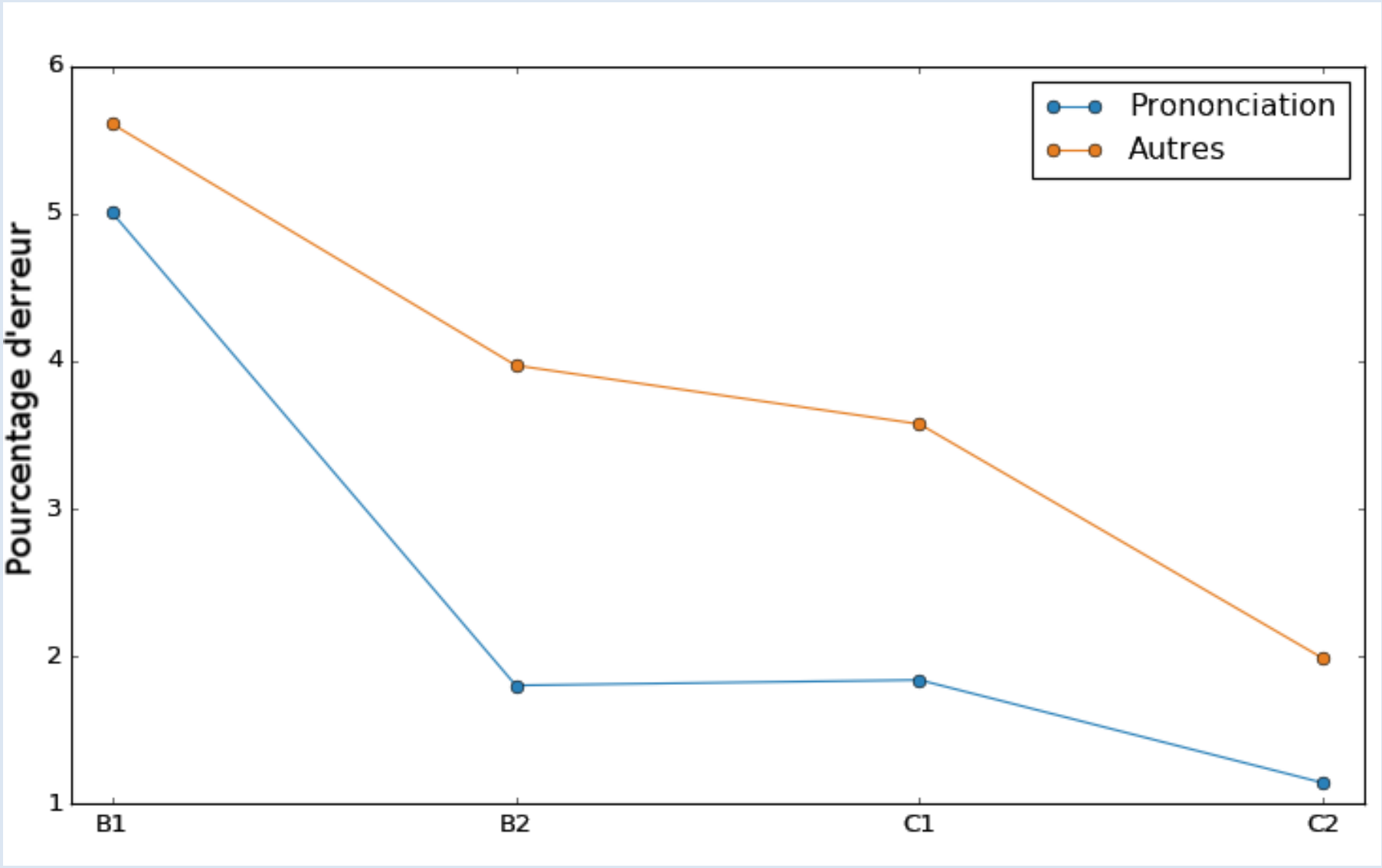
3) Perspectives d’analyse

Caractéristiques :

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

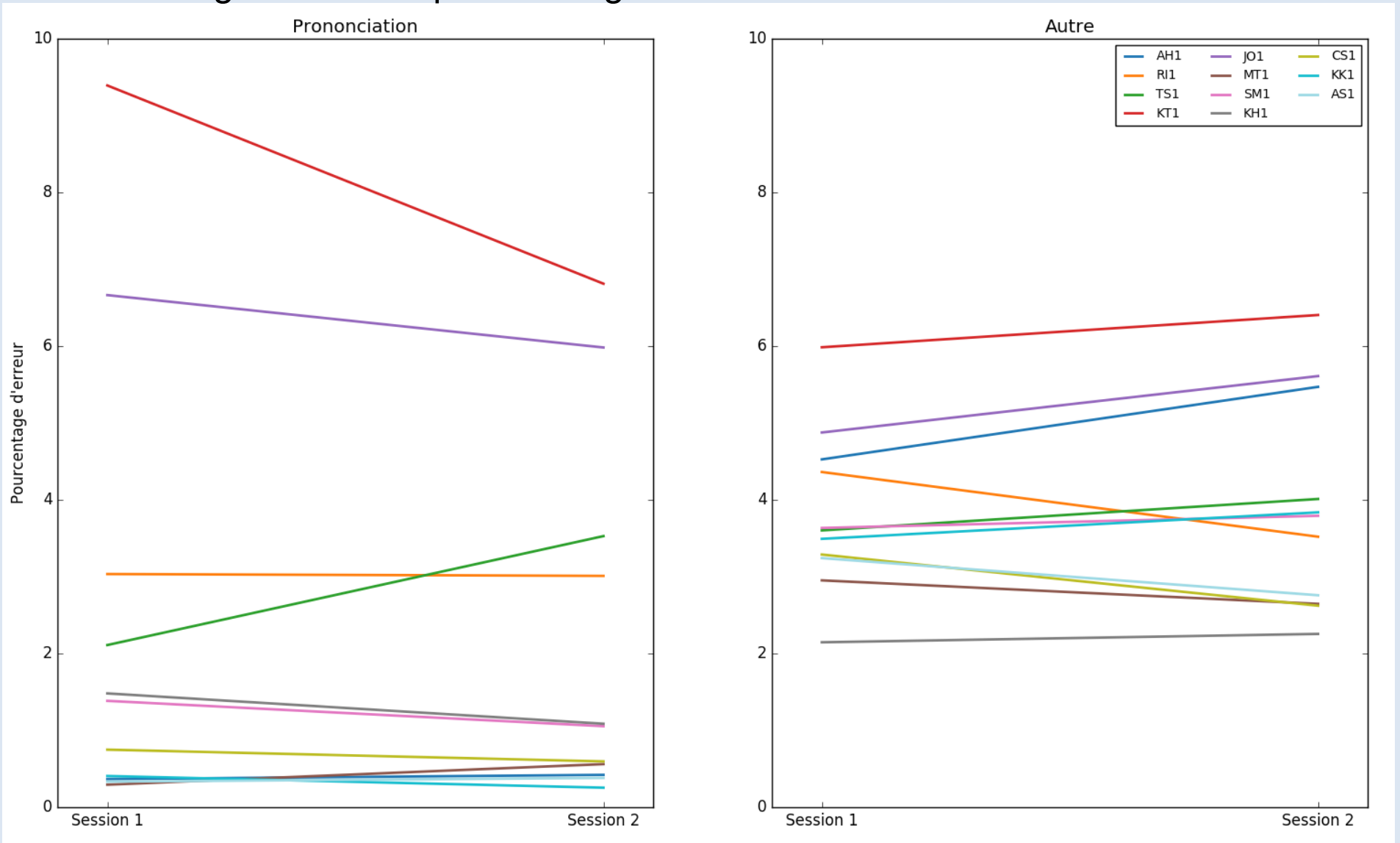
3) Perspectives d’analyse

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



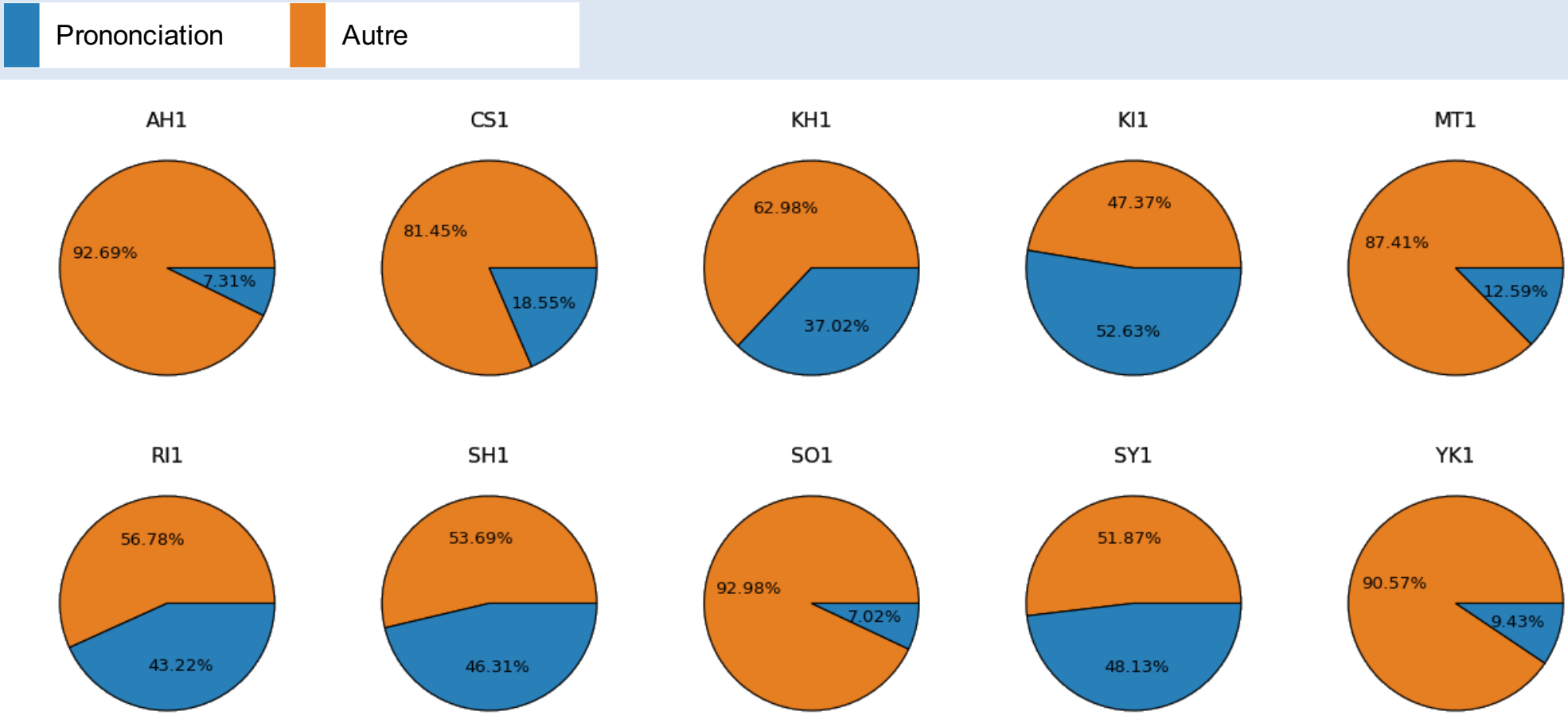
3) Perspectives d'analyse

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

<u>Au niveau phonético-phonologique:</u>
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)

<u>Défis:</u>

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

<u>Objectif ultime:</u> 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