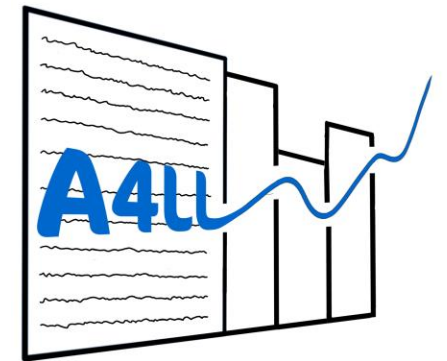

La linguistique de corpus à l'heure du code ouvert

LIDILE anr®



Consortium : Universités Rennes 2, Le Mans, Paris Cité et Galway (Irlande)

- **Coordinateur**

- GAILLAT Thomas - Associate Professor in linguistics and English teacher Université Rennes 2

- **Experts**

- SÉBILLOT Pascale - Professor of Computer Science - IRISA / INSA Rennes
- GRAVIER Guillaume - Senior Research Scientist - IRISA / CNRS

- **Membres**

- MALLART Cyriel – Research Engineer – Université Rennes 2
- BALLIER Nicolas - Professor of Linguistics - University of Paris
- SIMPKIN Andrew - Professor in Statistics – University of Galway
- STEARNS Bernardo - Research Associate & Ph.D candidate – University of Galway
- VENANT Rémi - Associate Professor in Computer Science - Le Mans University
- LI Jen-Yu - Ph.D. candidate - Université Rennes 2

Contexte: Evaluation écrite en langue étrangère





Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - Focalisation sur l'erreur



Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~Focalisation sur l'erreur~~
 - ~~Evaluation principalement sommative~~



Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~Focalisation sur l'erreur~~
 - ~~Evaluation principalement sommative~~
 - Feedback formatif



Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~◦ Focalisation sur l'erreur~~
 - ~~◦ Evaluation principalement sommative~~
 - Feedback formatif
 - Dimensions positives et négative du langage produit



Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~◦ Focalisation sur l'erreur~~
 - ~~◦ Evaluation principalement sommative~~
 - Feedback formatif
 - Dimensions positives et négative du langage produit
 - Comparaisons inter-sujets



Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~◦ Focalisation sur l'erreur~~
 - ~~◦ Evaluation principalement sommative~~
 - Feedback formatif
 - Dimensions positives et négative du langage produit
 - Comparaisons inter-sujets
- Etudiants du supérieur usagers de L2 en domaine professionnel

Analytics for Language Learning

- Evaluation automatique écrite en langue étrangère
 - ~~Focalisation sur l'erreur~~
 - ~~Evaluation principalement sommative~~
 - Feedback formatif
 - Dimensions positives et négative du langage produit
 - Comparaisons inter-sujets
- Etudiants du supérieur usagers de L2 en domaine professionnel
- **Systeme de *monitoring* par analytics linguistiques à destination des enseignants**

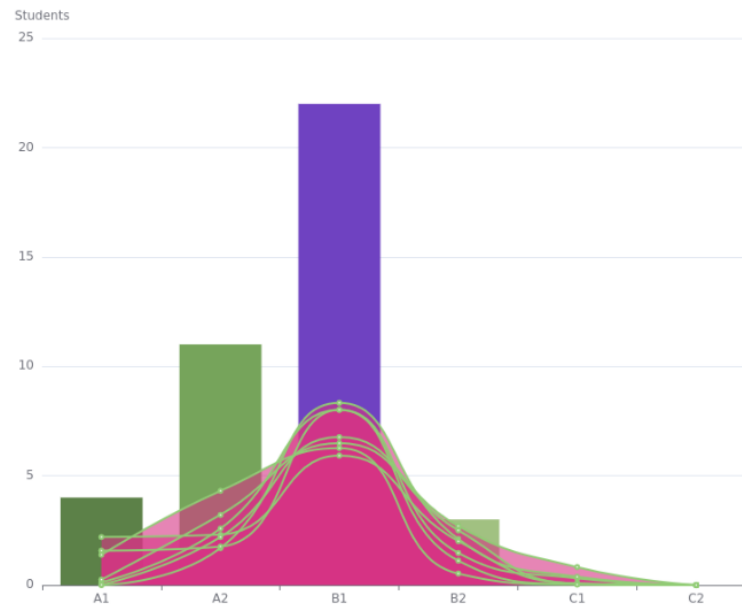
Un système EIAH de monitoring

Activity Dashboard



Student Name (Level)	Action
Cyriel Descartes (A1)	Text
Jean Jolivet (B2)	Text
Jean Velpeau (B1)	Text
Julie Jolivet (B1)	Text
Alice Jolivet (B1)	Text
Alice Fuye (B1)	Text
Alice Velpeau (B1)	Text
Jean Velpeau (B1)	Text
Alice Bellanger (B1)	Text
Alice Molasse (A2)	Text
Julie Fuye (B1)	Text
Julie Pisco (A2)	Text
Bob Pisco (B1)	Text
Bernardo Bellanger (A2)	Text
Henry Fuye (B1)	Text

Distribution of students by overall CEFR level



Indicator	genitive as sub...	word in clause	multinoun in cl...	compound as s...	existential
Value	↑	↓	↑	↓	↓
Impact	↓	↑	↓	↑	↑

CEFR Predictions distribution ▼

Explanations for Thomas Descartes, Remi Descartes, Alice Luengo... ^

The selected students make use of grammatical/lexical/text type-related indicator(s) in a notable manner that widely differs from the reference cohort (i.e., above or under 90 percent of ~1,200 students).

Diagnostics:

Indicator 1: genitive as subject

The indicator is a grouping of **12** measures which combine **genitive as subject**.

[Show the list of measures](#)

Interpretation:

The value of this indicator is high in the student's writing. The indicator is associated with worse proficiency.

🔍 Try to see why the use of genitive as subject drives the student to lower proficiency.

[Generate pedagogical activity](#)

Indicator 2: word in clause

The indicator is a grouping of **6** measures which combine **word in clause**.

[Show the list of measures](#)

Interpretation:

Un système EIAH

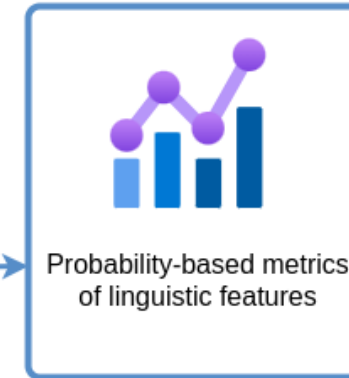
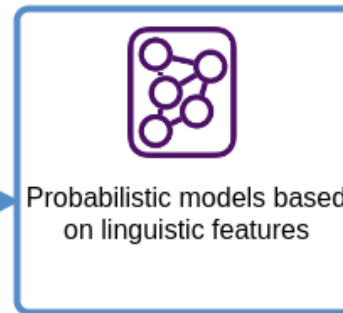
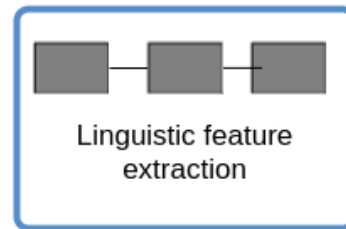


Mediated
interactions
Class + Moodle



Predictions
Explanations
Enriched texts

Text flow
Final production



Symbolic knowledge

Challenge scientifique

Questions de recherche

- Corpus annotées
- Mesures textuelles et modélisation
- Evaluations méthodiques

Code

- Expérimental
- Evolutif
- Protocole

Challenge valorisation

Ingénierie

- Logiciel
- Solution figée

Code

- Industriel

Enjeux

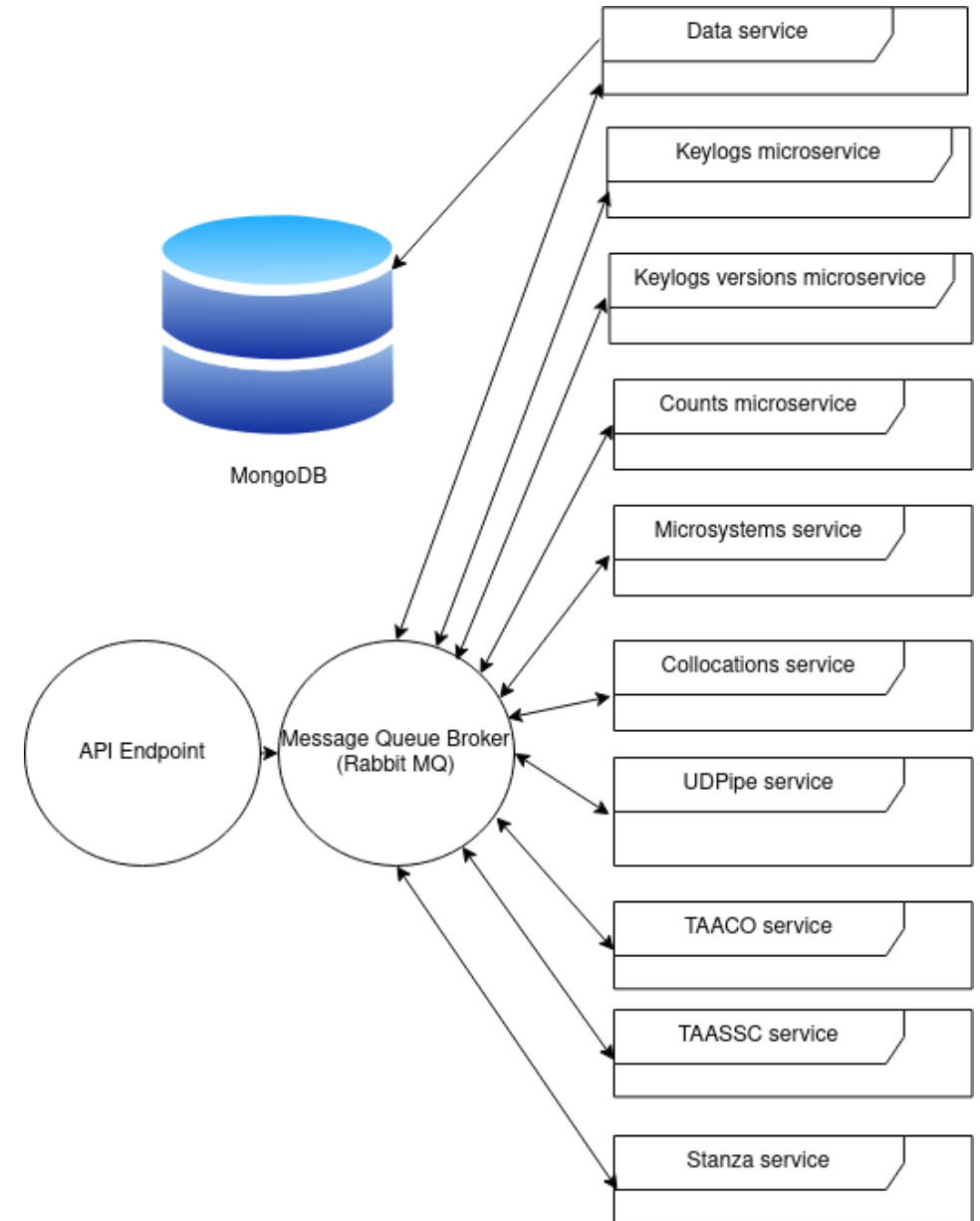
- Comment concilier recherche et valorisation en respectant les principes FAIR ?
 - Le code et de bonnes pratiques
 - Choix méthodologiques
- Comment concilier l'indépendance des QR dans un cadre unifié ?



Recherche ouverte

Modularité

- Chaque traitement = un module indépendant
 - Module nettoyage de données
 - Modules d'enrichissement des données
 - 1 module/famille de mesures
- 1 module = une validation ou justification scientifique
- Le code suit la démarche recherche et non l'inverse



Ouverture des données

Gitlab

- Code du nettoyage + anonymisation des données disponible
- Méthodologie transparente



Nakala

- Stockage ouvert des données
- Contribution à la recherche
- Diffusion des résultats du projet



Papiers de recherche

- Référencent le dépôt Nakala
- Justifient la méthodologie
- Résultats reproductibles



Ouverture du code

- Code des modules linguistiques indépendant du code de nettoyage des données (2 Gitlab distincts)
- La logique de code suit la méthodologie proposée
- Chaque module peut être modifié si la validation scientifique le requiert
- Contributions externes encouragées



Réutilisation/adaptation pour un autre projet de recherche

Reproductibilité des résultats

Validation indépendante de chaque contribution

Collaboration de recherche entre membres du projet



Valorisation du logiciel

Modularité : penser la valorisation dès le début

Développement
des modules en
parallèle

Ajout de modules
avec les résultats
de recherche

**Test/validation des
modules
indépendants**

Réflexion
d'architecture
logicielle en
amont

Choix
technologiques
délibérés et
maintenus



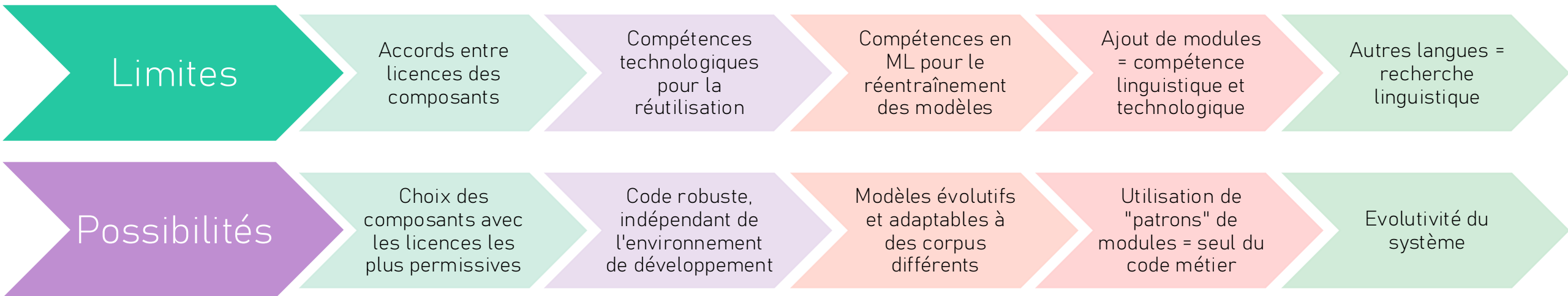
L'open source au cœur de la valorisation

- Système général sur Gitlab :
 - Disponible, ouvert, avec documentation
 - Transparence sur le cœur du traitement
 - Utilisable par la communauté -- mais connaissances techniques requises
- License importante
- Incubation en start-up basée sur un cœur open-source (début jan 2025)



Limites et possibilités

D'une limite vers une possibilité



Valorisation

Expérimentation

Ouvrir le code: des opportunités pour la recherche **et** la valorisation

