La linguistique de corpus à l'heure du code ouvert

LIDILE anr

A4LL

UNIVERSITÉ RENNES 2

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

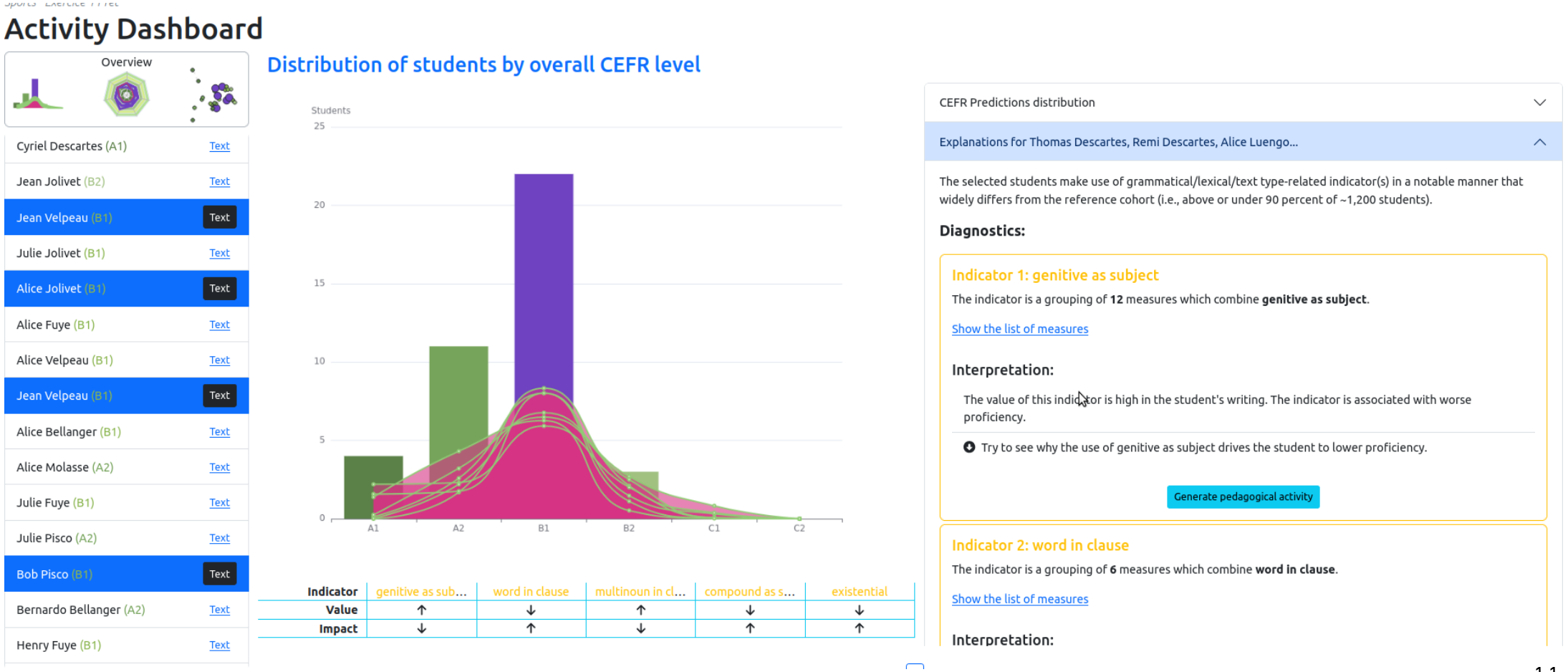
Analytics for Language Learning

- 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

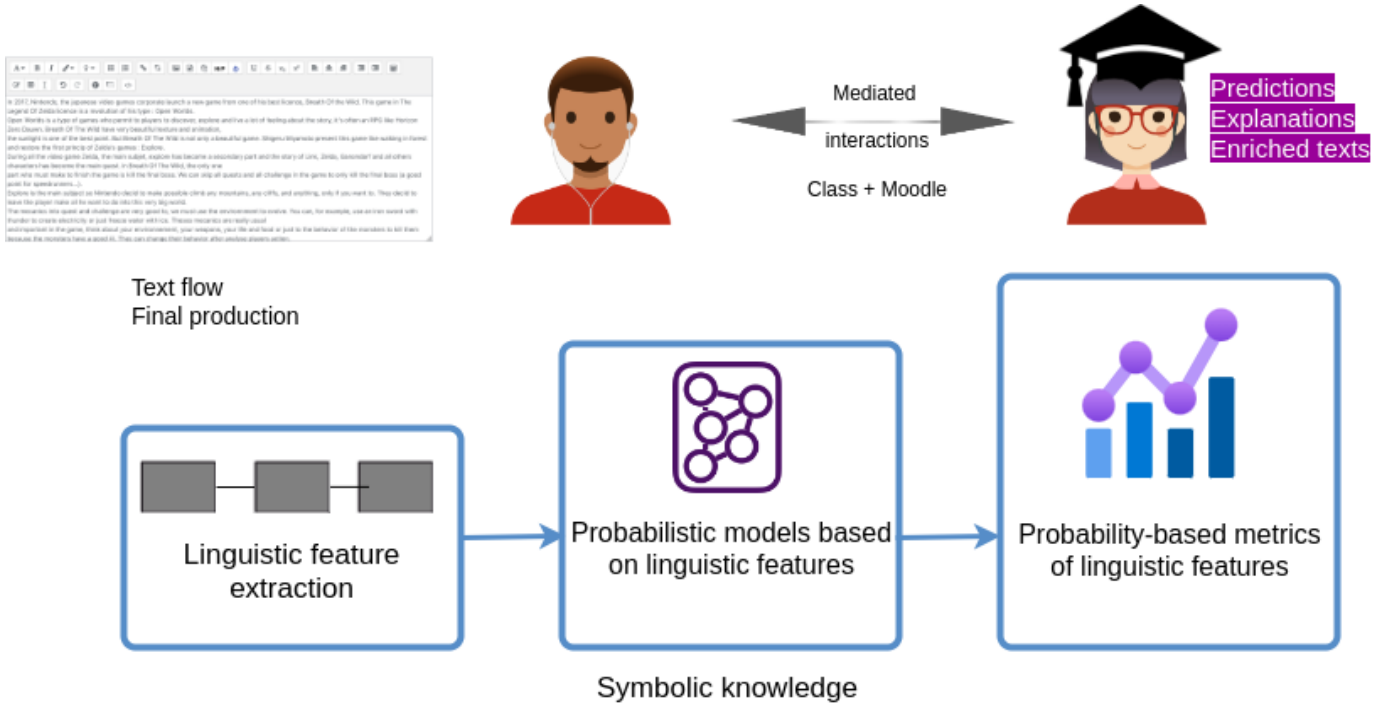
Analytics for Language Learning

- 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
- **Système de *monitoring* par analytics linguistiques** à destination des **enseignants**

Un système EIAH de monitoring



Un système EIAH



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

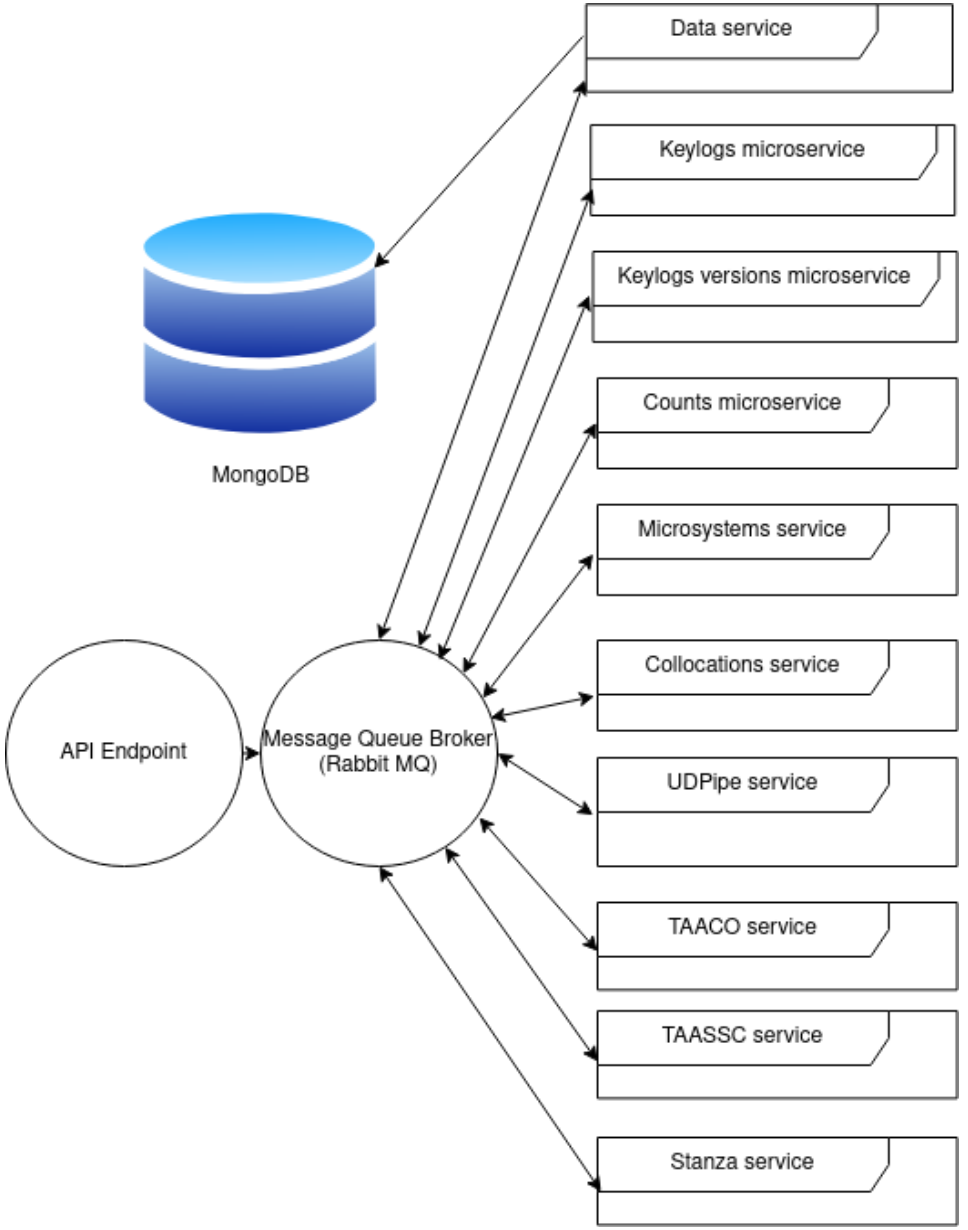
Enjeux

- 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

Huma-Num

Gitlab

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

GitLab

Nakala

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

nakala

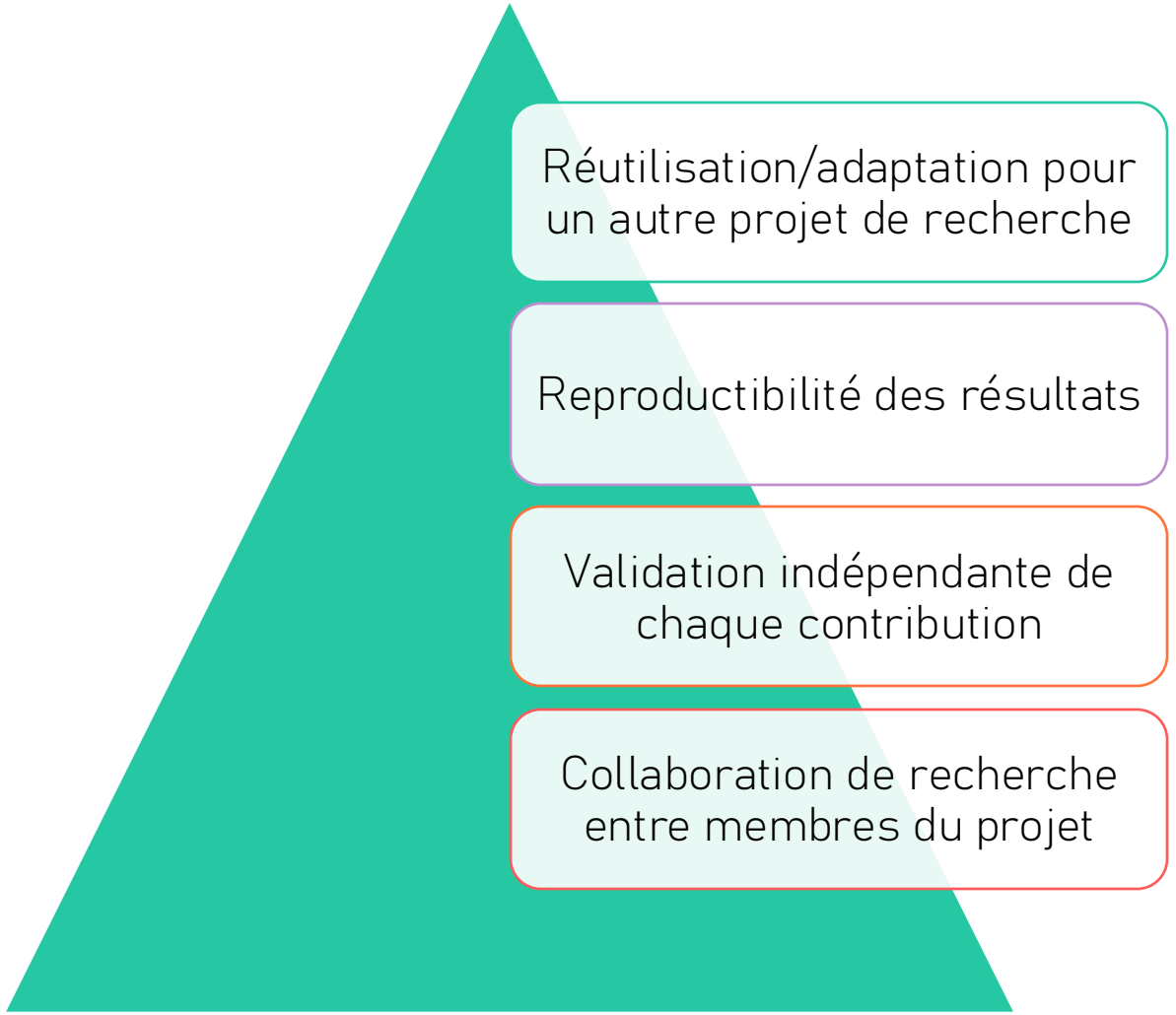
Papiers de recherche

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

HAL science ouverte

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



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

Limites	Accords entre licences des composants	Compétences technologiques pour la réutilisation	Compétences en ML pour le réentraînement des modèles	Ajout de modules = compétence linguistique et technologique	Autres langues = recherche linguistique
Possibilités	Choix des composants avec les licences les plus permissives	Code robuste, indépendant de l'environnement de développement	Modèles évolutifs et adaptables à des corpus différents	Utilisation de "patrons" de modules = seul du code métier	Evolutivité du système

Valorisation → Expérimentation

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