LIDILE Linguistique Ingénierie et Didactique des Langues

# ANR A4LL : Analytics for Language Learning

A4LL logo designed by Sidonie Tosser – Licence: CC-BY-NC 4.0

Pourquoi mon professeur d'anglais ne souligne-t-il jamais que les fautes dans mes rédactions ? Pourquoi la correction de ma rédaction prend-elle tant de temps ?

A4LL est maintenant accessible. Cliquez sur l'image pour accéder à la démo.



## Description

Le projet A4LL développera un système innovant d'analyse de l'apprentissage des langues conçu pour aider les enseignants et les apprenants grâce à des rapports objectifs reliant les compétences aux caractéristiques linguistiques. Thomas Gaillat, le coordinateur, propose une approche reposant sur des mesures textuelles opérationnalisant la complexité globale et structurelle, la phraséologie, la cohésion du discours et la fluidité. Ces mesures soutiendront la création automatique de rapports graphiques utilisés par les enseignants pour diagnostiquer les productions de leurs apprenants. L'ambition d'A4LL est de créer le premier système d'analyse L2 (langue seconde) entièrement automatisé au service des apprenants, des enseignants et des chercheurs universitaires via un flux de données intégré, de l'ingestion à l'analyse.

## Questions de recherche

Le projet A4LL fournira un système d'analyse de la L2 pour les apprenants et les enseignants d'anglais au niveau universitaire. Le projet abordera 3 questions de recherche principales visant à découvrir certaines des caractéristiques de l'interlangue, c'est-à-dire le système linguistique instable démontré par les apprenants d'une seconde langue : i) quelles sont les caractéristiques de la langue qui sont liées à des niveaux de compétence spécifiques ? ii) comment ces caractéristiques peuvent-elles être mesurées automatiquement ? iii) comment les mesures peuvent-elles être converties en analyses significatives pour un retour descriptif et des décisions d'enseignement ?

L'interlangue peut être considérée comme un système multifactoriel complexe qui rend difficile l'identification des critères de compétence. Avec le temps et la pratique, le système se stabilise progressivement. Cependant, il n'est pas évident de savoir quels facteurs sont en jeu à un moment donné. Pour comprendre comment l'interlangue se développe, les recherches actuelles montrent que les approches combinant des mesures linguistiques et des statistiques au sein de modèles informatiques permettent de mettre en évidence certaines caractéristiques de l'interlangue (Ballier et al., 2020 ; Yannakoudakis et al., 2018). Cependant, les métriques actuelles de l'état de l'art manquent de sens linguistique et nuisent ainsi à l'interprétabilité.

## Objectif

L'objectif est de développer un système informatique qui génère automatiquement des diagnostics linguistiques des écrits des apprenants. Ces diagnostics seront ainsi visualisés par les enseignants à travers MOODLE, l'un des principaux LMS open-source en France et dans le monde.Ces diagnostics aideront les enseignants à formuler des conseils à leurs élèves et à adapter leurs objectifs pédagogiques en fonction des profils de leurs groupes. Le développement du système impliquera un travail de recherche pour identifier les corrélations entre les caractéristiques linguistiques et les métadonnées, y compris les types de tâches, les compétences, les habitudes d'apprentissage et les capacités d'écriture.
Le système collectera, analysera automatiquement et fournira un retour d'information linguistique spécifique pour les écrits soumis dans MOODLE (voir figure 1).

En exploitant les métriques lexicales, syntaxiques et sémantiques, le système mettra en évidence les dimensions qui requièrent une attention particulière dans chaque texte.Des visualisations graphiques montreront les domaines linguistiques à améliorer pour atteindre un niveau de compétence donné.Le système s'appuiera sur une approche d'apprentissage supervisé avec des données d'apprenants collectées dans les deux centres de langues (en charge de 20 000 étudiants apprenant l'anglais à des fins spécifiques) des deux universités de Rennes. il sera modulaire pour permettre l'intégration ultérieure d'autres langues.

A4LL entend tirer parti de la force de deux prototypes précédemment développés auxquels le coordinateur a participé.

Le premier prototype, développé en 2019 (Gaillat, Simpkin, et al., 2021), permet une classification automatique des écrits des apprenants selon les niveaux du CECR.

Le second prototype, appelé VizLing (Gaillat, Knefati, et al., 2021), et développé en 2019, se concentre sur la génération automatique de graphiques pour visualiser la complexité linguistique dans les écrits.A4LL se développera dans la même voie, mais s'appuiera sur une sélection de métriques significatives et linguistiquement descriptives pour l'analyse des langues secondes. A4LL unifiera les tâches de traitement du langage naturel dans un cadre unique produisant des visualisations dans MOODLE.Il s'appuiera sur les métadonnées de l'apprenant afin de permettre aux enseignants d'établir le profil de leurs apprenants et de personnaliser le retour d'information.

L'objectif d'A4LL est donc :

1. d'offrir à la communauté des enseignants de langues des outils d'analyse de données permettant de positionner les apprenants en fonction de leurs compétences et des aspects de leur langue.
2. de modéliser le langage de l'apprenant afin de mettre en correspondance les caractéristiques linguistiques avec les compétences et, à terme, les étapes interlangues. A4LL entend apporter une solution aux centres de langues universitaires, en France et à l'étranger, qui ont en charge des millions d'étudiants qui étudient les langues à des fins professionnelles.

## Partenaires

| Établissement | Nom | Prénom | Fonction |
|---|---|---|---|
| *Rennes 2 University* | *GAILLAT* | *Thomas* | *PI & Associate Professor* |
| *Rennes 2 university* | MALLART | Cyriel | Research Engineer |
| *Rennes 2 University* | *LI* | *Jen-Yu* | *Ph.D. candidate* |
| *Rennes 2 University* | *FAUGERE* | *Anatole* | *Research Assistant and Computer programmer* |
| *University of Paris Cité* | *BALLIER* | *Nicolas* | *Professor of Linguistics* |
| *University of Paris Cité* | *LISSON* | *Paula* | *Research Engineer* |
| *University of Galway* | *SIMPKIN* | *Andrew* | *Associate professors in Statistics* |
| *University of Galway* | *STEARNS* | *Bernardo* | *Research Associate* |
| *Le Mans University* | *VENANT* | *Rémi* | *Associate Professor* |
| *IRISA / INSA Rennes* | *SÉBILLOT* | *Pascale* | *Professor of Computer Science* |
| *IRISA / CNRS* | *GRAVIER* | *Guillaume* | *Senior Research Scientist* |

## Projet Partenaire

Deep Learning for Language Assessment (DLLA)

## Annotateurs experts

### Annotation CEFR

| Établissement | Expert | Fonction | Structure |
|---|---|---|---|
| *Rennes 2 University* | Joanne Ward-Henry | English teacher | Centre de Langues |
| *Rennes 2 University* | Francoise Le Roux | English teacher | Centre de Langues |
| *University of Rennes* | Benedicte Dumont | English teacher | SCELVA |
| *University of Rennes* | Pascale Janvier | English teacher | SCELVA |

### Annotation linguistique

- Membres de l'équipe : Paula, Nicolas et Thomas
- Université Paris Cité - CLILLAC-ARP : Jessica Tayeh

## Conférences et publications

2026

Conference papers

titre
Validation procedures for structured learner data augmentation with the MatchIt algorithm
auteur
Thomas Gaillat, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Cyriel Mallart, Bernardo Stearns, Jen-Yu Li
article
*Learner Corpus Research Conference LCR2026*, Department of Linguistics at the Faculty of Arts, Charles University, Sep 2026, Praha ; Prague, Czech Republic
Accès au bibtex

titre
Automatic proficiency assessment of L2 writings using parameters of description of the CEFR
auteur
Thomas Gaillat, Nicolas Ballier, Rémi Venant, Andrew Simpkin, Jen-Yu Li, Bernardo Stearns, Cyriel Mallart
article
*Learner Corpus Research Conference 2026 (LCR 2026)*, Department of Linguistics at the Faculty of Arts, Charles University, Sep 2026, Prague (CZ), Czech Republic

Accès au bibtex

titre
A platform for teacher-centred analytics and diagnostics in learner English writings
auteur
Thomas Gaillat
article
*AI and Language Education*, INALCO Euniwell, Mar 2026, Paris, France
Accès au bibtex

2025

Journal articles

titre
Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English
auteur
Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Paula Lissón, Rémi Venant, Bernardo Stearns, Jen-Yu Li, Thomas Gaillat
article
*Research Methods in Applied Linguistics*, 2025, 4 (3), pp.100238. ⟨10.1016/j.rmal.2025.100238⟩
Accès au texte intégral et bibtex

titre
Exploring the cross-lingual influence of linguistic complexity in second language writing assessment
auteur
Sara Geremia, Thomas Gaillat, Nicolas Ballier, Andrew Simpkin
article
*Assessing Writing*, 2025, 66, pp.100951. ⟨10.1016/j.asw.2025.100951⟩
Accès au bibtex

Conference papers

titre
Visualiser les profils linguistiques des écrits d'étudiants en anglais L2
auteur
Thomas Gaillat, Rémi Venant, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns, Cyriel Mallart, Andrew Simpkin
article
*RANACLES*, Université de Nantes, Nov 2025, Nantes (France), France
Accès au bibtex

titre
Prédire et explorer les niveaux CECR dans les écrits d'étudiants en anglais L2
auteur
Thomas Gaillat, Bernardo Stearns, Rémi Venant, Cyriel Mallart, Nicolas Ballier, Andrew Simpkin
article
*CLES 25 ans*, Coordination Nationale CLES MESR, Oct 2025, Paris, France
Accès au bibtex

titre
Assessing the statistical validity of multi-noun alternation metrics as features of L2-English proficiency
auteur
Thomas Gaillat, Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Rémi Venant, Bernardo Stearns, Jen-Yu Li, Paula Lissón
article
*58th Annual Meeting of the Societas Linguistica Europaea*, Université Bordeaux Montaigne, Aug 2025, Bordeaux, France
Accès au bibtex

titre
Predicting CEFR levels for learners of English with keylogging metrics, an exploratory study
auteur
Ahood Al Sawar, Erin Pacquetet, Cyriel Mallart, Andrew J. Simpkin, Nicolas Ballier
article
*CORIA-TALN 2025*, Université d'Aix-Marseille et les UMR CNRS LIS et LPL, Jun 2025, Marseille, France.
https://talnarchives.atala.org/ateliers/2025/DYNTAL/index.html
Accès au texte intégral et bibtex

titre
Actionability in CALL: linking proficiency prediction models to interpretable indicators
auteur
Thomas Gaillat, Cyrielle Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen-Yu Li, Paula Lissón
article
*International Workshop on Foreign language learning and proficiency-rated reading materials: SLA research and AI methods supporting analysis and effective didactics in real-life education*, Universität Tübingen, Mar 2025, Tübingen, Allemagne, Germany
Accès au bibtex

titre
L'usage des collocations en anglais d'apprenants : une analyse croisée des L1 et des niveaux de compétence
auteur
Jen-Yu Li
article
*Approches interdisciplinaires des unités phraséologiques (UP) dans les langues du monde : Linguistique - TAL & IA - Traduction - Littérature*, Mar 2025, Paris, France
Accès au bibtex

Other publications

titre
Annotated English Verb Noun collocation dataset
auteur
Jen-Yu Li
article
2025
Accès au bibtex

titre
CELVA.Sp processed with A4LL metrics pipeline
auteur
Thomas Gaillat, Cyriel Mallart, Andrew J. Simpkin
article
2025, ⟨10.34847/nkl.3aba968r⟩
Accès au bibtex

Reports

titre
Project : A4LL Final report
auteur
Thomas Gaillat, Rémi Venant, Nicolas Ballier, Andrew Simpkin, Cyriel Mallart, Jen-Yu Li, Bernardo Stearns
article
ANR (Agence Nationale de la Recherche - France). 2025
Accès au texte intégral et bibtex

titre
A language-learning analytics system" project DMP
auteur
Nicolas Ballier, Thomas Gaillat, Jen-Yu Li, Cyriel Mallart, Andrew Simpkin, Bernardo Stearns, Rémi Venant
article
Opidor. 2025, https://dmp.opidor.fr/plans/13498
Accès au texte intégral et bibtex

Preprints, Working Papers, ...

titre
Assessing the validity of new paradigmatic complexity measures as criterial features for proficiency in L2 writings in English
auteur
Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Paula Lissón, Rémi Venant, Jen-Yu Li, Bernardo Stearns, Thomas Gaillat

article
2025
Accès au texte intégral et bibtex





2024

Conference papers

titre
La linguistique de corpus à l'heure du code ouvert
auteur
Cyrielle Mallart, Thomas Gaillat, Rémi Venant, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns
article
*Deuxième journée d'étude ARDoISE*, INRIA, Dec 2024, Rennes, France
Accès au bibtex



titre
Evaluating the Generalisation of an Artificial Learner
auteur
Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew J. Simpkin, John P. Mc Crae
article
*NLP4CALL2024 : Natural Language Processing for Computer-assisted Language Learning*, Université Rennes 2, France; University of Gothenburg, Sweden; Linköping University, Sweden, Oct 2024, Rennes, France
Accès au texte intégral et bibtex





titre
Overview of the linguistic features: creating measures – Joint presentation
auteur
Nicolas Ballier, Bernardo Stearns, Jen-Yu Li
article
*pre-conference workshop to NLP4CALL 2024*, Oct 2024, Rennes, France
Accès au bibtex



titre
Exploring learner knowledge with Large Language Models fine-tuned with the EFCAMDAT
auteur
Nicolas Ballier, Bernardo Stearns
article
*LCR2024 Learner Corpus Research conference*, University of Tartu; Learner Corpus Association, Sep 2024, Tartu (Estonie), Estonia
Accès au texte intégral et bibtex





titre
Assessing the validity of new structural complexity measures as features of proficiency in L2 English
auteur
Thomas Gaillat, Cyriel Mallart, Nicolas Ballier, Andrew Simpkin, Rémi Venant, Bernardo Stearns, Paula Lissón, Jen-Yu Li
article
*Learner Corpus Research Conference*, University of Tartu, Sep 2024, Tartu (Estonie), Estonia
Accès au bibtex



titre
Linguistic interoperability within a unified architecture
auteur
Thomas Gaillat, Cyrielle Mallart, Andrew J. Simpkin, Rémi Venant, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns
article
*Langues & Langage à la croisée des Disciplines - 1ère Rencontre annuelle LLcD*, Sorbonne Université; cnrs, Sep 2024, Paris, France
Accès au texte intégral et bibtex





titre

Analytics for Language Learning. Linguistic interoperability within a unified architecture

auteur

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen-Yu Li, Thomas Gaillat

article

*Langues & Langage à la croisée des Disciplines 1ère Rencontre annuelle LLcD*, Sep 2024, Paris, France

Accès au bibtex

Other publications

titre

Understanding Large Language Models

auteur

Cyriel Mallart

article

2024

Accès au bibtex

titre

Dictionary of Bigram-Score extracted from BNC with all association meausres by NLTK

auteur

Jen-Yu Li

article

2024

Accès au bibtex

Proceedings

titre

Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning

auteur

Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Griselda Drouet, Jen-Yu Li, David Alfter, Elena Volodina, Arne Jönsson

article

*The 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, Oct 2024, Rennes, France. LiU Electronic Press, 2024, Linköping electronic conference proceedings

Accès au texte intégral et bibtex

Software

titre

Analytics for Language Learning

auteur

Cyriel Mallart, Rémi Venant, Andrew Simpkin, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns, Thomas Gaillat

article

2024, ⟨swh:1:dir:d6ede95f5f25ec32fc400313c34d72a4fe12e9db;origin=https://gitlab.huma-num.fr/lidile/a4ll_mlpipeline.git;visit=swh:1:snp:a105a84ae8ae0f52445fc598870a45ed584c17e0;anchor=swh:1:rev:5c48738c7d9789988356711e248f300173ab95c6⟩

Accès au bibtex

2023

Scientific blog post

titre

CELVA.sp: A new learner language data set for the study of English for Specific Purposes at university level

auteur

Thomas Gaillat, Cyrielle Mallart, Rémi Venant, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns, Andrew Simpkin

article

2023

Accès au bibtex

Conference papers

titre

L'interopérabilité des corpus pour la modélisation des dynamiques d'acquisition de langue seconde

auteur

Thomas Gaillat, Cyrielle Mallart, Nicolas Ballier, Andrew Simpkin, Rémi Venant, Anatole Faugère, Bernardo Stearns, Jen-Yu Li, Paula Lissón

article

*Journée d'étude : « Corpus d'apprenants / corpus d'experts : Quels enseignements pour la caractérisation du discours scientifique ? »*, UR 3967 - CLILLAC-ARP : Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus - Atelier de Recherche sur la Parole; UFR EILA - Etudes Interculturelles de Langues Appliquées, Faculté Sociétés et Humanités d'Université Paris Cité, Dec 2023, Paris, France

Accès au bibtex

titre

Analytics for Language Learning: Interfacing MOODLE with A4LL via LTI

auteur

Thomas Gaillat, Cyrielle Mallart, Nicolas Ballier, Andrew Simpkin, Rémi Venant, Bernardo Stearns, Jen-Yu Li, Paula Lissón, Anatole Faugère

article

*Deep learning for language assessment closing event (DLLA Closing event 2023)*, UR 3967 - CLILLAC-ARP : Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole; UFR EILA de l'Université Paris Cité, Nov 2023, Paris, France

Accès au bibtex

titre

Exploring a New Grammatico-functional Type of Measure as Part of a Language Learning Expert System

auteur

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen-Yu Li, Thomas Gaillat

article

*Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Jul 2023, Toronto, Canada. pp.466-476, ⟨10.18653/v1/2023.bea-1.39⟩

Accès au texte intégral et bibtex

titre

Analytics for Language Learning : Transmettre aux enseignants les profils linguistiques de leurs apprenants

auteur

Thomas Gaillat, Cyrielle Mallart, Anatole Faugère, Andrew Simpkin, Bernardo Stearns, Paula Lissón, Jen-Yu Li, Nicolas Ballier, Rémi Venant

article

*Atelier GERAS @ 62e Congrès annuel de la SAES 2023*, Université Rennes 2; SAES La Sorbonne Nouvelle; GERAS (Groupe d'Etude et de Recherche en Anglais de Spécialité), Jun 2023, Rennes, France

Accès au bibtex

titre

Grammatical profiling with UD annotation (WiP)

auteur

Nicolas Ballier, Cyrielle Mallart, Thomas Gaillat

article

*Workshop on Profiling second language vocabulary and grammar*, University of Gothenburg, Humanisten., Apr 2023, Gothenburg, Sweden

Accès au bibtex

Poster communications

titre

Exploring Verb-Noun collocations in learner English

auteur

Jen-Yu Li, Cyriel Mallart, Thomas Gaillat, Elisabeth Richard

article

*Deep learning for language assessment (DLLA) closing event*, Nov 2023, Paris, France

Accès au texte intégral et bibtex

titre

Vers une grammaire probabiliste de microsystèmes fonctionnels en L2

auteur

Cyrielle Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen-Yu Li, Thomas Gaillat

article

*RéAL2: Grammaire(s) et acquisition des L2: Approches, trajectoires, interfaces,*, Oct 2023, Grenoble, France

Accès au texte intégral et bibtex

2022

Conference papers

titre

Language learning analytics : designing and testing new functional complexity measures in L2 writings

auteur

Thomas Gaillat

article

*11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)*, Dec 2022, Louvain la Neuve, Belgium. pp.55-60, ⟨10.3384/ecp190006⟩

Accès au texte intégral et bibtex





2018

Software

titre

CELVA.Sp corpus User Interface

auteur

Thomas Gaillat, Rémi Venant, Cyriel Mallart, Taylor Arnold, Anatole Fougère

article

2018, ⟨swh:1:dir:7405005eae86eb3f53662e5649f10f5c4f92e11a;origin=https://gitlab.huma-num.fr/lidile/celva.sp-ui;visit=swh:1:snp:198c7b3333fa18b5a721d36e06e8a5a0648600e3;anchor=swh:1:rev:95c370947852a8fe6ef9254069ca7812fd901188⟩

Accès au bibtex



## Livrables

### Logiciels

- Gitlab LIDILE: L'outil A4LL est disponible pour téléchargement et installation sur un serveur dédié (Technologie Docker). L'outil se décompose en deux paquets : A4LL_MLpipeline and A4LL_Tool
- Un outil pour la collecte de corpus d'apprenants : L'interface utilisateur CELVA.SP pour MOODLE
- Le Google Colab pour l'atelier «  Enrichissement linguistique des données textuelles » , une introduction à Python pour les experts en sciences humaines.

Soutenu par Rennes Métropole et ANR

### Ensembles de données et corpus

Corpus d'apprenants de la langue à des fins spécifiques : trois ensembles de données sur Nakala :

- Un avec l'annotation du CECR de Dialang
- Deux lots avec annotation CECR par des experts humains : 2018-2022 et 2023-2024

*Remerciements : Un grand merci aux enseignants de langues des universités de Rennes pour leur implication.*

18 décembre 2024