

Workshop on Low-Resource Languages and Domains (Toulouse)

17 novembre 2026 - 20 novembre 2026

campus Rangueil

Programme Scientifique

PROGRAMME AND HACKATHON

• 17 November 2026

8h45: Welcome coffee

8h55: Introduction

9h00-10h00: Damiano Spina, RMIT University (Australia)

Decolonising Information Retrieval Research

10h00-10h30: 3 short presentations

10h30-10h45: Coffee break

10h45-11h45: Jussi Karlgren, AMD Silo AI (Finland)

How to create culturally relevant instruction data sets and evaluation tests for generative language models

11h45-12h15: Mihaela Goga

12h15-14h00: Lunch

14h00-15h00: Djoerd Hiemstra, Radboud University (The Netherlands)

Blinder, it kin wier net! Language Identification on the Web for Low-Resource Languages

15h00-16h00: Tilahun Yeshambel, Addis Ababa University (Ethiopia)

Amharic in the Era of Language Models: Resources, Retrieval, and Open Challenges

16h00-16h15: Poster booster (3 mn each)

16h15-17h15: Coffee, discussions, posters

• 18 November 2026

9h00-10h00: Rodrigo Agerri, HiTZ Center - Ixa, University of the Basque Country UPV/EHU

10h00-10h30: 2 short presentations + poster booster

10h30-10h45 - Coffee and posters

10h45-11h45: Barbara Plank, LMU Munich (Germany)

11h45-12h15: 2 short presentations + Poster boosters

12h15-14h00 : Lunch

14h00-15h00 : Laurent Prévot, Aix Marseille Université (France)

What Can Existing Corpora Tell Us About Language Models? Linguistic and Conversational Sensitivity Across Languages

15h00-16h00: Adam Jatowt, University of Innsbruck (Austria)

Creating NLP Benchmarks for Low-resource Languages

16h00-16h15 - Coffee + poster

16h15-17h00 - Round table

• 19 November 2026

9h00-9h55: Shu Okabe, TUM Heilbronn (Germany)

Two case studies of NLP for endangered languages: parallel sentence mining and interlinear gloss generation.

9h55-10h05: Coffee break

10h05-11h00: Senja Pollak, Jožef Stefan Institute (Slovenia)

NLP for news analysis in Slovene

11h00-11h55: Eneko Agirre, University of the Basque Country UPV/EH (Spain)

Open LLMs for Your Language That Keep Pace with the Frontier: The Case for Basque

12h00-18h00 -Pizza - HACKATHON (including a break)

18h00 - Hackathon results

Posters presentations : Call for presentations is open until 13 October

--- HACKATHON

HACKATHON - L'IA et l'Occitan

Informaticien·ne ? Linguiste ? Vous avez besoin les uns des autres.

Si vous venez de l'informatique, vous pourrez mobiliser : Python, analyse de données, TAL, apprentissage automatique, évaluation, visualisation et expérimentation.

Si vous venez de la linguistique, vous pourrez mobiliser : syntaxe, morphologie, linguistique de

corpus, variation linguistique, annotation et analyse d'erreurs.

Les équipes les plus intéressantes seront celles qui croiseront ces deux regards.

LE DEFI

Les outils de traitement automatique des langues fonctionnent aujourd'hui très bien pour certaines grandes langues. Mais que se passe-t-il lorsqu'on les applique à une langue qui présente plusieurs variétés dialectales, des usages oraux et écrits, et beaucoup moins de données d'entraînement ?

Ce hackathon propose aux étudiant·es de Master et en école d'ingénieur, en informatique, TAL et linguistique, de travailler sur une vraie question de recherche : Dans quelle mesure les outils actuels de TAL sont-ils capables d'analyser l'occitan — et comment peut-on les améliorer ?

Vous travaillerez sur des corpus authentiques d'occitan représentant différents types de données : occitan écrit et oral, languedocien et gascon, parole spontanée et parole préparée. Ces données ont déjà été analysées automatiquement : segmentation en tokens, lemmatisation, étiquetage morphosyntaxique, traits morphologiques et dépendances syntaxiques.

Votre mission :

Mettre ces analyses automatiques à l'épreuve.

Où le système fonctionne-t-il bien ? Où se trompe-t-il ? Certaines variétés dialectales sont-elles plus difficiles à traiter ? La parole spontanée pose-t-elle davantage de problèmes que l'écrit ?

Peut-on identifier à l'avance les phrases qui risquent d'être mal analysées ? Et surtout : peut-on corriger automatiquement certaines erreurs récurrentes ?

Choisissez votre angle d'attaque !

Selon les compétences de votre équipe, vous pourrez par exemple :

Mesurer. Par exemple construire des scripts d'évaluation et comparer les performances de l'annotation automatique selon les corpus, les variétés et les types de données.

* Enquêter. Par exemple explorer les erreurs et identifier les phénomènes linguistiques qui mettent le système en difficulté.

* Visualiser. Par exemple transformer les résultats en visualisations permettant de comprendre où et pourquoi l'annotation automatique échoue.

* Améliorer. Par exemple proposer des règles de post-traitement, des heuristiques ou d'autres méthodes permettant de corriger certaines erreurs systématiques.

* Expérimenter. Par exemple tester un autre modèle ou une autre méthode de pré-annotation et vérifier si vous pouvez faire mieux que le système de référence.

Vous n'avez pas besoin d'être spécialiste de l'occitan ni des Universal Dependencies pour participer. Le hackathon est justement conçu comme une occasion d'apprendre en travaillant collectivement sur un problème concret.

QUE POURREZ-VOUS DECOUVRIR ?

À la fin du hackathon, votre équipe pourra par exemple produire :

- une évaluation quantitative d'une chaîne de traitement automatique de l'occitan ;
- une cartographie des erreurs montrant où le système rencontre le plus de difficultés ;
- une analyse de l'effet du dialecte, de l'oral et de l'écrit, ou du caractère spontané ou préparé des données ;
- des règles permettant de corriger automatiquement certaines erreurs ;
- des recommandations pour améliorer les futures données d'entraînement ;
- ou encore une comparaison montrant qu'un autre modèle fait mieux que la méthode de référence.

COMMENT PARTICIPER ?

Envoyez un mail à cimi26-organizers@irit.fr

Ou/Et inscrivez-vous <https://indico.math.cnrs.fr/event/16957/registrations/> et choisissez la seconde option. Vous pouvez vous inscrire sur le hackathon seulement, ou sur les différentes journées qui vous intéressent (programme <https://indico.math.cnrs.fr/event/16957/>)

POURQUOI PARTICIPER ?

Parce que vous ne travaillerez pas sur un exercice artificiel !

Vous vous attaquerez à une véritable question de recherche, à l'interface entre IA, traitement automatique des langues, linguistique et diversité linguistique.

Le hackathon sera l'occasion de :

- travailler sur de vraies données linguistiques annotées ;
- appliquer des compétences en programmation ou en analyse linguistique à un problème ouvert ;

- découvrir les Universal Dependencies et les méthodes d'évaluation en TAL ;
- collaborer avec des étudiant·es d'une autre discipline ;
- produire des résultats susceptibles de contribuer à l'amélioration des technologies linguistiques pour l'occitan.

1ier prix : un repas de groupe à l'Esplanade (Campus Rangueil, UT)

Pour tous :

12h00 Pizza partie et buffet

16h00 Pause café et mignardises

19 Novembre 2026 Université de Toulouse (Site Rangueil)