

ETAL : Rencontres Minute



ETAL

03/09/2025

Roscoff

Programme

1. Batch 1, AABID à Guichoux
2. Batch 2, Hernandez à Nazari
3. Batch 3, Nowakowski à Zhou
4. Hackathon time

Batch 1, AABID à Guichoux

Batch

	Cave-Laurène
1. Batch 1, AABID à Guichoux	Charpentier-Frederic
AABID-Mohsine	Clop-Cody
Adda-Pierre	ESCUDIE-Antony
Aubert-Julien	Guichoux-Teo
Azais-MarcAlexis	2. Batch 2, Hernandez à Nazari
Bezançon-Julien	3. Batch 3, Nowakowski à Zhou
Botcazou-Ivanhoe	4. Hackathon time
Caillard-Melusine	

Batch 1, AABID à Guichoux

AABID-Mohsine

Analyse des pratiques navigationnelle des utilisateurs

AABID Mohsine
ETAL 2025

En bref...

Analyser le comportement des utilisateurs à partir des fichiers logs dans une bibliothèque numérique:

Comment faire ?

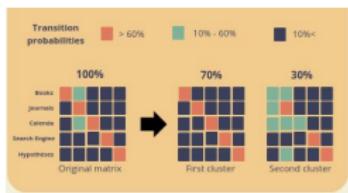
On peut générer des statistiques sur les requêtes individuelles , mais cela ne suffit pas à révéler des groupes de comportements collectifs distincts .

Problématiques clés :

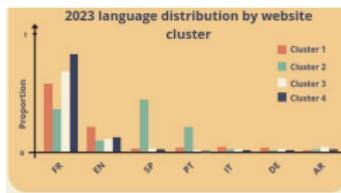
-  **Identification des sessions** : comment déterminer le début et la fin d'une session ?
-  **Filtrage des bots** : comment repérer les sessions de robots, dont l'usage diffère de celui des humains ?
-  **Extraction d'informations** : quels indicateurs ou connaissances pertinentes peut-on tirer de ces analyses ?

Qu'est-ce qu'on peut extraire ?

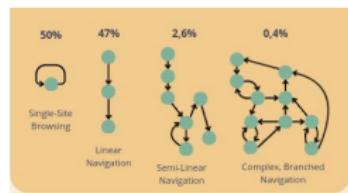
Matrice de transition



Extraction des préférences par groupes



Profil de navigation



Batch 1, AABID à Guichoux

Adda-Pierre

Thèse déjà soutenue en 2018:
physique des Plasmas au G2Elab (Grenoble) / UTC (Compiègne)

► 2024: Section R&D IA/Data du Groupe SII (ESN):

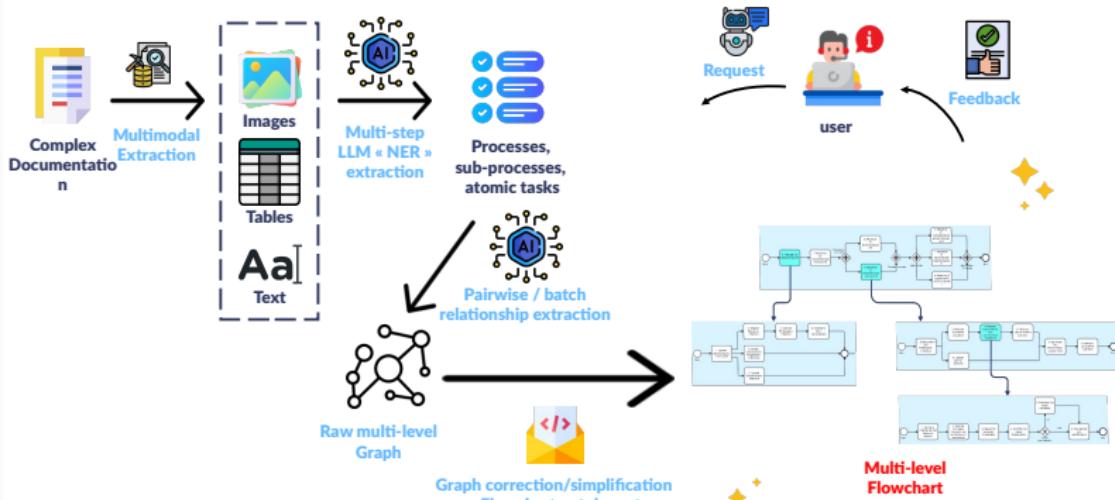


Travaux effectués depuis 2024:

- Extraction multimodale documentation complexe.
- Constitution de graphes de connaissances.
- RAG sur documentation d'entreprises.
- Classification de tickets utilisateurs (BERT).
- **Application IAgent**: génération de diagramme de flux à partir de texte

2

Automated Flowchart Generation: Principle



Batch 1, AABID à Guichoux

Aubert-Julien

Structurer une collection scientifique pour la montée en compétence.

Julien Aubert-Béduchaud - Doctorant (3ème année)



TALN

LS2N

Laboratoire des Sciences
du Numérique de Nantes



**Nantes
Université**



DGA



**AGENCE
INNOVATION
DéFENSE**

Liste de lecture

→ "Structured Generation of Technical Reading Lists" (Gordon et al., 2017)

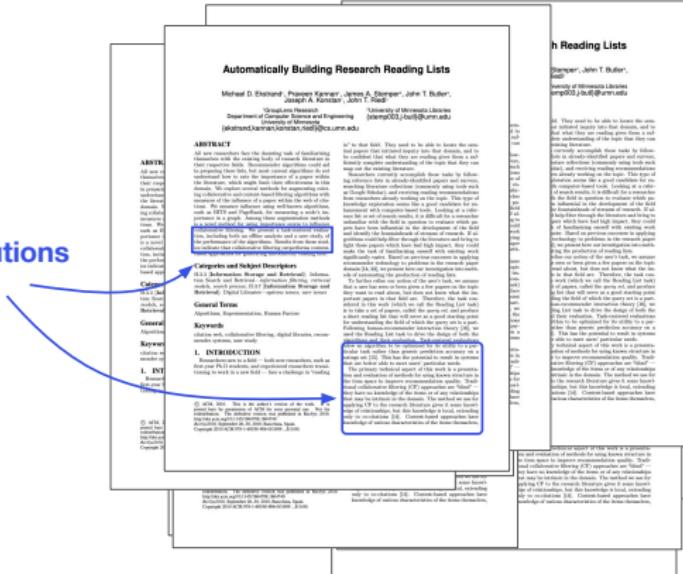
↳ PRE-REQUIS "Modeling Concept Dependencies in a Scientific Corpus" (Gordon et al., 2016)

→ "Automatically building research reading lists" (Ekstrand et al., 2010)

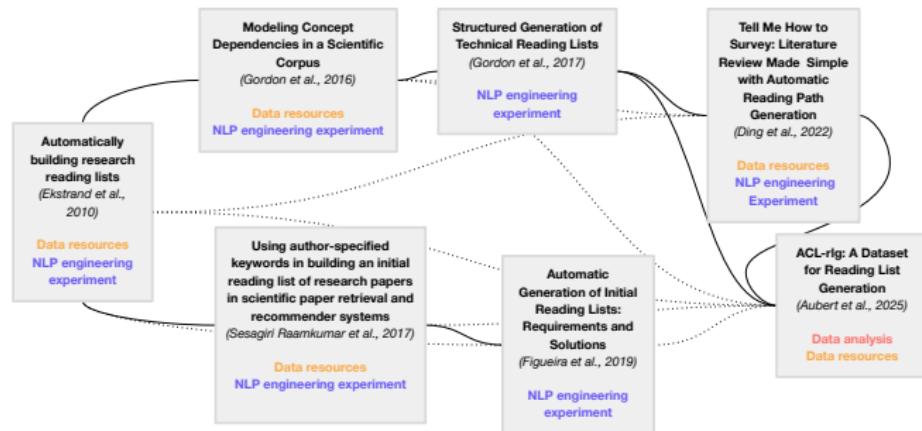
→ "Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems" (Sesagiri Raamkumar et al., 2017)

↳ ALTERNATIVE "Automatic Generation of Initial Reading Lists: Requirements and Solutions" (Figueira et al., 2019)

Contributions



Collection structurée



Batch 1, AABID à Guichoux

Azais-MarcAlexis



Regroupement Thématique des Commentaires Clients des Hôtels

Marc-Alexis Azais* Mickael Coustaty Jean-Loup Guillaumet
Laboratoire Informatique, Image et Interaction (L3i)
*RMD Technologies
[\[prenom.nom\]@univ-lr.fr](mailto:[prenom.nom]@univ-lr.fr)

Marc-Alexis Azaïs



Contexte

- On souhaite analyser l'**attractivité** des propositions hôtelières d'un secteur
 - L'**attractivité** d'un hôtel dépend de plusieurs critères :
 - Qualité des **chambres**
 - **Services** disponibles
 - **Environnement** alentours
 - La réputation de l'hôtel sur le web est un indicateur de son attractivité

Objectif

- **Objectif** : évaluer **forces et lacunes** de l'attractivité d'un hôtel à travers les avis clients.
 - Ces avis clients contiennent des aspects **explicites** et **implicites**.
 - **Aspect explicite** : directement mentionné dans l'avis
 - **Aspect implicite** : élément évalué mais non mentionné directement, inféré par le lecteur.
 - De manière pragmatique, il faut identifier des **thématisques** d'aspects et les assignés aux avis.
 - Nous développons **HospCSE**, un modèle qui projette les avis dans un espace vectoriel pour :
 - Regrouper les commentaires par thématiques et sentiments (positifs et négatifs).
 - Identifier les thématiques majeures.



Localisation sur une rue animée !

Nuisance Nocturne

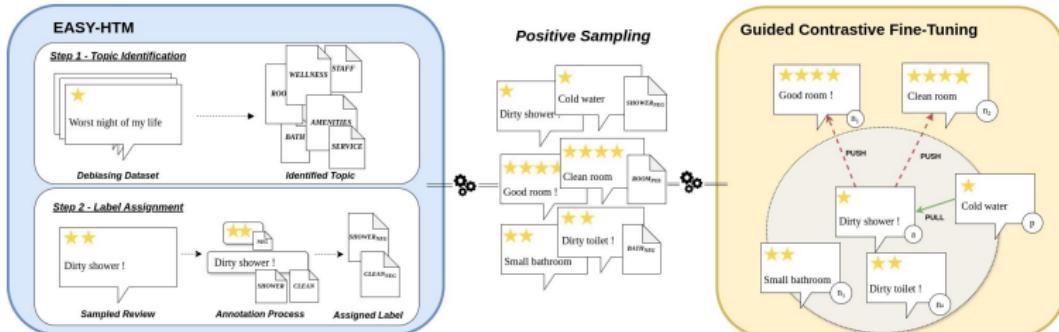
Quartier Svmpa

La localisation est un aspect **explicite**, mais le lecteur saisit **implicitelement** d'autres informations.



Aspects regroupés par thématique avec Word2Vec

Méthode : Hospitality Contrastive Sentence Embeddings



Embarrassingly Applicable Simple Yet - Hospitality Topic Model

- Identification des topics par clustering des aspects les plus fréquents

Batch 1, AABID à Guichoux

Bezançon-Julien

Extraction automatique d'expressions multi-mots défigées : une approche frugale, multilingue et explicable

Julien Bezançon

sous la direction de Gilles Siouffi, Gaël Lejeune et Antoine Gautier

03/09/2025

Sorbonne Université

Exemples de défigement capturés

travailler plus pour gagner plus

Candidat	Score	Fréquence
travailler plus pour gagner moins	0.90	307
travailler plus pour gagner plus	1.00	224
travailler moins pour gagner plus	0.90	141
travailler moins et gagner plus	0.69	37
travailler plus efficacement au quotidien	0.61	33
travailler plus pour payer plus	0.90	20
travailler pour gagner plus	0.76	16
travailler plus longtemps pour gagner moins	0.84	13
travailler plus pour vivre moins	0.78	13
travailler plus pour rembourser plus	0.90	12
travailler plus pour gagner autant	0.90	12

Table 1 – Candidats classés selon leur fréquence avec ASMR.

Notion de < moule > de défigement

Once upon a time in the West
The good, the bad and the ugly
a fistful of dollars
for a few dollars more

once upon a time in X	505
the good, the bad and the X	481
the good, the X and the X	356
the good, the bad & the X	150
a fistful of X	138
the good, the X and the ugly	92
the good, the X & the X	79
the good, the X & the ugly	24
for a few X more	20
for a fistful of X	12

Table 2 – Paternes de *snowclones* correspondant aux films de Sergio Leone avec leur nombre de *snowclones* affiliés.

Batch 1, AABID à Guichoux

Botcazou-Ivanhoe

Planning-constrained Language Models for the Abstract Summarization Task

Ivanhoé Botcazou

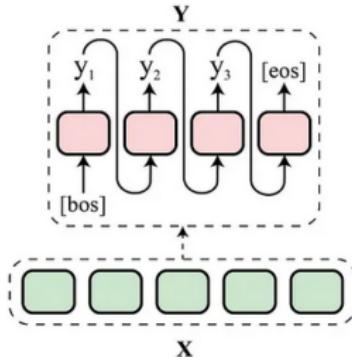
September 03, 2025

LERIA

Auto-regressive models, the current state of the art

Key Concepts:

- Transformer decoder architecture, widely used in generation tasks (**Radford et al., 2019**).
- Stochastic prediction of the next token based on previous content.



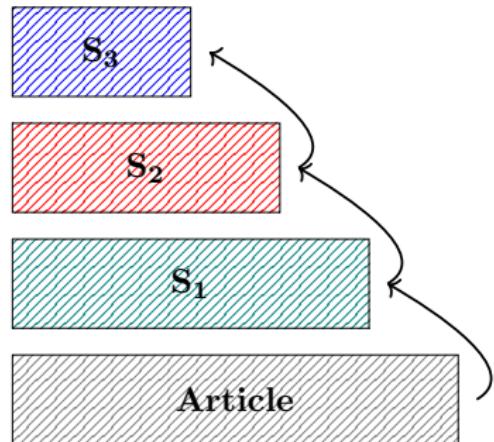
Current NLP systems face persistent challenges

- Lack of explainability in the generated planning which is typically handled via prompt instructions.
- Generation length control is primarily prompt-dependent (e.g., "*tl;dr in 100 words*").

The main aspect of this thesis research

Context & Motivations

- Generate autoregressively multiple summary drafts under planning constraints.
- Control the target lengths using an autoregressive model.
- Ensure control over factuality, faithfulness, and precision.



Research Question

- How can we build pyramidal summaries of an article using weak supervision?
- What are the benefits in terms of planning and explainability?

Batch 1, AABID à Guichoux

Caillard-Melusine



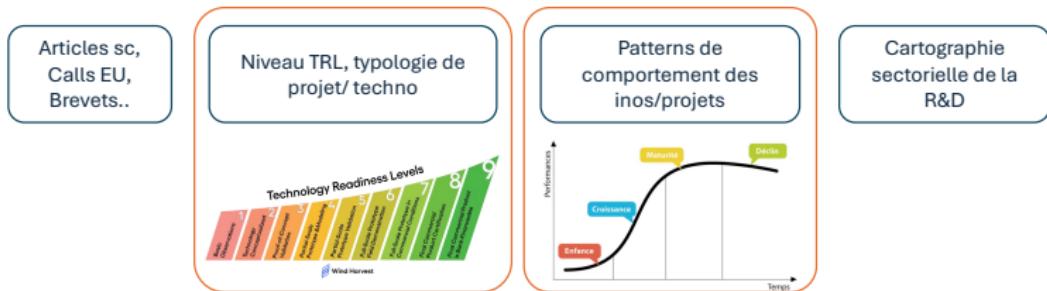
Modélisation sémantique et prédictive pour la détection des tendances émergentes et l'analyse prospective dans des corpus multi-sources

Encadrement:

Gaël Lejeune, Aoussat Améziane, Pierre-Emmanuel Fayemi

Mélusine Caillard – ETAL2025

Modélisation sémantique et prédictive pour la détection des tendances émergentes et l'analyse prospective dans des corpus multi-sources



Mélusine Caillard – ETAL2025

Batch 1, AABID à Guichoux

Cave-Laurène



Communs démocratiques - Aide à la rédaction

Laurène Cave

Sorbonne Université

Ma thèse

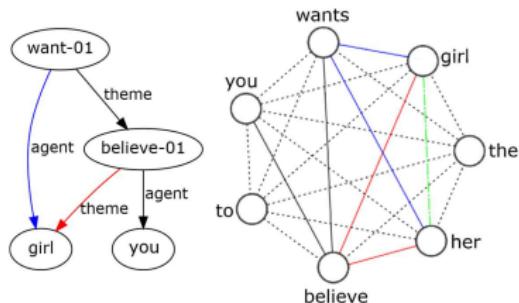
- Cadre des débats citoyens en ligne
- Aide à la rédaction (/ compréhension) de messages
- Biais de l'IA
- Pour le moment : travail sur l'ambiguité des termes

Batch 1, AABID à Guichoux

Charpentier-Frederic

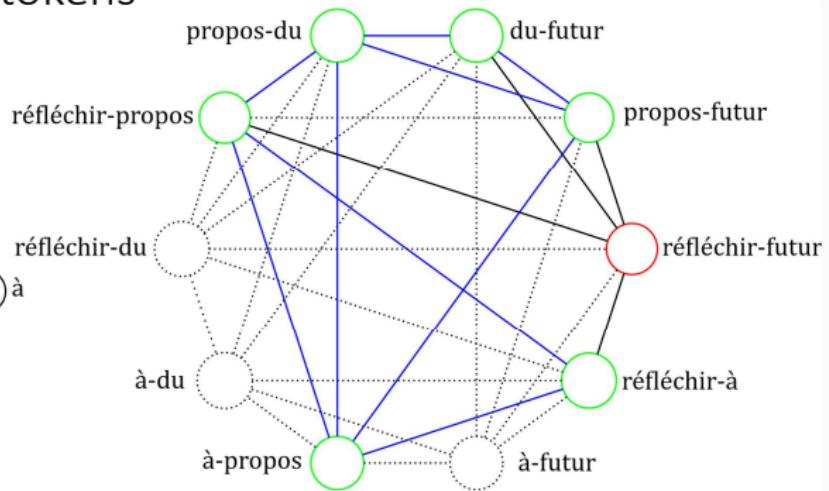
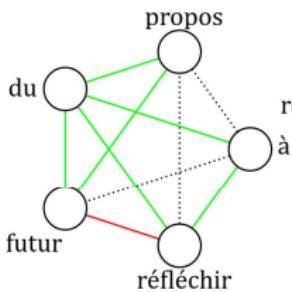
Frédéric Charpentier

AMRs



The girl wants you to believe her

Graphe de tokens



Batch 1, AABID à Guichoux

Clop-Cody

THALES

Sécurité des agents autonomes basés sur des LLMs

Doctorant : Cody Clop

Directeur : Benoit Favre

Co-directeur : Frédéric Béchet

Encadrant Entreprise : Yannick Teglia





Sécurité des agents autonomes basés sur des LLMs

- Systèmes d'agents
- Multimodalité Vision/Texte
- Retrieval Augmented Generation
- Attaques par injection de prompts
- Empoisonnement des données

clop.cody@gmail.com



Batch 1, AABID à Guichoux

ESCUDIE-Antony



Présentation

Antony Escudie, Dr. en physique des particules/astroparticules
Expert technique/Ingénieur R&D, Data et IA - DSN/DRCI du CHU d'Angers





RECHERCHE

Recherche - Maladies rares : P-AI

- **Besoin** : limiter l'errance diagnostique et optimiser la prise en charge d'individus porteurs de maladies rares dermatologiques
- **Solution** : Utilisation du traitement automatique du langage (NLP) pour analyser des centaines de milliers de comptes rendus d'hospitalisation et de consultation (urgences, dermatologie,...)
- **Objectifs** : pouvoir détecter rapidement des patients présentant une maladie rare pour une prise en charge médicale adaptée et précoce, sans être trop sélectif (risque de détecter que les patients déjà diagnostiqués) !
- Benchmark de plusieurs modèles de ML/DL
- Travail en cours avec un partenaire industriel



<https://www.turing.com/kb/natural-language-processing-function-in-ai/>



Guide Numérique Intelligent : Plateforme d'IA Gen CHU

Hébergement CHU



Une interface web qui permet à l'utilisateur d'interagir avec les modèles d'IA de manière graphique.



Possibilité « d'entrainer » un modèle sur les données du CHU (doc utilisateur, doc technique, ...)

Possibilité de faire une recherche sur internet et d'agrégier les résultats de différents moteurs.

SearXNG



Un orchestrateur qui est utilisé pour appeler et gérer des modèles d'IA, en local ou à distance.

Une infrastructure de calcul avec GPU





Data Quality - Example : StruQ (Structure Quality)

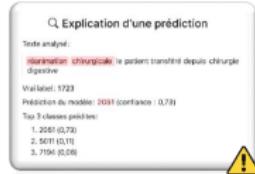
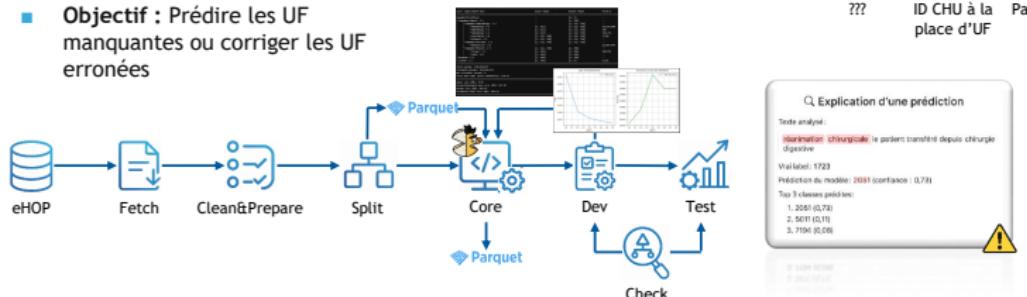
- Constat :** Données présentes dans notre entrepôt de données de santé sans UF associée ou avec UF erronée
- Objectif :** Prédire les UF manquantes ou corriger les UF erronées

ID_ENTREPOT	A/+ ID_DOC_SOURCE	123 ID_LOT	DATE_MAJ	A/+ UF	A/+ UF_EXEC	A/+ UM
2019-11-15 02:56:00.000	[NULL]	[NULL]	2022-10-17 16:10:00.000	10001561488	7013	[NULL]
2022-10-17 15:45:00.000	10002585692	7013				

???

ID CHU à la place d'UF

Pas d'UF



Encadrement d'une thèse sur le sujet de la qualité des données : Évaluation et amélioration de la qualité des données médicales structurées et non structurées pour le développement d'algorithmes d'intelligence artificielle en santé



Merci !

Antony Escudie, Dr. en physique des particules/astroparticules
Expert technique/Ingénieur R&D, Data et IA - DSN/DRCI du CHU d'Angers



Batch 1, AABID à Guichoux

Guichoux-Teo

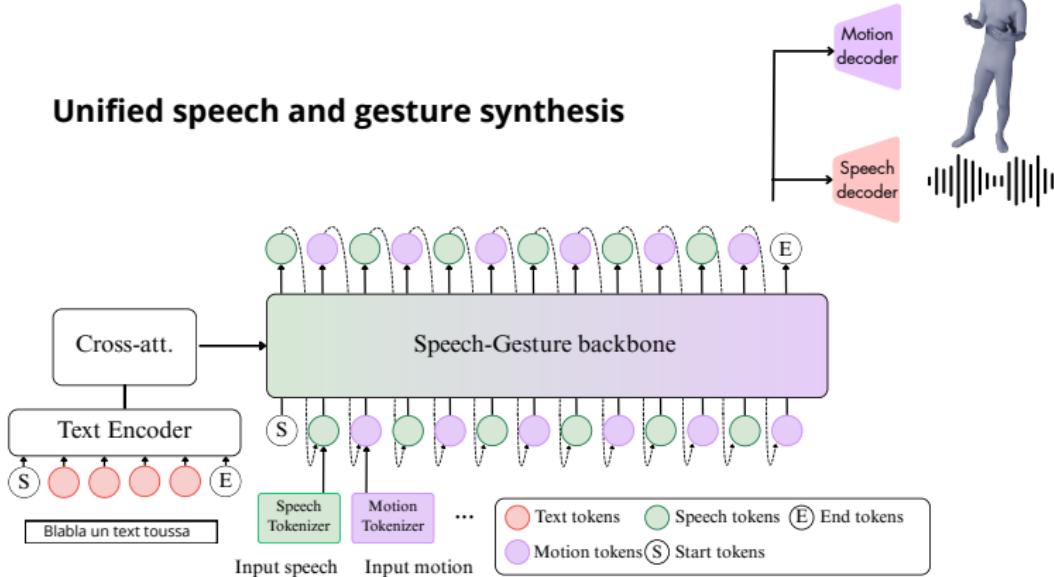
Multimodal Speech synthesis

Téo Guichoux

ISIR
IRCAM
Sorbonne Université



Unified speech and gesture synthesis



Batch 2, Hernandez à Nazari

Batch

	Lemesle-Quentin
1. Batch 1, AABID à Guichoux	Lepagnol-Pierre
2. Batch 2, Hernandez à Nazari	Liu-Jianying
HERNANDEZ-Marceau	Mahoudeau-Margot
Jara/aygalic	Maurel-Thibault
Jouitteau-Melanie	Nazari-Parisa
Kebir-Ahmed Rayane	3. Batch 3, Nowakowski à Zhou
Kerdraon-Gabriel	4. Hackathon time
Koudoro Parfait-Caroline	

Batch 2, Hernandez à Nazari

HERNANDEZ-Marceau

Analyse de la circulation de chansons par similarité multimodale dans des textes français des 17^{ème} et 18^{ème} siècles

Comité de Suivi Individuel

Marceau HERNANDEZ

1^{ère} année de doctorat

03/09/2025



En bref

- ▶ Étude de la **diffusion** des idées au travers des textes, parfois oralisés
- ▶ Premier axe : **Lien entre paroles et airs** de chansons

Quoi ? Détecter # de strophes, # de syllabes, # de vers et rimes

- ▶ Selon la métrique, correspondance possible avec un air

Pourquoi ? Lier les paroles à leurs airs ou entre elles

- ▶ La réutilisation d'airs n'étant pas au hasard

Comment ? Réseau neuronal passant d'un vers à sa suite de noyaux vocaliques

Comment ? II À partir des noyaux vocaliques, on a # de syllabes et rimes

Air à la coupe

Réutiliser la bibliothèque créée pour le passage d'**airs musicaux aux coupes** :

- ▶ **Annotation manuelle** d'un recueil d'airs -> Clé Du Caveau (CDC)
- ▶ Lier les airs aux paroles à partir des sorties finales

Expérimentation sur les **représentations intermédiaires** :

- ▶ Pouvoir calculer la **distance** entre les paroles
- ▶ Projection dans un **espace commun** des airs et paroles

Récupérer les **informations de strophes** :

- ▶ Méthode de **segmentation alternative** Clérice (2023)
 - ▶ Étiquetant les zones du document avec YOLOv5 Jocher et al. (2022)
- ▶ Déetecter **variations d'espacements** causés par les alinéas et sauts de lignes

Batch 2, Hernandez à Nazari

Jara-aygalic

Les hallucinations des LLMs

Aygalic Jara—Mikolajczak

1^{re} année de doctorat
Contrat CIFRE



Déetecter les hallucinations dans les compréhensions écrites

The screenshot shows a Wikipedia page for "Roscoff". The title is "Roscoff". Below it are tabs for "Article" and "Discussion". The main content starts with: "Roscoff (/ʁɔskɔf/ ; en breton : Rosko) est une commune française du Léon située sur la côte nord de la Bretagne, dans le département du Finistère." It continues with a paragraph about its history as a port for corsairs and контрабандистов, mentioning the Johnnies and their onions, and its preservation of architectural heritage from the XVth and XVIth centuries. It also notes its port for Brittany Ferries.

Below the text, a question is asked: "Où se situe la ville de Roscoff ?"

Réponse à la question



La ville de Roscoff se situe en Europe de l'Est

Modèles :

- Gemma 2
- Llama 2
- ...

Détection des tokens hallucinés



La ville de Roscoff se situe en **Europe de l'Est**

Modèles :

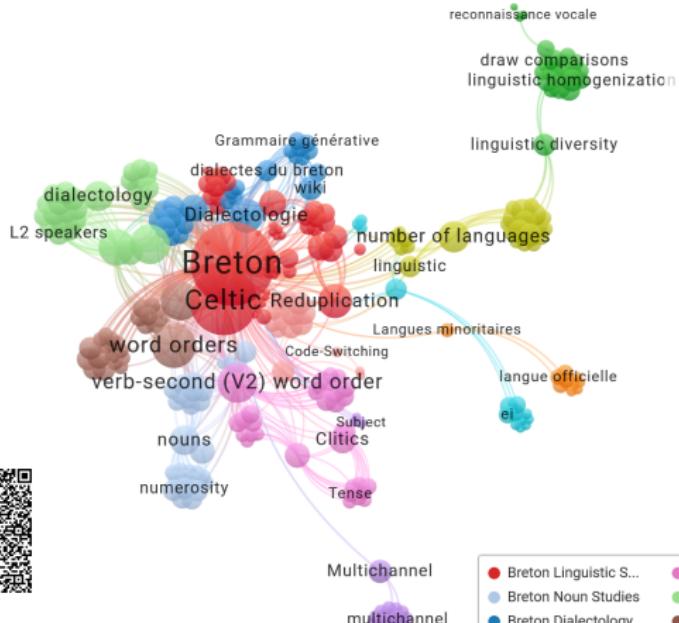
- Sparse AutoEncoders
- Linear Probes

Batch 2, Hernandez à Nazari

Jouitteau-Melanie

Si on ne s'occupe pas du TAL, le TAL s'occupe de nous

> "This is a visualization of a domains network from all publications."



- Breton Linguistic S...
- Breton Syntax Analy...
- Breton Noun Studies
- Breton Dialectology
- Breton Syntax and W...

73
29

CV

Sauver ce qui p/veut l'être

>Avant tout ne pas nuire

- Inventaire des langues de l'État français et de leurs ressources sociolinguistiques, linguistiques, numériques
<https://entrelangues.modyco.fr/index.php/>
- Outiller la mise en puissance des locuteurs
 - Wikigrammaire pour human
= corpus à diversité grammaticale construite
<https://arbres.iker.cnrs.fr>
 - ANR Yezh ar vRo (& Grobol, Millour, Antoine)
- appli de parcours sonores = collecte de son métadonnées
- plateforme de transcription collaborative toolkit pédagogique =
corpus aligné
<https://arbres.iker.cnrs.fr/YAR>

Batch 2, Hernandez à Nazari

Kebir-Ahmed Rayane

Clarifying User Information Needs in Conversational Search Systems

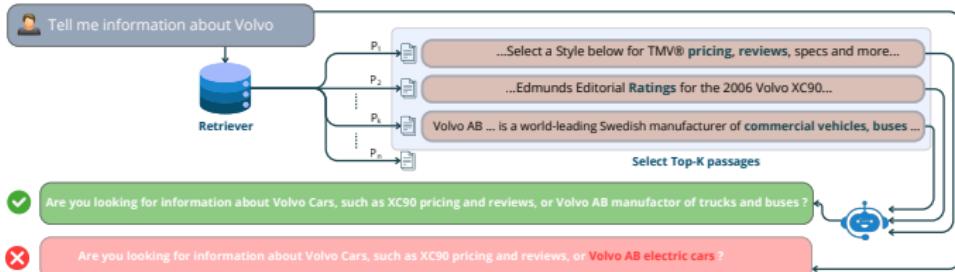


Figure 1 : Example of faithful versus unfaithful generated clarifying question, by a conversational search system, with the goal of answering user query based on retrieval results.

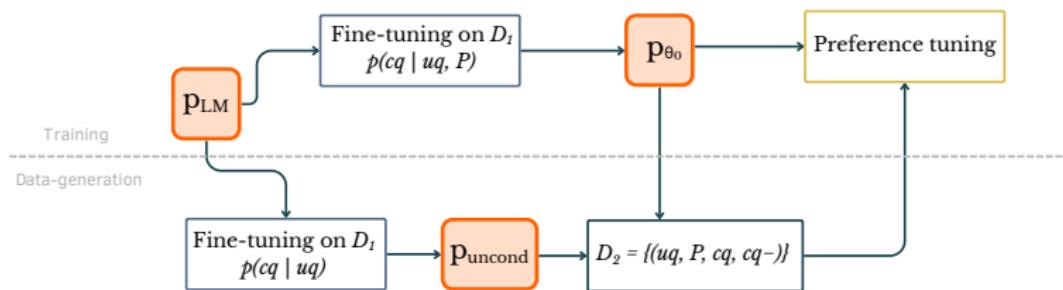


Figure 2 : RAC framework pipeline, for training LLMs to generate faithful clarifying questions conditioned by relevant retrieved passages.

Batch 2, Hernandez à Nazari

Kerdraon-Gabriel

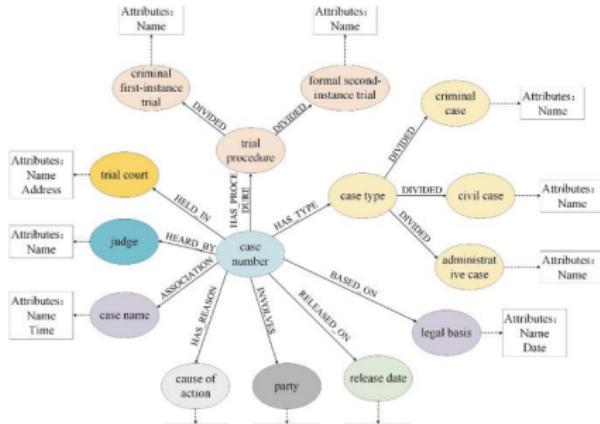
Qu'est-ce qu'une ontologie juridique ?

Une ontologie juridique est une description formelle et partagée des notions du droit et des liens qui les relient.

Elle sert à unifier des sources hétérogènes, à indexer les textes par concepts et à vérifier la cohérence des renvois, pour des résultats traçables.

Exemples :

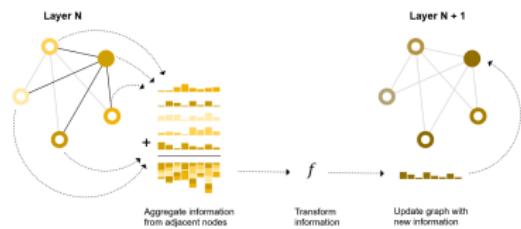
- Entités « Norme », « Institution », « Procédure »
- Relations « fait partie de », « cite », « modifie/abroge »
- Contraintes de compétence et de temporalité.



Construire des ontologies plus fiables : l'apport des réseaux de neurones de graphe

Dans la pratique, une ontologie est souvent représentée sous forme de graphe, enrichi de règles logiques qui imposent la validité des relations et permettent de raisonner sur les textes.

- Les réseaux de neurones de graphe exploitent la structure relationnelle (hiérarchie, citations, versions) pour désambiguer, typer et compléter les liens, en imposant une cohérence globale



Batch 2, Hernandez à Nazari

Koudoro Parfait-Caroline

Des IA au service de l'espace littéraire du XIX^e siècle : évaluation et analyse des outils de reconnaissance d'entités nommées spatiales

Caroline Koudoro-Parfait
caroline.parfait@sorbonne-universite.fr

6 Janvier 2025

Observatoire des Textes des Idées et des Corpus - Obtic,
Sorbonne Center for Artificial Intelligence - SCAI,
Sens Textes Informatique Histoire - STIH EA 4509, Sorbonne Université



Travaux de Thèse



Dépôt HAL



Épiméthée

Connecting Metropolitan and Colonial Societies through Notarial Archives : Digital Challenges for Information Extraction in Socio-Historical Research

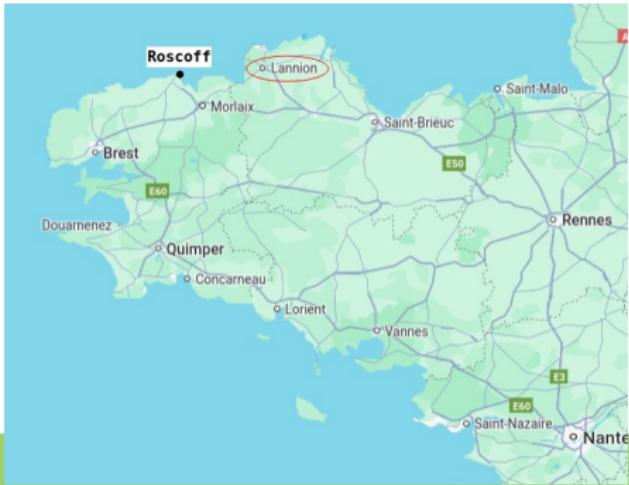
mots clés : Archives, HTR, Données bruitées, REN, Analyse de réseaux

- Université de Trèves, Centre des humanités numériques ;
- Damien Tricoire, Historien moderniste ;

Batch 2, Hernandez à Nazari

Lemesle-Quentin

Quentin LEMESLE



- Rentre en 3ème année



Sujet: Génération de paraphrases *(en raccourci)*

Je me suis intéressé à:

- La proximité sémantique à l'échelle phrase

En ce moment je m'intéresse à:

- L'affinage de *LLM* sans vérité terrain
- L'évaluation de génération de paraphrase par *LLM*



Batch 2, Hernandez à Nazari

Lepagnol-Pierre



Prompting de LM pour des tâches TAL

ETAL 2025

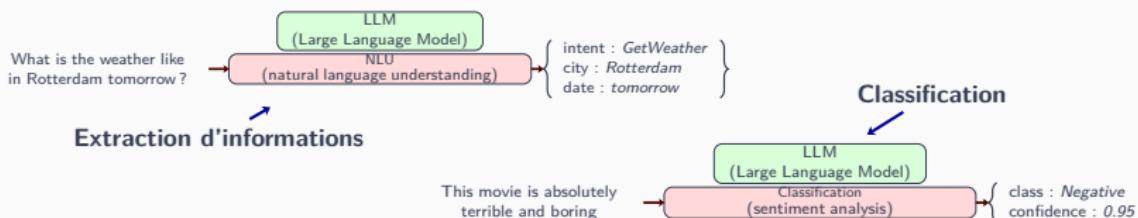
Pierre Lepagnol^{1,2} Sophie Rosset¹ Sahar Ghannay¹ Christophe Séjourné²

¹Université Paris-Saclay, CNRS, LISN

²SCIAM

2 septembre 2025

Tâches, Techniques et Méthodes



Pistes suivies dans la thèse :

- Benchmark de **petit modèle** pour la classification
- Génération contrainte par grammaire (GCD pour les intimes)
⇒ Filtrage des tokens par grammaire
- RI sélectionner des exemples
⇒ Sélection efficace d'exemples few-shot
- Impact sur les **formats de sortie** (JSON, KV, etc.)
⇒ Faire un bon choix en amont de la tâche



Batch 2, Hernandez à Nazari

Liu-Jianying



université
PARIS-SACLAY

Inria

CentralSupélec

Analyse Diachronique des Plongements Sémantiques dans un Corpus Scientifique et Technique : Mesures Quantitatives de la Nouveauté, de la Créativité et de l'Inventivité

Étape 1: Mesures des changements sémantiques dans la littérature scientifique avec SciBERT

Jianying LIU

02 Septembre 2025

Laboratoire Interdisciplinaire des Sciences du Numérique – LISN

jianying.liu@lisn.fr

Sujet et corpus

AI Articles evolution in 15 Years

Questions posées:

- La temporalité existe dans les embeddings des concepts scientifiques ? Comment l'inclure ?
- Comment mesurer la nouveauté avec ces embeddings ?

Corpus:

1. arXiv 2010-2024
2. 4 domaines: physique de la matière condensée (cond-mat), astrophysique (astro-ph), biologie quantitative, économie



Changements sémantiques dans la littérature scientifique

Table 1. Semantic shift & importance rank

subcorpus year	astro-ph (246 KW)		cond-mat (269 KW)	
	2010	2024	2010	2024
paper count	13 118	19 218	13 081	21 951
	Count	Dist	Count	Dist
KW only in top-200 of 2024	42	0.01305	64	0.01682
KW only in top-200 of 2010	46	0.00540	69	0.00932
KW in top-200 in both years				
all	154	0.00356	131	0.00587
same rank	5	0.00173	1	0.00505
$ \Delta\text{rank} \leq 5$	51	0.00309	29	0.00450
$ \Delta\text{rank} > 5$	103	0.00379	102	0.00626
representative KW examples	alma observation (0.099), radio burst (0.059), stochastic gravitational wave background (0.042)		topological semimetal(0.324), quantum material(0.083), chern insulator(0.058)	

Batch 2, Hernandez à Nazari

Mahoudeau-Margot

« On peut peut-être le faire avec de l'IA »

SEPTEMBRE 2025

**Développement d'un (modeste) projet de
recherche sur un corpus de presse par
une néophyte**



Document confidentiel –
ne peut être reproduit ni diffusé
sans l'accord préalable
de Sorbonne Université.

Contexte

- Ingénierie SHS à SUMMIT : soutien à des acteurs extérieurs aux SHS
- Proximité avec les métiers des maths et de l'informatique
- Evolution de la demande avec la popularisation des Humanités Numériques et des outils d'« intelligence artificielle »
- Remise en question au sein de la littérature d'un *a priori* technophobe

Original Article

Sociological Methods & Research
Volume 51 Number 2 June 2022
© The Author(s) 2022
Article first published online: 06 May 2022
DOI: 10.1177/0049177221104566
<https://journals.sagepub.com/doi/10.1177/0049177221104566>


The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy

Silene Di^{1,2} 
Étienne Ollion¹ 
and Rubin Shen^{1,2} 

Le projet « Eléphants »

• La demande

- Petit corpus (n=325) de presse écrite en langue française sur une période courte
- Relative liberté sur le plan analytique *mais* une demande portant sur de l'analyse de controverses
- Choix théoriques assez clairs dans la sociologie politique (« La dénonciation », Boltanski, Daré, Schiltz, 1984, notamment)

• Les méthodes

- Annotation à la main via Nvivo
- Analyse STM
- Application du package « Augmented Social Scientist »

• Pratичité des outils employés

- Ne remplacent pas la chercheuse mais la complètent/l'augmentent
- Gros questionnement sur la granularité de l'analyse
- Est-ce que ça vaut le coup ?

Le projet « Eléphants »

• La demande

- Petit corpus (n=325) de presse écrite en langue française sur une période courte
- Relative liberté sur le plan analytique *mais* une demande portant sur de l'analyse de controverses
- Choix théoriques assez clairs dans la sociologie politique (« La dénonciation », Boltanski, Daré, Schiltz, 1984, notamment)

• Les méthodes

- Annotation à la main via Nvivo
- Analyse STM
- Application du package « Augmented Social Scientist »

• Pratичité des outils employés

- Ne remplacent pas la chercheuse mais la complètent/l'augmentent
- Gros questionnement sur la granularité de l'analyse
- Est-ce que ça vaut le coup ?



Généré avec Paint

Batch 2, Hernandez à Nazari

Maurel-Thibault



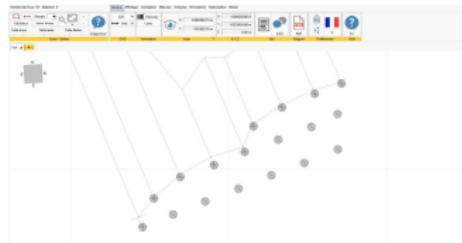
Assistant à la réalisation de plans de tir de mines

Contexte



Tir de mines

- Il y a de moins en moins d'experts → perte de connaissance
- La réalisation de plans de tir de mines est déléguée aux ingénieurs débutants → plan de tir mauvais



Plan de tir de mines

Problématique : Comment aider un ingénieur sans expérience à la réalisation d'un tir de mines ?

Approche



Logiciel de conception
et de simulation de tir
de mines

- **Documents** de conférences sur le domaine minier
- Données **client**



Expertise métier



Objectif : réaliser un chatbot capable de guider l'utilisateur dans le logiciel et de l'assister dans le paramétrage des fonctions



MERCI

Batch 2, Hernandez à Nazari

Nazari-Parisa

Parisa Nazari

1^{ère} année de thèse au LS2N

Sujet :

Sélection et intégration de connaissances linguistiques dans les modèles de langue avec contraintes de ressources : application au domaine de la santé

- Etat des lieux des connaissances linguistiques contenues dans les modèles de langue
- Approches innovantes en modélisation du langage:
 - Focus sur leur apprentissage et leur adaptation au moyen de connaissances linguistiques précises sélectionnées.
 - Contexte de ressources matérielles contraintes.
- Construction et entraînement de modèles de langue:
 - Connaissances linguistiques très limitées.
 - Enrichissement avec des connaissances linguistiques ciblées selon les objectifs visés.

Batch 3, Nowakowski à Zhou

Batch

	Rahman-Arafat
1. Batch 1, AABID à Guichoux	Rigal-Jacob
2. Batch 2, Hernandez à Nazari	Rodriguez-Ricardo
3. Batch 3, Nowakowski à Zhou	Ronzon-Mathis
Nowakowski-Nathan	Rousseau-Ismael
OZKAN-Beliz	Sauldubois-Christophe
Popovic-Senaid	YOUCEF KHODJA-Amine
Radola-Joanna	4. Hackathon time

Batch 3, Nowakowski à Zhou

Nowakowski-Nathan

Nathan Nowakowski

Début de thèse : Décembre 2024

Laboratoire : LIRIS – INSA Lyon

Financement: Mécénat - 

Directeur de thèse : Elöd Egyed-Zsigmond

Co-superviseurs : Diana Nurbakova & Sylvie Calabretto



Une IA confiante pour la prévision des risques dans les industries critiques

Objectif

Introduire des outils d'IA *digne de confiance* pour aider les experts en audits industriels

Travaux

Passé

- Challenge international sur la détection du sexisme sur les réseaux sociaux

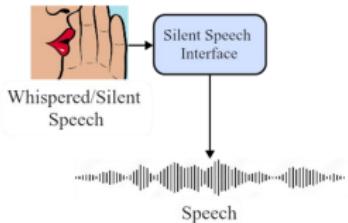
En cours

- Conception d'une jauge de confiance lors d'une classification – CNN / BERT

Batch 3, Nowakowski à Zhou

OZKAN-Beliz

Context



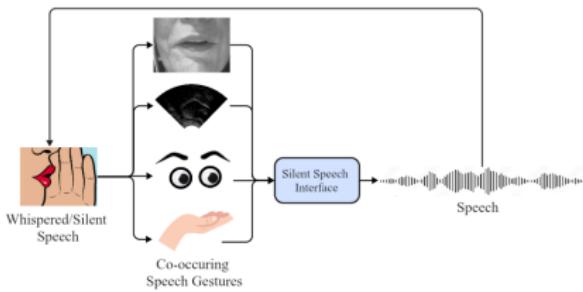
A silent speech interface (SSI) enables speech communication to take place when an audible acoustic signal is unavailable (B. Denby et al.¹).

This thesis proposes a deep learning-based prediction of prosodic functions from coverable speech gestures, as part of silent speech interface to allow for real-time speech reconstruction for individuals with laryngeal impairments.



Objectives

Real-time pitch prediction in speech reconstruction with interactive user-system adaptation



Is it possible to estimate pitch (F0) from speech gestures?

- According to Fant's source–filter model² : No.
- According to Tamás Grósz et al.³ : Yes, by using ultrasound tongue images.

Our hypothesis : It's possible to learn global correlation from data and correlations between articulatory behavior and prosodic functions can be exploited to improve F0 prediction (e.g., focus, question).



Bibliography

1. Denby, B. et al. “Silent speech interfaces”, *Speech Communication* (2010)
2. Fant, Gunnar. “Acoustic Theory of Speech Production”, *Mouton Edition* (1970)
3. Tamás Gábor Csapó, Mohammed Salah Al-Radhi, et al. “Ultrasound-based silent speech interface built on a continuous vocoder”, *Interspeech* (2019)



Batch 3, Nowakowski à Zhou

Popovic-Senaid

Main infos

Senaïd Popovic

Doctorant 1ère année

"Identification automatique et analyse de l'intention dans le contenu textuel" (pour l'instant, le pb = classification multi-classe) – use case: détection de phishing

Collaboration CIFRE entre LORIA (Nancy) et Hornetsecurity (cybersécurité)

Compétences: Python, pytorch, LLMs

Batch 3, Nowakowski à Zhou

Radola-Joanna



Génération de Textes Multilingues - Exploration de la capacité des modèles de langue à analyser et synthétiser des formes naturelles d'alternance codique

Projet de thèse

Joanna Radoła

supervised by François Yvon at ISIR, Sorbonne Université and Josep Crego at SYSTRAN

September 3, 2025

Quoi, pourquoi et comment

Code-switching (alternance codique): the phenomenon of plusieurs langues being used au sein de la même phrase.

Quoi, pourquoi et comment

Code-switching (alternance codique): the phenomenon of plusieurs langues being used au sein de la même phrase.

On s'intéresse à la capacité des LLM à traiter des entrées *code-switchées*.

Quoi, pourquoi et comment

Code-switching (alternance codique): the phenomenon of plusieurs langues being used au sein de la même phrase.

On s'intéresse à la capacité des LLM à traiter des entrées *code-switchées*.

Motivation : Améliorer la traduction en cas de mélange de langues, répondre aux besoins des communautés multilingues.

Quoi, pourquoi et comment

Code-switching (alternance codique): the phenomenon of plusieurs langues being used au sein de la même phrase.

On s'intéresse à la capacité des LLM à traiter des entrées *code-switchées*.

Motivation : Améliorer la traduction en cas de mélange de langues, répondre aux besoins des communautés multilingues.

Approche : Analyse des représentations multilingues, interventions pour mieux contrôler la langue de génération.

Batch 3, Nowakowski à Zhou

Rahman-Arafat

User study: Helping Academic Post-Editors

G M Arafat Rahman

Supervised by:
François Yvon, Marine Carpuat

Non-English Academics Need to Translate a Lot

The screenshot shows a web page from the ATALA (Association pour le Traitement Automatique des Langues) website. The header includes the logo, the name "ATALA", and the subtitle "Association pour le Traitement Automatique des Langues". Navigation links include Accueil, Association, Athélos, Revue TAL, Liste LN, Conférence TALN, Conférences, Prix de thèse, Prix TALN-RECTAL, Annuaire, Journées, Offres d'emploi, and Contact. A sidebar on the right has links for Mot d'utilisateur, Mot de passe, Se connecter, Créez un nouveau compte, and Réinitialiser votre mot de passe.

Évaluation de la qualité de rapport des essais cliniques avec des larges modèles de langue

Mathieu Lai-King¹ et Patrick Paroubek^{2*}
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France

Résumé
La qualité de rapport est un sujet important dans les articles de recherche sur les essais cliniques car elle peut avoir un impact sur les décisions cliniques prises. Nous testons la capacité des larges modèles de langage à évaluer la qualité de rapport de ce type d'article en utilisant les standards fusionnés pour la rédaction d'essais thérapeutiques (CONSORT). Nous créons un corpus d'évaluation à partir de deux études sur la vérification de la qualité de rapport de résultats d'articles avec les standards CONSORT définis pour les résultats. Nous évaluons ensuite la capacité de différents larges modèles de langue génératifs (du domaine général ou adaptés au domaine biomédical) à correctement évaluer chaque critère CONSORT avec différentes méthodes de requête (prompting) communes. Notre meilleure association de modèle et de méthode de requête obtient 85 % d'exactitude.

Résumé (en anglais)
Reporting quality is an important topic in clinical trial research articles, as it can have an impact on the clinical decisions made. In this article, we test the ability of large language models to assess the reporting quality of this type of article using the Consolidated Standards of Reporting Trials (CONSORT). We create an evaluation corpus from two studies on abstract reporting quality with CONSORT-abstract standards. We then evaluate the ability of different large generative language models (from the general domain or adapted to the biomedical domain) to correctly assess CONSORT criteria with different known prompting methods. Our best combination of model and prompting method achieves 85 % accuracy.

**Per son Scholartly Document Processing*

Document
[TAL_65_2_1.pdf](#)

Rank
1

Paper writing

The Cheapest Way: Machine Translate + Post Editing

anglais français

Reporting quality is an important issue in clinical trial research articles because it can impact clinical decisions. We test the ability of large language models to assess the reporting quality of this type of article using the Consolidated Standards of Reporting Trials (CONSORT) guidelines. We create an evaluation corpus from two studies on the verification of the reporting quality of article abstracts using the CONSORT guidelines defined for abstracts. We then evaluate the ability of different large generative language models (general domain or adapted to the biomedical domain) to correctly evaluate each CONSORT criterion using different known prompting methods. Our best combination of model and prompting method achieves 85% accuracy.

français Options

La qualité des rapports est un enjeu important dans les articles de recherche sur les essais cliniques, car elle peut avoir une incidence sur les décisions cliniques. Nous testons la capacité des grands modèles linguistiques à évaluer la qualité des rapports de ce type d'articles à l'aide des lignes directrices CONSORT (Consolidated Standards of Reporting Trials). Nous créons un corpus d'évaluation à partir de deux études sur la vérification de la qualité des rapports d'articles à l'aide des directives CONSORT définies pour les résumés. Nous évaluons ensuite la capacité de différents grands modèles linguistiques génératifs (de domaine général ou adaptés au domaine biomédical) à évaluer correctement chaque critère CONSORT à l'aide de différentes méthodes de prompting connues. Notre meilleure combinaison de modèle et de méthode de prompting atteint une précision de 85 %.

WRONG !

747 / 1500 Ouvrir dans Write

Up Down Like Dislike Share

Machine Translation makes Many Terminology Errors

The screenshot shows a machine translation interface with two columns: English on the left and French on the right. Both columns have language selection dropdowns at the top: 'anglais' and 'français'. A double-headed arrow icon is between the language dropdowns. On the far right, there is an 'Options' dropdown.

English Text:

Reporting quality is an important issue in clinical trial research articles because it can impact clinical decisions. We test the ability of large language models to assess the reporting quality of this type of article using the Consolidated Standards of Reporting Trials (CONSORT) guidelines. We create an evaluation corpus from two studies on the verification of the reporting quality of article abstracts using the CONSORT guidelines defined for abstracts. We then evaluate the ability of different large generative language models (general domain or adapted to the biomedical domain) to correctly evaluate each CONSORT criterion using different known prompting methods. Our best combination of model and prompting method achieves 85% accuracy.

French Translation:

La qualité des rapports est un enjeu important dans les articles de recherche sur les essais cliniques, car elle peut avoir une incidence sur les décisions cliniques. Nous testons la capacité des grands modèles de langue à évaluer la qualité des rapports de ce type d'articles à l'aide des directives CONSORT (Consolidated Standards of Reporting Trials). Nous créons un corpus d'évaluation à partir de deux études sur la vérification de la qualité des rapports d'articles à l'aide des directives CONSORT définies pour les résumés. Nous évaluons ensuite la capacité de différents grands modèles de langue génératifs (en domaine général ou adaptés au domaine biomédical) à évaluer correctement chaque critère CONSORT à l'aide de différentes méthodes de prompting connues. Notre meilleure combinaison de modèle et de méthode de prompting atteint une précision de 85 %.

Bottom Center:

Accuracy ≠ Précision

747 / 1500

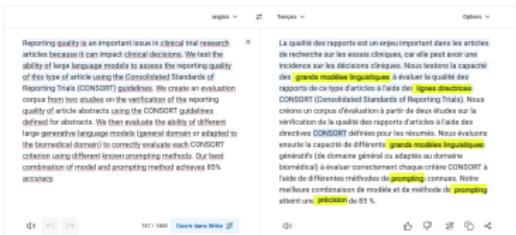
Navigation icons: back, forward, search, etc.

Can Automatic Error Detection help Post-Editors ?

anglais	français	Options
<p>Reporting quality is an important issue in clinical trial research articles because it can impact clinical decisions. We test the ability of large language models to assess the reporting quality of this type of article using the Consolidated Standards of Reporting Trials (CONSORT) guidelines. We create an evaluation corpus from two studies on the verification of the reporting quality of article abstracts using the CONSORT guidelines defined for abstracts. We then evaluate the ability of different large generative language models (general domain or adapted to the biomedical domain) to correctly evaluate each CONSORT criterion using different known prompting methods. Our best combination of model and prompting method achieves 85% accuracy.</p>	<p>La qualité des rapports est un enjeu important dans les articles de recherche sur les essais cliniques, car elle peut avoir une incidence sur les décisions cliniques. Nous testons la capacité des grands modèles linguistiques à évaluer la qualité des rapports de ce type d'articles à l'aide des lignes directrices CONSORT (Consolidated Standards of Reporting Trials). Nous créons un corpus d'évaluation à partir de deux études sur la vérification de la qualité des rapports d'articles à l'aide des directives CONSORT définies pour les résumés. Nous évaluons ensuite la capacité de différents grands modèles linguistiques génératifs (de domaine général ou adaptés au domaine biomédical) à évaluer correctement chaque critère CONSORT à l'aide de différentes méthodes de prompting connues. Notre meilleure combinaison de modèle et de méthode de prompting atteint une précision de 85 %.</p>	

Can Automatic Error Detection help Post-Editors ?

- How to identify error spans?
- What are the most severe?
- How to best display them?
- How to give control over the display?
- What if they are pressed by time (deadlines)
- ...



We need user studies

Interface for the User Study

Vous avez la possibilité de contrôler le nombre de zones surlignées en utilisant un curseur similaire à celui représenté sur l'image. En le déplaçant vers la droite, on augmente le nombre de zones; en le déplaçant vers la gauche on le diminue.



spans: 5 / 12

Une variable essentielle pour nous est **le temps de post-édition, que nous enregistrons**: une fois qu'une tâche est démarrée, **il est important de l'effectuer sans interruption**. Vous devrez vous concentrer uniquement sur le travail de correction de la traduction, et **ne pas consulter de sources d'information externes** (comme un dictionnaire ou sur un site web). Une fois que vous avez fini de post-éditer un résumé, vous serez redirigé(e) vers cette page avant de passer au résumé suivant. Vous pouvez faire une pause entre chaque texte.

Les textes sont présentés dans un ordre qui est fixé à l'avance : le prochain résumé à corriger apparaît dans une couleur plus foncée.

Assurez-vous que vous êtes dans un environnement calme, et disposez du matériel (écran, clavier) pour pouvoir travailler confortablement sur les textes.

Une fois que vous êtes prêt(e)s, cliquez sur OK pour débuter l'expérience. Elle durera environ 25 minutes.

OK

Article 1

Article 2

Article 3

Article 4

Interface for the User Study

00:03:41

Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)

We present the AGODA (Analyse sémantique et Graphes relationnels pour l'Ouverture des Débats à l'Assemblée nationale) project, which aims to create a platform for consulting and exploring digitised French parliamentary debates (1881-1940) available in the digital library of the National Library of France. This project brings together historians and NLP specialists: parliamentary debates are indeed an essential source for French history of the contemporary period, but also for linguistics. This project therefore aims to produce a corpus of texts that can be easily exploited with computational methods, and that respect the TEI standard. Ancient parliamentary debates are also an excellent case study for the development and application of tools for publishing and exploring large historical corpora. In this paper, we present the steps necessary to produce such a corpus. We detail the processing and publication chain of these documents, in particular by mentioning the problems linked to the extraction of texts from digitised images. We also introduce the first analyses that we have carried out on this corpus with "bag-of-words" techniques not too sensitive to OCR quality (namely topic modelling and word embedding).

Traduction automatique (cliquez pour ouvrir) ▾

Entre histoire et traitement du langage naturel: Etude, enrichissement et publication en ligne des débats parlementaires français du début de la Troisième République (1881-1899) Nous présentons le projet AGODA (Analyse sémantique et Graphes relationnels pour l'Ouverture des Débats à l'Assemblée nationale), qui vise à créer une plateforme de consultation et d'exploration des débats parlementaires français numérisés (1881-1940) disponible dans la bibliothèque numérique de la Bibliothèque nationale de France. Ce projet rassemble des historiens et des spécialistes de la PNL: les débats parlementaires sont en effet une source essentielle pour l'histoire française de l'époque contemporaine, mais aussi pour la linguistique. Ce projet vise donc à produire un corpus de textes qui peuvent être facilement exploités avec des méthodes de calcul, et qui respectent la norme TEI. Les débats parlementaires anciens constituent également une excellente étude de cas pour le développement et l'application d'outils d'édition et d'exploration de grands corpus historiques. Dans cet article, nous présentons les étapes nécessaires pour produire un tel corpus. Nous détaillons la chaîne de traitement et de publication de ces documents, notamment en mentionnant les problèmes liés à l'extraction de textes à partir d'images numérisées. Nous introduisons également les premières analyses que nous avons effectuées sur ce corpus avec des techniques de «bag-of-words» pas trop sensibles à la qualité de l'OCR (à savoir la modélisation des sujets et l'intégration de mots).

Traduction à post-éditer :

Nombre de zones (spans) erronées visibles:  spans: 6 / 12

Entre histoire et traitement du langage naturel: Etude, enrichissement et publication en ligne des débats parlementaires français du début de la Troisième République (1881-1899)

Nous présentons le projet AGODA (Analyse sémantique et Graphes relationnels pour l'Ouverture des Débats à l'Assemblée nationale), qui vise à créer une plateforme de consultation et d'exploration des débats parlementaires français numérisés (1881-1940) disponible dans la bibliothèque numérique de la Bibliothèque nationale de France. Ce projet rassemble des historiens et des spécialistes de la PNL: les débats parlementaires sont en effet une source essentielle pour l'histoire française de l'époque contemporaine, mais aussi pour la linguistique. Ce projet vise donc à produire un corpus de textes qui peuvent être facilement exploités avec des méthodes de calcul, et qui respectent la norme TEI. Les débats parlementaires anciens constituent également une excellente étude de cas pour le développement et l'application d'outils d'édition et d'exploration de grands corpus historiques. Dans cet article, nous présentons

To participate in the Study



<https://postedition.anr-matos.fr/tal/connexion.php>

Batch 3, Nowakowski à Zhou

Rigal-Jacob



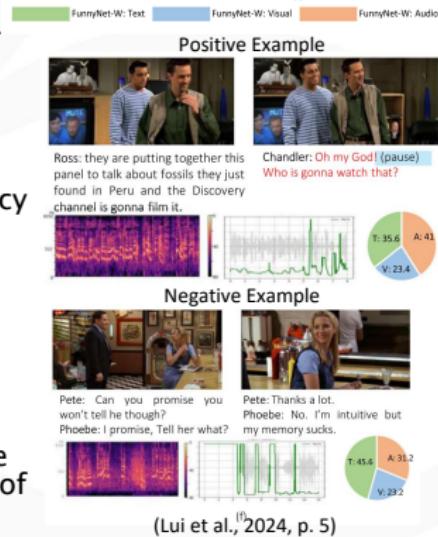
CNN-based tools for identifying humor

State of the art neural network image classifiers have achieved just over 91 % accuracy (Liu et al., 2022)

FunnyNet-W – Predicts funny moments (in 5 data sets including TED Talks) at 80 % accuracy using multi-modal cues (audio, visual) and time-aligned transcripts (Lui et al., 2024)

Limitations: Models trained on *laughter* as indicating humor, which can lead to false positives and missed funny moments (Mazzocconi et al., 2020).

Useful tool, but, false positives. Needs more detailed annotations (gesture type, presence of play cue, etc.) = A virtuous cycle





Université
Paris

OpenFace facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation

Frame-by-frame output (pics + data)

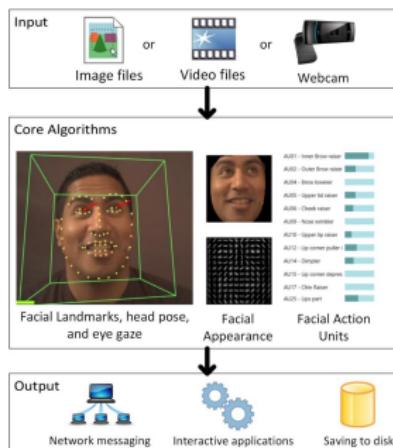
18 FACS Action Units (AUs) and sub-types

Binary (0 or 1 for absence/presence) and continuous values

3D Landmarks

Robust head pose estimation

Multimodal humor in action: A CNN-based approach towards a formal dialogic model



I extract 20-sec TED Talk clips with audience laughter in the middle. ..and non-humor control clips, pre-applause clips.

Run OpenFace 2.0 CNN to identify action units pre-humor, not.

Run Whisper.ai to align speech.

To do: Use a memory-augmented model to generate dialogue functions + reasoning patterns of different types of incongruities (humor, etc.) triggering laughter, fine tune with additional annotations (embodied play) to build robust, transparent multimodal incongruity recognizer

Batch 3, Nowakowski à Zhou

Rodriguez-Ricardo

Présentation Thèse ETAL 2025

Ricardo Rodriguez¹

¹Laboratoire Informatique d'Avignon, CERI, Avignon Université – LIA UPR 4128
ricardo.rodriguez@univ-avignon.fr

02 Septembre 2025



Présentation Thèse ETAL 2025



Ricardo Rodriguez
Doctorant en 1ère année

Uruguay
34 ans



Lab : Avignon Université – Laboratoire Informatique d'Avignon

Thèse : Utilisation de Grands Modèles de Langue dans le domaine médical, via une interaction vocale.

Co-encadrants : Mickael ROUVIER, Stéphane HUET, Benoît FAVRE.

Projet ANR MALADES (Avignon, Marseille, Nantes).

<https://lia.univ-avignon.fr/2024/06/04/projet-anr-malades/>

Avancements et Perspectives

- Prise en main LLMs texte
- 1er article accepté TALN/RANLP (QCM Humains vs LLMs) 
- État-de-l'art corpus médicaux, tâches, difficultés
- Prise en main LLMs audio/multimodaux
- Évaluation TTS dans le médical
- Choix de corpus et de tâches
- Synthèse vocale corpus médicaux
- Accès données CHU Nantes, Avignon, Rouan ?
- ...
-  Become a ninja

Batch 3, Nowakowski à Zhou

Ronzon-Mathis

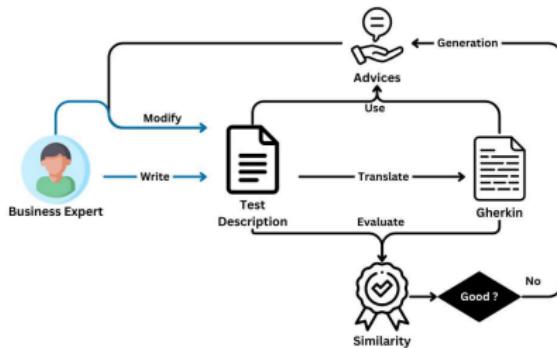


Sujet

Développement de logiciel en langage naturel dans différents domaines.

Vision

- Remplacer les langages informatiques et pouvoir coder uniquement grâce au langage naturel
- Aucune vue de la part de l'utilisateur du code produit. On interagit avec lui uniquement en langage naturel
- Volonté d'encadrer un utilisateur sans formation dans l'informatique au sein du processus de génération, comme les IDE actuels.



- Traduction d'un texte dans un langage spécifique
- Classification de la similarité entre deux textes
- Accompagnement de l'utilisateur

Batch 3, Nowakowski à Zhou

Rousseau-Ismael

Ismaël Rousseau



2020-2023 : Alternance axée Data Science / NLP

(encadrement : Géraldine Damnatil)

Travaux autour des thématiques de :

- Similarité sémantique
- Clustering
- Résumé automatique de conversations (BART)
- Tâches de classification diverses dans des cadres applicatifs
- Workflows LLMs & RAG

+ Création de maquettes interactives / Dev web

Ismaël Rousseau



Aix Marseille Université

2023-... : CDI à Orange

Contributions aux projets d'adaptation des LLMs

- Continuation du pré-entraînement, affinage sur les instructions, RL
- Création des corpus, mise en place de la chaîne d'évaluation etc.

2025-2028 : Thèse interne

(Encadrement: Géraldine Damnati & Frédéric Béchet)

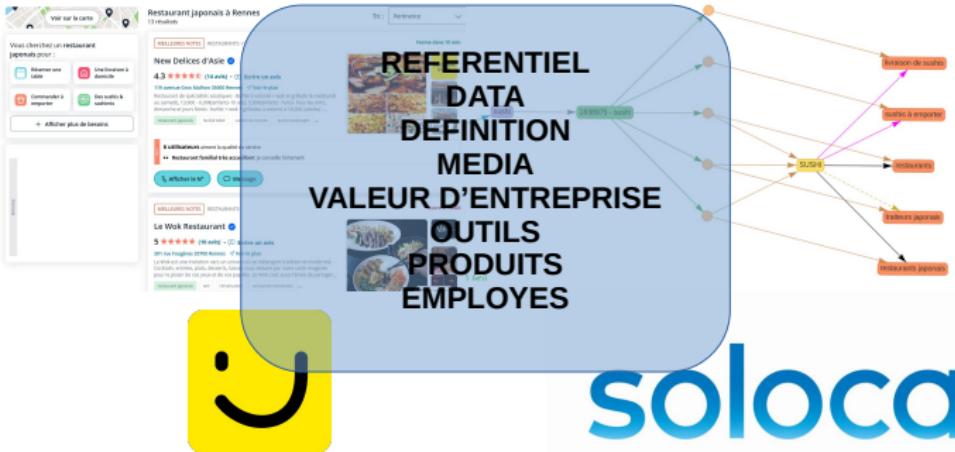
Utilisation des agents LLM pour l'exploration de bases documentaires

→ Systèmes type « RAG agentique » / « Deep Research » / « Computer use »

Batch 3, Nowakowski à Zhou

Sauldubois-Christophe

Acculturation de la CULTURE d'ENTREPRISE dans un LLM



Comment ?

- LLM Existant
- PEFT – Fine Tuning
- Embedding de Graph
- Adapter
- Optimisation des Tokens
- Prompting – RAG

Batch 3, Nowakowski à Zhou

YOUSSEF KHODJA-Amine

APPRENTISSAGE AUTO-SUPERVISÉ DE REPRÉSENTATIONS MULTIMODALES POUR LA DÉTECTION D'ANOMALIES

Directeur : Gaël Dias

Amine YOUSSEF KHODJA

¹Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

amine.yousef-khodja@etu.unicaen.fr



Sujet de thèse

Détection d'Anomalies Textuelles par Apprentissage Auto-Supervisé

Contexte Scientifique :

- ▶ Les anomalies textuelles (stylistiques, sémantiques, structurelles) sont fortement dépendantes du contexte et ne peuvent être cernées par des définitions strictes.

Objectif Central de la Thèse :

- ▶ Concevoir des architectures d'apprentissage profond capables d'acquérir des **représentations latentes robustes** pour la détection non supervisée d'anomalies, initialement dans le domaine textuel.
- ▶ Exploitation de l'**apprentissage auto-supervisé**, notamment l'**apprentissage contrastif**.
- ▶ Étendre l'approche contrastive pour capturer les dépendances intermodales et détecter les anomalies, initialement pour des paires texte-texte puis texte-image / texte-audio.

Opportunités Envisagées

Avancées Scientifiques et Applications



Applications en Santé Mentale et Validation sur Données Cliniques :

- ▶ Accès à des données multimodales (texte + audio) du **CHU de Brest** pour une première validation et exploration.
- ▶ Application des modèles développés pour l'analyse de données cliniques (ex: verbatim de patients, rapports médicaux).
- ▶ **XAI en contexte multimodal** : exploiter les interactions entre modalités (texte, image, etc.) pour expliquer les anomalies via des outils interactifs comme les heatmaps, afin de renforcer l'interprétabilité et la confiance utilisateur.

Hackathon time
