

Audition pour un poste de maître de conférences

IDMC - LORIA

Sophie Robert-Hayek
Post-doctorante à l'Université de Lorraine

6 Mai 2025 - Nancy

Présentation

Présentation du parcours: formation académique

- **Licence de Mathématiques** [Université Blaise Pascal, Clermont-Ferrand];
- **Master en Mathématiques Appliquées**, mention *Statistiques et Traitement des données* [Université Blaise Pascal, Clermont-Ferrand];
- **Doctorat en Informatique de l'Université Paris-Saclay:**
 - *Auto-tuning of I/O accelerators using black-box optimization*
 - Direction Pr. Zertal;
 - Financement CIFRE Bull - UPSaclay-UVSQ.

👉 7 conférences internationales (HPCS x 2 [B], MASCOTS [A], MEDI [C], OLA, META [B], SITA, [IDPS]); 2 papiers journaux (CCPE [Q2 - 84], Journal of Computational Science [Q2 - 71]); 3 brevets internationaux; Logiciel Open-Source déployé sur des super-calculateurs internationaux (<https://github.com/bds-ailab/shaman>);

Expérience R&D industrielle

2021-2023: Bull/Eviden (Data Engineer):

- **Pilotage technique** d'API pour l'instrumentation I/O de supercalculateurs.
- Projet ERC **IOSea** :
 - Placement I/O optimisé sur stockage hiérarchisé.
 - API REST pour l'orchestrateur **Slurm**.
- **Méthodologie Scrum** : cycles courts, coordination agile.

Encadrement doctorat/master

Placement intelligent des fichiers dans un **stockage hiérarchisé** par exploitation des cycles de vie pour le HPC.

- Nouvelle méthode basée sur la prédiction des comportements applicatifs pour placer efficacement les fichiers dans un stockage hiérarchisé;
- Simulation d'un stockage hiérarchisé (*burst buffer*) pour validation des résultats.

 Avec A. Khelili et S. Zertal:

Acceptées: 2 conférences internationales (SITA, SBAC-PAD [B]); 1 papier journal (Infocommunications journal [Q3]); 1 brevet international.

Soumis/en cours: 2 conférences (HPCC et AMMS).

ANR SHERBET

- **2022:** Montage d'un projet ANR avec Frédérique Rey (Ecritures, UL/MSH), Maxime Amblard (LORIA, UL) et Jacques Istas (LJK, UGA) [400.000€]
- **2022:** Montage d'un projet LUE BENTO avec Frédérique Rey et Maxime Amblard [110.000€].
- **2022:** Montage d'un projet Biblissima+ avec Frédérique Rey: projet SCRIBES (responsable technique) [40.000€].

Projet d'intégration

Intégration "Post-Sémagramme"

Apprentissage automatique pour l'analyse de la transmission des corpora multilingues.

Mon projet de recherche s'inscrit dans **l'équipe "post"-sémagramme** portée par Maxime Amblard [axe D4], [axe transverse "Traitement automatique des langues et intelligence artificielle"], dans **la continuité du projet SHERBET**.

Intégration "Post-Sémagramme"

Apprentissage automatique pour l'analyse de la transmission des corpora multilingues.

Mon projet de recherche s'inscrit dans **l'équipe "post"-sémagramme** portée par Maxime Amblard [axe D4], [axe transverse "Traitement automatique des langues et intelligence artificielle"], dans **la continuité du projet SHERBET**.

Intégration "Post-Sémagramme"

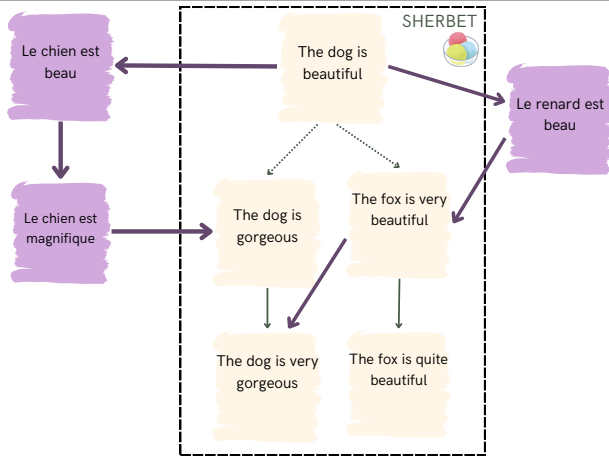
Apprentissage automatique pour l'analyse de la transmission des corpora multilingues.

Mon projet de recherche s'inscrit dans **l'équipe "post"-sémagramme** portée par Maxime Amblard [axe D4], [axe transverse "Traitement automatique des langues et intelligence artificielle"], dans **la continuité du projet SHERBET**.

En interaction avec:

- **SMarT (D4)** [*ABDUL: a new Approach to Build language models for Dialects Using formal Language corpora only*, Toughrai et al., 2025; *Knowledge Distillation for Efficient Algerian Dialect Processing: Training Compact BERT Models with DziriBERT*, Amina Laggoun et al., 2025]

Post-SHERBET: transmission des corpora multi-langues



Pas de prise en compte de:
Transmissions parallèles.
Contamination;

Problématiques de recherche

1. Comment **aligner un corpus multilingue** à l'aide des modèles probabilistes et neuronaux dans un contexte de langues peu dotées?

Problématiques de recherche

1. Comment **aligner un corpus multilingue** à l'aide des modèles probabilistes et neuronaux dans un contexte de langues peu dotées?
2. Comment mesurer **la proximité grammaticale et sémantique** entre textes en différentes langues?

Possibilité de financements du programme de recherche

ANR-DFG (Pr. Annette Weissenrieder / Pr. Hubert Mara): Transmission du texte latin dans l'Antiquité;

ERC Starting Grant (éligible jusqu'en 2029):

MaLAMuTe (Machine Learning for the Analysis of Multilingual Text Transmission).



Intégration avec le LORIA

Thématiques de recherche:

- Alignées avec les **objectifs post-sémaigramme**;
- Alignées avec le **dynamisme du laboratoire sur l'axe transverse NLP** et D4.

Intégration avec le LORIA

Thématiques de recherche:

- Alignées avec les **objectifs post-sémaigramme**;
- Alignées avec le **dynamisme du laboratoire sur l'axe transverse NLP** et D4.

Implication projets transverses:

- Contribution à **INSIGHT**;
- Contributions **ENACT**: Accès industriel à l'IA (HPC), montage de projet industriel.

Intégration avec le LORIA

Thématiques de recherche:

- Alignées avec les **objectifs post-sémaigramme**;
- Alignées avec le **dynamisme du laboratoire sur l'axe transverse NLP et D4**.

Implication projets transverses:

- Contribution à **INSIGHT**;
- Contributions **ENACT**: Accès industriel à l'IA (HPC), montage de projet industriel.

Vie de laboratoire:

- Participation au **DeepLorIa**;
- **Encadrements** d'étudiants.
- **Prise en charge** de séminaire d'axe/équipe.
- **Diffusion** de la recherche.

Projet d'intégration pédagogique

Expérience d'enseignement

- **Introduction à l'apprentissage automatique** [Université Grenoble Alpes, MIASHS, 24h de CM, 26h de TD/TP x 2, **Responsable d'UE**].
- **Introduction aux humanités numériques** [Université de Lorraine, M1 théologie-s, 12h de CM, 12h de TD/TP, **Responsable d'EC**].
- **Formations au design d'API** [Bull/Eviden].

Projet pédagogique

Objectifs pédagogiques

- Mettre **la réalité entreprise** au centre de l'enseignement (TDD, qualité du code, CI/CD, documentation);
- Former des **profils hybrides** à l'interface entre informatique et les besoins métiers;
- **Faire évoluer les contenus en phase avec les besoins du marché.**

Projet pédagogique

Objectifs pédagogiques

- Mettre **la réalité entreprise** au centre de l'enseignement (TDD, qualité du code, CI/CD, documentation);
- Former des **profils hybrides** à l'interface entre informatique et les besoins métiers;
- **Faire évoluer les contenus en phase avec les besoins du marché.**

Par delà l'enseignement...

- **Encadrement et accompagnement d'étudiants** (stages);
- **Ateliers professionnalisants** : simulations d'entretiens, rédaction de CV;
- Développement de **partenariats entreprises**;
- **Implication dans les responsabilités collectives.**

Projet d'enseignement

Unité d'Enseignement:

Le métier de LLMOps et les mutations des métiers liés à la gestion des Systèmes d'Information.

Niveau: M2 (SID);

Objectif: Former les étudiants au domaine émergent du déploiement d'agents LLM pour les entreprises.

Pré-requis: BDD [MIASHS], Développement Web [MIASHS], Conception des systèmes d'information et Algorithmes pour l'intelligence artificielle [UE 702], Analyse de données [UE 801].

Interaction avec les UE existantes: UE 904 [Big Data], UE 902 [Architecture Big Data];

Projet d'enseignement

Mise en place progressive d'une *stack* applicative déployant un chatbot capable d'utiliser un modèle LLM apte à requêter des données PDF propres à une entreprise.

Validation du cours

L'ensemble du travail réalisé sera parfaitement documenté d'un point de vue **théorique, administrateur et utilisateur** par un rapport accompagnant l'application Web. Un oral présentera au reste de la promotion l'application et son fonctionnement.

Licence MIASSH

Licence MIASSH:

✓ Mathématiques

UE 101 [Algèbre linéaire, Statistiques]

UE 201 [Probabilités et statistiques]

UE 301 [Statistiques, optimisation]);

✓ Informatique

UE 102 [Algorithmique - Programmation, Technologie du Web]

UE 202 [Algorithmique - Programmation, BDD]

UE 302 [OOP avancé]

UE 303 [BDD, Technologies du Web avancées]



UE 402 [Programmation en Python]

UE 403 [BDD Avancées, Technologies du Web avancées]




UE 405 [Python pour le TAL])

Master 1 MIAGE



UE 701:

-  Statistiques
-  Recherche opérationnelle;

UE 702

-  Conception des systèmes d'information (Méthodes);
-  Algorithmes pour l'intelligence artificielle;
-  Conception des systèmes d'information (Optimisation).

UE 801:

-  Analyse de données;
-  Structuration de documents.

UE 802: Génie logiciel.

UE 803: Programmation fonctionnelle.




UE 804: Management des équipes.

Master 2 MIAGE - SID


UE 901:

-  Modélisation, conception et mise en œuvre de SI à base de patrons;

UE 902:



-  Architectures orientées services/API;
-  Architectures Big Data;
-  Architecture à base de scripts.

UE 904:  Big data.



UE 905:  Mise en œuvre de cas industriels et mise en œuvre.

Master 2 MIAGE - ACSI

UE 902:

-  Principes fondamentaux des données massives;
-  Stockage et transmission des données massives.

UE 904:

-  Technologies du décisionnel;
-  Mise en œuvre de la business intelligence.

Merci pour votre attention!

Merci pour votre attention!

Slides supplémentaires

ANR SHERBET

SHERBET a pour objectif d'évaluer la performance de méthodes issues de la biologie (phylogénie) pour reconstruire la généalogie des manuscrits.

ANR SHERBET

SHERBET a pour objectif d'évaluer la performance de méthodes issues de la biologie (phylogénie) pour reconstruire la généalogie des manuscrits.

Avec SHERBET:

Accepté: 3 contributions dans des volumes [av F. Rey et D. D'Amico]. 5 conférences internationales sans actes. 7 stages.

En review: 2 journaux en cours de review (NTS [Q1], DSH [Q1]); 1 papier conférence [ACL [A], avec I. Stoupak et M. Amblard]; 2 volumes (Peeters, Brill).

En cours: 1 papier en cours de préparation (CHR 2025).

En humanités computationnelles:

1 journal [Q1];
2 contributions dans un volume;
1 papier conférence [CHR];
2 participations dans des projets internationaux;
Panel chair dans 2 conférences internationales.

WP 1
**Génération de
données synthétiques**

The dog is
gorgeous

The dog is
beautiful

The dog is very
gorgeous

The fox is very
beautiful

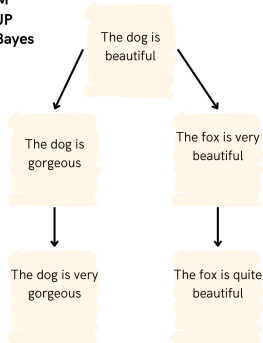
The fox is quite
beautiful

WP 1
Génération de
données synthétiques



WP 2
Reconstruction de l'arbre

Neighbor Joining
RHM
PAUP
MrBayes

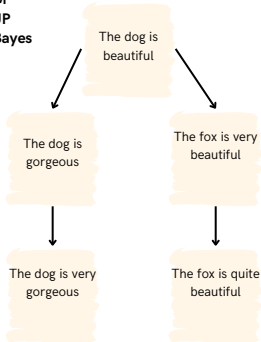


WP 1
Génération de
données synthétiques



WP 2
Reconstruction de l'arbre

Neighbor Joining
RHM
PAUP
MrBayes



WP 3
Comparaison avec
"ground truth" et sélection
de l'algorithme optimal

Comparaison des algorithmes

Neighbor Joining
RHM
PAUP
MrBayes

Selon plusieurs distances

GED
RF
Roos

Problématiques de recherche

1. Comment **aligner un corpus multilingue** à l'aide des modèles probabilistes et neuronaux dans un contexte de langues peu dotées?

Objectif:

- Texte source de longueur n : $(x_i)_{i \in \{1, \dots, n\}}$, texte traduit de longueur k : $(y_i)_{i \in \{1, \dots, k\}}$.
- Matrice d'alignement $\mathbf{A} \in \{0, 1\}^{n \times k}$, avec $a_{i,j} = 1$ si (x_i, y_j) sont alignés.

Problématiques de recherche

1. Comment **aligner un corpus multilingue** à l'aide des modèles probabilistes et neuronaux dans un contexte de langues peu dotées?

Objectif:

- Texte source de longueur n : $(x_i)_{i \in \{1, \dots, n\}}$, texte traduit de longueur k : $(y_i)_{i \in \{1, \dots, k\}}$.
- Matrice d'alignement $\mathbf{A} \in \{0, 1\}^{n \times k}$, avec $a_{i,j} = 1$ si (x_i, y_j) sont alignés.

Validation: Mise en place d'un gold standard d'alignement pour langues rares.

Problématiques de recherche

1. Comment **aligner un corpus multilingue** à l'aide des modèles probabilistes et neuronaux dans un contexte de langues peu dotées?

Objectif:

- Texte source de longueur n : $(x_i)_{i \in \{1, \dots, n\}}$, texte traduit de longueur k : $(y_i)_{i \in \{1, \dots, k\}}$.
- Matrice d'alignement $\mathbf{A} \in \{0, 1\}^{n \times k}$, avec $a_{i,j} = 1$ si (x_i, y_j) sont alignés.

Validation: Mise en place d'un gold standard d'alignement pour langues rares.

Résultat: Analyse des techniques de traduction.

Problématiques de recherche

2. Comment mesurer **la proximité grammaticale et sémantique** entre textes en différentes langues?

Objectif:

- Classification de la nature de $a_{ij} \in J$ avec J à déterminer ($J \in \{polygenetic, morphological, lexical...\}$);
- Détermination d'une fonction de score $s : J \rightarrow \mathbb{R}$.

Problématiques de recherche

2. Comment mesurer **la proximité grammaticale et sémantique** entre textes en différentes langues?

Objectif:

- Classification de la nature de $a_{ij} \in J$ avec J à déterminer ($J \in \{polygenetic, morphological, lexical...\}$);
- Détermination d'une fonction de score $s : J \rightarrow \mathbb{R}$.

Validation: Génération de données synthétiques à l'aide de *stemmabench*.

Problématiques de recherche

2. Comment mesurer **la proximité grammaticale et sémantique** entre textes en différentes langues?

Objectif:

- Classification de la nature de $a_{ij} \in J$ avec J à déterminer ($J \in \{polygenetic, morphological, lexical...\}$);
- Détermination d'une fonction de score $s : J \rightarrow \mathbb{R}$.

Validation: Génération de données synthétiques à l'aide de *stemmabench*.

Résultat: Analyse de la dispersion géographique des textes.

Projet d'enseignement

Compétences à acquérir:

- 1 Fondements théoriques des LLM.

Projet d'enseignement

Compétences à acquérir:

- ❶ Fondements théoriques des LLM.
- ❷ Déploiement et intégration.

Projet d'enseignement

Compétences à acquérir:

- ❶ Fondements théoriques des LLM.
- ❷ Déploiement et intégration.
- ❸ Gestion des données.

Projet d'enseignement

Compétences à acquérir:

- ❶ Fondements théoriques des LLM.
- ❷ Déploiement et intégration.
- ❸ Gestion des données.
- ❹ Sécurité et éthique.

Projet d'enseignement

Compétences à acquérir:

- ❶ Fondements théoriques des LLM.
- ❷ Déploiement et intégration.
- ❸ Gestion des données.
- ❹ Sécurité et éthique.
- ❺ Surveillance et maintenance.