Language, Inclusion, Remediation, Interculturality & Communication | LiRiC
The interdisciplinary thematic institutes of the University of Strasbourg & cnrs & Inserm

Intitulé du stage : **Reconnaissance automatique de la langue alsacienne dans les métadonnées bibliographiques**

La Bibliothèque nationale et universitaire de Strasbourg conserve un fonds régional de premier plan, les **Alsatiques**, constitué notamment dans le cadre du dépôt légal imprimeur. La qualité et l'homogénéité des métadonnées bibliographiques associées à ces collections constituent un enjeu majeur pour leur accessibilité et leur valorisation.

Un état des lieux met en évidence une **hétérogénéité importante dans l'usage du critère de langue** pour les documents rédigés en alsacien (par exemple : « langues germaniques », « français », « allemand », « alémanique », voire langue indéterminée). Cette situation est liée à la diversité des pratiques de signalement mais aussi aux limites de la norme ISO 639-2 actuellement en vigueur dans le catalogue universitaire (SUDOC) et peu adaptée aux langues régionales, dans la mesure où moins de 500 langues disposent d'un code dans cette norme. La transition vers la norme ISO 639-3, prévue à partir de 2026, sera l'occasion d'uniformiser et améliorer les pratiques de signalement des documents rédigés en alsacien.

Objectif du stage :

Le stage vise à **utiliser les méthodes de traitement automatique des langues** pour améliorer l'identification des documents en alsacien à partir des métadonnées bibliographiques du fonds des Alsatiques, et à préparer leur future caractérisation selon la norme ISO 639-3.

Activités du stage :

1. Analyse des données et constitution du corpus d'apprentissage

- Analyse exploratoire des métadonnées existantes (titres, sous-titres, notes, résumés, etc.).
- Étude des pratiques de codage de la langue au sein du fonds des Alsatiques.
- Constitution d'un jeu de données annoté à partir d'un échantillon de notices validées (documents en alsacien / autres langues).

2. Entraînement d'un modèle prédictif

- Mise en œuvre de méthodes de classification automatique de langue adaptées à des données textuelles courtes et hétérogènes.
- Entraînement et évaluation d'un modèle supervisé destiné à détecter les documents en alsacien à partir des métadonnées seules.
- Analyse des performances et des limites du modèle.

3. Détection et caractérisation des documents en alsacien

- Application du modèle à un corpus élargi de notices afin d'identifier les documents en alsacien non ou mal codés.
- Proposition d'une première caractérisation des documents détectés en vue de leur future description selon la norme ISO 639-3.

Profil recherché
- Étudiant·e en Master 2 TAL, Humanités numériques, Linguistique computationnelle ou Informatique appliquée aux langues.
- Compétences techniques : Python (HuggingFace, Transformers, spaCy, PyTorch), prétraitement de corpus, évaluation de modèles.
- Connaissances linguistiques souhaitées : morphosyntaxe et dialectologie alsacienne (ou à défaut, connaissance de l'allemand), intérêt pour les langues régionales.
- Autonomie, rigueur, curiosité scientifique.

Encadrement : Le/La stagiaire est accueilli(e) au sein du Service des bibliothèques et données numériques de la Bnu et placé(e) sous l'autorité de Mme Perrine Hamann, adjointe du service et également Responsable qualité et usage des métadonnées. Le /la stagiaire pourra s'appuyer sur son encadrante Bnu pour les aspects techniques du stage et sur Mme Delphine Bernhard, responsable pédagogique du master Technologies des Langues. Sur le plan scientifique, il/elle pourra être accompagné(e) par le responsable des fonds Alsatiques M. Daniel Bornemann.
Durée et calendrier : Le stage peut s'effectuer entre le mois d'avril et le mois de juillet 2025. La durée du stage est de trois mois. A ce titre il fait l'objet d'une gratification financée par l'ITI LiRiC.

Conditions du stage
- Durée : 3 mois (avril – juillet 2026)
- Lieu : Strasbourg (Université de Strasbourg / Bnu)

- Télétravail partiel possible selon les besoins.
- Gratification de stage financée par l'ITI LiRiC

Bibliographie

- Bernhard, D., Vergez-Couret, M., & Dupuy, E. (2024). *Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues.* Cahiers du plurilinguisme européen, 16.
- Binot, J., Werner, C., & Bernhard, D. (2024). *Mistral sur les Vosges : L'IA souffle-t-elle dans la bonne direction pour l'alsacien ?* Faculté des Langues, Université de Strasbourg.
- Kargaran, A. H. et al. (2023). *GlotLID: Language Identification for Low-Resource Languages.* EMNLP.

Candidature (date limite 15/02/2026) :

Envoi des CV et lettre de motivation à :

perrine.hamann@bnu.fr
rosanne.wingert@bnu.fr
dbernhard@unistra.fr