OFFRE DE STAGE **TRAITEMENT AUTOMATIQUE DES LANGUES** (6 mois, niveau Master)

Durée du stage : 6 mois (début : février 2025)
Date limite de candidature : 25 janvier 2025
Gratification : montant légal en vigueur (environ 600€ / mois).
Laboratoire d'accueil : Équipe PASTIS du *LIASD* (EA 4383), Université Paris 8, 2 rue de la liberté, 93 526 Saint-Denis
Document au format PDF : https://alicemillour.github.io/assets/stage_LIASD2025.pdf

Sujet de Stage : Développement d'un système de *Speech-to-Text* (STT) pour la langue bretonne

Contexte

Ce sujet de stage s'inscrit dans le contexte de la **langue bretonne** : en danger de disparition selon les critères de l'UNESCO, ses locuteurs et locutrices disposent de peu d'outils logiciels adaptés à leurs pratiques (Jouitteau 2023, Jouitteau & Bideault 2023). Le développement d'un **système de reconnaissance vocale automatique** (*Automatic Speech Recognition*, ASR, ou *Speech-to-text,* STT) **adapté au contexte linguistique du breton** constitue notamment un besoin urgent. Les locuteur·ices, créateur·ices de contenu en ligne et journalistes demandent explicitement à pouvoir utiliser les mêmes facilités de dictée, sous-titrage et transcription automatiques qu'ils et elles utilisent déjà en français et en anglais.

Objectifs du stage :

L'objectif du stage est de développer et d'évaluer un prototype de système *speech-to-text* capable de convertir la parole en breton en texte écrit. Ce système pourra être utilisé pour des applications telles que la transcription de discours, la création de sous-titres automatiques, ou la reconnaissance vocale dans des interfaces utilisateurs en breton.

Le ou la stagiaire étudiera les ressources et systèmes existants (Inventaire des ressources numériques pour le breton, dont Duval-Guennoc 2022 & Vangberg & Farhat 2023). Après un état de l'art des méthodes récentes de transcription automatique en contexte peu doté, l'étudiant·e exploitera les jeux de données existants, tels que Common Voice (Ardila et al., 2019) afin d'entraîner et d'évaluer un modèle adapté.

Ce stage s'adresse aux étudiant·es de Master 2 d'informatique ou de traitement automatique des langues.

Candidature

Les candidatures sont à adresser à :

- Alice Millour, LIASD, Université Paris 8 Vincennes Saint-Denis, am@up8.edu
- Loïc Grobol, MoDyCo, Université Paris Nanterre, lgrobol@parisnanterre.fr

Le ou la stagiaire sera accueilli·e dans les locaux de l'Université Paris 8 au sein du laboratoire LIASD. Ce travail fait l'objet d'une collaboration avec l'Université Paris Nanterre,

le laboratoire CNRS IKER et l'université de Bangor (UK), inclura la participation à un workshop collectif prévu en mars 2025 à Brest.

Références

- Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers and Gregor Weber. 2019"Common Voice: A Massively-Multilingual Speech Corpus." *ArXiv* abs/1912.06670.
- Duval-Guennoc Gweltaz. 2022-présent. *Anaouder, a VOSK model for the Breton language*.
- Jouitteau Mélanie. 2023. *Community Internally-driven Corpus Buildings. Three Examples from the Breton Ecosystem*. Proc. of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), 103-107, doi: 10.21437/SIGUL.2023-22.
- Jouitteau Mélanie et Bideault Reun. 2023. *Outils numériques et traitement automatique du breton*, Annie Rialland, Michela Russo (dir.), Langues régionales de France: nouvelles approches, nouvelles méthodologies, revitalisation, Éditions de la Société de Linguistique de Paris, 37-74.
- Vangberg, Preben & Leena Farhat. 2023. 'Speech-to-text for Breton', présentation à la *Celtic Student Conference*, 30 Mars 2023, Glasgow, United Kingdom.