Suhaïla ABARKAN

Mini Rapport

Stage linguistique (6 juin 2024 – 12 juillet 2024)

En tant que stagiaire de première année de Master en CMI (Cursus Master Ingénierie) ISI (Ingénieur de la Statistique et Informatique) d'une durée de six semaines (du 6 juin 2024 au 12 juillet 2024) encadrée par Alexandre Genadot, et Nicolas Guilliot, à l'INRIA, je souhaite remercier chaleureusement mes tuteurs de stage pour m'avoir permis de vivre cette expérience enrichissante au sein d'une équipe de recherche dans le monde du travail.

Le but de ce stage était de mettre en valeur des données de l'enquête Bourciez et de l'équipe PaRL, mais aussi de participer à la conception d'une base de données. En effet, construire une méta-base incluant des méta-données et proposer des outils de visualisation de celles-ci était l'une des missions proposées. En somme, prendre connaissance des méthodes de transcription, création de bases de données et analyses statistiques formaient mes principaux intérêts.

Une transcription permet de recopier un texte depuis une image ou un son qu'on nous fournit (donc existant sous plusieurs formes) et diffère selon son type, bien que souvent il s'agit d'une intelligence artificielle qui transcrit le texte dans le monde informatique. Concernant les types de transcriptions étudiées, la première est l'application Transkribus utilisable gratuitement et sûrement la plus efficace pour ma part afin de transcrire un fichier historique : le texte de l'enfant prodigue, du canton de Caraman. La seconde est la mise à disposition des bibliothèques Python, mais malheureusement inefficace sur ce texte historique (bien qu'efficace sur un texte de fichier plus "lisible" traditionnellement dans le français actuel). Par manque de temps, je n'ai pas pu établir plus de recherches concernant une transcription du style Python, bien qu'une méthode proposée par M. Genadot était d'entrainer plusieurs fois le texte/paragraphe en blanc et noir et de corriger si besoin (style réseaux de neurones), en procédant ligne par ligne pour plus d'efficacité. Globalement, cet outil informatique qu'est la transcription automatique me semble très utile à l'avenir si, pour une création de base de données par exemple, il y a une nécessité de récupérer ces archives sous un format de texte, comme c'était le cas au sein de cette enquête, pour pouvoir ensuite réaliser des analyses statistiques ou des visualisations de données.

La création de la base de données sous un format csv depuis des réponses à l'enquête PaRL sur la traduction du conte de l'âne triste dans l'idiome de la commune sous un format excel, était l'une des missions qui représentait un grand challenge pour moi. En effet, n'ayant jamais créé de base de données sous un format csv, et en ayant le libre-arbitre sur comment procéder et organiser cette méta-base fut très encourageant et m'a permis une nouvelle compréhension des structures de données pour la manipulation et l'organisation des méta-données. Cela explique le fait que j'ai créé deux versions assez distinctes de bases de données selon les analyses qu'on voudrait élaborer dessus. Pour compléter cette base de données qui contient en plus les coordonnées géographiques et les informations du locuteur, j'ai réalisé des découpages par mots pour y introduire la traduction correspondante, depuis la traduction de la phrase globale. Ce fut l'une des missions les plus complexes car, en effet, les méta-bases ne sont pas parfaites en

raison de la pertinence pour certaines traductions de mots. Pour finir sur cette partie, des analyses statistiques pour étudier les variations linguistiques (telles que les fréquences) selon Levenshtein ont été réalisées.

L'exploitation de la base de données de l'enquête Bourciez est constituée d'analyses statistiques sur les variations linguistiques et géographiques, du clustering, des méthodes de calculs des plus proches voisins, des matrices, des recherches de données sur la population, des modèles de régression... Sans aucun doute la partie la plus fructueuse en termes de nouveaux concepts et apprentissages dans le cadre de ce stage linguistique.

L'analyse des variations linguistiques est appliquée à non seulement des mots choisis, mais aussi sur les plus proches voisins (à la méthode de la triangulation de Delaunay) de chaque mot (78 au total) de cette base de données. Grâce à un processus automatisé, cette analyse crée un csv pour chaque mot de la base de données qui étudie le type d'opérations (selon Levenshtein) pour passer de la traduction du mot à la traduction des voisins du mot, s'il y a. Concernant les matrices sur les distances géographiques et linguistiques (que ce soit sur tous les mots, ou sur des mots particuliers), des corrélations sont étudiées ainsi qu'une étude entre le nombre de mots pris en compte dans la corrélation des ces deux matrices pour capter une possible évolution. Des courbes d'ajustement ainsi que des nuages de points selon une fonction logarithmique sont aussi étudiées. Du côté des méthodes de clustering, selon la méthode du coude, des dendrogrammes de la CAH avec la méthode de Ward ainsi que des projections de clusters sur des cartes sont réalisées avec les matrices créées précédemment.

Afin de compléter la 3e base de données de l'enquête Bourciez, le nombre de population par communes qui ont répondu à l'enquête va être récupéré grâce à des fichiers excel que dispose l'INSEE en 1896 et un nouveau fichier de base de données en sera créé pour inclure cette information. Des valeurs manquantes seront enregistrées dans quelques communes (516 sur 3392) et les procédés suivants seront faits pour y remédier : gérer les communes se terminant par "bis" en leur associant le même nombre de population qu'une commune ayant le même nom sans "bis", et associer le même nombre de population d'une commune que le nombre dans la commune la plus proche géographiquement grâce au triangle de Delaunay. Une autre façon de faire est d'utiliser le web scraping pour récupérer les informations sur des pages internet grâce à du code "intelligent", mais par manque de temps je n'ai pas pu explorer cette idée. Pour le clustering sur cette population, on retrouve : des matrices sur la population, des corrélations entre ces matrices de population créées et les matrices de distances linguistiques (selon Levenshtein) et distances géographiques, des modèles de régression avec un calcul du R^2 pour notamment voir si l'inclusion de la population augmente ce score, et enfin des dendrogrammes de la CAH de Ward.

En conclusion, ce stage m'a permis de suivre une suite logique du projet réalisé durant mon semestre avec M. Genadot comme encadrant et d'acquérir de nouveaux concepts informatiques et statistiques tout en étant dans un environnement professionnel afin de consolider mes compétences pratiques. Ce stage a formé une étape importante pour ma formation et les compétences acquises durant ces six semaines me seront précieuses pour la suite.

Un grand merci à mes tuteurs
Alexandre Genadot et Nicolas Guilliot !