ተቋም национален שפה 文化 شرقية

inalco

Institut national
des langues
et civilisations orientales

Jennifer Lewis-Wong

Enjeux et méthodes pour la création de corpus en langues peu dotées. Application à la classification de textes pour l'apprentissage du birman.

Directeurs de thèse : M. Mathieu Valette et M^me^ San San Hnin Tun

Date de soutenance : le 27 janvier 2023

Mathieu Valette

San San HNIN TUN

Résumé

Trouver du matériel de lecture adapté aux apprenants de langues peu enseignées est un problème courant, tant pour les apprenants que pour les enseignants. Le traitement automatique offre des méthodes prometteuses pour faciliter ce processus. Comme leur mise en œuvre nécessite des corpus d'entraînement spécifiques à la langue, et que ces langues sont également peu dotées, la qualité des corpus est encore plus importante. Il nous a semblé nécessaire de considérer les particularités de la langue et de l'informatisation de son système d'écriture et le contexte d'utilisation du corpus, les études en linguistique et en lexicographie, les aspects culturels et même la tradition d'enseignement, car les apprenants sont probablement davantage influencés par les ressources existantes lorsqu'elles sont peu nombreuses. Cette thèse porte sur une méthode d'évaluation lexicale de textes pour le birman langue étrangère. D'abord la création de deux types de corpus : des textes authentiques et des ressources didactiques, ce dernier renseignant comment segmenter en unités minimales d'analyse ou « mots », prétraitement nécessaire car le birman ne les délimite pas par des espaces. Nous prenons également en compte les aspects culturels et la fréquence conjointe des syllabes dans l'entraînement d'un outil de segmentation. Les textes authentiques sont utilisés pour créer une liste de fréquences lexicales, utilisant la méthode de la fréquence réduite moyenne pour tenir compte de la dispersion. Cette liste est utilisée pour entraîner une SVM afin de classer les textes par difficulté croissante, méthode purement lexicale et prometteuse pour les langues peu dotées.