Motor, un outil de segmentation accessible en ligne

Guillaume de Malézieux[1] Jennifer Lewis-Wong[1,2] Vincent Berment[1, 3]

(1) Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM - EA 2520), INALCO, 2 rue de Lille, 75343 Paris Cedex 07, France
(2) Langues et Civilisations à Tradition Orale - CNRS / Paris III / INALCO (LACITO - UMR 7107), 7 rue Guy Môquet (bât. D), 94801 Villejuif Cedex, France
(3) GETALP, LIG-campus, CS 40700 - 38058 Grenoble cedex 9, France

guillaume.de-malezieux@protonmail.com, jennifer.wong@inalco.fr, vincent.berment@inalco.fr

RÉSUMÉ

Dans cette démonstration, nous montrons le fonctionnement des segmenteurs disponibles en ligne pour diverses langues (birman, khmer, lao, thaï et tibétain) et réalisés avec l'outil Motor.

ABSTRACT

Motor, a segmentation tool accessible online.
In this demonstration, we present the use of segmenters available online for several languages (Burmese, Khmer, Lao, Thai, Tibetan), and developed with a tool called Motor.

MOTS-CLÉS : Segmentation, tokenization, langues peu dotées.

KEYWORDS: Segmentation, tokenization, under-resourced languages.

La plupart des langues d'Asie utilisent des systèmes d'écriture dits non segmentés, car ne séparant pas les mots par des espaces. La segmentation en mots est une question centrale pour ces langues, et la performance des segmenteurs est déterminante pour tous les traitements avals. Dans cet article, nous présentons les segmenteurs réalisés avec Motor, qui est un outil permettant de développer des segmenteurs à partir d'une liste des mots de la langue. L'algorithme mis en œuvre dans Motor est un algorithme de « plus petit nombre de mots ». Il a été utilisé pour plusieurs langues d'Asie du Sud-Est : birman, khmer, lao, thaï (siamois) et tibétain (avec XXX et son équipe à XXX), ainsi que pour le japonais. L'état courant de ces segmenteurs est sur www.lingwarium.org/motor/Segmentation.

Le développement des segmenteurs à partir de Motor se fait a partir d'une liste de mots fournie a Motor sous forme d'une table dans une base de données sqlite3. Cette base est dotée d'une unique table avec une colonne ≪ Cle ≫ (simple numéro d'ordre) et une colonne « Article » pour les mots. Pour optimiser la vitesse de segmentation, un index doit être mis sur la colonne des mots.

L'environnement de développement est constitué de l'interface de test de la page publique citée plus haut, et d'une zone permettant de mettre à jour la base de données (voir figures 1 et 2).

Les segmenteurs réalisés ont été utilisés comme première étape dans des systèmes de traduction automatique (cf. projet « Petit Prince » : lingwarium.org/heloise/index.php?Ref=&ws=LittlePrince). Dans ce cadre, des informations ont été ajoutées à la base lexicale, la même base ayant ainsi pu produire la segmentation, mais aussi l'analyse morphologique et le transfert lexical.

Motor, comme les autres outils disponibles sur lingwarium.org, est disponible en ligne, de manière à permettre des développements collaboratifs. Un service d'API est aussi disponible pour permettre l'utilisation des segmenteurs par des sites ou outils externes (appels en cURL...).

Figure 1: Interface de test

Figure 2: Mise à jour de la base de données