# Translittération automatique pour une paire de langues peu dotée

Ngoc Tan Le[1] Fatiha Sadat[1] Lucie Ménard[2]
(1) Université du Québec à Montréal, 201 avenue du Président-Kennedy, H2X 3Y7, Montréal, Canada
(2) UQÀM, Laboratoire de phonétique, 320 Sainte-Catherine Est, H2X 1L7, Montréal, Canada
le.ngoc_tan@courrier.uqam.ca, sadat.fatiha@uqam.ca, menard.lucie@uqam.ca

RÉSUMÉ

La translittération convertit phonétiquement les mots dans une langue source (*i.e. français*) en mots équivalents dans une langue cible (*i.e. vietnamien*). Cette conversion nécessite un nombre considérable de règles définies par les experts linguistes pour déterminer comment les phonèmes sont alignés ainsi que prendre en compte le système de phonologie de la langue cible. La problématique pour les paires de langues peu dotées lie à la pénurie des ressources linguistiques. Dans ce travail de recherche, nous présentons une démonstration de conversion de graphème en phonème pour pallier au problème de translittération pour une paire de langues peu dotée, avec une application sur français-vietnamien. Notre système nécessite un petit corpus d'apprentissage phonétique bilingue. Nous avons obtenu des résultats prometteurs, avec un gain de +4,40% de score BLEU, par rapport au système de base utilisant l'approche de traduction automatique statistique.

MOTS-CLÉS : Translittération, graphème, phonème, traduction automatique, langue peu dotée, français-vietnamien.

KEYWORDS: Transliteration, grapheme, phoneme, machine translation, under-resourced language, French-Vietnamese.

## 1 Introduction

La translittération consiste en un processus de transformation d'un mot d'un système d'écriture (appelé mot source) vers un mot, phonétiquement équivalent, d'un autre système d'écriture (appelé mot cible) (Knight & Graehl, 1998). Beaucoup d'entités nommées (*i.e. les noms de personne, de location, d'organisation, les termes techniques, etc.*) sont souvent translittérées d'une langue source vers une langue cible quand la traduction est difficile ou impossible. La translittération peut être considérée comme une sous-tâche de la traduction automatique (TA).

Par ailleurs, avec l'évolution de hautes technologies, les gens ont tendance à inventer de nouveaux mots. Il est très difficile de définir toutes les règles possibles de conversion phonétique entre la langue source et la langue cible. Nous nous intéressons à résoudre les mots hors vocabulaire (MHV) considérés comme noms propres ou termes techniques issus d'un système de traduction automatique (TA) pour une paire de langues peu dotée, français-vietnamien.

## 2 Approche proposée

Notre approche se déroule en trois étapes principales : *(1) prétraitement, (2) classification et (3) re-classement de la liste des k-meilleurs résultats*. Tout le processus est illustré dans la Figure 2.

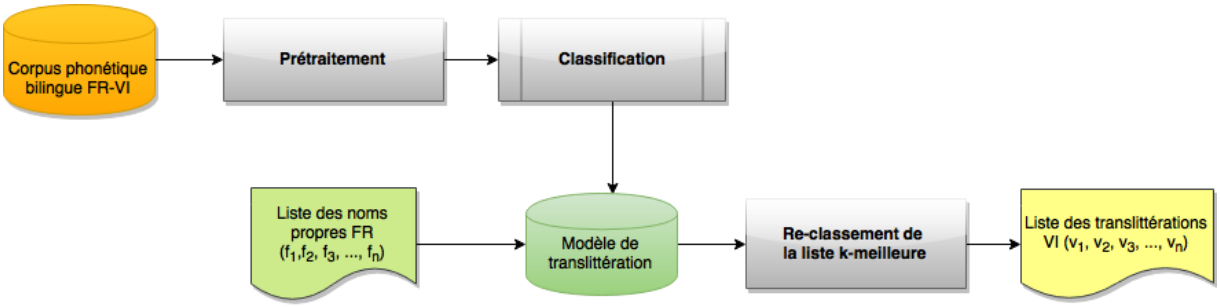


FIGURE 1 – Architecture de translittération des entités nommées bilingues pour une paire de langues peu dotée

## 3 Expérimentation

Nous utilisons un corpus phonétique bilingue qui a été collecté depuis les sites Web d'actualités comme présentés dans (Cao *et al.*, 2010). Ce corpus d'apprentissage possède 4 259 paires de noms propres bilingues français-vietnamien. Le corpus phonétique bilingue d'apprentissage est découpé en deux ensembles d'apprentissage et de test avec un ratio de 90% et 10% respectivement. Pour évaluer notre approche proposée, nous implémentons trois systèmes, notamment le système de base (*pbSMT sans distorsion*), le système 1 (*pbSMT avec distorsion*) et le système 2 (*notre approche proposée*) (Table 1).

| **Métrique** | **Système** | **Moyenne** | $\overline{s}_{sel}$ | $s_{Test}$ | $p$-**valeur** |
|---|---|---|---|---|---|
| BLEU ↑ | Système de base (*pbSMT sans distorsion*) | 61,3 | 1,7 | - | - |
| | Système 1 (*pbSMT avec distorsion*) | 61,6 | 1,7 | - | 0,79 |
| | Système 2 (*Notre approche*) | **65,7** | 1,5 | - | 0,01 |
| TER ↓ | Système de base (*pbSMT sans distorsion*) | 24,8 | 1,2 | - | - |
| | Système 1 (*pbSMT avec distorsion*) | 24,5 | 1,2 | - | 0,13 |
| | Système 2 (*Notre approche*) | **20,5** | 1,0 | - | 0,00 |
| PER ↓ | Système de base (*pbSMT sans distorsion*) | 4,42 | - | - | - |
| | Système 1 (*pbSMT avec distorsion*) | 4.05 | - | - | - |
| | Système 2 (*Notre approche*) | **3,80** | - | - | - |

TABLE 1 – Évaluation des scores pour tous les systèmes : **BLEU**, **TER** et **PER**.
$p$-**valeurs** sont relatives au système de base et indiquent si une différence de cette magnitude entre le système de base et le système à comparer. $\overline{s}_{sel}$ indique la variance due à la sélection du test.

# 4 Conclusion et perspective

Dans cet article, nous avons présenté une méthode originale supervisée pour pallier au problème de translittération pour une paire de langues peu dotée, avec une application sur la paire de langues français-vietnamien. Nous avons obtenu des résultats prometteurs, avec un gain de +4,40% de score BLEU, en comparant notre approche au système de base. Ce système de translittération des entités nommées bilingues possède la capacité d'apprendre, de manière automatique, les régularités linguistiques à partir du corpus phonétique bilingue.

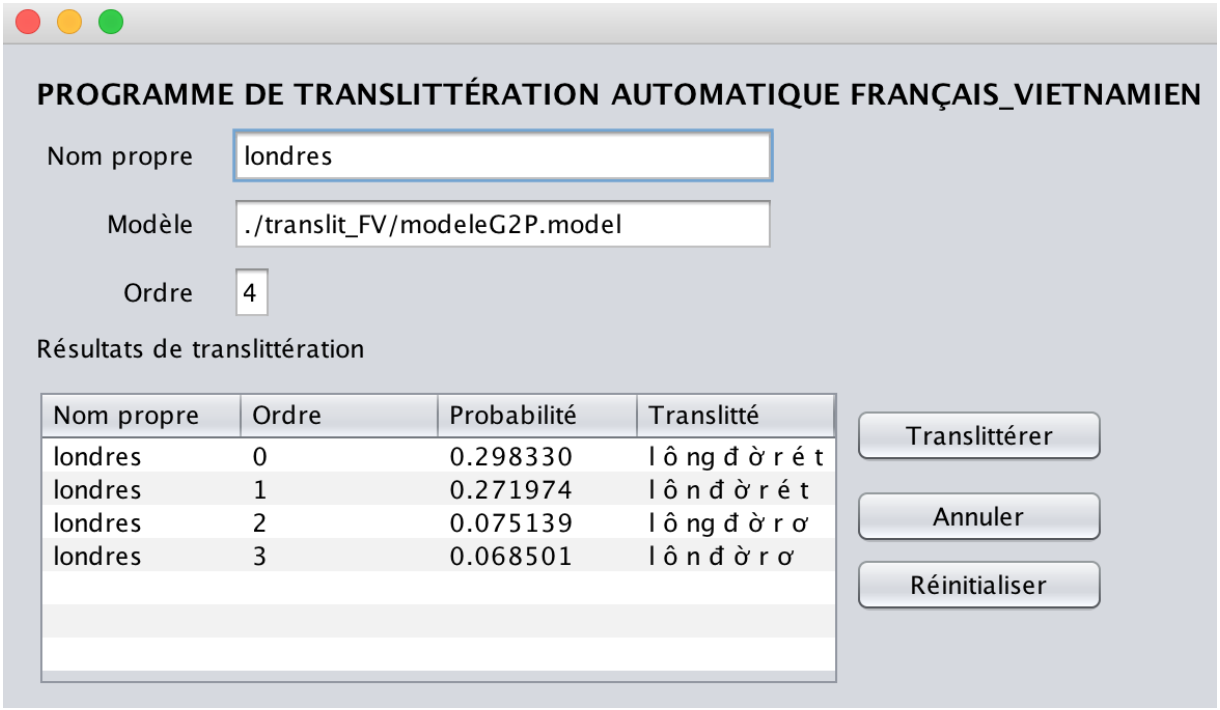


FIGURE 2 – Interface de translittération des entités nommées bilingues pour une paire de langues peu dotée

# Références


BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, **50**(5), 434–451.

CAO N. X., PHAM N. M. & VU Q. H. (2010). Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model. In *Proceedings of the 2010 Symposium on Information and Communication Technology*, p. 59–63 : Association for Computing Machinery.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, p. 177–180 : Association for Computational Linguistics.

NGO H. G., CHEN N. F., NGUYEN B. M., MA B. & LI H. (2015). Phonology-augmented statistical transliteration for low-resource languages. In *Interspeech*, p. 3670–3674.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

THU Y. K., PA W. P., SAGISAKA Y. & IWAHASHI N. (2016). Comparison of grapheme–to–phoneme conversion methods on a myanmar pronunciation dictionary. *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing 2016*, p. 11–22.