UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TRADUCTION AUTOMATIQUE POUR UNE PAIRE DE LANGUES PEU DOTÉE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

NGOC TAN LE

NOVEMBRE 2019

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.07-2011). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

« Success consists of going from failure to failure without loss of enthusiasm. »

Winston Churchill

REMERCIEMENTS

Tout d'abord, je tiens à particulièrement remercier ma directrice de recherche principale, Professeure Fatiha Sadat, qui a été mon guide, pour m'avoir apporté un soutien étendu et de la patience tout au long de mon doctorat, et sans qui l'achèvement de cette thèse n'aurait été possible. Ses références pertinentes et ses nombreux commentaires avisés ont grandement contribué à réaliser cette thèse d'envergure. Je voudrais remercier également ma co-directrice de recherche, Professeure Lucie Ménard, pour son soutien et ses commentaires à travers les aspects cognitifs linguistiques.

Je tiens également à remercier les membres de mon comité : Professeur Yllias Chali, Professeur Grégoire Winterstein, Professeur Vladimir Makarenkov, pour leurs encouragements, leurs commentaires et leurs suggestions constructifs et perspicaces.

Mes sincères remerciements s'adressent au Gouvernement Canadien pour m'avoir octroyé une bourse de la Francophonie d'études supérieures pour leur soutien financier dans la poursuite de mes études de doctorat à l'Université du Québec à Montréal, Québec, Canada. D'ailleurs, mon doctorat a été enrichi grâce aux diverses interactions au sein du GDAC-CRIA, avec des membres de notre équipe (passé et présent) Alexsandro, Ange, Fatma, Billal, Mickael, Thomas, Rahma, Pamela, Ghaith, Yacine, etc. Je voudrais adresser un grand remerciement à tout le monde dans le laboratoire du GDAC-CRIA ainsi qu'au personnel de notre DIC - Doctorat en Informatique Cognitive.

Finalement, je voudrais adresser mes profonds remerciements à mes parents, à la famille de mon grand frère, au Vietnam, pour leur soutien indéfectible, leur encouragement en tout temps, afin que je puisse réaliser mon rêve d'effectuer une thèse au Canada.

TABLE DES MATIÈRES

LISTE DES TABLEAUX viii
LISTE DES FIGURES xi
LISTE DES ABRÉVIATIONS xii
RÉSUMÉ xvi
INTRODUCTION 1
CHAPITRE I FONDEMENTS THÉORIQUES 7
1.1 Phonie et graphie 7
1.2 Caractéristiques du vietnamien 10
1.3 Étude comparative entre les systèmes phonologiques du français et du vietnamien 12
1.3.1 Systèmes vocaliques 13
1.3.2 Systèmes consonantiques 16
1.4 Traitement phonologique des emprunts 20
1.4.1 Traitement des voyelles 21
1.4.2 Traitement des consonnes 23
1.5 Réflexion sociolinguistique sur les emprunts directs du français dans la langue vietnamienne 30
1.5.1 Contexte historique 30
1.5.2 Problématiques 32
1.6 Apprentissage automatique 34
1.6.1 Champs aléatoires conditionnels 34
1.6.2 Séparateurs à vaste marge 36
1.6.3 Plongements de mots 38
1.6.4 Réseaux de neurones récurrents 40

1.6.5 Mécanisme d'attention 41
CHAPITRE II ÉTAT DE L'ART 43
2.1 Noms propres 43
2.2 Reconnaissance des entités nommées 44
2.3 Translittération automatique des noms propres 46
2.4 Traduction automatique des noms propres 49
2.5 Traduction automatique statistique 52
2.5.1 Modèle de langue 54
2.5.2 Modèle de traduction à base de segments 55
2.5.3 Décodeur 56
2.6 Traduction automatique pour les langues peu dotées 57
2.7 Évaluations 60
2.7.1 Performances de transcription et de translittération 60
2.7.2 Précision, Rappel, score de F1 61
2.7.3 Performance de traduction automatique (BLEU) 62
2.7.4 Taux d'erreurs de traduction (TER) 63
2.7.5 Métrique d'évaluation de traduction avec l'ordonnancement explicite (METEOR) 63
CHAPITRE III EXTRACTION DES ENTITÉS NOMMÉES PAR PROJECTION TRANS-LINGUISTIQUE 65
3.1 Introduction 65
3.2 Approche proposée 68
3.3 Expérimentations 71
3.3.1 Préparation des données 71
3.3.2 Configuration 71
3.4 Évaluations 73
3.4.1 Résultats obtenus 73

3.4.2 Analyse des erreurs 76
3.5 Conclusion 77
CHAPITRE IV MODÈLE DE RECONNAISSANCE DES ENTITÉS NOMMÉES POUR UNE LANGUE PEU DOTÉE 78
4.1 Introduction 78
4.2 Approche proposée 80
4.2.1 Extraction des traits linguistiques 80
4.2.2 Modèle neuronal bidirectionnel 81
4.3 Expérimentations 83
4.3.1 Préparation des données 84
4.3.2 Configuration 85
4.4 Évaluations 87
4.4.1 Résultats obtenus 87
4.4.2 Analyse des erreurs 89
4.5 Conclusion 91
CHAPITRE V MODÈLE DE TRANSLITTÉRATION DES ENTITÉS NOMMÉES POUR UNE PAIRE DE LANGUES PEU DOTÉE 92
5.1 Introduction 92
5.2 Méthodologie 94
5.2.1 Approche proposée 94
5.2.2 Alignements graphémiques pour une langue peu dotée 96
5.3 Expérimentations 98
5.3.1 Préparation des données 98
5.3.2 Configuration 98
5.4 Évaluations 100
5.4.1 Résultats obtenus 100
5.4.2 Analyse des erreurs 104

5.5 Conclusion 107

CHAPITRE VI SYSTÈME DE TRADUCTION ET DE TRANSLITTÉRATION AUTOMATIQUE DES ENTITÉS NOMMÉES (TAEN) 108

6.1 Introduction 108

6.2 Approches proposées 109

6.2.1 Approche à base de statistiques 110

6.2.2 Approche à base de réseaux de neurones 111

6.3 Expérimentations 113

6.3.1 Préparation des données 113

6.3.2 Configuration 115

6.4 Évaluations 119

6.4.1 Résultats obtenus 119

6.4.2 Analyse des erreurs 121

6.5 Conclusion 125

CONCLUSION 127

ANNEXE A 133

RÉFÉRENCES 137

LISTE DES TABLEAUX

Tableau Page

1.1 Les tons du vietnamien standard. La notion a été présenté dans le travail de recherche de Ferlus (1999). *Source : Huynh (2010).* 12

1.2 Inventaire des voyelles en français. *Source : Kang* et al. *(2016)* 13

1.3 Inventaire des voyelles en vietnamien. *Source : Kang* et al. *(2016)* . . . 13

1.4 Systèmes vocaliques du français et du vietnamien : Points communs - Le lieu d'articulation. *Source : Xuan (2002).* 14

1.5 Systèmes vocaliques du français et du vietnamien : Points communs - Le degré d'ouverture. *Source : Xuan (2002).* 15

1.6 Systèmes vocaliques du français et du vietnamien : Points différents. *Source : Xuan (2002)* 16

1.7 Inventaire des consonnes en français. *Source : Kang* et al. *(2016)* 17

1.8 Inventaire des consonnes en vietnamien. *Source : Kang* et al. *(2016)* . . 17

1.9 Systèmes consonantiques du français et du vietnamien : Points communs. *Source : Xuan (2002)* 18

1.10 Systèmes consonantiques du français et du vietnamien : Points différents. *Source : Xuan (2002)* 19

1.11 Résumé et exemples des correspondances de voyelles entre le français et le vietnamien. *Source : Kang* et al. *(2016)* 23

1.12 Résumé et exemples des correspondances de consonnes entre le français et le vietnamien. *Source : Huynh (2010)* 29

3.1 Statistiques des données d'apprentissage pour la tâche de l'extraction des entités nommées bilingues pour la paire de langues peu dotée en français-vietnamien. 72

3.2 Résultats de l'expérimentation de l'extraction des entités nommées bilingues, avec un corpus de test de 1 060 paires de phrases en français-vietnamien. 74

3.3 Quelques exemples montrant une similarité de syllabes entre le terme source, en français, et le terme cible, en vietnamien. 75

3.4 Résultats des expérimentations de traduction automatique des entités nommées pour la paire de langues peu dotée français-vietnamien. 75

4.1 Illustration des traits linguistiques, à savoir surface de mot, lemme, grappe de mot (*word cluster*), étiquette morpho-syntaxique, *chunk* (segment syntaxique) et ENs (entités nommées). Ici, *en vietnamien : « Chỉ_có Pháp và Anh ủng_hộ đề_xuất của Fischler. », en français : « Seules la France et la Grande-Bretagne ont soutenu la proposition de Fischler. »* 81

4.2 Statistiques du corpus annoté par chaque classe d'entités nommées. . . . 84

4.3 Comparaison de performances entre différents systèmes de reconnaissance des entités nommées, notamment *SVM : Support Vector Machine* (en français : Machine à vecteurs de support), *CRF : Conditional Random Fields* (en français : Champs aléatoires conditionnels) et *Bi-LSTM : Bidirectional Long-Short Term Memory* (Mémoire à long-court terme bidirectionnel) 87

4.4 Quelques exemples des prédictions produites par notre modèle de reconnaissance des entités nommées pour la langue peu dotée vietnamienne. 88

4.5 Illustration des erreurs de prédiction de nouvelles entités nommées. . . . 90

4.6 Illustration des erreurs de prédiction d'annoter un nom d'organisation au lieu d'un nom de personne. 90

5.1 Quelques illustrations d'alignements graphémiques. Le graphème G s'aligne sur un graphème simple (1-à-1), un graphème composé (1-à-2), deux graphèmes sur un graphème simple (2-à-1) et plusieurs graphèmes sur plusieurs graphème (m-à-m). La lettre 's' est alignée sur un graphème nul '_' qui n'est pas prononcé, dans l'exemple de 'PARIS'. . . 96

5.2 Évaluation des expérimentations avec les métriques d'évaluation telles que *BLEU (BiLingual Evaluation Understudy)* (Papineni *et al.*, 2002), *taux d'erreurs de traduction (TER - Translation Error Rate)* (Snover *et al.*, 2009) et *taux d'erreurs de graphèmes (GER - Grapheme Error Rate)*. 102

5.3 Évaluation sur l'architecture d'encodeur-décodeur de LSTM bi-directionnel, avec différentes dimensions de 64 et de 128. 103

5.4 Exemples de résultats de prédiction de translittération par tous les systèmes d'évaluation, hypothèses et références pour six noms propres, tels que *PARIS, TIGRANE, TOULOUSE, TOURS, TRUFFAUT et ZURICH*. 105

5.5 Exemples de résultats de prédiction de translittération par deux systèmes d'évaluation, hypothèses et références pour six noms propres tels que *PARIS, NICE, JACQUES, VINCENTE, YUKON et ZIMBABWE*. 106

6.1 Statistiques des données d'apprentissage pour la tâche de traduction automatique des entités nommées pour la paire de langues peu dotée en français-vietnamien. 115

6.2 Évaluation des performances de différents systèmes de traduction automatique des entités nommées pour la paire de langues peu dotée français-vietnamien, en appliquant nos deux méthodes. 120

6.3 Quelques illustrations de prédiction de la traduction automatique des entités nommées produites par les quatre systèmes d'expérimentation, avec un ensemble d'entités nommées en français {*Indonésie, Cambodge, Asie du Sud-Est, Mékong, khmer*}. 122

6.4 Évaluation de la performance de notre modèle, le système 3 - TAS, pour la tâche de traduction automatique de 100 noms propres de la Bible bilingue française-vietnamienne, ainsi que 100 termes techniques de chimie en général. 122

6.5 Quelques exemples de prédiction de traduction automatique des noms propres, extraits de la Bible bilingue française-vietnamienne, produits par notre approche (système 3 - TAS). 123

6.6 Quelques exemples de prédiction de la traduction automatique de termes techniques de chimie, extraits d'un dictionnaire bilingue de mots d'emprunt du français dans la langue vietnamienne, produits par notre approche (système 3 - TAS). 124

LISTE DES FIGURES

Figure		Page
1.1	Alphabet latin (26 lettres), ayant 6 voyelles et 20 consonnes versus alphabet vietnamien (29 lettres), ayant 12 voyelles et 17 consonnes avec 18 uni-grammes, 9 digrammes et 1 tri-gramme. *Source : Do (2011).* . .	11
1.2	Modèle de sac de mots continus (CBOW) et modèle de Skip-gramme. Source : Mikolov *et al.* (2013).	39
2.1	Architecture générale d'un système de traduction automatique statistique, sans prendre en considération les deux phases de pré-traitement et de post-traitement	53
3.1	Architecture d'un système d'extraction des entités nommées bilingues par projection translinguistique, avec application pour une paire de langues peu dotée, telle que le français et le vietnamien.	70
4.1	Architecture du modèle de reconnaissance des entités nommées en vietnamien à base de réseaux de neurones (*Traduction de l'illustration en français : « Voici l'université des sciences naturelles »*).	82
5.1	Architecture du système de translittération à base de réseaux de neurones récurrents pour une paire de langues peu dotée, traitant des entités nommées bilingues.	94
5.2	Notre architecture de modèle à base de réseaux de neurones récurrents avec un encodeur-décodeur de LSTM bidirectionnel et la représentation d'alignement sur les séquences d'entrée. Nous utilisons les marqueurs <s> et </s>, <os> et </eos> pour remplir les séquences graphémiques à une longueur fixe.	95
5.3	Illustration de la phase de pré-traitement du corpus d'apprentissage avec une normalisation en minuscule, une suppression des traits d'union des syllabes et une segmentation de toutes les syllabes au niveau des caractères.	98

5.4 Illustration d'un alignement du mécanisme d'attention généré par le système de translittération à base de réseaux de neurones récurrents pour une paire de langues peu dotée, traitant des entités nommées bilingues, avec la source en français est « *risorgimento* » et la cible en vietnamien est « *rigiótghimântô* ». 103

6.1 Architecture générale de notre système de traduction automatique statistique des entités nommées bilingues du français vers le vietnamien . . 112

6.2 Architecture générale de notre système de traduction automatique neuronale des entités nommées bilingues. Illustration inspirée de (Koehn, 2017), avec un graphe de computation entièrement déroulé pour un exemple d'apprentissage ayant 7 *tokens* d'entrée <*s*> *Đây là Đại_học khoa_học tự_nhiên* </*s*> et 7 *tokens* de sortie <*s*> *Voici l'Université des Sciences* </*s*>. 114

LISTE DES ABRÉVIATIONS

ABB Abréviation

Bi-LSTM *Bidirectional Long-Short Term Memory* (en français : Mémoire à long-court terme bidirectionnelle)

BLEU *BiLingual Evaluation Understudy*

BRN Branche

CBOW *Continuous Bag-Of-Words* (en français : Sac de mots continus)

CLC Centre de linguistique computationnel

CLP Cercle Linguistique de Prague

CNN *Convolutional Neural Network* (en français : Réseau neuronal convolutif)

CPU *Central Processing Unit* (en français : Unité centrale de traitement)

CRF *Conditional Random Fields* (en français : Champs aléatoires conditionnels)

DES Désignation

DTM Date et heure

EN Entité Nommée

G2P *Grapheme to Phoneme* (en français : Graphème en Phonème)

GER *Grapheme Error Rate* (en français : Taux d'erreur de graphèmes)

GPU *Graphics Processing Unit* (en français : Unité de traitement graphique)

HMM *Hidden Markov Model* (en français : Modèle de Markov caché)

LOC Nom de lieu

LSTM *Long-Short Term Memory* (en français : Mémoire à long-court terme)

MEA Mesures

MERT *Minimum Error Rate Training* (en français : Apprentissage de taux d'erreur minimal)

METEOR *Metric for Evaluation of Translation with Explicite ORdering* (en français : Métrique d'évaluation de traduction avec l'ordonnancement explicite)

MHV Mots Hors Vocabulaire

MUC *Message Understanding Conference* (en français : Conférence de compréhension de message)

NIST *National Institute of Standards and Technology*

NLP *Natural Language Processing* (en français : Traitement automatique de langues naturelles)

NMT *Neural Machine Translation* (en français : Traduction automatique neuronale)

NPs Noms propres

NUM Numérique

ORG Nom d'organisation

P Précision

PBSMT *Phrase Based Statistical Machine Translation* (en français : Traduction automatique statistique à base de segments)

R Rappel

REN Reconnaissance des Entités Nommées

RNN *Recurrent Neural Network* (en français : Réseau de neurones récurrents)

SPE *The Sound Pattern of English*

SRILM *The SRI Language Modeling Toolkit* (en français : Outil de création de modèle de langage de SRI)

SVM *Support Vector Machine* (en français : Machines à vecteurs de support)

TA Traduction Automatique

TAEN Traduction automatique des entités nommées

TALN Traitement Automatique de Langues Naturelles

TAN Traduction Automatique Neuronale

TAS Traduction Automatique Statistique

TER *Translation Error Rate* (en français : Taux d'erreur de traduction)

TRM Terminologie

TLL Titre

URL *Uniform Resource Locator* (en français : Localisateur uniforme de ressource)

WER *Word Error Rate* (en français : Taux d'erreur de mot)

RÉSUMÉ

Durant ces dernières décennies, le Traitement Automatique des Langues Naturelles (TALN) a sans nul doute fait des progrès considérables sur le plan de la diversité des outils linguistiques mis à la disposition de la communauté et celui de la qualité des résultats obtenus. Néanmoins, ces progrès sont, jusqu'à une date récente, restés limités à un nombre relativement restreint de langues, majoritairement occidentales.

Les langues peu dotées, ou langues-π, langues moins bien informatisées que les grandes langues véhiculaires (l'anglais, l'espagnol, le français, etc.), souffrent du manque de ressources linguistiques dont les corpus parallèles, ressources importantes pour le développement de systèmes de TALN, notamment les systèmes de traduction automatique, et ainsi présentent plusieurs défis dans le domaine du TALN.

Ce projet de recherche se concentre sur la paire de langues peu dotée français-vietnamien afin de développer un système de traduction automatique performant tout en se focalisant sur la reconnaissance des entités nommées et leurs translittérations. En premier, une analyse théorique, du point de vue cognitif, est abordée sous l'angle des théories de la linguistique et de la traduction. Par la suite, du point de vue informatique, l'objectif principal de notre projet de recherche consiste à proposer une méthode originale et fiable ainsi que des solutions aux problèmes rencontrés lors de la traduction automatique des entités nommées. Aussi, nous déterminons les sous-objectifs de ce projet de recherche, dont la reconnaissance des entités nommées pour la langue vietnamien et également la tâche de la translittération des entités nommées bilingues pour la paire de langues français-vietnamien. En plus des approches statistiques, nous adaptons l'approche d'apprentissage profond (en anglais, deep learning) au sein de nos différents systèmes afin d'améliorer davantage la qualité et l'efficacité de la traduction automatique des entités nommées. Les contributions à la traduction automatique et à la translittération automatique des entités nommées bilingues ont permis non seulement de réduire le taux des mots hors vocabulaires, mots non traduits et/ou incorrectement traduits, mais aussi d'améliorer la qualité du système de la traduction automatique.

Mots clés : Traduction automatique, corpus bilingue, paire de langues peu dotée, français-vietnamien, entités nommées, reconnaissance des entités nommées, translittération.

INTRODUCTION

1. Contexte et motivation

Les sociétés humaines sont multilingues, et la mondialisation appelle à l'utilisation et à l'acquisition accrues de plusieurs langues pour la communication, médiatisée ou non par les technologies de l'information. La Traduction Automatique (TA) désigne des systèmes informatisés qui peuvent produire des traductions avec ou sans assistance humaine. Il s'agit d'un sous-domaine de la linguistique computationnelle qui tente d'automatiser le processus de traduction d'une langue naturelle source (texte ou parole) à une autre langue cible, en utilisant l'ordinateur. Le défi de la TA réside dans la façon de programmer un ordinateur pour comprendre un texte comme un homme le fait et aussi pour créer un nouveau texte dans la langue cible comme s'il était écrit par un humain (Do, 2011).

Les langues peu dotées, ou langues-π, langues moins bien informatisées que les grandes langues véhiculaires (l'anglais, l'espagnol, le français, etc.), souffrent du manque de ressources linguistiques dont les corpus parallèles, ressources importantes pour le développement d'un système de traduction automatique, et ainsi présentent plusieurs défis dans le domaine du traitement automatique de langues naturelles (TALN) (Berment, 2004).

La polysémie, « pluralité de significations au sein d'un continuum sémantique pour un même mot » (Grass, 2010), constitue le problème le plus souvent signalé au sein des systèmes de traduction automatique. Les mots ne fonctionnent pas tant comme des unités discrètes, c'est-à-dire bien délimitées et séparées les unes des autres, mais

comme des occurrences sujettes à un certain nombre de variations sémantiques au sein d'un champ spécifique.

D'après notre recherche dans les revues de littérature (voir Chapitre 2), en regard de la langue vietnamienne, il existe très peu, voire aucune recherche concernant la traduction automatique des entités nommées bilingues depuis et vers cette langue. C'est la raison pour laquelle le présent projet de thèse focalise sur ce thème.

Ainsi, nous nous concentrons sur les points suivants :

(i) la reconnaissance des entités nommées,

(ii) leurs traductions et translittérations vers une langue cible peu dotée,

(iii) l'amélioration de la qualité et de la performance de la traduction automatique pour les paires de langues peu dotées.

2. Objectifs de recherche

Nous présentons les objectifs de recherche de notre projet en les divisant en deux composantes cognitive et informatique.

2.1. Composante cognitive

Ce projet de recherche aborde une analyse théorique du point de vue cognitif, spécialement sous l'angle des théories liées à la phonétique et à la phonologie, ainsi que sous l'angle de la théorie de la traduction.

Dans le cadre de cette thèse, nous nous concentrons sur (i) la transcription phonétique, particulièrement en utilisant les règles de conversion graphémique propres aux deux langues française et vietnamienne, et sur (ii) les entités nommées, spécifiquement sur les noms propres, comme présentées dans la section 2.1.

L'objectif principal de notre projet de recherche consiste à traiter les particularités linguistiques pour une paire de langues peu dotée de ressources linguistiques, en l'occurrence, *français-vietnamien*. Nous présentons une étude comparative entre les systèmes phonologiques du français et du vietnamien, dans la section 1.3, ainsi qu'une autre étude du traitement phonologique des emprunts du français dans la langue vietnamienne, dans la section 1.4.

En ce qui concerne la théorie de la traduction automatique, il y a deux approches fondamentales : approche symbolique et approche statistique. En réalité, l'approche symbolique se base sur les règles de grammaire formelle, de formalismes logiques. Du point de vue des sciences cognitives, on distingue l'approche symbolique de l'approche connexionniste. Tandis que, du point de vue de la traduction automatique, les approches sont à base de règles, à base de statistiques et à base de réseaux de neurones. Dans le cadre de cette thèse, nous nous concentrons sur les approches à base de statistiques et à base de réseaux de neurones au sein de notre architecture. Nous allons expliquer davantage ces paradigmes dans le chapitre 2.

2.2. Composante informatique

L'objectif principal de notre projet de recherche, du point de vue informatique, consiste à proposer une méthode originale et fiable ainsi que des solutions aux problèmes rencontrés lors de la traduction automatique afin d'améliorer la qualité et la performance des systèmes de traduction automatique en utilisant une paire de langues peu dotée, avec une application sur la traduction automatique des entités nommées du français vers le vietnamien.

Par ailleurs, nous déterminons les sous-objectifs impliqués dans notre projet de recherche. Il s'agit d'une tâche de la reconnaissance des entités nommées pour la langue

vietnamien et également d'une autre tâche de la translittération des entités nommées bilingues pour la paire de langues français-vietnamien.

Nous détaillons notre projet de recherche en deux phases comme suit :

- La première phase consiste à effectuer les tâches suivantes :
  (i) La construction des ressources linguistiques bilingues pour la paire de langues français-vietnamien.
  (ii) Le développement d'un module de reconnaissance des entités nommées ainsi qu'un module de translittération des entités nommées. La recherche consiste à identifier et à classifier des entités nommées bilingues pour la paire de langues français-vietnamien et leurs translittérations.
  (iii) La construction d'un système de traduction automatique statistique (TAS) en incluant les deux modules de reconnaissance et de translittération des entités nommées, et ce pour une paire de langues peu dotées.
- La deuxième phase consiste à adapter l'approche d'apprentissage profond (*deep learning*) au sein de notre système de TA afin d'améliorer davantage la qualité et l'efficacité de la traduction automatique.

Les évaluations sont effectuées afin de comparer entre les différents systèmes en terme de performance.

3. Organisation de la thèse

La suite de cette thèse s'organise en deux parties. La première partie comporte les trois premiers chapitres. Tout d'abord, l'introduction est présentée. Puis, les fondements théoriques sont abordés dans le chapitre 1. Ensuite, l'état de l'art est présenté dans le chapitre 2, en détaillant davantage les langues-π ou langues peu dotées, les caractéristiques linguistiques concernant les noms propres et/ou les entités nommées ainsi que toutes les approches dans la traduction automatique des noms propres.

La deuxième partie comporte les chapitres 3 à 6 dans lesquels nos principales contributions sont présentées en détail. Chaque chapitre est présenté sous forme d'une structure spécifique. Chacun commence par une introduction, des approches proposées, des expérimentations avec des évaluations appropriées, et termine par une conclusion.

De plus, chaque chapitre a mené à diverses publications qui sont présentées ci-contre de manière non exhaustive. Une liste complète de nos publications est exposée dans la section d'annexe.

- Le chapitre 3 aborde une méthode d'extraction des entités nommées par projection translinguistique. Un corpus parallèle d'entités nommées est construit pour des fins de modéliser un système de traduction automatique des entités nommées.

  Ngoc Tan Le, Fatiha Sadat (2018). Cross-linguistic Projection for French-Vietnamese Named Entity Recognition. *Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, Springer LNAI (Lecture Notes in Artificial Intelligence) Series, Lecture Notes in Computer Science, DOI : 10.1007/978-3-319-93782-3_29. In book : Human Language Technology. Challenges for Computer Science and Linguistics, 407-419, June 2018.*

- Le chapitre 4 présente la modélisation d'un module de reconnaissance des entités nommées pour une langue peu dotée, avec l'application au vietnamien.

  Ngoc Tan Le, Fatiha Sadat (2018). Data Adaptation for Named Entity Recognition in Twitter with Features-Rich CRF. *Dans Proceedings of NAACL HLT 2018, Widening Natural Language Processing workshop, New Orleans, Louisiana, USA, 1 to 6 June 2018.*

  Ngoc Tan Le, Long Nguyen, Fatiha Sadat (2018). Neural Network-based Model for Named Entity Recognition in Low-Resource Settings. *Dans Proceedings*

of NAACL HLT 2018, Widening Natural Language Processing workshop, New Orleans, Louisiana, USA, 1 to 6 June 2018.

- Le chapitre 5 présente la modélisation d'un module de translittération des entités nommées avec différentes approches telles que l'approche à base de statistique et l'approche à base de réseaux de neurones. Par la suite, ce module sera incorporé dans le même système de traduction automatique.

  Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2018). Low Resource Machine Transliteration Using Recurrent Neural Network. *Dans Proceedings of ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), ISSN 2375-4699, 2018.*

- Le chapitre 6 présente notre système de traduction et de translittération automatique des entités nommées pour une paire de langues peu dotée, avec l'application du français vers le vietnamien.

Finalement, la conclusion récapitule les principales contributions de la thèse, ainsi que les recherches potentielles à venir.

CHAPITRE I

FONDEMENTS THÉORIQUES

1.1 Phonie et graphie

Ce chapitre aborde les fondements théoriques en présentant les points en commun et les différences du point de vue de la phonie et de la graphie, particulièrement sur les règles de conversion graphémique propres aux deux langues française et vietnamienne.

Tout d'abord, nous souhaitons présenter des notions de base selon le dictionnaire de la linguistique et de la phonétique (Crystal, 2011).

La *phonétique* étudie les caractéristiques de la fabrication du son humain, en particulier les sons utilisés dans la parole, et fournit des méthodes pour leur description, leur classification et leur transcription. Trois branches sont généralement reconnues :

(a) la *phonétique articulatoire* étudie la façon dont les sons de la parole sont produits (« articulés ») par les organes vocaux ;

(b) la *phonétique acoustique* étudie les propriétés physiques du son de la parole, telles qu'elles sont transmises entre la bouche et l'oreille ;

(c) la *phonétique auditive* étudie la réponse perceptuelle aux sons de la parole, telle que par l'oreille, le nerf auditif et le cerveau.

Le terme de « *phonétique instrumentale* » est utilisé pour l'étude de n'importe lequel de ces aspects du domaine à l'aide d'appareils physiques, tels que des dispositifs de

mesure du débit d'air ou d'analyse des ondes sonores. Les personnes engagées dans l'étude de la phonétique sont appelées *phonéticiens*.

Les travaux en phonétique peuvent donc être classés en deux grandes catégories :

- études générales sur l'articulation, l'acoustique ou la perception de la parole,
- études des propriétés phonétiques de langues spécifiques.

Dans la deuxième catégorie, il est évident qu'une autre dimension sera nécessaire pour étudier la manière dont les sons sont utilisés dans le système de prononciation d'une langue. Cette approche « fonctionnelle » de la phonétique est généralement utilisée sous le titre de la phonologie.

La *phonologie* est une branche de la linguistique qui étudie les systèmes sonores des langues. Sur un large éventail de sons que l'appareil vocal humain peut produire, et qui sont étudiés par la phonétique, seul un nombre relativement petit est utilisé distinctement dans une seule langue. Les sons sont organisés en un système de contrastes, qui sont analysés en termes de phonèmes, de traits distinctifs ou d'autres unités phonologiques similaires, en fonction de la théorie utilisée.

La phonologie a pour objectif de démontrer les patrons de sons distinctifs trouvés dans une langue et de faire des déclarations aussi générales que possible sur la nature des systèmes sonores dans les langues du monde. En d'autres termes, la phonologie s'intéresse à l'étendue et à la fonction des sons dans des langues spécifiques (appelées souvent « *phonétique fonctionnelle* »), ainsi qu'aux règles qui peuvent être écrites pour montrer les types de relations phonétiques qui lient et contrastent des mots et d'autres unités linguistiques.

Les unités de base de la phonétique et de la phonologie sont le *phone* et le *phonème*, respectivement.

Le *phonème* est l'unité minimale dans le système sonore d'une langue, selon les théories phonologiques traditionnelles. La motivation initiale du concept provenait du souci d'établir des modèles d'organisation au sein de la gamme infiniment large de sons entendus dans les langues. Les spécifications phonétiques des sons (ou des phones) entendus dans la parole contiennent bien plus de détails qu'il n'en faut pour identifier la manière dont les langues donnent des contrastes. La notion de phonème permettait aux linguistes de regrouper des ensembles de phones semblables en tant que variantes ou « membres » de la même unité sous-jacente.

Le *graphème* est l'unité contrastive minimale dans le système d'écriture d'une langue ; généralement entre parenthèses. Le graphème <a>, par exemple, se présente sous la forme de plusieurs allographes A, a, a, etc., qui peuvent être vus comme des unités de distribution complémentaire (par exemple, les majuscules limitées à la position initiale de la phrase, les noms propres, etc.), ou en variation libre (comme dans certains styles d'écriture manuscrite), tout comme en analyse phonémique. L'analyse des graphèmes est l'activité principale de la graphie (ou graphologie).

Le *français* comporte 37 phonèmes tandis que le *vietnamien* comporte 67 phonèmes (Xuan, 2002). L'alphabet français possède 26 graphèmes tandis que l'alphabet vietnamien en possède 29. Le nombre de phonèmes du vietnamien est presque deux fois plus celui du français. Cela peut être dû à diverses adaptations phonétiques de plusieurs langues à travers l'histoire du Vietnam, notamment le *chinois*, le *français*, le *portugais* et également l'*anglais*. Les caractéristiques du vietnamien sont présentées dans la section 1.2.

Dans le cadre de notre thèse, nous souhaitons observer les phénomènes de translittération des noms propres du point de vue phonologique, notamment avec réflexions sociolinguistique sur des emprunts de mots du français vers le vietnamien tout au long de l'Histoire du Vietnam.

1.2 Caractéristiques du vietnamien

À la différence de l'alphabet français, qui est issu de l'alphabet latin, l'alphabet vietnamien résulte d'une adaptation au vietnamien de l'alphabet latin tel qu'il était utilisé pour les langues romanes que parlaient les missionnaires (et surtout le *portugais*). Il s'agit de la romanisation du vietnamien. Ces missionnaires ont ajouté quelques lettres et pris, pour la représentation des tons, des lettres utilisées en grec, auxquels ils ont ajouté un point souscrit et un petit point d'interrogation – signe marquant dans les langues romanes une courbe mélodique particulière.

Le vietnamien est l'une des langues du sud-est asiatique avec l'alphabet latin. Par contre, le khmer, le laotien et le thaï ont gardé leur système graphique traditionnel (Hirst et Di Cristo, 1998). Le mandarin possède un système graphique avec un alphabet latin modifié, le *pinyin*, utilisé dans l'enseignement de la langue chinoise simplifiée aux étrangers.

Grâce à la romanisation du vietnamien, les relations phono-graphématiques simplifiées facilitent l'intégration de nouveaux mots étrangers, en particulier ceux d'origine française. D'où un grand nombre de mots français ont été prononcés à la vietnamienne, notamment dans les domaines scientifique, économique et social.

L'alphabet vietnamien tel qu'il est utilisé actuellement contient 29 lettres, soit, par rapport à l'alphabet standard de 26 lettres, 4 lettres en moins (f, j, w, z) et 7 lettres en plus, qui consistent, sur le plan graphique, en l'adjonction d'une marque supplémentaire (voir Figure 1.1).

La structure des syllabes en français est très riche, avec une variété de structures telles que *CV*, *CVC*, *CCVCC*, etc. où *C* est une consonne et *V* est une voyelle. En revanche, la structure de syllabes en vietnamien est très simple. Phe (1997) a constaté environ 10 000 syllabes pour le vietnamien. Une syllabe en vietnamien se constitue avec la

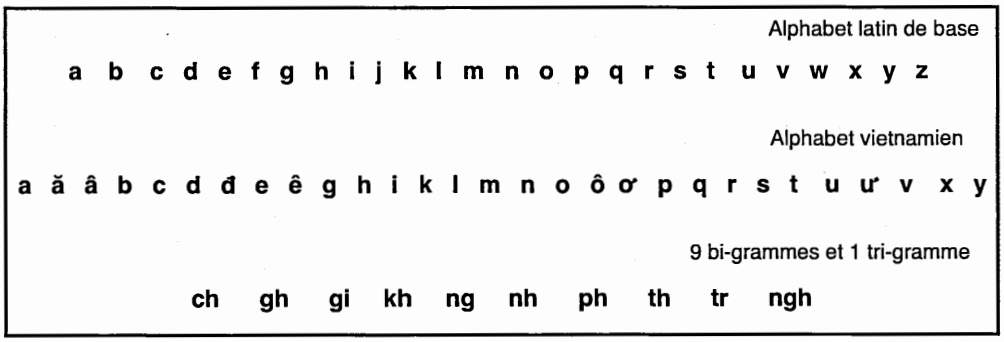

Figure 1.1 Alphabet latin (26 lettres), ayant 6 voyelles et 20 consonnes versus alphabet vietnamien (29 lettres), ayant 12 voyelles et 17 consonnes avec 18 uni-grammes, 9 digrammes et 1 tri-gramme. *Source : Do (2011).*

structure 1.1 suivante :

$$Syllabe = Attaque + Noyau + Coda + Ton\ lexical \tag{1.1}$$

La frontière d'une syllabe dépend des groupes (*clusters*) de consonnes, des voyelles et des tons lexicaux. Le *noyau* se constitue des voyelles telles que $Noyau = \{a, ă, â, e, ê, i, o, ô, ơ, u, ư, y\}$. Les consonnes initiales (*attaque*) sont $Attaque = \{b, ch, c, d, đ, gi, gh, g, h, kh, k, l, m, ngh, ng, nh, n, ph, q, r, s, th, tr, t, v, x, p\}$. Parmi ces consonnes finales (*coda*), il y en a 8 en coda telles que $Coda = \{c, ch, n, nh, ng, m, p, t\}$. Du point de vue linguistique, le ton lexical est une marque phonologique différenciant des morphèmes de structure segmentale identique. Il remplit donc une fonction distinctive. Le vietnamien possède six tons lexicaux. Ils sont ainsi décrits dans le Tableau 1.1 ci-dessous.

Tableau 1.1 Les tons du vietnamien standard. La notion a été présenté dans le travail de recherche de Ferlus (1999). *Source : Huynh (2010).*

Ton	**ngang (1)**	**huyền (2)**	**sắc (3)**		**nặng (4)**		**hỏi (5)**	**ngã (6)**
Notation	A1	A2	B1	D1	B2	D2	C1	C2
finale +/- obstruante	-	-	-	+	-	+	-	-
Registre	moyen haut	bas	haut		bas		moyen bas	haut
Courbe mélodique	égal	traînant, descendant	ascendant		descendant, bref		descendant puis remonté	descendant puis remonté bien
Autres caractéristiques	sans constriction	sans constriction, voix soufflée	voix tendue		voix tendue, glottalisée		cassé, voix tendue	cassé, voix glottalisée

Quelques exemples illustrent les six tons lexicaux :

(1) ngang A1 (*i.e. ba = père, ma = fantôme*),

(2) huyền A2 (*i.e. trà = thé, rồng = dragon*),

(3) sắc B1 et D1 (*i.e. có = posséder, tóc = cheveux*),

(4) nặng B2 et D2 (*i.e. lại = venir, bạn = ami*),

(5) hỏi C1 (*i.e. thuỷ = eau, tủ = armoire*),

(6) ngã C2 (*i.e. mỹ = américain, sĩ = intellectuel*).

1.3 Étude comparative entre les systèmes phonologiques du français et du vietnamien

Cette section vise à présenter une étude comparative des inventaires vocaliques et consonantiques du français et du vietnamien. En ce qui concerne l'inventaire des

phonèmes dans les deux systèmes, il est nécessaire de les présenter sous forme de tableaux comparatifs. Cela facilitera l'observation des convergences et des divergences des deux systèmes.

1.3.1 Systèmes vocaliques

En français standard, le système vocalique comporte des voyelles nasales et des voyelles orales (voir Tableau 1.2). Le vietnamien possède 9 voyelles monophtongues et 3 voyelles diphtongues (voir Tableau 1.3).

Tableau 1.2 Inventaire des voyelles en français. *Source : Kang* et al. *(2016)*

	Avant	**Central**	**Arrière**
Voyelles orales	i y		u
	e ø		o
	ɛ œ	ə	ɔ
		a	(ɑ)
Voyelles nasales	ɛ̃ (œ̃)		ɔ̃
			ɑ̃

Tableau 1.3 Inventaire des voyelles en vietnamien. *Source : Kang* et al. *(2016)*

	Avant	**Central**	**Arrière**
Monophtongues	i	ɯ	u
	e	ɣ	o
	ɛ (ɛ :)	ɤ̆	ɔ (ɔ :)
		ă	
		a	
Diphtongues	iə	ɯə	uə

Dans les deux systèmes, Xuan (2002) a constaté des points communs et des points différents pour deux langues en étude sur les mêmes critères de classement : *(i) le lieu d'articulation* et *(ii) le degré d'ouverture.*

(i) Le lieu d'articulation permet de différencier les voyelles antérieures des voyelles postérieures sur le plan articulatoire ou les voyelles aiguës des voyelles graves sur le plan acoustique. Selon ce critère, on distingue cinq séries de voyelles telles que *(1) antérieure, (2) mi-antérieure, (3) médiane, (4) mi-postérieure et (5) postérieure* (voir Tableau 1.4).

Tableau 1.4 Systèmes vocaliques du français et du vietnamien : Points communs - Le lieu d'articulation. *Source : Xuan (2002).*

	En français	**En vietnamien**
Antérieure	/ i / : Paris [pari]	/ i / : đi [đi] : aller
mi-antérieure	/ e / : été [ete]	/ e / : vê [ve] : rentrer
Médiane	/ ə / : petit [pəti]	/ a / : ba [ba] : papa
mi-postérieure	/ o / : mot [mo]	/ o / : một [mot] : un
Postérieure	/ u / : loup [lu]	/ u / : tủ [tu] : armoire

(ii) Le degré d'ouverture permet de différencier les voyelles fermées des voyelles ouvertes. On distingue cinq séries de voyelles telles que *(1) fermées, (2) mi-fermées, (3) moyenne, (4) mi-ouvertes et (5) ouvertes* (voir Tableau 1.5).

En ce qui concerne les points différents, on distingue en français comme en vietnamien sept séries comme suit :

(1) Autres critères de classement : En français, on constate la nasalisation tandis qu'en vietnamien, on constate l'absence de voyelles nasales et la stabilité du timbre. Aussi, on observe la labialisation qui est relative aux lèvres en français. Mais en vietnamien, l'entourage consonantique peut avoir une incidence sur la durée des voyelles. Une syllabe accentuée est plus longue qu'une syllabe non accentuée, en finale de phrase, les voyelles sont donc plus longues.

Tableau 1.5 Systèmes vocaliques du français et du vietnamien : Points communs - Le degré d'ouverture. *Source : Xuan (2002).*

	En français	**En vietnamien**
Fermées	/ i /, / u / : vie [vi], roue [ʀu]	/ i /, / u / : tim [tim], xu [su]
Mi-fermées	/ e /, / ø / : thé [te], bleu [blø]	/ iʌ /, / uʌ / : tiên [tiʌ n], uống [uʌ ŋ]
Moyenne	/ ə / : degré [dəgre]	/ ɣ /, / ɣ̆ / : mở [mɣ], mới [mɣ̆i]
Mi-ouvertes	/ ɛ /, / œ /: gai [gɛ], fleur [flœr]	/ ɛ /, / ĕ / : mẹ [mɛ], anh [ĕŋ]
Ouvertes	/ ɑ /, / ɑ̃ / : patte [pɑt], an [ɑ̃]	/ ɑ / ă / : tá [tɑ], tặng [tă̆ŋ]

(2) Plan distributionnel : En français, on constate que les voyelles se trouvent dans toutes les positions initiale, inter-consonantique et finale. Elles peuvent apparaître en contiguïté. Alors qu'en vietnamien, les voyelles se trouvent uniquement en deux positions, inter-consonantique et finale.

(3) Début vocalique : En français, un mot commence par un début vocalique arrondi. Tandis qu'en vietnamien, un mot (ou un syllabe) commence par une *attaque* brusque dite *coup de glotte*.

(4) Voyelles composées : En français, il existe des voyelles antérieures arrondies telles que /y/ (*i.e. **nu***), /ɸ/ (*i.e. **jeu***) et /œ/ (*i.e. **fleur***).

(5) Voyelles nasales : En français, il existe des voyelles nasales telles que /ɔ̃/ (*i.e. **montre***), /œ̃/ (*i.e. **brun***), /ɛ̃/ (*i.e. **brin***) et /ɑ̃/ (*i.e. **ancre***). En vietnamien, il n'en existe pas.

(6) Diphtongues : On constate qu'il existe des diphtongues uniquement en vietnamien tels que /iʌ/ (*i.e. ia, iê-* dans *« **chia** » (diviser), « **viên** » (boule)*), /ɯʌ/ (*i.e. ua-* dans *« **mua** » (vendre)*) et /uʌ/ (*i.e. uô-* dans *« **buôn** » (marchander)*).

(7) Semi-voyelles (ou semi-consonnes) : En français, il existe des semi-voyelles ou semi-consonnes telles que /j/ (*i.e. **ail***), /w/ (*i.e. **fouet***) et /ɥ/ (*i.e. **lui***). Tandis

qu'en vietnamien, il existe /j/, /w/ comme deux semi-voyelles. Mais /ɥ/ est inconnu.

Le Tableau 1.6 récapitule sept séries de points différents concernant les systèmes vocaliques du français et du vietnamien.

Tableau 1.6 Systèmes vocaliques du français et du vietnamien : Points différents. *Source : Xuan (2002)*

	En français	**En vietnamien**
1. Autres critères de classement	- la nasalisation - la labialisation	- la stabilité du timbre - la durée des voyelles
2. Plan distributionnel	- toutes les positions : initiale, interconsonantique, finale. - peuvent apparaître en contiguïté, ex : « aéroport »	- deux positions : inter-consonantique et finale. Jamais en position initiale. - inexistantes
3. Début vocalique	- arrondi	- commence toujours par une attaque brusque dite « coup de glotte » ou occlusion glottale, notée [ʔ]
4. Voyelles composées	- / y /, / ø /, / œ /	- inexistantes
5. Voyelles nasales	- / ɔ̃ /, / œ̃ /, / ɛ̃ /, / ɑ̃ /	- inexistantes
6. Diphtongues	- inexistantes	- / iʌ /, / ɯʌ /, / uʌ /
7. Semi-voyelles (ou semi-consonnes)	- / j /, / ɯ /, / ɥ /	- / ɯ / et / j / (en position finale) - / ɥ / inconnu

1.3.2 Systèmes consonantiques

En ce qui concerne le système consonantique, le français comporte 20 consonnes, dont 9 consonnes occlusives, 11 consonnes constrictives parmi lesquelles se trouvent 3 semi-consonnes (voir Tableau 1.7). Alors que le système consonantique du vietnamien comporte 23 consonnes dont le coup de glotte et deux demi-consonnes /w, j/ (voir Tableau 1.8).

Tableau 1.7 Inventaire des consonnes en français. *Source : Kang* et al. *(2016)*

	Labiale	**Dentale**	**Post Alvéolaire**	**Palatale**	**Labio Palatale**	**Vélaire**	**Uvulaire**
Occlusive	p b	t d				k g	
Nasale	m	n		ɲ		(ŋ)	
Fricative	f v	s z	ʃ ʒ				ʀ
Approxi-mative	w	l		j	ɥ		

Tableau 1.8 Inventaire des consonnes en vietnamien. *Source : Kang* et al. *(2016)*

	Labiale	**Dentale/Alvéolaire**	**Palatale**	**Vélaire**	**Glottale**
Occlusive	(p) ɓ	t^h t ɗ	tç	k	ʔ
Nasale	m	n	ɲ	ŋ	
Fricative	f v	s z		x ɣ	h
Approximative	w	l (r)			

En ce qui concerne le système consonantique du français et du vietnamien, on constate des points communs (voir Tableau 1.9) ainsi que des points différents (voir Tableau 1.10).

- Les points communs sont basés sur les critères de classement suivants :

  (1) Le mode d'articulation : En français comme en vietnamien, on constate que les occlusives, caractérisées par une fermeture totale, s'opposent aux constrictives (*fricatives* ou continues) pour lesquelles le chenal buccal est resserré. On observe des consonnes tels que /b, d, g, p, t, k/ versus /f, s, ʃ, v, z, ʒ/ en français et /b, d, p, t/ versus /f, s, v, z/ en vietnamien.

  (2) Le lieu d'articulation : On constate les mêmes lieux d'articulation concernant les bilabiales, les labiodentales, les alvéolaires, les post-alvéolaires, les

Tableau 1.9 Systèmes consonantiques du français et du vietnamien : Points communs.

Source : Xuan (2002)

Criteres	En français	En vietnamien
1. Le mode d'articulation : Les occlusives (= explosives = momentanées) caractérisées par une fermeture totale : s'opposent aux constrictives (= fricatives = continues) pour lesquelles le chenal buccal est resserré :	/p/, /t/, /k/, /b/, /d/, /g/ ≻≺ /f/, /s/, /ʃ/, /v/, /z/, /ʒ/	/p/, /t/, /b/, /d/ ≻≺ /f/, /s/ /v/, /z/
2. Le lieu d'articulation Les bilabiales, les labiodentales Les alvéolaires, Les post-alvéolaires, les palatales, les vélaires	/p/, /b/, /m/, /f/ et /v/ /t/, /d/, /n/, /s/ et /z/ /ʃ/, /ʒ/, /ŋ/, /k/, /g/	/p/, /b/, /m/, /f/ et /v/ /t/, /d/, /n/, /s/ et /z/ /ʂ/, /ʐ/, /ŋ/, /k/, /ɣ/
3. La nasalite Selon que le voile du palais est relevé ou non, on oppose les orales aux nasales :	/b/, /d/, /g/ ≻≺ /m/, /n/, /ŋ/	/b/, /d/ ≻≺ /m/, /n/
4. La sonorite Selon que le cordes vocales vibrent ou non, on distingue : - les sonores (= voisées) - des sourdes (= non-voisées)	/b/, /d/, /v/, /z/, /g/ ≻≺ /p/, /t/, /f/, /s/, /k/	/b/, /d/, /v/, /z/, /ɣ/ ≻≺ /p/, /t/, /f/, /s/, /x/

palatales et les vélaires, *i.e* consonnes telles que /b, d, f, k, g, m, n, p, t, s, z, ʃ, ʒ, ŋ/ en français et /b, d, f, k, m, n, p, s, t, v, z, ŋ, ʂ, ʐ, ɣ/ en vietnamien.

(3) La nasalité : On distingue les orales des nasales en fonction du relèvement du voile du palais, *i.e.* consonnes telles que /b, d, g/ versus /m, n, ŋ/ en français et /b, d/ versus /m, n/ en vietnamien.

(4) La sonorité : On distingue les sonores des sourdes selon la vibration des cordes vocales, *i.e.* consonnes telles que /b, d, g, v, z/ versus /f, k, p, t, s/ en français et /b, d, v, z, ɣ/ versus /f, p, t, s, x/ en vietnamien.

- En ce qui concerne les points différents, on distingue en français comme en vietnamien sept séries comme suit :

Tableau 1.10 Systèmes consonantiques du français et du vietnamien : Points différents.

Source : Xuan (2002)

	En français	**En vietnamien**
1. Nombre	- 17	- 23
2. Plan distributionnel	- Toutes positions : initiale, intervocalique, finale	- Initiale, sauf /p/ - Seulement 6 en position finale : /p/, /t/, /k/, /m/, /n/, /ŋ /
3. Mode articulatoire	Les consonnes finales présentent toujours 2 phases : Explosion et implosion. Elles sont caractérisées par une détente vers la fin de leur réalisation.	- Toujours implosives
4. Consonnes occlusives sourdes	- /p/, /t/, /k/	Quatre de plus : - /tʰ/ : alvéodentale - /ʈ/ : post-alvéolaire - /c/ : palatale - /ʔ/ : glottale
5. Consonne fricative glottale	- Inexistante	- /h/
6. Groupes consonantiques	[pr], [br], [pl], [pr], [str], etc.	- Inexistants
7. Consonnes géminées (ou consonnes doubles)	- honnêteté [ɔnɛtte]	- Inexistantes

(1) Nombre de consonnes : Le français en comporte 17 tandis que le vietnamien en comporte 23.

(2) Plan distributionnel : En français, les consonnes se trouvent dans toutes les positions initiale, inter-vocalique et finale. Alors qu'en vietnamien, elles se trouvent à la position initiale et seulement à la position finale concernant 6 consonnes telles que /k, m, n, p, t, ŋ/.

(3) Mode articulatoire : En français, on constate que les consonnes finales se présentent toujours en deux phases : *explosion* et *implosion*. Alors qu'en vietnamien, on les trouve toujours *implosives*.

(4) Consonnes occlusives sourdes : En français, il existe trois consonnes telles que /k, p, t/. Mais en vietnamien, il en existe quatre de plus /t^h, ʈ, c, ʔ/.

(5) Consonne fricative glottale : Il existe uniquement la consonne /h/ en vietnamien.

(6) Groupes consonantiques : Il existe des groupes de consonnes telles que /br, pl, pr, str, *etc.*/ en français. En vietnamien, il n'en existe pas.

(7) Consonnes géminées (ou consonnes doubles) : Il existe des consonnes géminées telles que /ɔnɛtte/ (*i.e. honnêteté*) en français. En vietnamien, il n'en existe pas.

1.4 Traitement phonologique des emprunts

Cette section présente l'étude du traitement phonologique des emprunts, ainsi que les changements apportés aux consonnes et aux voyelles d'une langue à une autre. Le Tableau 1.11 récapitule et illustre des correspondances de voyelles entre le français et le vietnamien (voir la page 23). Le Tableau 1.12 récapitule et illustre des correspondances de consonnes entre le français et le vietnamien (voir la page 29).

La langue vietnamienne possède une orthographe phonétique, dans laquelle les mots d'emprunt provenant non seulement du chinois, mais aussi du français ont généralement été l'objet d'une assimilation morphophonologique et graphique. L'impact d'une langue sur une autre langue, appelé *interférence* par Martinet (1970), se manifeste à différents niveaux notamment lexical, syntaxique et sémantique. Dans cette section, la question d'interférence sera examinée sur le plan phonétique, ainsi que les traitements des voyelles et des consonnes de correspondance entre le français et le vietnamien. Elle consiste en la manière de percevoir et de reproduire les sons d'une langue. Ainsi, nous présentons, par la suite, la phonologie du français et celle du vietnamien. En ce qui

concerne le vietnamien, nous décidons de choisir les informations phonologiques qui sont basées sur le dialecte de Hanoï pour deux raisons principales suivantes :

- La langue vietnamienne n'a pas vraiment de prononciation standard. Il existe deux dialectes dominants, celui de Hanoï et celui de Saigon, qui possèdent 6 tons et 5 tons, respectivement.
- Le dialecte de Hanoï peut être considéré comme le standard du pays parce qu'il est utilisé dans les informations télévisées et radiodiffusées et qu'il est largement répandu, de façon générale (Thompson, 1987). Celui de Saigon est le plus parlé au Vietnam après celui de Hanoï. À noter que Ho Chi Minh ville (Saigon) est considérée comme la capitale économique du pays.

1.4.1 Traitement des voyelles

En ce qui concerne les voyelles orales, le français standard en comporte 7 d'avant /i,e, ɛ, a, y, ø, œ/, une centrale /ə/ et 4 d'arrière /u, o, ɔ, ɑ/. Les deux voyelles /i/ et /u/ sont directement associées à leurs voyelles correspondantes en vietnamien, mais toutes les autres voyelles présentent des variations ou nécessitent une explication.

- Règles :
  - ◇ /i/ > /i/ *e.g.* li**t**re *lít*
  - ◇ /u/ > /u/ *e.g.* co**u**pe *cúp*

Les voyelles avant françaises arrondies n'ont pas de voyelles correspondantes en vietnamien et /y/ est adaptée en /wi/, /u/ ou /i/. La forme /wi/ est la règle d'adaptation la plus courante en raison de sa forme qui conserve à la fois les traits phonétiques arrondi et frontal de /y/, tandis que /u/ et /i/ ne retiennent qu'une de ces deux traits.

- Règles :
  - ◇ /y/ > /wi/ *e.g.* b**u**s *buýt*
  - ◇ /y/ > /u/ *e.g.* vir**u**s *vi rút*

◇ /y/ > /i/ *e.g.* lég**u**me *lê ghim*

Les voyelles arrondies centrales /ə/, /ø/ et /œ/, d'autre part, sont adaptées en tant que /ɤ/.

- Règles :
  - ◇ /ə/ > /ɤ/ *e.g.* chemise *sơ mi*
  - ◇ /ø/ > /ɤ/ *e.g.* b**leu** *lơ*
  - ◇ /œ/ > /ɤ/ *e.g.* doct**eur** *đốc tơ*

En ce qui concerne les voyelles nasales, le français standard en comporte quatre /ɑ̃, ɛ̃, œ̃, ɔ̃/. /œ̃/ tend à disparaître au profit de /ɛ̃/. Ces voyelles nasales sont une des caractéristiques les plus originales du système vocalique du français. Puisqu'en vietnamien, il n'en existe pas, les voyelles nasales du français vont à l'encontre des contraintes phonologiques de la langue vietnamienne. Elles sont dénasalisées dans la plupart des cas et une consonne nasale est introduite. Concrètement, elles sont adaptées par une séquence d'une voyelle et une coda nasale dans la langue vietnamienne.

- Règles :
  - ◇ /ɑ̃/, /ɛ̃/, /ɔ̃/ > *une séquence d'une voyelle et une coda nasale*
    e.g. c**an**tine *căng tin*, satin *sa tanh*, savon *sà bong*

Lorsque la voyelle nasale est suivie d'une consonne labiale (/p/ ou /b/) dans l'entrée française, la nasale finale est réalisée sous la forme /m/. Lorsqu'il s'agit d'une consonne vélaire, l'adaptation est /ŋ/.

- Règles :
  - ◇ /ɛ̃/ > /m/ *e.g.* t**im**bre *tem*
  - ◇ /ɑ̃/ > /ŋ/ *e.g.* br**an**card *băng ca*

La voyelle, en français, ouverte d'arrière /ɑ/ n'existe pas en vietnamien. Elle est réinterprétée en /a/, comme suit :

- Règle :
  - ◇ /ɑ/ > /a/ *e.g.* câble *cáp,* gâteau *ga-tô*

Tableau 1.11 Résumé et exemples des correspondances de voyelles entre le français et le vietnamien. *Source : Kang* et al. *(2016)*

	En français		**En vietnamien**	
/i/ > /i/	*litre*	/litʁ/	*lít*	/litˊ/
/u/ > /u/	*coupe*	/kup/	*cúp*	/kupˊ/
/y/ > /wi/, /u/, /i/	*(auto)bus*	/bys/	*buýt*	/ɓwitˊ/
	virus	/viʁys/	*vi rút*	/viˀ zutˊ/
	légume	/legym/	*la ghim*	/laˀ ɣimˀ/
			lê ghim	/leˀ ɣimˀ/
/ə/, /ø/, /œ/ > /ɤ/	*bleu*	/blø/	*lơ*	/lɤˀ/
	docteur	/dɔktœʁ/	*đốc tơ*	/ɗokˊ tɤˀ/
	chemise	/ʃəmiz/	*sơ mi*	/ʂɤˀ miˀ/
/ɑ̃/ > /ãŋ/, /an/, /aŋ/	*satan*	/satɑ̃/	*xa tăng*	/saˀ tăŋˀ/
	cantine	/kɑ̃tin/	*căng tin*	/kăŋˀ tinˀ/
	brancard	/bʁɑ̃kaʁ/	*băng ca*	/ɓăŋˀ kaˀ/
> /ăm/ (before /p, b/)	*jambon*	/ʒɑ̃bɔ̃/	*giăm bông*	/ʐămˀ ɓoŋˀ/
			dăm bông	/ʐămˀ ɓoŋˀ/
			dăm bon	/ʐămˀ ɓɔnˀ/
/ɛ̃/ > /ɛŋ/ [ăŋ], /ɛːŋ/	*satin*	/satɛ̃/	*xa tanh, sa tanh*	/saˀ tɛŋˀ/ [saˀ tăŋˀ]
			xa teng	/saˀ tɛːŋˀ/
> /ɛm/ (before /p, b/)	*timbre*	/tɛ̃bʁ/	*tem*	/tɛmˀ/
/ɔ̃/ > /oŋ/, /ɔːŋ/, /ɔŋ/	*savon*	/savɔ̃/	*xà ông*	/saˬ oŋˀ/
			xà phòng	/saˬ fɔŋˬ/
			sà bong, xà bong	/saˬ ɓɔŋˀ/
			xà vông	/saˬ voŋˀ/
	pont	/pɔ̃/	*boong*	/ɓɔːŋˀ/
> /ɔm/, /om/ (before /p b/)	*trompette*	/tʁɔ̃pɛt/	*tom bét*	/tɔmˀ ɓɛtˊ/
			trôm pét	/comˀ pɛtˊ/
			trom pét	/cɔmˀ pɛtˊ/

1.4.2 Traitement des consonnes

En comparant le système consonantique du français, ayant seulement 20 consonnes, celui du vietnamien en comporte 23. En revanche, il n'existe pas de groupes consonan-

tiques en vietnamien. En plus, la langue vietnamienne ne connaît pas les sons français /ʃ, ʒ, ʁ, ɥ/.

En principe, la plupart des consonnes en français sont directement représentées par leurs correspondances de consonnes en vietnamien. Par contre, il existe beaucoup de variantes de prononciation des mots d'emprunt du français parce que les locuteurs vietnamiens ont imité la prononciation française. Concrètement, nous allons détailler une étude de cas du traitement des consonnes du français dans la langue vietnamienne, selon l'inventaire consonantique du français (voir Tableau 1.7) :

(a) Consonnes labiales :

Le /p/ final français est maintenu en vietnamien. En règle générale, la consonne *p* ne se rencontre normalement qu'en position finale en vietnamien. Par conséquent, elle s'adapte par la consonne voisée correspondante /ɓ-/ en position initiale. Les groupes consonantiques, tels que *pl*, ne sont pas tolérés en vietnamien. Il y a une réduction, d'où la règle de conversion phonétique /p/ > /ø/.

La règle /p/ > /pɓ/ se base sur la resyllabification : /p/ est compatible avec la position en finale de mot. /ɓ/ est compatible avec la position en initiale. Donc la séquence /pɓ/ serait la réfraction de /p/ à la frontière entre les deux syllabes.

◇ Règles :

⋆ /p/ > /p-/	*e.g.* chalou**pe**	*xà lúp*
⋆ /p/ > /ɓ-/	*e.g.* **p**agaille	***b**a gai*
⋆ /p/ > /ø/	*e.g.* **pl**astique	*lát tich*
⋆ /p/ > /pɓ/	*e.g.* pou**pe**e	***búp bê***

Pareillement, si la consonne *b* ne se trouve pas en position initiale, elle est adaptée en *p* , d'où la règle /-b/ > /-p/. Si elle fait partie d'un groupe consonantique, tel que *bl* ou *br*, qui n'est pas toléré en vietnamien, elle est également adaptée en *p*.

En ce qui concerne *chambre* et *timbre*, il y a une dénasalisation et une introduction de la consomme /m/. Ainsi, il n'y a plus de place pour d'autres consonnes parce qu'en vietnamien, la structure CVCC, avec C désigne consonne et V désigne voyelle, n'est pas concevable. Donc on aura la structure CVC, *e.g. săm*, *tem*, au lieu de *săm*p, *temp* (n'existant pas en vietnamien).

◇ Règles :

⋆ /-b/ > /-p/ *e.g.* ab**c**ès ***áp** xe*

⋆ /bl-/ > /p-/ *e.g.* câ**bl**e ***cáp***

⋆ /br-/ > /p-/ *e.g.* li**br**e ***líp***

(b) Semi-consonne labiale :

La semi-consonne /w/ n'apparaît ni devant les voyelles arrondies /u/, /o/, /ɔ/ (car leur timbre est bas), ni devant la voyelle fermée /ɯ/.

Nous avons une exception de règle de conversion phonétique, /w/ > /ø/, *e.g.* p**oi**ds-lourd > ***ba**-lua*. Il s'agit donc d'une opposition phonologique parce qu'elle permet d'opérer la distinction lexicale entre *poids* et *boire*.

◇ Règles :

⋆ /w/ > /w/ *e.g.* pourb**oi**re ***boa***

⋆ /w/ > /ø/ *e.g.* p**oi**ds-lourd ***ba**-lua*

(c) Consonnes labiodentales :

Le système phonologique du vietnamien ne tolère le phonème /f/ qu'en position initiale, d'où la règle /f/ > /f-/. Sinon, en position finale de syllabe, il est remplacé par une occlusive orale bilabiale, donc /p/.

En ce qui concerne la consonne *v*, en principe, elle est prononcée /v/ en vietnamien. Par contre, il y a quelques exceptions, notamment /v/ > /f-/ ou /v/ > /b-/. Il y a une assimilation progressive avec une désonorisation du /v/ : /v/ est devenu sourd et interprété comme /f/ sous l'influence de la consonne sourde /s/.

Les variantes de prononciation de savon en vietnamien correspondraient à des variations dialectales. La forme *xà **bông*** est employée dans le sud du Vietnam. Il n'y a pas de désonorisation. Le voisement de la consonne est maintenu, mais son articulation est passée à occlusive bilabiale /b/.

◇ Règles :

⋆ /f/ > /f-/ *e.g.* café *cà **phê***

⋆ /f/ > /p/ *e.g.* chef *sếp*

⋆ /v/ > /f-/ *e.g.* savon *xà **phòng***

⋆ /v/ > /b-/ *e.g.* savon *xà **bông***

(d) Consonnes dentales :

Si la consonne *t* se trouve en groupes consonantique en français, tels que /tl/, /tʀ/, /st/, elle est réinterprété en /k/ en vietnamien. Il y a d'autres exceptions, dans la même règle /-t/ > /-k/, par exemple, car**te** > cá**c**, contac**t** > *công tắc*. Le /ʀ/ est tombé dans le cas de *carte* parce que la forme de CVʀ n'est pas tolérée en vietnamien. Le /-t/ est devenu /-k/ sous l'influence du /k-/ avant, il y a eu une assimilation de gauche à droite. Dans le cas de *contact*, le /-t/ en finale est tombé en raison de la forme de CVCC qui n'est pas admise en vietnamien. Donc c'est la consonne *t* est en trop.

En ce qui concerne la règle de conversion phonétique /-d/ > /-t/, /d/ n'est réalisée qu'en initiale en vietnamien, par exemple, bi**d**on > *bi **đ**ông* ou docteur > *đốc tơ*. Donc il y a une désonorisation et une distribution automatique du phone /t/ qui peut être réalisée en finale et en initiale.

◇ Règles :

⋆ /-t/ > /-k/ *e.g.* ma**te**lot *mạ**ch** lô*, fil**tre** *phí**ch***, turfi**ste** *tuyê**t** phí**ch***

⋆ /-d/ > /-t/ *e.g.* aci**de** *a xí**t***

(e) Consonnes alvéolaires :

La consonne *s* n'est tolérée qu'en position initiale en vietnamien. Elle est représentée en /t/ en finale, en vietnamien. De même, le système phonologique vietnamien ne tolère pas /z/ en finale absolue, d'où la règle /z/ > /ø/.

◇ Règles :

⋆ /-s/ > /-t/ *e.g.* as *át*, bus *buýt*

⋆ /z/ > /ø/ *e.g.* ga**z** *ga*, vali**se** *va li*

(f) Consonnes post-alvéolaires :

/ʃ/, n'existant pas en vietnamien, est interprété en /s/. De même, /ʒ/, n'existant pas en vietnamien, est interprété en /z/ ou tend à disparaître de la forme vietnamisée s'il se trouve en finale.

◇ Règles :

⋆ /ʃ/ > /s/ *e.g.* **ch**aland *sà lan*, **ch**aîne *sên*

⋆ /ʒ/ > /z/ *e.g.* **j**ambon ***d**ăm bông*

⋆ /ʒ/ > /ø/ *e.g.* solfè**ge** *sôn phe*

(g) Consonnes palatales :

En français, à cause de la semi-voyelle /j/, la voyelle qui suit /j/ est plus ouverte. Le groupe /jV/ devient la diphtongue /ie/ après la vietnamisation. En ce qui concerne /ɲ/, en vietnamien, en fin de syllabe, la nasale n'est pas relâchée.

◇ Règles :

⋆ /jV/ > diphtongue /ie/ *e.g.* bi**è**re *bia*

⋆ /ɲ/ > /ŋ/ *e.g.* champa**gne** *sâm ba**nh***

(h) Le *h* aspiré :

En français, la consonne *h* demeure muette. Par contre, en vietnamien, elle est toujours prononcée. Il y a donc une réactivation.

◇ Règle :

⋆ /ø/ > /h/ *e.g.* **h**ydrogène ***h**ýt rô*

(i) Consonne labio-palatale : La sonante labio-palatale /ɥ/ n'existe pas en vietnamien. L'articulation très fermée de la voyelle antérieure arrondie /y/ donne un /ɥ/ en français. Or, la voyelle /y/ n'existe pas en vietnamien, elle devient en générale /wi/ en syllabe ouverte ou fermée, ou bien /u/ (voir la sous-section 1.4.1).

◇ Règles :

⋆ /ɥ/ > /u/ *e.g.* c*ui*llère *c**ù**i dìa*

⋆ /ɥ/ > /ø/ *e.g.* aig*ui*lle *ghi*

(j) Consonnes vélaires :

En ce qui concerne le /k/, en initiale en français, il est directement représenté en vietnamien. S'il se trouve en finale, il se peut qu'il est interprété soit en /ø/ soit en /t/. Dans le premier cas, il y a une dénasalisation et une introduction de la consonne /ŋ/. Dans le deuxième cas, il y a une assimilation de gauche à droite. Le /k/ est devenu /t/ sous l'influence du /t/ qui précède.

En ce qui concerne le /g/, en groupe consonantique en français, il est tombé et la consonne qui suit a été maintenue en initiale en vietnamien. Et le /ŋ/ a été interprété comme /ŋ/, non relâché, en vietnamien.

◇ Règles :

⋆ /k/ > /ø/ *e.g.* **h**ydrogène ***h**ýt rô*

⋆ /k/ > /t/ *e.g.* biftec**k** *bíp tết*

⋆ /g/ > /ø/ *e.g.* **g**laïeul *lay ơn*

⋆ /ŋ/ > /ŋ/ *e.g.* dancin**g** *đăng xin**h***

(k) Consonnes uvulaires :

En français, le /ʀ/ est grasseyé, mais il n'existe pas en vietnamien. C'est peut-être

la raison pour laquelle il est tombé, /ʀ/ > /ø/, lors de la vietnamisation des mots d'emprunt. En revanche, il existe quelques exceptions. Il se peut être réinterprété en /k/, par exemple *course*, ou en /t/, par exemple *verni*. Dans le cas de *course*, il y a une assimilation sous l'influence de l'occlusive vélaire initiale /k/. Dans le cas de *verni*, il y a un ajout d'un /t/ épenthétique à la frontière des mots.

◇ Règles :

⋆ /ʀ/ > /k/ *e.g.* course *cuốc*

⋆ /ʀ/ > /t/ *e.g.* verni *vẹt ni*

Tableau 1.12 Résumé et exemples des correspondances de consonnes entre le français et le vietnamien. *Source : Huynh (2010)*

	En français	**En vietnamien**	
/p/ > /p-/	chaloupe	xà lúp	/salup-/
/p/ > /ɓ-/	pagaille	ba gai	/ɓagai/
/p/ > /ø/	plastique	lát tich	/lat-tic/
/p/ > /pɓ/	poupée	búp bê	/ɓup- ɓe/
/-b/ > /-p/	abcès	áp xe	/apsɛ/
/bl-/, /br-/ > /p-/	câble, libre	cáp, líp	/kap/, /lip-/
/w/ > /w/	pourboire	boa	/bwa/
/f/ > /f-/ ou /p-/	café ou chef	cà phê ou sếp	/kafe/ ou /sep-/
/v/ > /f-/ ou /b-/	savon	xà phòng ou xà bông	/safɔŋ/ ou /saɓoŋ/
/-t/ > /-k/	carte, contact	các, công tắc	/kak-/, /koŋm-tak-/
/-d/ > /-t/	acide	a xít	/asit-/
/ɥ/ > /u/ ou /ø/	cuillère ou aiguille	cùi dìa ou ghi	/kuijie/ ou /ɣi/
/ʀ/ > /k/ ou /t/	course ou verni	cuốc ou vẹt ni	/kuok/ ou /vetni/

En somme, les principes généraux reliés au traitement phonologique des emprunts se basent sur les restrictions phonotactiques dans la langue cible ainsi que des comparai-

sons de la similarité phonétique entre deux langues. Nous avons constaté divers phénomènes phonétiques. Les linguistes vietnamiens semblent s'appuyer sur leur connaissance de la phonologie et de la tendance phonotactique du français, à savoir la *Loi de Position* (Féry, 2003; Storme, 2015), afin de parvenir à l'adaptation des emprunts. Le contraste de la phonologie du français (/ʀ/ versus /k/) est neutralisé en raison d'une restriction phonologique du vietnamien (c'est-à-dire, aucun /ʀ/ en *coda* vietnamien), mais l'adaptation vietnamienne conserve systématiquement le contraste dans le texte, la différence de qualité et de longueur dans la voyelle précédente. Ces analyses ont illustré la complexité du processus d'adaptation des emprunts, dans lequel divers facteurs, notamment les restrictions phonologiques de la langue cible, la similarité phonétique et les connaissances phonologiques de la langue source, interagissent pour influer l'adaptation.

1.5 Réflexion sociolinguistique sur les emprunts directs du français dans la langue vietnamienne

1.5.1 Contexte historique

Tout au long de l'histoire du Vietnam, la langue vietnamienne a évolué. Dans cette partie, nous abordons l'évolution du vietnamien sous les aspects socio-historiques et sociolinguistiques dans différentes classes sociales et notamment par les emprunts directs du français dans la langue vietnamienne.

Tout d'abord, il est important de rappeler qu'à travers l'histoire de la langue vietnamienne, il existe toujours un phénomène indéniable d'emprunt de mots à la langue chinoise, ensuite à la langue française et à la langue anglaise. Le chinois a joué un rôle crucial, pour la langue vietnamienne, similaire à celui du grec et du latin pour les langues européennes (DeFrancis, 1977). L'apport du chinois consiste en un vocabulaire des sciences humaines, de la politique et de la philosophie. Alors que le français

a donné beaucoup de termes dans le domaine des sciences, des techniques, des noms propres historiques et de la géographie.

Dans ce contexte historique, Minh (2011) a résumé trois périodes importantes comme suit :

(1) Période 1 : Jusqu'en 1858, le contact linguistique entre les communautés française et vietnamienne est, sans conteste, l'œuvre de la colonisation française. Dans l'objectif de remplacer la langue officielle par la langue français, le gouvernement colonial a cherché à évincer la langue chinoise écrite, la langue des lettrés, en la remplaçant par le vietnamien « *quốc ngữ* » dans l'enseignement du français. On constate que la première génération des emprunts du français a touché, en grande partie, le statut social des emprunteurs, notamment les coolies, les porteurs, les soldats mercenaires, les valets de chambre, les filles de joie, etc. Ces collaborateurs, recrutés par les Français d'autrefois sur le territoire vietnamien, étaient au bas de l'échelle sociale, d'où des mots d'emprunts dans l'armée et dans les services. *Par exemple : ca-pô-ran (caporal), ca-pi-ten (capitaine), ma-nớp / man-nớp (manœuvres), la mát / la mác (la marche), ba-tai-dông (bataillon), etc.*

(2) Période 2 : Entre 1858 et 1884, avec sa présence imposée sur l'ensemble du territoire du Vietnam, la France a créé une multitude de situations linguistiques entre trois régions : en Cochinchine (Sud), au Tonkin (Nord) et en Annam (Centre). Cela rajoutait une complexité extraordinaire pour de nombreux Vietnamiens vivant sous différents régimes. On observe que les emprunts directs du français ont été non seulement confinés dans l'armée, mais aussi dans la société, voire dans la vie quotidienne. *Par exemple : bót / bốt (poste), sà lim / xà lim / xà linh / xen luyn (cellule), xa vông / xà vồng / xà phòng / xà bông (savon), tuy-nen / tu-nền / tu-nen (tunnel), etc.*

(3) Période 3 : De la fin du XIX^{ème} siècle au début du XX^{ème} siècle, on constate une croissance des emprunts directs du français dans la plupart des domaines, notamment :

- ◇ les nouvelles industries, leurs équipements et leurs produits, *par exemple : ê-kíp / kíp (équipe), ba-lông (ballon), cao su / cao xu (caoutchouc)*
- ◇ la gastronomie, *par exemple : bơ (beurre), pho mai / phó mát / phô mai / phó mách (fromage), dăm bông / ram bông (jambon), ba tê (pâté)*
- ◇ la mode, *par exemple : ca-vát / cà vạt / ca-ra-vát / cà vát / cà là hoách / cà ra hoách/ ca la oách (cravate), voan (voile), măng tô (manteau)*
- ◇ le luxe, *par exemple : cô lôn / ô-đờ-cô-lôn (eau de Cologne), xì gà (cigare), đi văng / đi quăng (divan)*
- ◇ les nouveaux services, *par exemple : băng (banque), séc (chèque), tem (timbre), tê-lê-phôn / tô-lô-phôn / phôn (téléphone), buya rô / bu rô / buy rô (bureau), đốc / đốc tơ / đốc tờ / đốc tưa (docteur), công cua (concours)*
- ◇ les sports européens, *par exemple : ten nít / tơ-nít (tennis), banh (balle), bi a / bi da (billard)*
- ◇ les arts, *par exemple : xi nê ma / xi nê / xi la ma (cinéma), phim (film), bồ gam (programme), xen (scène)*
- ◇ les moyens de transport, *par exemple : la ga / ga (gare), sếp ga / xếp ga (chef de gare), sếp tanh / xếp tanh (chef de train), ray / rầy (rail), sốp phơ / sớp phơ (chauffeur), ô tô buýt / buýt (autobus), tích kê (ticket)*

1.5.2 Problématiques

Parmi les multiples travaux de recherche au sujet des emprunts de mots du français dans la langue vietnamienne, nous pouvons en citer quelques uns très pertinents. Phuong (1973) a étudié, dans sa thèse, trois facteurs clés qui ont contribué à l'expansion et au

maintien des termes empruntés du français : (1) le snobisme linguistique des intellectuels et des citadins, (2) l'habitude de l'usage des termes français appris à l'école ou utilisés au travail et (3) l'imitation de la mode européenne. Toan (1992) a abordé, dans son livre, les conditions historico-sociales et la colonisation française au Vietnam ainsi que les besoins de communication et d'échanges culturels ayant pu entraîner les emprunts du français en vietnamien. Huynh (2010) a présenté une description approfondie de la façon dont les mots français ont été vietnamisés depuis l'époque de la colonisation française jusqu'à présent, en vietnamien moderne. Minh (2011) a présenté une réflexion sociolinguistique sur les emprunts directs du français et a construit un dictionnaire des mots communs d'emprunt du français dans la langue vietnamienne.

Les emprunts de mots du français se sont bel et bien intégrés dans la langue vietnamienne. Par contre, ces mots empruntés du français se voient menacés en raison de leurs consonances étrangères (*i.e. croa xăng / croát xăng* pour *croissant*), surtout quand ils se trouvent en conflit synonymique avec d'autres emprunts plus réussis (*i.e. quát xăng*) et de nouvelles créations purement vietnamiennes (*i.e. bánh sừng bò*).

Modard et Vignes (2006) ont démontré l'insécurité linguistique pour une longue suite de parfaits synonymes ainsi que les mécanismes concernant une adaptation phonologique (*i.e. xa vông / xà vồng / xà phòng / xà bông* pour *savon*). Par ailleurs, il faut chercher des explications à ce phénomène, en dehors de considérations purement linguistiques, dans les rapports entre la langue et les usagers (*i.e. politiciens, docteurs, ingénieurs, gens riches et cultivés* versus *soldats, gens pauvres et non-cultivés*), entre la langue et l'usage (*i.e. au sein de l'université, de l'école* versus *dans l'armée, à la maison, etc.*).

On peut constater des causes susceptibles de nous éclairer sur ces faits :

- Il n'existait pas de règles linguistiques consistantes pour faire face à la diversité des synonymies des mots empruntés du français dans la langue vietnamienne.

- Les variantes ne sont ni formalisées, ni normalisées. Les nouveaux intellectuels, formés à l'école française, se plaisaient à recourir aux emprunts du français au détriment des mots vietnamiens.
- Une des causes importantes demeure une société fortement hiérarchisée au Vietnam à l'époque (fin du XIX^{ème} siècle jusqu'au début du XX^{ème} siècle). L'apprentissage de la langue et l'école étaient uniquement réservés aux familles de l'élite et aux riches.
- Une autre cause consiste en la complexité de transcription. Certaines combinaisons de sons/graphèmes sont moins compétitives que d'autres dans le conflit synonymique.
  - ◇ Les mots pluri-syllabiques tendent très souvent à se réduire à une seule syllabe. Par exemple : *băng-rôn* au détriment de *băng-đơ-rôn* / *băng-đờ-rôn* (*banderole*).
  - ◇ Les groupes de consonnes complexes et longs sont remplacés au profit de consonnes simples et courtes. Par exemple : *phi-bô xi-măng* au détriment de *phi-brô xi-măng* (*fibrociment*).
  - ◇ Il existe des formes inexistantes en vietnamien. Par exemple : *tuy-nen (tunnel)* dont la seconde syllabe de cette forme « *nen* » n'existe nulle part ailleurs en vietnamien.

1.6 Apprentissage automatique

1.6.1 Champs aléatoires conditionnels

L'algorithme de champs aléatoires conditionnels (en anglais : *Conditional Random Fields* ou CRFs), proposé par Lafferty *et al.* (2001), n'impose aucune hypothèse d'indépendance parmi les données observées et modélise directement la probabilité de vraisemblance d'une hypothèse d'annotation discriminatoire. Cet algorithme modélise éga-

lement la compatibilité de contiguïté entre les classes des entités nommées qui peuvent capturer des contraintes fortes, comme une classe « I-classe » ne suivant pas une classe O (comme *Out*, hors de classe) sans une classe « B-classe ». Dans notre modèle de reconnaissance des entités nommées, les CRFs sont formulés comme suit.

Pour une séquence de mots $X = (x_1, x_2, ..., x_N)$, soit E une matrice de dimension (k, N) où k est le nombre de classes uniques et N est la longueur de la séquence. Soit T la matrice de transition carrée de taille $(k+2, k+2)$ qui capture le score de transition entre les k classes, le symbole de début et le symbole de fin. Soit $Y = (y_1, y_2, ..., y_N)$ la séquence de classes associée à la séquence de mots d'entrée, chaque y_i constituant un index dans l'ensemble ordonné de classes uniques. Soit y_0 le symbole de début et y_{N+1} le symbole de fin. Le score de cette séquence est évalué comme suit :

$$S(X,Y) = \sum_{i=1}^{N} E_{y_i,i} + \sum_{j=0}^{N} T_{y_j,y_{j+1}} \tag{1.2}$$

Soit χ indiquant l'espace exponentiel de toutes les classes possibles de cette séquence X. La fonction de partition associée est ensuite évaluée comme suit :

$$Z = \sum_{Y \in \chi} e^{S(X,Y)} \tag{1.3}$$

La probabilité d'un étiquetage spécifique $Y \in \chi$:

$$P(Y|X) = \frac{e^{S(X,Y)}}{Z} \tag{1.4}$$

L'objectif de la phase d'apprentissage consiste à maximiser la probabilité logarithmique conditionnelle de la séquence d'étiquetage correcte Y :

$$log(P(Y^*|X)) = S(X,Y^*) - log(Z) \tag{1.5}$$

Le critère de décodage pour une séquence d'entrée X est :

$$Y^* = \underset{Y \in \chi}{argmax}\, S(X,Y) \tag{1.6}$$

Normalement, l'évaluation de la fonction de partition sur l'espace exponentiel de toutes les classes possibles serait impossible. Cependant, comme décrit dans (Lafferty *et al.*, 2001), cette évaluation peut être efficacement effectuée en appliquant les CRFs de chaîne linéaire à l'aide de l'algorithme de retour en arrière (Della Pietra *et al.*, 1997).

1.6.2 Séparateurs à vaste marge

Les séparateurs à vaste marge (en anglais : *Support Vector Machines* ou SVMs) (Cortes et Vapnik, 1995), faisant partie des méthodes d'apprentissage automatique puissantes, ont été appliqués, avec succès, dans diverses tâches du traitement automatique des langues naturelles. L'objectif de l'algorithme de SVM consiste à trouver un hyperplan dans un espace à N dimensions (N - le nombre de caractéristiques) qui classe distinctement les points de données.

Le SVM construit un classifieur binaire qui produit $+1$ ou -1 en sortie étant donné un vecteur d'exemple $x \in \mathbb{R}^n$. La décision est basée sur l'hyperplan séparé comme suit :

$$c(x) = \begin{cases} +1 & si\ (w \cdot x + b) > 0, w \in \mathbb{R}^n, b \in \mathbb{R} \\ -1 & autre \end{cases} \tag{1.7}$$

Étant donné un ensemble d'échantillons d'apprentissage étiquetés $(x_1, y_1), ..., (x_i, y_i)$, $x_i \in \mathbb{R}^n$, $y_i \in \{+1, -1\}$, le SVM essaie de trouver l'hyperplan optimal, c'est-à-dire l'hyperplan avec la marge maximale. La marge est définie comme la distance entre l'hyperplan et les échantillons d'apprentissage les plus proches de l'hyperplan. Pour maximiser la marge, ces échantillons les plus proches, appelés les vecteurs de support, se trouvent des deux côtés de l'hyperplan de séparation et que celui-ci se situe exactement au milieu de ces vecteurs de support. Cette maximisation de la marge est étroitement liée à la puissance de généralisation des SVMs.

En supposant que $|w \cdot x + b| = 1$ au niveau des vecteurs de support, sans perte de généralité, l'apprentissage du SVM peut être formulé comme le problème d'optimisation suivant.

- minimiser $\frac{1}{2} \parallel w \parallel^2$
- sujet à

$$y_i(w \cdot x_i + b) \geq 1, i = \{1, ..., N\} \tag{1.8}$$

La solution de ce problème, en utilisant seulement des vecteurs de support SV et des poids α_i correspondants, est reformulée comme suit :

$$f(x) = w \cdot x + b = \sum_{i \in SV} y_i \alpha_i x \cdot x_i + b \tag{1.9}$$

où le produit interne $(x \cdot x_i)$ est appelée une fonction de noyau.

Afin de créer un système de reconnaissance des entités nommées, nous avons besoin d'un classifieur SVM à multi-classes. Parmi plusieurs méthodes pour construire un SVM à multi-classes, nous décrivons ici la méthode d'un-contre-tout (en anglais : *one-versus-rest*) et la méthode de paire mutuelle (en anglais : *pairwise*). Les deux méthodes construisent un classifieur à classes multiples en combinant plusieurs SVM binaires.

(1) Il s'agit de construire K classifieurs SVMs binaires. Chacun détermine si l'échantillon doit être classé dans la classe i ou dans les autres. Le résultat de sortie est la classe ayant la valeur maximal de $f(x)$ dans l'équation 1.9.

(2) Il s'agit de construire $K(K-1)/2$ classifieurs SVMs binaires. Chacun détermine si l'échantillon doit être classé dans la classe i ou dans la classe j. Chaque SVM binaire obtient une vote et le résultat de sortie consiste en la classe ayant le nombre maximal de votes.

1.6.3 Plongements de mots

Inspiré par l'hypothèse distribuée (Harris, 1954; Firth, 1957) et les modèles de langue neuronaux, Mikolov *et al.* (2013) a proposé *word2vec*, un algorithme simple et efficace pour déduire des représentations vectorielles denses de mots à partir d'un grand corpus non étiqueté. Il s'agît d'un modèle à base des réseaux de neurones peu profond qui apprend les représentations de mots en optimisant la fonction objective impliquant à la fois un mot cible et un mot de contexte. *Word2vec* construit le vocabulaire à partir du corpus et apprend les représentations de mot en formant un réseau de neurones à trois couches. *Word2vec* propose deux modèles, à savoir le sac de mots continus (en anglais : *Continuous Bag Of Words* ou *CBOW*) et le skip-gramme.

(1) Le modèle de *CBOW* apprend les représentations en prédisant le mot cible en fonction de ses mots de contexte. Le modèle de *CBOW* est constitué de trois couches, à savoir la couche d'entrée, la couche cachée et la couche de sortie (voir Figure 1.2). Les couches sont reliées par deux matrices de poids W et W'. La couche d'entrée prend les vecteurs *one-hot* (dont la représentation binaire n'a qu'un seul chiffre 1) des mots de contexte en entrée et la couche de sortie appliquant la fonction de *softmax* prédit le vecteur *one-hot* du mot cible. La fonction objective du modèle CBOW est comme suit :

$$J = \frac{1}{V}\sum_{i=1}^{V} logp(w_i|w_{i-n},...,w_{i-2},w_{i-1},w_{i+1},w_{i+2},...,w_{i+n}) \tag{1.10}$$

où V est la taille du vocabulaire et n est la taille du fenêtre contextuelle.

(2) Le modèle de skip-gramme fonctionne exactement à l'opposé du modèle de *CBOW*. Il apprend les plongements en prédisant les mots du contexte en fonction du mot cible (voir Figure 1.2). La fonction objective du modèle de skip-gramme

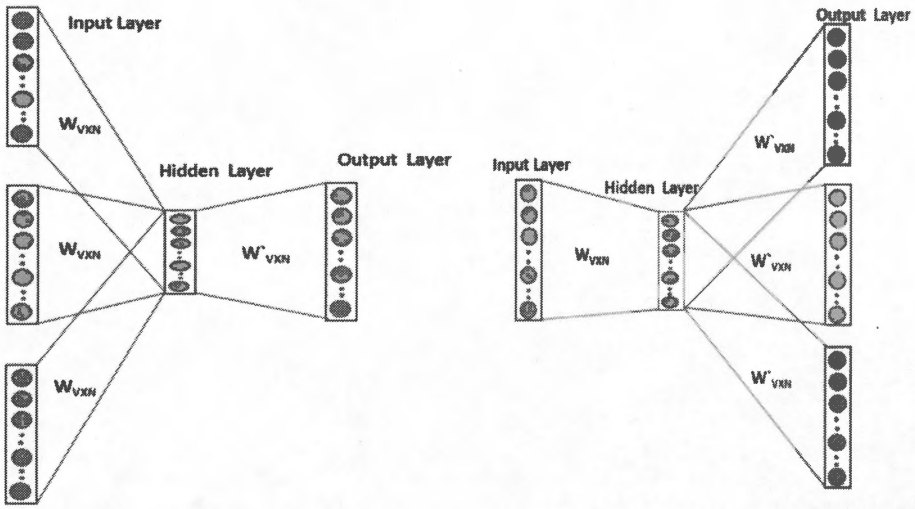

Figure 1.2 Modèle de sac de mots continus (CBOW) et modèle de Skip-gramme. Source : Mikolov *et al.* (2013).

est comme suit :

$$J = \frac{1}{V}\sum_{i=1}^{V}\sum_{-n\leq j\leq n,\neq 0}^{V} logp(w_{i+j}|w_j) \tag{1.11}$$

où V est la taille du vocabulaire et n est la taille du fenêtre contextuelle.

Ces deux modèles sont des modèles log-linéaires et la phase d'apprentissage se compose en deux étapes. La principale différence réside dans la fonction de perte utilisée pour mettre à jour le modèle. Par ailleurs, dans les deux modèles, l'apprentissage des vecteurs de sortie est coûteux en calcul. Pour résoudre ce problème, Mikolov *et al.* (2013) a proposé deux méthodes, à savoir l'échantillonnage négatif (en anglais : *negative sampling*) et les couches de *softmax* hiérarchiques. L'échantillonnage négatif fonctionne bien avec les mots peu fréquents, tandis que les couches de *softmax* hiérarchiques fonctionnent bien avec les mots fréquents.

1.6.4 Réseaux de neurones récurrents

Dans les tâches du TALN, les données d'entrée sont généralement séquentielles. Les réseaux de neurones artificiels traditionnels, avec des entrées de taille fixe, ne conviennent pas pour ces tâches. Les réseaux de neurones récurrents (*en anglais : Recurrent Neural Network ou RNN*) sont une sorte de réseaux de neurones qui sont conçus pour fonctionner sur des séquences de données. Ils possèdent un mécanisme de gestion de mémoire qui peut capter l'information contextuelle des états calculés antérieurement.

En théorie, les RNN peuvent mémoriser, de manière arbitraire, le contexte de longues séquences. Cependant, en pratique, ils ne comprennent que quelques étapes en arrière (Bengio *et al.*, 1994). Une architecture de RNN couramment utilisée, dite mémoire à long-court terme (*en anglais : Long-Short Term Memory ou LSTM*), qui fournit une autre manière de calculer les états cachés, est capable de mieux capturer le contexte de longues séquences.

Les architectures de LSTM ont été conçus pour incorporer une cellule de mémoire qui peut protéger et contrôler l'état de la cellule. Ils utilisent plusieurs portes pour contrôler la quantité d'informations des états précédents, qui doivent être oubliés, et les informations d'entrées, qui doivent être mises à jour dans la cellule de mémoire (Hochreiter et Schmidhuber, 1997).

Les architectures de LSTM, ainsi que celles de RNN, prennent comme entrée une séquence de vecteurs $(x_1, x_2, ..., x_n)$ et produisent comme sortie une séquence de vecteurs $(h_1, h_2, ..., h_n)$ pour représenter les informations à chaque étape d'entrée. Formellement, la suite d'équations suivantes décrit comment les blocs de mémoire sont mis à jour à

chaque étape t :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{1.12}$$

$$c_t = (1 - i_t) \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{1.13}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{1.14}$$

$$h_t = o_t \odot tanh(c_t) \tag{1.15}$$

où les matrices de W_k sont les matrices de poids entre deux couches cachées consécutives, et entre les couches d'entrée et les couches cachées; b_k est le vecteur de biais correspondant. k est le port (*en anglais : gate*) concernant l'ensemble d'entrée, de sortie, d'oubli et de cellule de mémoire. σ est la fonction sigmoïde par élément, et $\odot$ est le produit par élément de deux vecteurs. c_t et o_t sont l'état de la cellule et la sortie à l'étape t, respectivement.

1.6.5 Mécanisme d'attention

Soit $w_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ indiquant le vecteur du mot bi-directionnel LSTM en sortie (de dimension d_1) à l'étape t correspondant au $t^{ème}$ mot (X_t) dans la séquence d'entrée X. Soit M_t étant la matrice contenant les sorties en caractère bi-directionnel LSTM pour chaque caractère de X_t. Cette matrice a des dimensions (l_t, d_2) où d_2 est la dimension des couches cachées du caractère bi-directionnel LSTM et l_t est le nombre de caractères dans X_t. Soit m_t^i dénote la $t^{ème}$ rangée de la matrice M_t. Soit P étant une matrice de paramètres de dimension (d_1, d_2) et p étant un vecteur de biais de longueur d_2. Nous suivons (Bahdanau *et al.*, 2014) dans la formulation du vecteur de contexte d'attention

a_t :

$$w_t^{'} = tanh(w_t \cdot P + p) \tag{1.16}$$

$$\alpha_i = \frac{exp(w_t^{'} \cdot m_t^i)}{\sum_{j=1}^{l_t} exp(w_t^{'} \cdot m_t^j)} \tag{1.17}$$

$$a_t = \sum_{i=1}^{l_t} exp(\alpha_i * m_t^i) \tag{1.18}$$

Il est à noter que chaque mot possède un vecteur de contexte d'attention distinct obtenu à l'aide des états cachés de caractère LSTM générés par ses constituants au niveau de caractères.

CHAPITRE II

ÉTAT DE L'ART

2.1 Noms propres

Cette section présente les différentes classifications de noms propres, du point de vue du traitement automatique des langues naturelles (TALN). De nombreux travaux de recherches traitent les noms propres étant comme les entités nommées en TALN, spécifiquement dans le domaine de la reconnaissance des entités nommées. Ce domaine a un grand impact pour diverses applications du TALN telles que la traduction automatique, l'extraction de l'information interlangue, la recherche de l'information, etc. Nous détaillons davantage différentes approches de ce domaine dans la section 2.2.

La notion d'entités nommées couvre non seulement les noms propres, mais aussi des entités plus complexes comme les expressions multimots. Ehrmann (2008) définit « entité nommée », étant donnée un modèle applicatif et un corpus, toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Les entités nommées sont, en général, typées selon des taxonomies plus ou moins vastes et fortement dépendantes du domaine d'application ou des besoins considérés. Elles recouvrent classiquement des noms désignant des personnes, des lieux ou des organisations, mais peuvent aussi se rapporter à des notions plus techniques comme les gènes,

les expressions temporelles et monétaires (Chinchor *et al.*, 1998; Chinchor, 1998; Sekine et Isahara, 2000; Tjong Kim Sang et De Meulder, 2003; Doddington *et al.*, 2004).

Les problèmes les plus souvent rencontrés, en ce qui concerne les noms propres en TALN, consistent à déterminer correctement ses frontières, afin de mieux les classifier. Par conséquent, nous proposons de modéliser notre propre système du traitement automatique des entités nommées, y compris l'extraction des entités nommées et la reconnaissance des entités nommées pour la paire de langues française-vietnamienne. Nous abordons également la translittération automatique et la traduction automatique liées aux noms propres, dans les sections 2.3 et 2.4, respectivement.

2.2 Reconnaissance des entités nommées

La tâche de reconnaissance des entités nommées (REN) est apparue la première fois dans la sixième conférence de MUC (Message Understanding Conference) (Sundheim, 1996). La REN consiste en un processus d'identification d'un mot ou d'un groupe de mots qui se réfèrent à une entité particulière dans un texte.

Les approches principales dans la reconnaissance des entités nommées sont basées sur (i) la linguistique informatique (TALN) et (ii) l'apprentissage automatique. L'approche linguistique informatique utilise les modèles à base de règles. Un ensemble de règles est défini et corrigé manuellement par des experts en linguistique afin d'identifier les entités nommées dans un texte. Un des inconvénients majeurs de cette approche est que les systèmes de REN nécessitent beaucoup d'expériences et de bonnes connaissances dans différentes langues ou différents domaines.

De plus, ces systèmes ne peuvent pas être appliqués à d'autres langues en raison des caractéristiques spécifiques de chaque langue. Par ailleurs, l'approche basée sur l'apprentissage automatique utilise un large corpus qui a déjà été annoté comme corpus

d'apprentissage dans le but de construire un système basé sur les connaissances linguistiques. De nombreuses techniques dans l'apprentissage automatique sont appliquées, telles que le modèle de Markov caché (HMM), l'entropie maximale (Maximum Entropy), l'arbre de décision, les machines à vecteurs de support (Support Vector Machines ou SVM) et les champs aléatoires conditionnels (Conditional Random Fields ou CRF) (Nadeau et Sekine, 2007).

En général, les deux approches se servent des informations lexicales comme d'une base de connaissance pour l'apprentissage, afin de construire des systèmes de REN, car cela permet d'améliorer la performance de ces systèmes.

En français ou en anglais, la reconnaissance des entités nommées est plus facile qu'en d'autres langues en raison de la capitalisation des entités dans les corpus. Parmi les approches symboliques autour des tâches liées à la détection des entités nommées concernant le français, plusieurs utilisent des transducteurs à nombre fini d'états, en citant le système de CasEN (Friburger, 2002) ayant 73,9% de F1, dans la campagne d'ESTER2, le système de TagEn (Poibeau, 2003) ayant 90% de F1 dans la campagne de MUC-6 corpus [1], le système de mXS (Nouvel *et al.*, 2013), à l'utilisation de techniques de fouille de textes à base de détection de motifs hiérarchiques pour l'évolution semi-automatique de nos bases de connaissances, ayant 79% de F1 dans la campagne d'évaluation d'Étape qui dépend de la campagne d'ESTER2. Récemment, Azpeitia *et al.* (2014) ont proposé un système de NERC-fr pour le français, en appliquant l'approche d'apprentissage automatique, avec l'algorithme d'entropie maximale, ayant 80,58% de F1, dans la campagne d'ESTER. Les campagnes d'évaluation d'ESTER et d'ESTER2 se basent sur des corpus constitués d'émissions radiophoniques transcrites.

En ce qui concerne la reconnaissance des entités nommées dans la langue vietnamienne, Tu *et al.* (2005) ont proposé un système de REN supervisé, en utilisant l'algorithme de

1. Évaluation de MUC-6 : `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

CRF, qui a atteint une performance de 85,51% de F1. Tran *et al.* (2007) ont présenté un autre modèle de reconnaissance des entités nommées supervisé en appliquant l'algorithme de SVM ayant une meilleure performance de 87,75% de F1. Nguyen *et al.* (2010) ont proposé un système de reconnaissance des entités nommées à base de règles qui a atteint une performance de 83% de F1. Jusqu'à maintenant, il y a très peu, voire aucune publication sur la reconnaissance des entités nommées appliquée au corpus bilingue français-vietnamien.

Récemment, les modèles à base de réseaux de neurones ont prouvé leur efficacité dans la tâche de REN. Chiu et Nichols (2015); Lample *et al.* (2016) et Yao et Huang (2016) ont proposé une amélioration du modèle à base de l'architecture de LSTM bidirectionnelle en combinant le modèle avec une couche de CRF ou en utilisant une pile des couches de LSTM. Ces modèles LSTM ont montré de très bons résultats par rapport aux approches traditionnelles, à savoir SVM et CRF, même s'ils n'ont pas besoin d'utiliser des dictionnaires, des *gazetteers* ou des informations additionnelles. Le principe consiste à construire un modèle de plongements de mots (Mikolov *et al.*, 2013) entraîné sur une grande quantité de textes non annotés. En utilisant des plongements de mots, deux mots, qui sont similaires, seront exprimés comme deux vecteurs proches dans un espace vectoriel.

2.3 Translittération automatique des noms propres

La translittération peut être considérée comme une sous-tâche de traduction automatique, lorsque nous avons besoin de traduire des graphèmes/phonèmes sources en graphèmes/phonèmes cibles. En d'autres termes, un modèle d'alignement doit d'abord être construit et un modèle de traduction est construit à partir de ces alignements. La translittération d'un mot de la langue de son origine vers une langue étrangère est appelée *forward transliteration* (translittération avant), tandis que la translittération d'un

mot d'emprunt écrit dans une langue étrangère vers la langue de son origine est appelée *backward transliteration* (translittération arrière) (Karimi *et al.*, 2011).

Les approches de génération de translittérations existantes peuvent être classées selon deux catégories : le niveau auquel le processus de mappage bilingue est modélisé et l'approche adoptée pour apprendre les règles de mappage. Dans l'approche à base de graphèmes, la translittération est considérée comme un processus de mise en correspondance d'une séquence de graphèmes d'une langue source vers une langue cible en ignorant les processus impliqués au niveau du phonème. En revanche, dans l'approche à base de phonèmes, l'ensemble du processus de mappage des phonèmes de la langue source vers les phonèmes de la langue cible est modélisé. Dans l'approche hybride, les modèles utilisant les deux approches ci-dessus sont combinés soit en les interpolant, soit en conditionnant les modèles de graphèmes sur les phonèmes correspondants (Oh *et al.*, 2006). Les approches à base de graphèmes et de phonèmes nécessitent des règles de mappage entre les entités correspondantes. Ces règles peuvent être manuellement créées ou peuvent être apprises en utilisant des techniques statistiques. Lorsque ces règles sont manuellement créées afin de modéliser un système de translittération, il s'agit de l'approche à base de règles.

Les techniques statistiques, basées sur de grands corpus parallèles de translittération, fonctionnent bien pour les langues dotées ou riches en ressources linguistiques. Par contre, pour les langues peu dotées n'ayant pas de telles ressources linguistiques, la translittération à base de règles est une option fiable (Tsuji *et al.*, 2002; Bhalla *et al.*, 2013).

De 2009 à 2016, divers systèmes de translittération ont été proposés lors des campagnes d'évaluation de l'atelier sur les entités nommées [2] (Duan *et al.*, 2016). Ces campagnes consistent à translittérer de l'anglais vers des langues ayant une grande variété

2. `http://workshop.colips.org/news2016/`

de systèmes d'écriture, comme l'hindou, le tamoul, le russe, le kannada, le chinois, le coréen, le thaï et le japonais. Nous pouvons constater que la romanisation des systèmes d'écriture non latins demeure une tâche complexe, qui dépend fortement de la langue source.

Grâce à cet atelier, beaucoup de progrès ont été réalisés impliquant des méthodologies pour résoudre la translittération des noms propres. Nous trouvons une émergence de différentes approches, notamment la conversion de graphème en phonème, l'approche à base de statistiques comme la traduction automatique (Laurent *et al.*, 2009; Nicolai *et al.*, 2015) ainsi que l'approche à base de réseaux de neurones (Finch *et al.*, 2016; Shao et Nivre, 2016; Thu *et al.*, 2016). D'autres travaux, Yao et Zweig (2015), utilisent des modèles de séquence-à-séquence sans le mécanisme d'attention pour la tâche de translittération. Rao *et al.* (2015) a utilisé un modèle LSTM bidirectionnel avec des délais d'entrée pour la conversion de graphème en phonème.

Par ailleurs, la variété des systèmes d'écriture pose d'importants défis dans l'extraction des entités nommées et la translittération automatique. Toutes ces difficultés sont aggravées par le manque de dictionnaires de prononciation bilingues des noms propres, l'ambiguïté des transcriptions ainsi que les variantes orthographiques au sein d'une langue. Un autre type de système de génération de translittération consiste en un système de forage de translittération dans lequel des paires de translittération parallèles sont extraites à partir de corpus comparables (Klementiev et Roth, 2006; Udupa *et al.*, 2009; Kumaran *et al.*, 2010; Tran *et al.*, 2016; Sajjad *et al.*, 2017).

Dans la revue de la littérature, nous constatons que la tâche de translittération, par rapport à la traduction automatique, n'est pas beaucoup étudiée alors qu'elle pourrait résoudre plusieurs problèmes liés aux mots hors vocabulaire, aux entités nommées, aux expressions polylexicales, etc. D'ailleurs, nous avons pu trouver un nombre très limité de recherches dans lesquelles la langue vietnamienne a été étudiée pour la tâche de

translittération. Cao *et al.* (2010) ont appliqué l'approche à base de statistiques, comme traduction automatique, dans la tâche de translittération pour une paire de langues peu dotée anglais-vietnamien. Leur système a obtenu une performance de 63 points de BLEU (en anglais, *BiLingual Evaluation Understudy*) (Papineni *et al.*, 2002).

De même, Ngo *et al.* (2015) ont proposé un autre modèle statistique pour une paire de langues peu dotée anglais-vietnamien, avec une contrainte phonologique sur les syllabes. Leur système a obtenu une meilleure performance que le système basé sur des règles, avec une réduction de 70% du taux d'erreurs. Le et Sadat (2017) ont exploré les réseaux de neurones récurrents, en particulier, la mémoire à long-court terme (*LSTM*), dans la tâche de translittération pour une paire de langues peu dotée française-vietnamienne. Les résultats expérimentaux ont montré que le système à base de réseaux de neurones récurrents était plus performant que le système basé sur l'approche statistique.

2.4 Traduction automatique des noms propres

La présente partie de ce chapitre présente la traduction automatique des entités nommées (TAEN) d'une langue étrangère en vietnamien. Bien qu'il existe plusieurs études portant sur la TAEN pour différentes paires de langues, la TAEN concernant le français-vietnamien n'a jamais été abordée jusqu'à ce jour. Ainsi notre recherche contribue essentiellement à l'étude de la TAEN pour cette paire de langues peu dotée.

Il existe plusieurs études portant sur la TAEN pour différentes paires de langues, par exemple, anglais-chinois (Cheung et Chen, 1998), espagnol-anglais (Hirschman *et al.*, 2000), français-anglais (Moore, 2003), anglais-arabe (Kashani *et al.*, 2007), anglais-japonais (Kumano *et al.*, 2004), etc. En revanche, nous n'avons trouvé aucune recherche concernant la TAEN pour le français-vietnamien. Les systèmes de TA doivent faire face à de nombreux problèmes concernant cette paire de langues, tels que les caractéristiques

des entités nommées (y compris des noms propres), ainsi que l'inconsistance de leur écriture et de la transcription ou la translittération en vietnamien.

En ce qui concerne la traduction des noms propres d'une langue étrangère en vietnamien, nous discuterons de trois questions en particulier :

1. Les principes d'écriture des noms propres en vietnamien.
2. La transcription/translittération et la traduction des noms propres d'une langue étrangère en vietnamien.
3. Les défis de la TAEN de l'anglais et du français vers le vietnamien.

Nous pouvons diviser en deux catégories les noms propres (NPs) vietnamiens : **NPs d'origine vietnamienne** et **NPs d'origine étrangère**. Dans la première catégorie, nous présentons quelques principes d'écriture liés à la mise en majuscule. En ce qui concerne l'écriture des noms des habitants du Vietnam, de la Chine (prononcés en sino-vietnamien), ou les habitants d'autres pays qui sont translittérés en chinois, les premières lettres de chaque syllabe doivent être en majuscule et sans traits d'union, *par exemple* : *Trần Văn Tạo ; Đinh Tiên Hoàng ; Lý Bạch ; Lỗ Tấn ; Thành Cát Tư Hãn.*

Il existe deux façons d'écrire les noms de personnes étrangères qui sont translittérés phonétiquement ou via une langue indo-européenne :

- Capitaliser la première lettre, les syllabes côte à côte et sans traits d'union, *par exemple* : *Frăngxoa (François), Ivan (Ivan), Napôlêông (Napoléon), Ađam (Adam).*
- Capitaliser la première lettre, les syllabes étant séparées par des traits d'union, *par exemple* : *Na-pô-lê-ông (Napoléon), Frăng-xoa (François), Huy-gô (Hugo).*

En ce qui concerne l'écriture des noms de lieu, ceux qui sont translittérés en chinois ou en sino-vietnamien sont capitalisés en premières lettres de chaque syllabe et sans traits d'union, *par exemple* : *Sầm Sơn (Sam Son), Hạ Long (Baie d'Ha Long), Bắc Kinh (Pékin), Thụy Sĩ (Suisse), Nhật Bản (Japon), Phần Lan (Finlande).* Les noms de lieu,

prononcés en sino-vietnamien ou en vietnamien indiquant une direction ou une position faisant partie de ces noms, doivent être capitalisés en premières lettres de chaque syllabe et sans traits d'union, *par exemple* : *Đông Nam Á (Asie du Sud-Est), Tây Âu (Europe de l'Ouest), Bắc Mỹ (Amérique du Nord), Bắc Triều Tiên (Corée du Nord), Đông Âu (Europe de l'Est).*

Il existe deux façons d'écrire les noms de lieux étrangers qui sont translittérés phonétiquement ou via une langue indo-européenne :

- Capitaliser la première lettre, les syllabes côte à côte et sans traits d'union, *par exemple* : *Chilê (Chili), Braxin (Brésil), Oasinhtơn (Washington), Mátxcơva (Moscou), Tôkyô (Tokyo).*
- Capitaliser la première lettre, les syllabes étant séparées par des traits d'union, *par exemple* : *Béc-lin (Berlin), Pa-ri (Paris), Mát-xcơ-va (Moscou), Tô-ky-ô (Tokyo).*

En ce qui concerne l'écriture des noms d'organisations, des devises des pays, des événements, il existe deux manières suivantes :

- Capitaliser la première lettre de la première syllabe. Si un nom de ce type inclut d'autres noms propres, il est nécessaire de capitaliser les noms propres en vertu des principes mentionnés, *par exemple* : *Cộng hòa xã hội chủ nghĩa Việt Nam (République socialiste du Vietnam), Chính phủ Việt Nam (Gouvernement du Vietnam); Quốc hội Hoa Kỳ (Assemblée des États-Unis); Đại học bách khoa (École polytechnique).*
- Capitaliser la première lettre de chaque mot, *par exemple* : *Cộng_hòa Xã_hội Chủ_nghĩa Việt Nam (République Socialiste du Vietnam); Trường Đại_học Bách_khoa Hà Nội (École polytechnique de Hanoi); Bộ Giao_thông Vận_tải (Ministère de Transport).* À noter que le souligné bas (_) représente le mot comme unité lexicale.

Quant à la deuxième catégorie, nous examinons la question de la transcription et/ou de la translittération des NPs, dont l'écriture est basée sur les caractères latins ou sur d'autres caractères. Puis nous proposons une méthode de traitement de la traduction des NPs du français en vietnamien (voir Chapitre 6).

2.5 Traduction automatique statistique

Au lieu de construire manuellement des dictionnaires et des règles de transformation pour l'approche de TA à base de règles, l'approche de traduction automatique statistique se base sur des modèles statistiques dans le but de développer automatiquement des dictionnaires et des règles de traduction. Pour ce faire, le système nécessite l'utilisation de corpus bilingues de grande taille, qui sont alignés au niveau des phrases. Le système calcule des statistiques et des probabilités de traduction de mots / termes ou de structure syntaxique entre les deux langues, des probabilités de ré-ordonnancement entre ces deux langues, ainsi que des probabilités d'occurrence de mots / termes étant donné un certain contexte.

En réalité, le problème de la TAS consiste à trouver une meilleure traduction $\hat{e}$ d'une phrase en langue cible $e = \{t_1, t_2, t_3, t_4, ...t_n\}$ étant donnée une phrase en langue source $f = \{s_1, s_2, s_3, s_4, ...s_n\}$, d'après le théorème de Bayes :

$$p(e|f) = \frac{p(f|e)\ p(e)}{p(f)} \tag{2.1}$$

Il s'agit de trouver la probabilité maximale $p(e|f)$. Puisque la probabilité de $p(f)$ est indépendante de e, le problème de TAS peut déduire de l'équation 2.1 :

$$\hat{e} = argmax_{e \in e^*}\ p(e|f) = argmax_{e \in e^*}\ p(f|e)\ p(e) \tag{2.2}$$

où :

- $p(f|e)$: modèle de traduction, permettant d'évaluer la probabilité de la traduction de la chaîne source f sachant la chaîne cible e,

- $p(e)$: modèle de langue, permettant d'évaluer la probabilité d'une séquence de mots dans la langue cible e.

La figure 2.1 illustre les différentes composantes d'une implémentation d'un système de TAS.

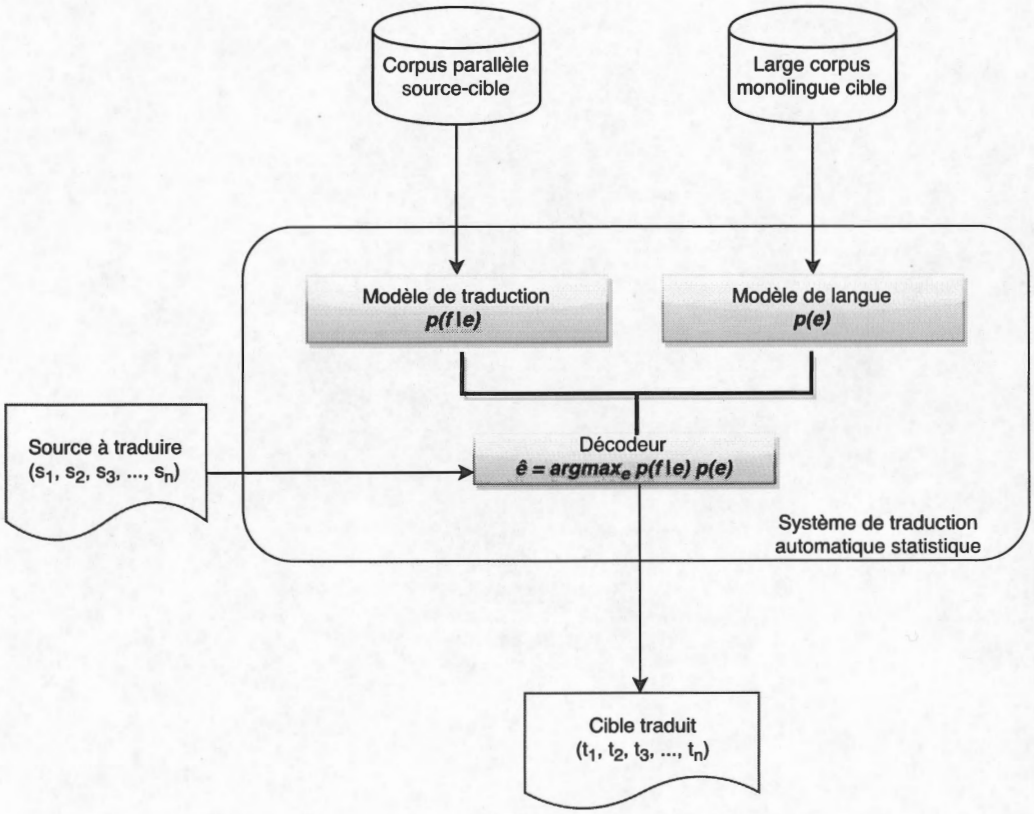

Figure 2.1 Architecture générale d'un système de traduction automatique statistique, sans prendre en considération les deux phases de pré-traitement et de post-traitement

2.5.1 Modèle de langue

L'un des modèles de langue standards utilisés dans la traduction automatique est le modèle de N-grammes. Ce modèle représente la probabilité de générer le mot à la position n en tenant compte des mots précédents $n-1$, ce que l'on appelle l'historique.

Par l'utilisation de la probabilité des règles en chaîne (Jurafsky et Martin, 2000), nous pourrions calculer la probabilité d'une séquence de mots ou d'une phrase $P(W)$ avec l'équation 2.3 suivante :

$$\begin{aligned} P(W) = P(w_1^n) &= P(w_1, w_2, w_3, ..., w_n) \\ &= P(w_1) P(w_2|w_1) P(w_3|w_1^2) ... P(w_n|w_1^{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1^{i-1}) \end{aligned} \tag{2.3}$$

où $W = (w_1, w_2, w_3, ..., w_n) = w_1^n$, avec w_i est le mot à la position i dans la phrase W, $1 \leq i \leq n$.

La règle de chaîne ci-dessus est servie à calculer la probabilité d'une phrase et également à estimer la probabilité conditionnelle d'un mot à la position i étant donnés tous les mots précédents.

Cependant, il n'est pas possible de calculer la probabilité d'un mot étant donnée une séquence de mots précédents, $P(w_n|w_1^{n-1})$. Ainsi, nous utilisons l'hypothèse de *Markov* afin d'évaluer la probabilité d'un mot dépendant uniquement d'une courte histoire (entre 2-grammes à 5-grammes). Par exemple, dans l'hypothèse bi-gramme, la probabilité de chaque mot de l'équation 2.3 est calculée comme suit :

$$P(W) = P(w_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{2.4}$$

où la probabilité d'un bi-gramme $P(w_i|w_{i-1})$ est calculée :

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{2.5}$$

où C est une fonction de comptage. Par conséquent, en utilisant l'hypothèse N-gramme pour calculer la probabilité de chaque mot, nous avons l'équation 2.6 comme suit :

$$\begin{aligned} P(W) = P(w_1^n) &= P(w_1, w_2, w_3, ..., w_n) \\ &= P(w_1)P(w_2|w_1)P\left(w_3|w_1^2\right)...P\left(w_n|w_1^{n-1}\right) \\ &\approx P(w_1)P(w_2|w_1)...P\left(w_{N-1}|w_1^{N-2}\right)\prod_{i=N}^{n} P\left(w_i|w_{i-N+1}^{N-1}\right) \quad (2.6) \end{aligned}$$

Le modèle de langage est généralement généré à partir d'un ensemble donné de mots appelés mots de vocabulaire. Les autres mots sont considérés comme des mots hors du vocabulaire (MHV). Un problème s'expose lorsque le modèle de langage peut attribuer une probabilité nulle à des mots de type MHV. Afin d'éviter ce problème, une variété de méthodes de lissage permettent de générer des probabilités pour des mots non observés dans le corpus d'apprentissage, en citant quelques unes telles que lissage additif, estimation de Good-Turing (Good, 1953), lissage de Jelinek-Mercer (interpolation) (Jelinek et Mercer, 1980), lissage de Katz (recul) (Katz, 1987), lissage de Witten-Bell (Witten et Bell, 1991), lissage de Kneser-Ney (Kneser et Ney, 1995).

2.5.2 Modèle de traduction à base de segments

Un des enjeux des modèles de traduction à base de mots consiste à ne pas tenir compte de l'information contextuelle. Ils sont uniquement basés sur les analyses statistiques de chaque mot dans une phrase donnée. Par contre, les modèles de traduction à base de segments sont plus performants que ceux à base de mots. Cette approche à base de segments permet aux systèmes de TAS d'être en mesure de traduire des segments ou des groupes de mots au lieu de traduire mot par mot.

En ce qui concerne un système de TAS à base de segments (Koehn *et al.*, 2003), la phrase de langue source f est découpée en plusieurs segments f_i'. Chaque segment f_i' est traduit en chaque segment e_i' de la langue cible en se basant sur la distribution de

probabilité $\varphi\left(f_i^{'}|e_i^{'}\right)$. Ensuite, chaque segment $e_i{}'$ sera converti en fonction du modèle de ré-ordonnancement $d(start_i, end_{i-1})$. Ainsi, les meilleures traductions seront calculées avec l'équation 2.7 suivante :

$$p(f|e) = \prod_i \varphi\left(f_i^{'}|e_i^{'}\right) d(start_i - end_{i-1} - 1) \tag{2.7}$$

où

- $\varphi\left(f_i^{'}|e_i^{'}\right)$ signifie la probabilité de traduction pour produire le segment d'hypothèse i^{th} (séquence de mots) $e_i^{'}$ étant donné le segment $f_i^{'}$ de la phrase source f.
- $d(start_i - end_{i-1} - 1)$ signifie la probabilité de distribution relative de la distorsion, $start_i$ et end_{i-1} sont la position du premier mot et la position du dernier mot de chaque segment $f_i^{'}$ de la langue source, respectivement.

2.5.3 Décodeur

Le fonctionnement d'un décodeur permet de générer de meilleures hypothèses de traduction. Sa modélisation nécessite plusieurs éléments, à savoir :

- *un modèle de langue*, permettant de s'assurer de la structure grammaticale des hypothèses de traduction produites,
- *un modèle de traduction à base de segments*, contenant une table de traduction de segments entre la langue source et la langue cible,
- *un modèle de ré-ordonnancement* (ou *de distorsion*) *de segments*, permettant de gérer l'alignement des segments.

Les scores associés à chaque modèle vont permettre au décodeur de sélectionner les segments cibles les plus probables étant donnée une phrase source. Chaque paramètre peut être défini par un ensemble de fonctions ayant comme paramètre les phrases source S et cible T, ainsi que leurs alignements au niveau des mots.

En partant du modèle de traduction, la table de traduction permet de générer des hypothèses de traduction en calculant les probabilités de chaque paire de segments source-cible. Ensuite grâce au modèle de langage, les scores de ces hypothèses sont pondérés par la probabilité d'occurrence de ces segments dans la langue cible en se basant sur le modèle de ré-ordonnancement. Finalement, le décodeur sélectionne une hypothèse de traduction ayant la probabilité maximale générée par le produit du modèle de langue $p(e)$ et du modèle de traduction $p(f|e)$.

Les algorithmes utilisés pour cette tâche sont souvent basés sur la recherche de A*, la recherche à faisceaux (*beam search*) ou la recherche gourmande (*greedy hill-climbing search*) (Koehn, 2009).

2.6 Traduction automatique pour les langues peu dotées

Berment (2004) a défini dans sa thèse la notion d'une langue peu dotée (ou une langue-π) qui consiste, dans le contexte du traitement automatique de langues naturelles (TALN), en une langue qui n'a pas une grande quantité de données et de ressources linguistiques informatisées disponibles. L'approche statistique de la TA, tout en étant extrêmement efficace dans la pratique, est déjà vulnérable à la pénurie de données, même avec de grands ensembles de données d'apprentissage. En général, les modèles de langue et les modèles de traduction sont capables de générer correctement des traductions, ayant exposées pendant la phase d'apprentissage, à une certaine capacité à déduire des traductions d'expressions similaires.

En revanche, les mots qui ne sont pas vus en phase d'apprentissage ne seront pas reconnus, d'où la problématique de mots hors vocabulaires (en anglais, *out-of-vocabulary* ou *OOV*). Le manque de ressources linguistiques, y compris des outils de TALN, pour des paires de langues peu dotées rend la tâche de construction des systèmes de traduction

automatique statistique plus difficile. Généralement, les données disponibles ne sont pas suffisantes pour couvrir toutes les variétés de la langue source ou/et cible.

En Europe, les langues, qui font partie de cette catégorie, sont les langues *catalane, basque, galloise, turque, norvégienne* et en Europe de l'Est, les langues : *tchèque, slovaque, ukrainienne, croate, roumaine, estonienne, lettonne, lituanienne, etc.* ; en Asie, les langues *thaïe, khmère, indonésienne, vietnamienne, mongole, etc.* En Afrique, il y a plusieurs langues officielles, dont le *zoulou* et le *xhosa* en Afrique du Sud (Mangeot et Sadat, 2014).

En ce qui concerne les langues peu dotées et la TA, nous constatons que la plupart des recherches concernées traitent de la collection des corpus comparables et parallèles bilingues et multilingues. Autrement dit, il s'agit de la phase de création des ressources linguistiques servant par la suite à des fins de la TA. Les approches utilisées sont basées sur **(1) les similitudes entre les langues proches, (2) une langue de pivot** et **(3) la projection interlinguistique, avec l'apprentissage par transfert, c'est-à-dire à partir des langues riches en ressources linguistiques.**

Dans la première approche, les systèmes explorent les similarités entre les langues proches reposant sur des techniques d'analyse de bas niveaux et des ressources ayant peu d'annotations linguistiques. Nous observons les langues de même famille, par exemple : *ukrainien-russe* (Babych *et al.*, 2007), *catalan-espagnol* (Alonso et Bocsák, 2005), *mongol-chinois* (Chen *et al.*, 2014), etc. Pinnis *et al.* (2012) ont présenté un flux de travail permettant d'extraire des termes à partir de documents textuels, d'annoter les termes dans les documents et d'aligner les termes bilingues dans les corpus comparables pour quatre langues peu dotées : *croate, letton, lituanien* et *roumain.*

Dans la deuxième approche, Babych *et al.* (2007) ont proposé une méthode pour comparer la performance entre la traduction directe concernant la paire de langues *ukrainien-anglais* et la traduction grâce à une langue pivot par le biais de l'évaluation automatisée

des segments avec différentes difficultés de traductions concernant les langues *slave-russe-anglais*, dont la langue pivot est le *russe*. Sharwood (2013) a présenté un rapport concernant la construction d'un système de traduction automatique *xhosa-anglais-zoulou* dont la langue pivot est l'*anglais*.

Dans la troisième approche, Luong *et al.* (2010) et Wen *et al.* (2010) ont montré que l'application de la segmentation morphologique des langues riches en morphologie était en mesure d'améliorer la qualité des systèmes de traduction automatique statistique (TAS). Ainsi, il y a des publications qui étudient les interactions entre les paires de langues peu dotées et riches en ressources, par exemple, *grec-anglais* (Mastropavlos et Papavassiliou, 2011), *quechua-espagnol* (Sánchez, 2003), *azéri-russe* en Azerbaijan (Zuercher, 2009), *tagalog-anglais* aux Philippines (Bautista et Bolton, 2008), *zoulou-anglais* en Afrique du Sud (Ramsay-Brijball, 1999), *hmong-anglais* (Lewis et Yang, 2012) pour la langue *hmong-mien* en Asie du Sud. Notre méthodologie concerne la troisième approche, en appliquant sur la paire de langues du *français*, langue riche en ressources linguistiques, vers le *vietnamien*, langue peu dotée en ressources linguistiques.

Par ailleurs, avec l'émergence de l'approche profonde, la traduction automatique neuronale (TAN) ayant peu de ressources est connue pour facilement sur-apprendre et obtenir une performance inférieure lorsque les données d'entraînement sont limitées (Koehn et Knowles, 2017). En général, il existe deux manières de traiter le problème de la traduction ayant peu de ressources : (1) en utilisant les ressources de données monolingues non étiquetées pour entraîner les plongements et (2) en partageant les connaissances entre des paires de langues à ressources faibles et élevées. De nombreux efforts de recherche ont été consacrés à l'intégration des corpus monolingues dans la traduction automatique, tels que l'apprentissage de multi-tâche (Gulcehre *et al.*, 2015; Zhang et Zong, 2016), la traduction en arrière (Sennrich *et al.*, 2015), l'apprentissage en arrière et en avant ou à double tâche (He *et al.*, 2016) et la traduction automatique non su-

pervisée avec uniquement des corpus monolingues pour les deux côtés source et cible (Artetxe *et al.*, 2017; Lample *et al.*, 2018; Yang *et al.*, 2018).

En ce qui concerne les partages des connaissances entre des paires de langues à ressources faibles et élevées, des recherches antérieures ont travaillé sur des méthodes permettant d'exploiter les connaissances des traductions auxiliaires, voire des tâches auxiliaires. Par exemple, Cheng *et al.* (2016); Chen *et al.* (2018) étudient l'utilisation d'un pivot pour créer un chemin de traduction entre deux langues, même sans aucune ressource dirigée. Le pivot peut être une troisième langue ou même une image dans des domaines multimodaux. Puisque les pivots ne sont pas faciles à obtenir, Firat *et al.* (2016); Johnson *et al.* (2016) ont montré que la structure de la TAN est adaptée à la traduction automatique multilingue. Gu *et al.* (2018) ont également montré qu'un tel système multilingue de TAN pourrait améliorer les performances de la traduction automatique à faibles ressources en utilisant une représentation lexicale universelle afin de partager des informations de plongements à travers plusieurs langues.

2.7 Évaluations

Dans cette section, nous présentons les métriques utilisées afin d'évaluer, de manière automatique, les performances des modèles de langue, des modèles de translittération ainsi que des modèles de traduction automatique.

2.7.1 Performances de transcription et de translittération

L'une des métriques d'évaluation les plus utilisées pour la transcription est le calcul du taux d'erreurs de mots (en anglais : *Word Error Rate* ou *WER*), définie comme suit :

$$\text{WER} = \frac{\#Ins + \#Sub + \#Del}{\#Total\ des\ tokens\ dans\ la\ référence} \times 100\% \tag{2.8}$$

où *#Ins* désigne le nombre des *tokens* alignés, qui sont ajoutés dans l'hypothèse, *#Sub* est le nombre de mots en référence, qui sont remplacés par les mots alignés dans l'hypothèse, *#Del* est le nombre de mots, qui sont manqués de la référence.

En ce qui concerne la performance de translittération, nous utilisons la méthode de calcul du taux d'erreurs de graphèmes (en anglais : *Grapheme Error Rate* ou *GER*). Cette méthode est similaire à celle de calcul du taux d'erreurs de mots.

2.7.2 Précision, Rappel, score de F1

Afin d'évaluer, de manière automatique, les prédictions d'annotation produites par les systèmes de reconnaissance des entités nommées, nous utilisons les métriques d'évaluation de performance telles que la *précision* (P), le *rappel* (R) et le *score de F1*.

$$P = \frac{|\{tokens\ pertinents\} \cap \{tokens\ trouvés\}|}{\{tokens\ trouvés\}} \times 100\% \tag{2.9}$$

$$R = \frac{|\{tokens\ pertinents\} \cap \{tokens\ trouvés\}|}{\{tokens\ pertinents\}} \times 100\% \tag{2.10}$$

où $\{tokens\ trouvés\}$ est le nombre de *tokens* prédits, $\{tokens\ pertinents\}$ désigne le nombre de *tokens* qui sont correctement annotés.

Par conséquent, la *précision* représente le nombre d'étiquettes trouvées correctes, tandis que le *rappel* représente le nombre d'étiquettes pertinentes renvoyées par le modèle.

Généralement, utilisée comme une moyenne de ratios, la *moyenne harmonique*, appelée H_{mean}, est définie comme suit :

$$H_{mean} = \frac{N}{\sum_{i=1}^{N} \frac{1}{m_i}} \times 100\% \tag{2.11}$$

Avec $N = 2$, $m_1 = P$ et $m_2 = R$, nous avons :

$$\begin{aligned} H_{mean} &= \frac{2}{\frac{1}{P}+\frac{1}{R}} \times 100\% \\ &= \frac{2 \times P \times R}{P+R} \times 100\% \end{aligned} \tag{2.12}$$

H_{mean} aussi connue comme *F1* ou *F-measure* (Rijsbergen, 1979) devrait satisfaire la contrainte. Donc, pour évaluer *F1*, nous pourrions utiliser la formule suivante :

$$F1 = \frac{2 \times P \times R}{P+R} \times 100\% \tag{2.13}$$

2.7.3 Performance de traduction automatique (BLEU)

La métrique d'évaluation de BLEU (en anglais : *Bilingual Evaluation Understudy*) consiste en une méthode qui sert à évaluer la qualité d'une hypothèse issue d'un système de traduction automatique. Cette métrique d'évaluation automatique, proposée par Papineni *et al.* (2002), représente une corrélation correcte avec l'évaluation humaine.

Nous définissons $p_1, p_2, p_3, ...p_N$ comme précision d'unigramme (proportion de mots corrects parmi tous les mots candidats), précision de bigramme (proportion de bigrammes corrects parmi tous les bigrammes candidats), précision de trigramme (proportion de trigrammes corrects parmi tous les trigrammes candidats), etc.

Ensuite, nous mettons toutes les précisions ensemble en calculant la *moyenne géométrique*, appelée G_{mean}, des ratios donnés comme suit :

$$\begin{aligned} G_{mean} &= \left(\prod_{i=1}^{N} p_i \right)^{1/N} \\ &= \exp\left(\log_e \left(\left(\prod_{i=1}^{N} p_i \right)^{1/N} \right) \right) \\ &= \exp\left(\frac{1}{N} \sum_{i=1}^{N} \log_e (p_i) \right) \end{aligned} \tag{2.14}$$

Puis, nous introduisons la pénalité de brièveté (en anglais : *brevity penalty* ou *BP*), qui calcule la corrélation entre la longueur de référence r et la longueur d'hypothèse c, comme :

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{otherwise} \end{cases} \tag{2.15}$$

Finalement, la métrique d'évaluation de BLEU est définie comme suit :

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log_e(p_i)\right) \tag{2.16}$$

2.7.4 Taux d'erreurs de traduction (TER)

Le taux d'erreurs de traduction (en anglais : *Translation Error Rate* ou *TER*) (Snover *et al.*, 2009) est une métrique d'évaluation automatique, qui compte le nombre d'opérations d'édition (substitution (*#SUB*), suppression (*#DEL*), insertion (*#INS*) d'un mot et décalage (*#SHF*) d'un mot ou des mots adjacents, nécessaires pour transformer une hypothèse, issue d'un système de la traduction automatique, en une référence. Ce taux est calculé par l'équation suivante :

$$\mathrm{TER} = \frac{\text{nombre d'opérations d'édition}}{\text{longueur moyenne des références}} = \frac{\#SUB + \#DEL + \#INS + \#SHF}{\text{longueur moyenne des références}} \tag{2.17}$$

2.7.5 Métrique d'évaluation de traduction avec l'ordonnancement explicite (METEOR)

La métrique d'évaluation de traduction avec l'ordonnancement explicite (en anglais : *Metric for Evaluation of Translation with Explicit ORdering* ou *METEOR*), proposée par Banerjee et Lavie (2005a), consiste à mieux corréler les hypothèses de traduction avec les jugements humains, en utilisant davantage des alignements de mots à mots

entre une hypothèse et quelques références. L'alignement est considéré selon trois modules : le premier utilise une correspondance exacte entre les formes de surface des mots, le second compare les racines des mots (en anglais : *stems*) et le troisième utilise les synonymes d'une ressource lexicale, telle que *WordNet* (Miller, 1998).

Formellement, le score de *METEOR* utilise la moyenne harmonique $F_{measure}$ et un facteur additionnel, appelé *Pénalité* (en anglais : *Penalty*), comme suit :

$$F\text{-}measure = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R} \tag{2.18}$$

$$Penalty = \gamma \times \left(\frac{\text{number of chunks}}{\text{number of matches}} \right)^{\beta} \tag{2.19}$$

$$METEOR\ score = (1 - Penalty) \times F\text{-}measure \tag{2.20}$$

CHAPITRE III

EXTRACTION DES ENTITÉS NOMMÉES PAR PROJECTION TRANS-LINGUISTIQUE

3.1 Introduction

En raison de multiples significations des mots ou de polysémie, des expressions qui entraîne une ambiguïté, ainsi que des métaphores dans les langues, les systèmes de traduction automatique présents n'offrent pas toujours de traductions correctes. Ils peuvent refléter un nom commun écrit en majuscule comme s'il s'agissait d'un nom propre et vice versa ; ils traduisent un nom ayant une signification dans un dictionnaire bilingue comme dans le cas d'un nom commun.

La traduction des entités nommées, en particulier, permet d'identifier correctement à travers les noms propres, notamment les noms de personne, les noms d'organisation et aussi les noms de lieu. Le traitement des entités nommées demeure un sujet très important, qui peut se retrouver au sein de plusieurs domaines de recherche tels que la traduction automatique statistique, la recherche d'information interlinguistique, l'extraction d'information et les systèmes de questions & réponses, parce que les entités nommées - en particulier les noms de personne, les noms de lieu et les noms d'organisation - contiennent une signification essentielle dans le traitement automatique du langage naturel.

Nous pouvons citer les exemples suivants concernant les traductions d'entités nommées correctes et incorrectes du français vers le vietnamien. Ici, nous utilisons *Google Translate*[1] pour illustrer.

(1.1) (fr) hier soir , j'ai mangé avec **m. michel chevalier** .
(1.2) (vi) đêm qua, tôi đã ăn với **m. Hiệp sĩ Michael**. [*traduction erronée, par Google Translate*]
(1.3) (vi) đêm qua, tôi đã ăn với **ông Mi-sên Sơ-va-li-ê**. [*traduction et translittération correcte*]

Dans le premier exemple ci-dessus, l'entité nommée qui désigne un nom de personne a été incorrectement traduite, « m. michel chevalier » dans la phrase (1.1) par « m. Hiệp sĩ Michael » (*littéralement : m. Chevalier Michel*) dans la phrase (1.2) au lieu de « với ông Michel Chevalier » dans la phrase (1.3). On constate une tendance autrefois par rapport aux noms propres concernant leur provenance. Ceux-ci possèdent un sens lié à la personne qui le porte, le plus souvent en relation avec soit des titres de noblesse, des origines géographiques (*i.e.* Chevalier, de France), soit des métiers (*i.e.* Lumière, Pelletier) (Nuessel, 1992).

(2.1) (fr) Ma famille voyage dans **le delta du fleuve Rouge**.
(2.2) (vi) Gia đình tôi đi ở **đồng bằng sông Đỏ**. [*littéralement traduction incorrecte*]
(2.3) (vi) Gia đình tôi du lịch ở **đồng bằng sông Hồng**. [*traduction correcte par Google Translate*]

1. `https://translate.google.ca/`, consultée Samedi 28 Septembre 2019

Dans le second exemple, l'entité nommée qui désigne un nom de lieu a été incorrectement traduite « fleuve Rouge » dans la phrase (2.1) par « sông Đỏ » dans la phrase (2.2) au lieu de « sông Hồng » dans la phrase (2.3). L'erreur de traduction concerne ici la synonymie entre deux mots « Đỏ » et « Hồng » qui ont la même signification en sino-vietnamien.

(3.1) (fr) Il est en train de lire **L'Observateur**.
(3.2) (vi) Hiện anh đang đọc **The Observer**. [*traduction incorrecte par Google Translate*]
(3.3) (vi) Anh ta đang đọc **L'Observateur**. [*traduction correcte*]

Nous pouvons observer dans le troisième exemple ci-dessus, parfois un échec de discrimination entre le nom propre et le nom commun. Dans ce cas, l'entité nommée, qui désigne le nom d'une organisation, « L'Observateur » (3.1), a été incorrectement traduite du français vers le vietnamien. Sa traduction a été empruntée à l'anglais par « The Observer » (3.2) au lieu de garder son nom intact comme « L'Observateur » (3.3).

Une solution possible consiste à créer un dictionnaire ou un lexique bilingue d'entités nommées. Un dictionnaire d'entités nommées ou un lexique de paires d'entités nommées est une base de connaissances pour la traduction à base des règles et la méthode de translittération à base de statistiques. Cependant, cette approche nécessite, tout d'abord, un large corpus bilingue avec des entités nommées annotées. L'annotation manuelle de corpus bilingues est fastidieuse et le processus est coûteux. Par ailleurs, il y a, malheureusement, très peu ou aucune recherche concernant la traduction d'entités nommées pour la paire de langues peu dotée française-vietnamienne.

Dans ce chapitre, nous proposons une approche itérative pour l'extraction des entités nommées bilingues par projection translinguistique, avec application pour la paire de langues peu dotée français-vietnamien.

Le reste de ce chapitre est organisé comme suit : La section 3.2 présente l'approche proposée. Ensuite, les sections 3.3 et 3.4 rapportent les expérimentations et l'évaluation. Enfin, la section 3.5 conclut le chapitre.

3.2 Approche proposée

Nous présentons une approche d'extraction d'entités nommées pour le français-vietnamien. Nous discutons ici d'une apparence morpho-syntaxique entre les entités nommées dans la langue source, le *français*, et la langue cible, le *vietnamien*. Notre approche repose sur les deux postulats suivants :

Postulat 1. Nous supposons qu'une entité nommée dans la langue source et son entité nommée traduite dans la langue cible possèdent la même catégorie, telle que le nom de personne, le nom de lieu ou le nom d'organisation.

Postulat 2. Considérant que les noms de personne et de lieu sont souvent phonétiquement traduits et que leurs formes écrites sont similaires à leurs prononciations, nous pouvons ajouter une caractéristique supplémentaire de cohérence syllabique comme un critère de sélection. Il s'agit d'une mesure de syllabification qui permet de comparer le nombre de syllabes dans les mots ou le groupe de mots entre la langue source et la langue cible.

Par conséquent, l'approche proposée est composée de trois étapes principales :

- Étape 1 : Extraire une liste de candidats d'entités nommées en français et une liste de candidats de noms propres en vietnamien à partir du corpus bilingue français-vietnamien qui a été aligné au niveau des phrases.

- Étape 2 : Filtrer des entités nommées traduites du français en vietnamien en utilisant un modèle statistique et un modèle d'alignement de mots.
- Étape 3 : Calculer la similitude entre des paires de candidats d'entités nommées bilingues traduits. Ce score est basée sur l'équation de Levenshtein (3.1) comme suit :

$$similarity(S_i, T_j) = 1 - \frac{edit_distance(S_i, T_j)}{maxlength(|S_i|, |T_j|)} \tag{3.1}$$

La fonction de Levenshtein ou *minimum edit distance* (la distance d'édition minimale) est largement utilisée pour calculer la distance de similarité entre deux chaînes de caractères. Elle retourne des ensembles de poids minimal des opérations d'édition qui transforment un mot source S_i en un mot cible T_j lié à l'insertion, la suppression ou la substitution.

Nous nous servons d'un corpus d'apprentissage bilingue français-vietnamien qui a été aligné au niveau des phrases. Nous divisons ce corpus bilingue en deux corpus monolingues comme le montre la Figure 3.1.

1. En premier lieu, nous appliquons un module de reconnaissance d'entités nommées pour le corpus français. Ensuite nous obtenons une liste de candidats d'entités nommées en français, appelée *liste_FR_EN*. Pour le corpus vietnamien, nous appliquons un module d'annotation des catégories grammaticales (*part-of-speech*). Une liste de candidats de noms propres en vietnamien, appelée *liste_ref_VI_NP*, a été également extraite.
2. En deuxième lieu, un modèle statistique est appliqué afin de traduire la liste de candidats d'entités nommées en français *liste_FR_EN* vers le vietnamien. Nous obtenons une liste de candidats d'entités nommées traduites en vietnamien, appelée *liste_traduite_VI_EN*.
3. En troisième lieu, nous calculons les scores de similarité entre la liste d'entités nommées traduites en vietnamien *liste_traduite_VI_EN* comme source S_i et la liste de candidats de noms propres en vietnamien *liste_ref_VI_NP* comme cible

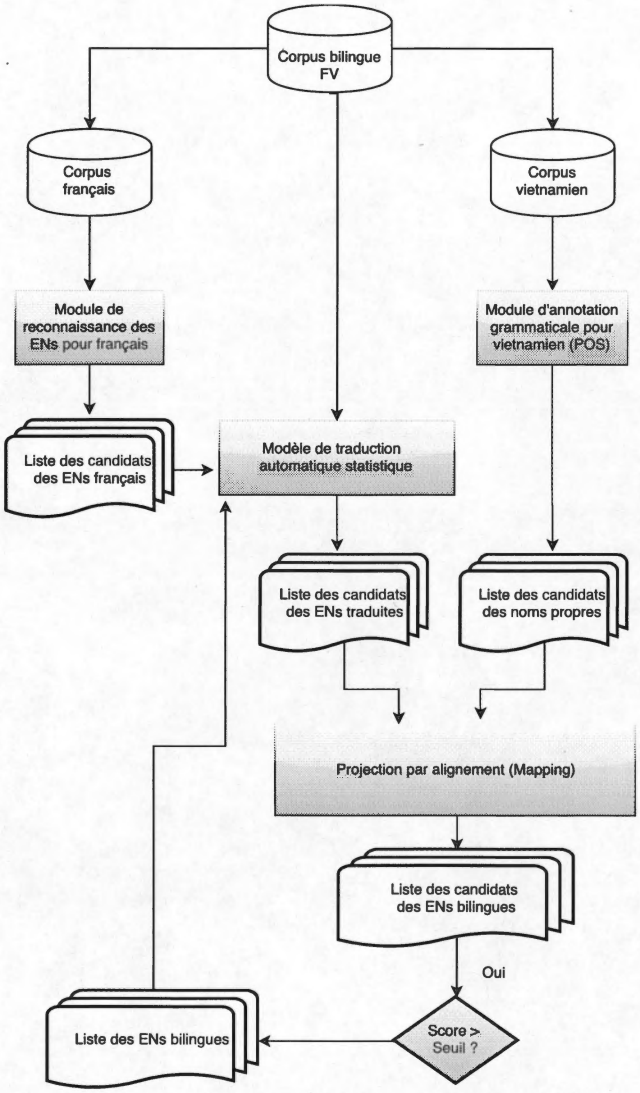

Figure 3.1 Architecture d'un système d'extraction des entités nommées bilingues par projection translinguistique, avec application pour une paire de langues peu dotée, telle que le français et le vietnamien.

T_j. Si la valeur de score d'une paire de candidats est supérieure à la valeur du seuil prédéfini, cette paire est sélectionnée et stockée dans une autre liste de paires d'entités nommées bilingues, appelée *liste_EN_bilingue_FR_VI*.

Après avoir analysé les erreurs susceptibles de la traduction des entités nommées bilingues, nous ré-entraînons cette liste *liste_EN_bilingue_FR_VI* post-éditée dans le modèle statistique dans le but d'observer des impacts possibles sur la qualité de la traduction des entités nommées bilingues.

3.3 Expérimentations

3.3.1 Préparation des données

Le corpus d'apprentissage bilingue français-vietnamien a été collecté avec 14 963 paires de phrases provenant des pages Web multilingues[2] pour l'actualité et 5 284 paires de phrases extraites de la conférence de Tam Dao, en 2009[3] (Lagrée *et al.*, 2009). Il s'agit d'une conférence économique internationale organisée annuellement au Vietnam, dans laquelle un grand nombre d'entités nommées ont été remarqués. En raison de contraintes de temps, nous avons seulement extrait la version 2009. Ultérieurement, nous pourrions en extraire davantage. Dans nos expérimentations, nous utilisons un petit corpus de test ayant 1 060 paires de phrases bilingues français-vietnamiens. Le Tableau 3.1 montre les statistiques pour les trois expérimentations à procéder.

3.3.2 Configuration

En ce qui concerne le corpus vietnamien, toutes les phrases sont segmentées au niveau des mots. L'outil CLC_VN_WS (Vu, 2011) du Centre de linguistique computationnel[4]

---

2. `https://fr.vietnamplus.vn`, `http://www.vietnamplus.vn`, `https://vn.ambafrance.org/-Francais-`, `http://vovworld.vn/fr-CH.vov`

3. `http://www.tamdaoconf.com`

4. `http://www.clc.hcmus.edu.vn/?lang=fr`

Tableau 3.1 Statistiques des données d'apprentissage pour la tâche de l'extraction des entités nommées bilingues pour la paire de langues peu dotée en français-vietnamien.

Données	**Description**	**#paires de phrases**	**#tokens fr - vi**
Apprentissage	Expérimentation 1 (Exp1)	14 963	**232 541 - 274 093**
	Expérimentation 2 (Exp2)	14 963	**232 541 - 274 093**
	Expérimentation 3 (Exp3)	20 247	**354 577 - 422 982**

(CLC) à l'Université des Sciences, au Vietnam est utilisé pour cette segmentation de phrases en mots. De plus, le corpus vietnamien a été également annoté en utilisant l'outil d'annotation en catégories grammaticales en vietnamien CLC_VN_POS (Nguyen et Dinh, 2011) basé sur l'approche d'entropie maximale. Il est à noter que la performance du l'outil CLC_VN_WS a obtenu 97,33% de la moyenne harmonieuse F-score. Et que l'outil de CLC_VN_POS a obtenu 97,08% de la précision globale et 85,78% de la précision des mots inconnus. Le corpus d'apprentissage consiste en 87 000 phrases en vietnamien annotées, appelé VietTreeBank (Thi *et al.*, 2013), avec les 36 étiquettes en catégories grammaticales.

En ce qui concerne le corpus français, un outil de reconnaissance des entités nommées, développé par Nouvel *et al.* (2011), a été utilisé. Nous appliquons l'outil *Moses*[5] (Koehn, 2009) pour construire notre système de traduction automatique statistique. *Moses* offre un ensemble d'outils nécessaires dans le but de construire un modèle de traduction statistique.

5. `http://www.statmt.org/`

Nous avons effectué les trois expérimentations suivantes :

- Expérimentation 1 (Exp1) : Dans cette première expérimentation, qui est considérée comme système de base, nous utilisons un corpus d'apprentissage et un corpus de test composés de 14 963 et de 1 060 paires de phrases bilingues, respectivement.
- Expérimentation 2 (Exp2) : Dans cette deuxième expérimentation, nous utilisons un corpus d'apprentissage de 14 963 paires de phrases bilingues et un corpus de test ayant 340 entités nommées extraites de 1 060 paires de phrases bilingues.
- Expérimentation 3 (Exp3) : Dans cette troisième expérimentation, nous utilisons un corpus d'apprentissage de 14 963 paires de phrases bilingues, combiné à 5 284 paires de phrases bilingues extraites de la conférence Tam Dao 2009 et un corpus de test de 340 entités nommées extraites de 1 060 paires de phrases bilingues.

3.4 Évaluations

3.4.1 Résultats obtenus

Afin d'évaluer la performance de notre système de l'extraction des entités nommées bilingues, nous utilisons les métriques d'évaluation telles que la précision, le rappel et le score de F1. Dans cette tâche, nous nous intéressons à la construction d'un lexique des entités nommées bilingues de types tels que nom de personne, nom d'organisation et nom de lieu. Le Tableau 3.2 montre les résultats obtenus lors de l'expérimentation.

Avec un petit corpus d'apprentissage, nous avons obtenu 340 entités nommées bilingues correctement trouvés, y compris 86 noms de personne, 237 noms de lieu et 17 noms d'organisation (Tableau 3.2).

Nous constatons que le système de l'extraction des entités nommées a eu la difficulté d'identifier celles concernant les noms d'organisation, ayant un faible nombre d'entités

Tableau 3.2 Résultats de l'expérimentation de l'extraction des entités nommées bilingues, avec un corpus de test de 1 060 paires de phrases en français-vietnamien.

	#Trouvé Correct	**#Bruit**	**#Référence**	**Précision (%)**	**Rappel (%)**	**F1 (%)**
Lieu	237	4	246	98,34	96,34	**97,33**
Organisation	17	11	55	60,71	30,91	**40,96**
Personne	86	4	96	95,56	89,58	**92,47**

nommées correctement trouvées 17 sur 55. Son score de F1 n'a atteint que 40,96% (Tableau 3.2). En revanche, nous observons que les entités nommées concernant les noms de personne et les noms de lieu possèdent une grande similitude entre le français et le vietnamien. Le système de l'extraction des entités nommées les a très bien identifiées, avec 92,47% de F1 et 97,33% de F1, respectivement (Tableau 3.2). Les mots en vietnamien ont tendance à être transcrits de la même manière que les noms étrangers, avec une prononciation similaire. C'est une des raisons pour lesquelles nous pouvons mesurer la similarité syllabique en comptant le nombre de syllabes entre le terme source et le terme cible. Quelques exemples sont illustrés dans le Tableau 3.3.

Tableau 3.3 Quelques exemples montrant une similarité de syllabes entre le terme source, en français, et le terme cible, en vietnamien.

En français	→	**En vietnamien**	**Commentaire**
Phillipines	→	Phi-lip-pin	*nom de lieu ayant 3 syllabes*
Vietnam	→	Việt Nam	*nom de lieu ayant 2 syllabes*
Singapour	→	Sing-ga-po	*nom de lieu ayant 3 syllabes*
Nicolas	→	Ni-cô-la	*nom de personne ayant 3 syllabes*
Joseph	→	Giô-sép	*nom de personne ayant 2 syllabes*

Afin d'évaluer la performance du système de la traduction automatique des entités nommées, nous utilisons une métrique d'évaluation automatique, le score de BLEU (*Bilingue Evaluation Understudy*) (Papineni *et al.*, 2002), ainsi qu'un taux des mots hors vocabulaires (Taux MHV) sur le corpus de test. Le Tableau 3.4 montre les résultats obtenus des expérimentations.

Tableau 3.4 Résultats des expérimentations de traduction automatique des entités nommées pour la paire de langues peu dotée français-vietnamien.

Expérimentations	**BLEU (%)**	**Taux MHV (%)**
Exp1(*système de base*)	21,68	29,20
Exp2	24,05 (*+2,37*)	25,90 (*-3,30*)
Exp3	**26,66** (*+4,98*)	**21,40** (*-7,80*)

Les trois évaluations ci-dessus montrent que la performance du système de traduction automatique des entités nommées bilingues s'améliore nettement avec 21,68% (Exp1) → 24,05% (Exp2, *+2,37*) → 26,66% (Exp3, *+2,61*). Nous remarquons qu'il reste encore des mots inconnus dans la première expérimentation. Par conséquent, la métrique de BLEU est atteinte seulement à 21,68% en raison du bruit des mots inconnus. Ensuite, les deuxième et troisième expérimentations ont été uniquement effectuées avec les entités nommées extraites du corpus de test, ayant 359 entités nommées. Les métriques d'évaluation de ces deux systèmes d'Exp2 et d'Exp3 montrent des améliorations de performance avec des gains de +2,37 BLEU ET de +4,98 BLEU par rapport au système de base, respectivement.

De plus, nous observons que les taux des mots hors vocabulaires sont également réduits avec l'augmentation des données d'apprentissage. Les résultats obtenus sont encourageants, avec 29,20% (Exp1) → 25,90% (Exp2, *-3,30*) → 21,40% (Exp3, *-4,50*). Pour mieux comprendre les résultats obtenus lors des expérimentations, nous effectuons une analyses des erreurs.

3.4.2 Analyse des erreurs

À travers différentes expérimentations citées ci-dessus, nous constatons qu'il y a encore des lacunes et des erreurs dans la traduction des entités nommées bilingues. Un inconvénient majeur d'un système basé sur le modèle de traduction automatique statistique implique la quantité de données d'apprentissage. Les données d'apprentissage doivent être aussi grandes que possible afin de couvrir toutes les variantes linguistiques de traduction.

Les erreurs de traduction sont classées selon trois critères concernant différents niveaux de (1) lexique, de (2) syntaxe et de (3) transcription / translittération.

(1) Les erreurs lexicales concernent le manque de mots. Le système traite les mots hors vocabulaires, les mots manquants ou les mots incorrects.
Par exemple :
(fr) fleuve Rouge → (vi) sông Đỏ # au lieu de « sông Hồng »
(fr) Mékong → (vi) Mékong # au lieu de « Cửu Long »
(fr) Long Biên → (vi) Dài Biên # au lieu de « Long Biên »

(2) Les erreurs syntaxiques concernent la traduction erronée des structures de noms ou l'ordre des mots dans les syntagmes nominaux.
Par exemple :
(fr) Asie du Sud-Est → (vi) Á của Đông Nam # au lieu de « Đông_Nam_Á»
(fr) Afrique → (vi) Phi Châu # au lieu de « Châu Phi »
(fr) Asie orientale → (vi) Châu Đông_Á # au lieu de « Á Đông »

(3) Les erreurs de transcription / translittération concernent des noms propres qui sont mal transcrits ou mal translittérés par le système de traduction automatique en raison de l'influence des mots en anglais durant la traduction manuelle.
Par exemple :
(fr) Singapore → (vi) Singapore # au lieu de « Singapour → Sing-ga-po »

(fr) Algeria → (vi) An-giê-ri # au lieu de « Algérie »

(fr) Californie → (vi) California # au lieu de « Ca-li-pho-ni-a »

(fr) Malaysie → (vi) Malaysia # au lieu de « Malaisie → Ma-lai-xi-a » or « Mã Lai »

3.5 Conclusion

Dans ce chapitre, nous avons présenté une approche itérative de l'extraction des entités nommées bilingues françaises-vietnamiennes par projection translinguistique. Cela a donné lieu à un petit corpus bilingue annoté qui servirait à la traduction automatique des entités nommées bilingues en français-vietnamien.

Dans le chapitre suivant, nous allons concevoir un modèle de reconnaissance des entités nommées pour une langue peu dotée de ressources linguistiques informatisées, avec une étude de cas en vietnamien.

CHAPITRE IV

MODÈLE DE RECONNAISSANCE DES ENTITÉS NOMMÉES POUR UNE LANGUE PEU DOTÉE

4.1 Introduction

Les entités nommées (ENs), en particulier les noms de personne (PER), les noms de lieu (LOC) et les noms d'organisation (ORG), jouent un rôle important dans de nombreuses applications de TALN, notamment dans l'extraction d'informations, la recherche d'informations, le résumé automatique de texte et la traduction automatique. La reconnaissance des entités nommées (REN) est une tâche fondamentale du TALN. Depuis 1995, un grand nombre de systèmes de REN ont été développés pour de multiples langues (Palmer et Day, 1997). Malgré sa longue histoire, la tâche de REN demeure encore l'une des tâches les plus difficiles à résoudre.

D'ailleurs, de nombreuses approches récentes, en citant quelques travaux de recherche (Chiu et Nichols, 2015; Lample *et al.*, 2016; Yao et Huang, 2016) ont démontré que l'architecture de LSTM bidirectionnel, qui peut représenter un mot par son contexte dans les deux directions d'une même phrase, convient parfaitement pour les tâches d'annotation telles que l'étiquetage des catégories grammaticales, la reconnaissance des entités nommées, etc.

La plupart des approches doivent utiliser plusieurs données supplémentaires pour améliorer la performance de reconnaissance. Comme il n'y a pas beaucoup de ressources linguistiques annotées en vietnamien, notre objectif consiste à concevoir et à implémenter un modèle de REN, qui utilise des informations supplémentaires provenant de ressources linguistiques non annotées. En fait, ce type de ressources linguistiques (par exemple, textes formels) peuvent être facilement collectées sur Internet. De plus, le modèle devrait avoir la capacité de rajouter davantage de données d'apprentissage, qui seront recueillies plus tard. À partir de données textuelles non étiquetées, le principe consiste à entraîner un modèle de plongements de mots (Mikolov *et al.*, 2013) permettant d'apprendre la représentation vectorielle de mots. Les vecteurs de mots résultants peuvent ensuite être utilisés en tant que caractéristiques additionnelles dans une architecture à base de réseaux de neurones pour améliorer les performances de cette architecture.

Ce chapitre présente notre méthodologie de conception d'un modèle à base de réseaux de neurones bidirectionnels, en combinant une petite quantité de ressources linguistiques annotées avec des plongements de mots pré-entraînés pour avoir un meilleur résultat dans la tâche de REN d'une langue peu dotée. Par la suite, nous allons comparer notre modèle avec plusieurs techniques de REN de l'état de l'art.

Le reste de ce chapitre est organisé comme suit : La section 4.2 présente l'approche proposée. Ensuite, les expérimentations et l'évaluation sont rapportées dans les sections 4.3 et 4.4, avec une analyse des erreurs susceptibles de notre modèle. Enfin, la section 4.5 conclut le chapitre.

4.2 Approche proposée

4.2.1 Extraction des traits linguistiques

Notre système se base sur l'architecture de LSTM, en combinant des traits linguistiques et des plongements de mots pré-entraînés. Nous allons aborder les traits linguistiques à extraire, et par la suite les plongements de mots au sein de notre système. Ils constituent les facteurs linguistiques additionnels à incorporer au niveau des mots, notamment :

- Surface de mot : Il s'agit de l'orthographe d'un mot, ici *token*, contenant différents constituants, à savoir des lettres en majuscule ou en minuscule, au singulier ou au pluriel, des ponctuations, des chiffres ou d'autres caractères spécifiques (par exemple, URL - adresse Web).
- Lemme : En linguistique, nous nous intéressons au lemme, qui est considéré comme une unité autonome constituante du lexique d'une langue.
- *Word cluster* (grappe de mots) : Les représentations de mots non supervisées, par exemple, *clusters de Brown* (Brown *et al.*, 1992), ont montré qu'elles amélioraient les performances des modèles supervisés du TALN (Turian *et al.*, 2010; Ritter *et al.*, 2011; Owoputi *et al.*, 2013). Dans ce travail, nous avons construit des traits à partir du regroupement hiérarchique de mots sur un large corpus non annoté, dans un ensemble de base de *clusters* fixé à 1 000. Ce trait est généré en utilisant l'algorithme *k-means*. Nous utilisons une boîte d'outils de *word2vec* [1].
- Étiquettes morpho-syntaxiques : Il s'agit d'étiquettes de catégories grammaticales, à savoir adjectif, adverbe, verbe, nom, etc.
- *Chunk* (segment syntaxique) : Ce sont des étiquettes de segment syntaxique, à savoir syntagme nominal, syntagme verbal, syntagme adverbial, etc.

1. `https://code.google.com/archive/p/word2vec/`

- Entités nommées : Il s'agit de classes manuellement annotées concernant chaque *token*, à savoir B-classe (le début d'une classe), I-classe (l'intérieur d'une classe) et O (hors classe).

Tous les traits ont été sélectionnés en raison de leur pertinence dans plusieurs tâches du TALN, telles que l'étiquetage de catégories grammaticales, l'annotation de segments syntaxiques et la reconnaissance des entités nommées (Le *et al.*, 2016). Le Tableau 4.1 illustre différents traits linguistiques dans une phrase en vietnamien.

Tableau 4.1 Illustration des traits linguistiques, à savoir surface de mot, lemme, grappe de mot (*word cluster*), étiquette morpho-syntaxique, *chunk* (segment syntaxique) et ENs (entités nommées). Ici, *en vietnamien : « Chỉ_có Pháp và Anh ủng_hộ đề_xuất của Fischler. », en français : « Seules la France et la Grande-Bretagne ont soutenu la proposition de Fischler. »*

Surface de mot	**Lemme**	**Grappe de mot**	**Étiquette morpho-syntaxique**	***Chunk***	**ENs**
Chỉ_có	chỉ_có	011000111	RB	I-NP	O
Pháp	pháp	111001	NNP	I-NP	I-LOC
và	và	0011110	CC	I-NP	O
Anh	anh	111001	NNP	I-NP	I-LOC
ủng_hộ	ủng_hộ	0110011011	VBD	I-VP	O
đề_xuất	đề_xuất	01010111	NN	B-NP	O
của	của	001110	POS	I-NP	O
Fischler	fischler	0110100	NNP	I-NP	I-PER
.	.	0010	.	O	O

4.2.2 Modèle neuronal bidirectionnel

Notre modèle de REN se base sur une architecture de LSTM bidirectionnelle. Elle opère sur une séquence de données dans les deux directions, avant et arrière, respectivement.

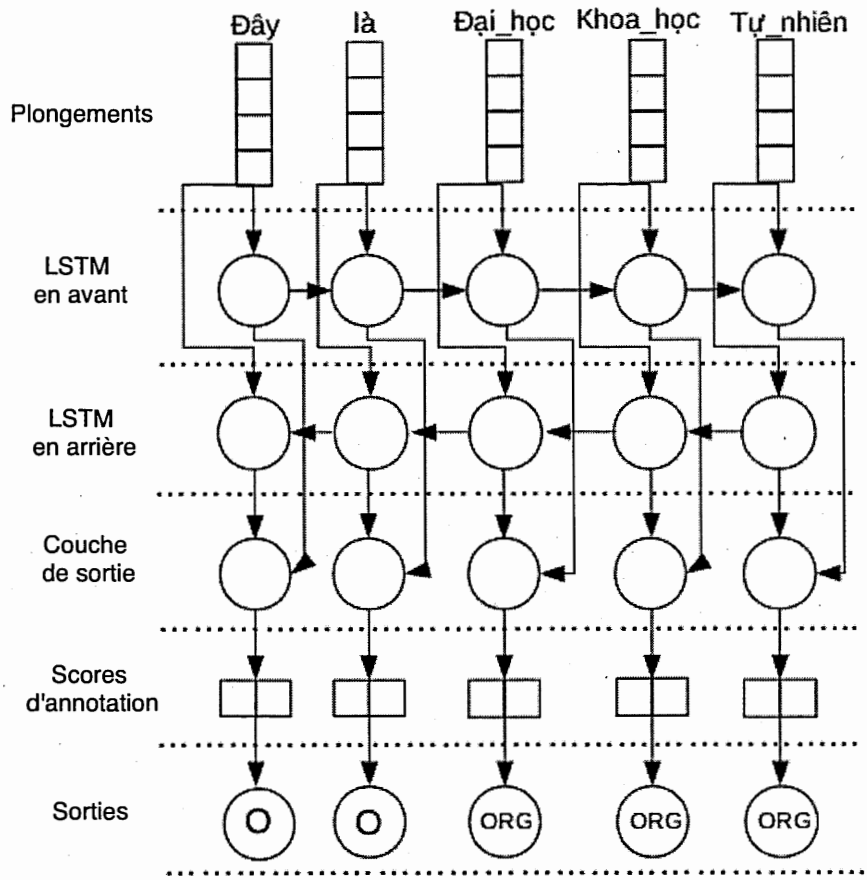

Figure 4.1 Architecture du modèle de reconnaissance des entités nommées en vietnamien à base de réseaux de neurones (*Traduction de l'illustration en français : « Voici l'université des sciences naturelles »*).

La Figure 4.1 illustre l'architecture de notre modèle pour la tâche de REN pour une langue peu dotée.

Tout d'abord, un modèle de plongement de mots est entraîné en utilisant un large corpus d'apprentissage monolingue et non annoté.

Ensuite, chaque mot dans les séquences d'entrée est représenté comme un des vecteurs de mots individuels appliquant ce modèle de plongement de mots pré-entraîné. Puis ces séquences de vecteurs de mots sont injectées dans notre architecture de LSTM bidirectionnel. Si $\overrightarrow{h_j}$ est la représentation récapitulative d'un mot à la position j-ème à partir de la gauche d'une séquence, $\overleftarrow{h_j}$ sera la représentation récapitulative de ce mot à partir de la droite de la séquence. Cela signifie que la combinaison de $h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}]$ résume la représentation du mot entier j. Il s'agit donc des vecteurs de mots contextuels.

Enfin, ces derniers sont projetés dans la couche de sortie, qui constitue une couche de réseau de neurones, avec une fonction d'activation de *softmax*, afin de produire des scores d'annotation.

4.3 Expérimentations

Cette section présente la configuration des expérimentations, les résultats obtenus avec les évaluations en comparant notre méthode avec diverses méthodes dans la tâche de REN pour une langue peu dotée, dans un cas d'étude sur la langue vietnamienne. Puis nous présentons également une analyse des erreurs susceptibles d'être produites par notre modèle.

4.3.1 Préparation des données

Pour entraîner un modèle de plongement de mots, nous avons utilisé un large corpus monolingue non étiqueté (extrait depuis les sites Web d'actualité), contenant 17 095 994 phrases en vietnamien (équivalent à 372 720 876 *tokens*). Ce corpus a été automatiquement segmenté au niveau des mots par la boîte à outils CLC_VN_WS [2].

En ce qui concerne la tâche de REN, nous avons utilisé un autre corpus d'apprentissage, contenant 12 000 phrases en vietnamien (équivalent à 311 255 *tokens*). Ce corpus a été manuellement segmenté et annoté. Le tableau 4.2 décrit les statistiques des entités nommées appartenant à chaque classe réciproque.

Tableau 4.2 Statistiques du corpus annoté par chaque classe d'entités nommées.

Classe	**Quantité**	**Classe**	**Quantité**
ABB	1 137	DTM	2 684
ABB_DES	84	LOC	4 299
ABB_LOC	372	MEA	2 618
ABB_ORG	686	NUM	4 754
ABB_TRM	69	ORG	3 279
ABB_TTL	8	PER	4 161
BRN	220	TRM	190
DES	1 029	TTL	988

En l'occurrence, nous avons obtenu 26 578 entités nommées réparties en plusieurs classes, à savoir personne (PER), organisation (ORG), lieu (LOC), abréviation (ABB), branche (BRN), désignation (DES), date et heure (DTM), mesures (MEA), terminologie (TRM) et titre (TTL).

---

2. http ://www.clc.hcmus.edu.vn/

Dans nos expérimentations, nous divisons le corpus d'apprentissage en 500 phrases pour le corpus de développement et 500 autres phrases pour le corpus de test. Le reste du corpus d'apprentissage sert à la phase d'entraînement de notre modèle.

4.3.2 Configuration

Dans ce travail, nous adaptons l'implémentation de (Lample *et al.*, 2016) [3], dans un cas d'étude pour une langue peu dotée dont le vietnamien.

Pour entraîner un modèle de plongement de mots, nous avons utilisé la boîte d'outils de *word2vec* [4]. L'entraînement de ce modèle a été réalisé avec le modèle de *skip-gram* qui est plus lent que le modèle de *CBOW*, mais préférable pour traiter des mots peu fréquents. La taille de dimension du modèle est fixée à 100 avec une taille de fenêtre de contexte de 40 mots.

Dans le cas d'étude des données d'apprentissage limitées, un réseau neuronal engendre généralement des problèmes de surapprentissage (*en anglais : overfitting*) en raison de leur capacité à apprendre des relations compliquées entre les données d'entrée et les données de sortie. Lample *et al.* (2016) ont déduit que l'utilisation d'un taux d'abandon, qui consiste en une technique de régularisation pour réduire le surapprentissage dans les réseaux neuronaux, pourrait, à la couche finale, améliorer de manière significative les performances d'un système de REN. Ainsi, le taux d'abandon (*en anglais : dropout rate*) est fixé à 0,5 dans nos expérimentations.

Nous avons entraîné notre modèle en utilisant l'algorithme du gradient stochastique (*en anglais : stochastic gradient descent*) (Rumellhart, 1986) avec un taux d'apprentissage

---

3. `https://github.com/glample/tagger/`

4. `https://code.google.com/p/word2vec/`

fixe de 0,01 et une taille d'échantillon de 32. Les pondérations du modèle sont initiées de manière aléatoire et le nombre d'itérations est fixé à 100. Nous avons également essayé la descente en gradient mini-batch mais la performance ne semble pas changer. Nous avons décidé de choisir les hyper-paramètres selon notre étude de l'état de l'art et après avoir testé d'autres configurations.

Dans nos expérimentations, nous utilisons les boîtes à outils de Yamcha[5] et de CRF++[6], qui permettent d'implémenter les algorithmes de SVM et de CRF, respectivement, à appliquer à l'ensemble des corpus d'apprentissage. Les quatre systèmes de REN ont été configurés comme suit :

(1) Système 1 : Il s'agit d'un système de base, avec l'algorithme d'apprentissage de SVM (*Support Vector Machine*).

(2) Système 2 : Il s'agit d'un système de base, avec l'algorithme d'apprentissage de CRF (*Conditional Random Fields*).

(3) Système 3 : Il s'agit d'un système de base, dit Bi-LSTM, qui se base sur l'architecture de LSTM bidirectionnelle (Lample *et al.*, 2016), seulement avec les plongements de mots et de caractères.

(4) Système 4 : Il s'agit de notre approche, qui se base également sur l'architecture de Bi-LSTM, avec tous les traits linguistiques additionnels.

---

5. `http://chasen.org/~taku/software/yamcha/`

6. `https://taku910.github.io/crfpp/`

4.4 Évaluations

4.4.1 Résultats obtenus

En ce qui concerne l'évaluation des résultats obtenus, nous appliquons le script *conlleval.pl*[7], qui est souvent utilisé dans les campagnes d'évaluation de la tâche de REN, afin d'évaluer, de manière automatique, les prédictions d'annotation produites par les systèmes de REN, avec des métriques d'évaluation de performances telles que la précision (P), le rappel (R) et le score de F1.

Nous avons effectué diverses expérimentations sur la tâche de REN pour la langue peu dotée vietnamienne. Nous les avons ensuite comparées avec le travail de recherche de Pham *et al.* (2007), dans lequel ils ont appliqué les méthodes de SVM et CRF, et ils ont rapporté que la méthode de CRF a été plus performante que la méthode de SVM dans la tâche de REN pour la langue vietnamienne.

Tableau 4.3 Comparaison de performances entre différents systèmes de reconnaissance des entités nommées, notamment *SVM : Support Vector Machine* (en français : Machine à vecteurs de support), *CRF : Conditional Random Fields* (en français : Champs aléatoires conditionnels) et *Bi-LSTM : Bidirectional Long-Short Term Memory* (Mémoire à long-court terme bidirectionnel)

Expérimentations	**Précision (%)**	**Rappel (%)**	**F1 (%)**
Système 1 (SVM)	85,23	78,02	81,46
Système 2 (CRF)	**86,70**	79,54	82,97
Système 3 (Bi-LSTM)	81,08	83,50	82,27
Système 4 (notre approche)	84,53	**87,93**	**86,20**

7. `https://www.clips.uantwerpen.be/conll2000/chunking/output.html`

Nous constatons que les performances des systèmes de SVM, de CRF et de Bi-LSTM sont comparables en terme de scores de F1. Tandis que notre modèle a réalisé une amélioration significative en terme de scores de F1, avec des gains de $+3,23\%$, de $+4,74\%$, de $+3,93\%$ par rapport aux systèmes de CRF, de SVM et de Bi-LSTM, respectivement (voir Tableau 4.3). Le Tableau 4.4 illustre quelques exemples des prédictions produites par notre modèle de reconnaissance des entités nommées pour la langue peu dotée vietnamienne.

Tableau 4.4 Quelques exemples des prédictions produites par notre modèle de reconnaissance des entités nommées pour la langue peu dotée vietnamienne.

Trao_đổi với VnExpress , ông Trần_Quý_Tường , Cục phó Cục khám_chữa_bệnh (Bộ Y_tế) cho rằng , mức chuẩn được điều_khiển xe_máy , ô_tô vừa ban_hành , thấp hơn nhiều so với chiều cao , cân nặng trung_bình của người VN .
Trao_đổi__O với__O **VnExpress__B-ORG** ,__O **ông__B-TTL Trần_Quý_Tường__B-PER** ,__O **Cục__phó__B-DES Cục__B-ORG** khám_chữa_bệnh__I-ORG (__O **Bộ__B-ORG Y_tế__I-ORG**)__O cho__O rằng__O ,__O mức__O chuẩn__O được__O điều_khiển__O xe_máy__O ,__O ô_tô__O vừa__O ban_hành__O ,__O thấp__O hơn__O nhiều__O so__O với__O chiều__O cao__O ,__O cân__O nặng__O trung_bình__O của__O người__O **VN__B-ABB_LOC** .__O
Bộ GD-ĐT cho_biết ngày thi thứ hai của kì thi tốt_nghiệp THPT năm 2009 có thêm 842 thí_sinh bỏ thi .
Bộ__B-ABB_ORG GD-ĐT__I-ABB_ORG cho_biết__O ngày__O thi__O thứ__O hai__O của__O kì__O thi__O tốt_nghiệp__O **THPT__B-ABB năm__B-DTM 2009__I-DTM** có__O thêm__O **842__B-NUM** thí_sinh__O bỏ__O thi__O .__O
Ngày 4-5 , các trường THCS tại TP. HCM bắt_đầu phát các ấn_phẩm thông_tin về quy_chế , chỉ_tiêu tuyển_sinh và mẫu .
Ngày__B-DTM 4-5__I-DTM ,__O các__O trường__O **THCS__B-ABB** tại__O **TP.__B-ABB_LOC HCM__I-ABB_LOC** bắt_đầu__O phát__O các__O ấn_phẩm__O thông_tin__O về__O quy_chế__O ,__O chỉ_tiêu__O tuyển_sinh__O và__O mẫu__O .__O

De plus, nous examinons également comment ces approches ont impacté sur la qualité de la reconnaissance de nouvelles entités nommées, qui n'apparaissent pas encore dans les données d'apprentissage.

Nous constatons la pertinence de l'utilisation d'un modèle pré-entraîné de plongement de mots sur un large corpus monolingue non annoté sur les séquences d'entrée.

4.4.2 Analyse des erreurs

Les résultats obtenus, dans la sous-section 4.4.1 ci-dessus, ont montré une amélioration significative de la tâche de REN, dans un cas d'étude sur la langue peu dotée vietnamienne. Cependant, nous constatons des erreurs de prédiction d'annotation produites par notre modèle.

Premièrement, la performance du système de REN a été impactée par des entités nommées peu fréquentes. Le Tableau 4.5 montre que l'entité nommée « *khoa Kế_hoạch gia_đình* » a été manuellement annotée avec l'étiquette de LOC, pour un lieu. Alors que le modèle a produit une erreur de prédiction en l'annotant avec une autre étiquette «O», pour hors de classe. Une des raisons pour lesquelles le modèle a échoué, consiste en une nouvelle entité nommée qui n'est jamais apparue dans la phase d'apprentissage. Par contre, le modèle a bien produit une prédiction pour l'entité nommée « *Bệnh_viện phụ_sản Từ_Dũ* » en l'annotant avec l'étiquette de ORG, pour une organisation.

Deuxièmement, certaines catégories d'entités nommées (par exemple ORG, LOC) sont difficiles à être correctement annotées parce qu'elles peuvent être composées d'un nom de personne et d'un nom commun. D'ailleurs, il se peut que le texte utilise un nom de personne pour désigner un lieu ou une organisation. Un autre exemple est illustré dans le Tableau 4.6. Nous constatons que les deux entités nommées « *Jetstar* » et « *Simon_Westaway* » ont été manuellement annotées avec les étiquettes de ORG (pour une

Tableau 4.5 Illustration des erreurs de prédiction de nouvelles entités nommées.

Theo Vv O O		
thống_kê Vv O O	**613 X B-NUM B-NUM**	
của Nn O O	em Nn O O	tại Cm O O
riêng Aa O O	dưới Nn O O	**khoa Nn B-LOC O**
Bệnh_viện Nn B-ORG B-ORG	**18 X B-MEA B-MEA**	**Kế_hoạch Nn I-LOC O**
phụ_sản Nn I-ORG I-ORG	**tuổi Nn I-MEA I-MEA**	**gia_đình Nn I-LOC O**
Từ_Dũ X I-ORG I-ORG	đến Vd O O	của Nn O O
, PU O O	xin Vv O O	bệnh_viện Nn O O
năm Nt B-DTM B-DTM	bỏ Vv O O	. PU O O
2007 X I-DTM I-DTM	thai Nn O O	
có Ve O O		

organisation) et de PER (pour une personne), respectivement. Le modèle a produit des erreurs de prédiction, en annotant un nom d'organisation au lieu d'un nom de personne.

Tableau 4.6 Illustration des erreurs de prédiction d'annoter un nom d'organisation au lieu d'un nom de personne.

	một Nq B-NUM B-NUM	để Vv O O
« PU O O	trong Nn O O	dập Vv O O
Khi Nt O O	số Nn O O	lửa Nn O O
khói Nn O O	các Nq O O	» PU O O
bốc Vv O O	phi_công Nn O O	, PU O O
lên Vd O O	được Vv O O	**phát_ngôn_viên Nn B-DES B-DES**
trong Nn O O	lệnh Nn O O	**Jetstar X B-ORG I-ORG**
buồng_lái Nn O O	dùng Vv O O	**Simon_Westaway X B-PER I-ORG**
, PU O O	bình Nn O O	cho_biết Vv O O
	cứu_hoả Vv O O	. PU O O

4.5 Conclusion

Dans ce chapitre, nous avons étudié l'importance des traits linguistiques additionnels dans un modèle à base de réseaux de neurones, dans la tâche de reconnaissance des entités nommées pour une langue peu dotée, avec une application sur le vietnamien.

Par rapport au travail de Lample *et al.* (2016), notre contribution consiste à extraire différents traits linguistiques additionnels et à les combiner au sein de notre modèle de reconnaissance des entités nommées. Cela a pu augmenter considérablement la performance du modèle de reconnaissance des entités nommées. Par ailleurs, nous avons démontré l'utilité de l'intégration des plongements de mots pré-entraînés permettant de contourner le manque de ressources linguistiques annotées. En outre, ces traits utilisés dans notre étude sont faciles à établir grâce aux boîtes à outils existantes et en libre-service.

Dans le chapitre suivant, nous allons concevoir un modèle de translittération des entités nommées pour une paire de langues peu dotées, avec une étude de cas du français-vietnamien.

CHAPITRE V

MODÈLE DE TRANSLITTÉRATION DES ENTITÉS NOMMÉES POUR UNE PAIRE DE LANGUES PEU DOTÉE

5.1 Introduction

La translittération consiste en un processus de transformation d'un mot d'un système d'écriture (appelé *mot source*) en un mot phonétiquement équivalent dans un autre système d'écriture (appelé *mot cible*) (Knight et Graehl, 1998). Cette transformation nécessite un ensemble important de règles pré-définies par des linguistes experts pour déterminer comment les phonèmes sont alignés et tenir compte du système phonologique de la langue cible. De nombreuses paires de langues ont adopté des règles différentes concernant la tâche de translittération au fil du temps et la plupart des systèmes de translittération dépendent de l'origine d'un mot à translittérer (Waxmonsky et Reddy, 2012).

Dans de nombreux domaines, y compris la traduction automatique, les modèles d'apprentissage profond sont devenus une bonne alternative à l'approche plus traditionnelle à base de statistiques (Koehn, 2017; Wu *et al.*, 2016). Ainsi notre motivation consiste à adopter une approche similaire ayant pour objectif de construire un système de translittération pour une paire de langues peu dotée.

Nous observons que la méthode d'état de l'art de conversion de graphème en phonème (Bisani et Ney, 2008) est basée sur l'utilisation d'alignement (ou mappage) de graphème en phonème. Mais les approches de réseaux de neurones récurrents n'utilisent aucune information d'alignement. Ainsi nous proposons une méthode de construction d'un système de translittération pour une paire de langues peu dotée en utilisant des modèles à base de réseaux de neurones récurrents et les informations d'alignement pour les séquences d'entrée. Ce système permet de prédire automatiquement la représentation graphémique d'un mot dans la langue cible étant donné un nouveau mot dans la langue source, qui n'existe pas dans le dictionnaire de prononciation bilingue. Nous nous intéressons à la résolution de mots hors vocabulaire, qui sont considérés comme des noms propres ou des termes techniques issus d'un autre système de traduction automatique pour une paire de langues peu dotée, avec une étude de cas du français-vietnamien.

Dans le cadre de ce travail de recherche, nos contributions consistent à :

(1) Explorer l'architecture des réseaux de neurones récurrents dans la tâche de translittération.

(2) Appliquer les informations d'alignement afin de développer d'autres représentations pour les séquences d'entrée et utiliser une intégration des plongements de source / cible pré-entraînés dans la couche d'entrée ayant pour objectif de construire des modèles de translittération neuronaux en se basant sur ces nouvelles formes de représentation.

(3) Appliquer plusieurs techniques de la traduction automatique neuronale telles que l'architecture d'encodeur-décodeur, le mécanisme d'attention afin de construire un système de translittération dans des configurations possédant peu de ressources linguistiques.

(4) Évaluer les différentes expérimentations sur la paire de langues peu dotée français-vietnamien. Les expérimentations ont montré des améliorations consis-

tantes de la performance des systèmes de translittération malgré de petites quantités de données d'apprentissage.

Le reste de ce chapitre est organisé comme suit : La section 5.2 présente l'approche proposée. Ensuite, la section 5.3 aborde les expérimentations. La section 5.4 montre les performances de notre système versus d'autres systèmes, ainsi qu'une analyse des erreurs probables. Enfin, la section 5.5 conclut le chapitre.

5.2 Méthodologie

5.2.1 Approche proposée

Notre approche consiste en trois étapes principales : *(1) pré-traitement, (2) modification des séquences d'entrée basée sur la représentation d'alignement et (3) création d'un système de translittération à base de réseaux de neurones récurrents.* L'ensemble du processus est illustré dans la Figure 5.1.

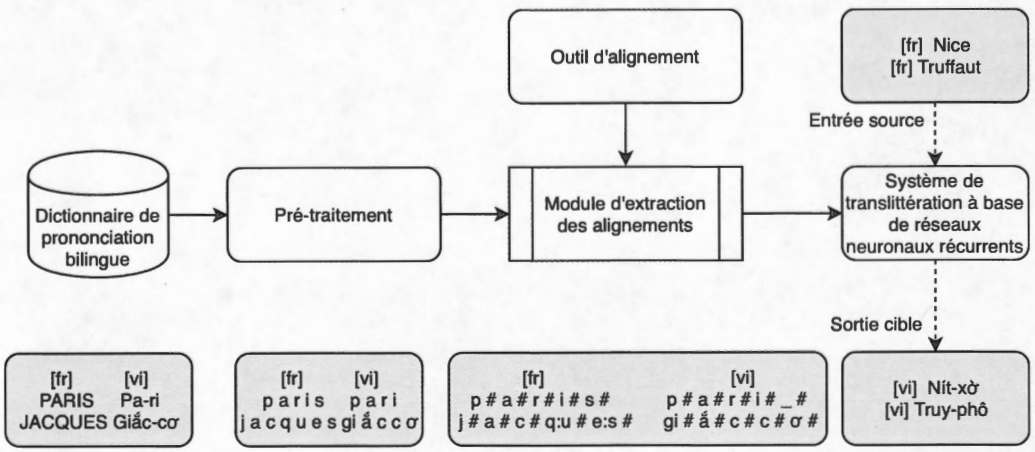

Figure 5.1 Architecture du système de translittération à base de réseaux de neurones récurrents pour une paire de langues peu dotée, traitant des entités nommées bilingues.

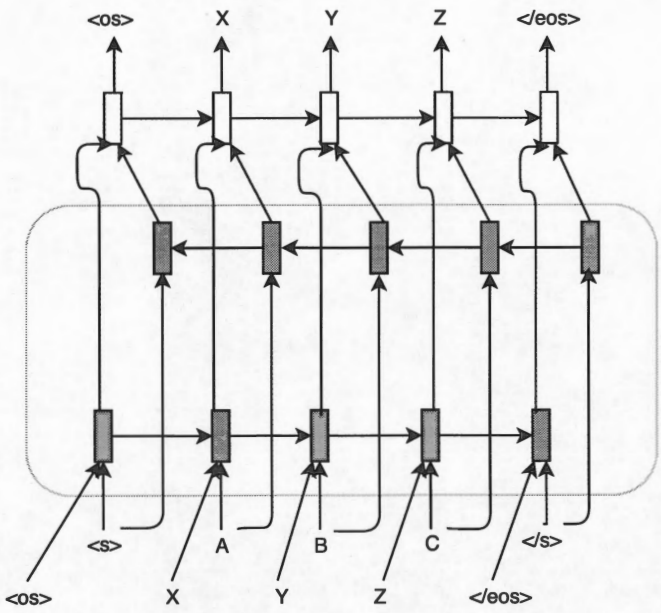

Figure 5.2 Notre architecture de modèle à base de réseaux de neurones récurrents avec un encodeur-décodeur de LSTM bidirectionnel et la représentation d'alignement sur les séquences d'entrée. Nous utilisons les marqueurs <s> et </s>, <os> et </eos> pour remplir les séquences graphémiques à une longueur fixe.

(1) Tout d'abord, nous collectons un dictionnaire de prononciation bilingue pour une paire de langues peu dotée, une étude de cas du français-vietnamien. Ensuite, ces données d'apprentissage sont pré-traitées avec une normalisation en minuscule, une suppression des traits d'union des syllabes et une segmentation de toutes les syllabes au niveau des caractères.

(2) Nous créons un module d'extraction des alignements à partir du dictionnaire de prononciation bilingue. Et les séquences d'entrée sont modifiées en fonction des résultats d'alignement (Figure 5.2).

(3) Nous entraînons, par la suite, un système de translittération à base de réseaux de neurones récurrents.

5.2.2 Alignements graphémiques pour une langue peu dotée

Étant donné une séquence de graphèmes d'un mot et une autre séquence correspondant à sa prononciation, une stratégie d'alignement consiste à associer chaque graphème à son correspondant, souvent de longueur 1. En d'autres termes, les séquences graphémiques ont été alignées pour former des unités conjointes. Inspirant de (Yao et Zweig, 2015), en ce qui concerne ce type d'alignements, un graphème peut correspondre à un graphème nul ou un graphème simple ou un graphème composé, ayant au moins deux graphèmes. Quelques illustrations sont données dans le Tableau 5.1.

Tableau 5.1 Quelques illustrations d'alignements graphémiques. Le graphème G s'aligne sur un graphème simple (1-à-1), un graphème composé (1-à-2), deux graphèmes sur un graphème simple (2-à-1) et plusieurs graphèmes sur plusieurs graphème (m-à-m). La lettre 's' est alignée sur un graphème nul '_' qui n'est pas prononcé, dans l'exemple de 'PARIS'.

Alignements	**1-to-1**	**1-to-2**	**2-to-1**	**m-to-m**
Source	p | a | r | i | s |	n | i | c | e |	j | a | c | qu | es |	tr | u | ff | au | t |
Cible	p | a | r | i | _	n | ít | x | ɔ̃ |	gi | ǎ | c | c | ɑ |	tr | uy | ph | ô | _ |

Étant donné une séquence de graphèmes source $G = \{g_1, g_2, ..., g_N\}$, une autre séquence de graphèmes cible correspondante $P = \{p_1, p_2, ..., p_N\}$ et un alignement A, la probabilité postérieure $p(P|G,A)$ est estimée comme suit :

$$p(P|G,A) \approx \prod_{n=1}^{N} p(p_n | p_{n-1}^{n-k}, g_{n-k}^{n+k}) \tag{5.1}$$

où k est la taille de la fenêtre de contexte, n est l'index de position dans les alignements et deux séquences P et G possèdent la même longueur de N éléments.

Une stratégie d'alignement graphémique peut être automatiquement apprise à partir d'un dictionnaire de prononciation en utilisant une entropie maximale (Chen, 2003;

Barros et Weiss, 2006), dérivée de l'équation 5.1 ou une maximisation d'expectation (Jiampojamarn *et al.*, 2007). Aussi Bisani et Ney (2008) ont utilisé une telle stratégie d'alignement dans un modèle de séquences multi-jointes pour la conversion de graphème en phonème. Leur système a obtenu la meilleure précision en termes de taux d'erreurs de graphèmes.

En revanche, dans les modèles à base de réseaux de neurones récurrents, ces informations d'alignement ne sont pas disponibles pendant la phase de décodage, dans laquelle la séquence de graphèmes donnée n'indique aucun groupe aligné sur un graphème nul ou simple ou composé et vice versa. Nous observons que la combinaison des stratégies d'alignement de 1-à-2 et de 2-à-1, appelées *alignement bigramme*, peut résoudre le défi de la pénurie de ressources linguistiques. L'utilisation des alignements bigrammes permet de couvrir l'alphabet de sortie dans les données d'apprentissage.

Dans ce chapitre, nous utilisons les informations additionnelles d'alignement pour modifier les séquences d'entrée, parce que dans les configurations à faibles ressources, la représentation est cruciale pour les modèles à base de réseaux de neurones. D'autre part, les plongements de source et de cible seront pré-entraînés en se basant sur les données d'entraînement. Ces plongements sont considérés, pour la tâche de translittération, comme des mappages vectoriels linéaires entre la source et la cible. Ensuite, ils deviennent l'un des traits de la couche d'entrée. Nous nous attendons à ce que l'exploitation de cette représentation d'alignement améliore les performances du système de translittération.

5.3 Expérimentations

5.3.1 Préparation des données

En ce qui concerne les données d'apprentissage, nous collectons un dictionnaire de prononciation bilingue français-vietnamien, tel que présenté dans Cao *et al.* (2010). Les données d'apprentissage se composent de 4 259 paires d'entités nommées bilingues. Ce dictionnaire bilingue est filtré à partir d'un corpus de 146 millions de mots, provenant des journaux électroniques franco-vietnamiens disponibles entre Avril 2008 et Octobre 2009. Nous traitons les entités nommées, identifiées comme noms de personne, noms de lieu et noms d'organisation. Pour pallier le problème de la dispersion des données d'apprentissage, nous avons effectué la première étape du pré-traitement (Figure 5.3).

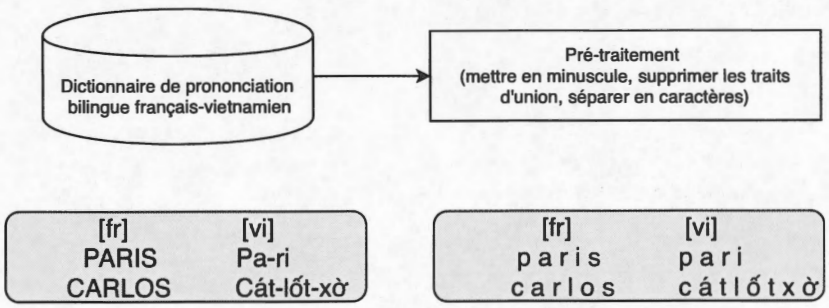

Figure 5.3 Illustration de la phase de pré-traitement du corpus d'apprentissage avec une normalisation en minuscule, une suppression des traits d'union des syllabes et une segmentation de toutes les syllabes au niveau des caractères.

5.3.2 Configuration

Pour traiter la représentation de l'alignement, nous avons utilisé l'outil de *m-2-m aligner*[1] (Jiampojamarn *et al.*, 2007) qui permet d'aligner les données d'apprentissage

1. `https://github.com/letter-to-phoneme/m2m-aligner/`

au niveau des caractères. Nous avons choisi $m = 2$ (*bigramme*), ce qui signifie qu'un maximum de 2 graphèmes du côté source seront alignés sur un maximum de 2 graphèmes du côté cible. Pour les plongements de source et de cible pré-entraînés, nous avons appliqué l'outil de *word2vec*[2] avec une dimension de 64, une taille de la fenêtre contextuelle de 5 grammes et l'option de *skip-gram*.

Afin d'entraîner notre modèle de translittération, nous avons appliqué la boîte à outils de *nmt-keras*[3] (Peris, 2017) dans une étude de cas d'une paire de langues peu dotée, le français-vietnamien. Cette boîte à outils permet d'implémenter l'architecture d'encodeur-décodeur, avec un mécanisme d'attention, l'intégration des plongements et surtout construire un système de translittération neuronal. Dans la configuration de notre système de translittération neuronal, nous avons utilisé deux couches de l'architecture d'encodeur-décodeur LSTM bi-directionnel, avec une couche d'entrée à 64 dimensions pour encoder les séquences d'entrée et 128 nœuds dans la couche cachée. Nous avons utilisé l'optimiseur « *Adam* » (Kingma et Ba, 2014) pour apprendre les poids du réseau avec un taux d'apprentissage par défaut de 0,001. Pour le décodage, la recherche en faisceau est configurée à la taille de fenêtre contextuelle de 6. Nous avons décidé de choisir les hyper-paramètres selon notre étude de l'état de l'art et après avoir testé d'autres configurations.

---

2. `https://code.google.com/archive/p/word2vec/`

3. `https://github.com/lvapeab/nmt-keras/`

5.4 Évaluations

5.4.1 Résultats obtenus

Le dictionnaire de prononciation bilingue, constituant le corpus d'apprentissage, est divisé en trois parties, notamment corpus d'apprentissage, corpus de développement et corpus de test avec un ratio de 90%, 5% et 5% respectivement.

Dans ce travail, nous allons construire un système de translittération neuronal qui s'inspire de la traduction automatique neuronale. Par conséquent, nous appliquons différentes métriques d'évaluation telles que *BLEU (BiLingual Evaluation Understudy)* (Papineni *et al.*, 2002), *taux d'erreurs de traduction (TER - Translation Error Rate)* (Snover *et al.*, 2009) et *taux d'erreurs de graphèmes (GER - Grapheme Error Rate)*. Ces métriques sont automatiquement évaluées avec le script de *multi-bleu.pl*[4] de *Moses*[5].

Afin d'évaluer l'approche proposée, nous avons implémenté cinq systèmes (Tableau 5.2) :

(1) Système de base A : Il s'agit d'un système de base appliquant l'approche de traduction automatique statistique à base de segments (appelé *pbSMT - phrase-based statistical machine translation*).

Nous avons implémenté le système de base A avec la boîte des outils *Moses* (Koehn *et al.*, 2007). Nous avons utilisé *mGIZA* (Gao et Vogel, 2008) pour aligner le corpus au niveau du caractère, et *SRILM* (Stolcke, 2002) pour créer un modèle de langage de 5 grammes basé sur des caractères pour la langue cible.

---

4. `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl`

5. `https://github.com/moses-smt`

(2) Système de base B : Il s'agit d'un autre système de base appliquant le modèle de séquences multi-jointes. Nous appliquons l'outil de *Sequitur-G2P* [6] pour entraîner un modèle de translittération.

(3) Système 1 : Il s'agit d'un système d'*encodeur-décodeur bidirectionnel LSTM* avec *un mécanisme d'attention.*

(4) Système 2 : Il s'agit d'un système d'*encodeur-décodeur bidirectionnel LSTM* avec *un mécanisme d'attention* et *une représentation d'alignement pour les séquences d'entrée.*

(5) Système 3 : Il s'agit d'un système d'*encodeur-décodeur bidirectionnel LSTM* avec *un mécanisme d'attention, une représentation d'alignement pour les séquences d'entrée* et *deux plongements de sources et de cibles pré-entraînés.*

Entre les deux systèmes de base A et B, la différence de performance est mineure. Le système de base B semble légèrement plus efficace que le système de base A, avec un gain de +4.40 points BLEU, ainsi que des réductions avec -3.58 points et -6.20 points d'erreurs de traduction (TER) et d'erreurs de graphèmes (GER) respectivement (Tableau 5.2).

Par ailleurs, en comparant les deux systèmes de base avec les trois systèmes 1, 2 et 3 (*l'approche proposée*), nous constatons des résultats significatifs avec des scores jusqu'à 95,96 points BLEU, des réductions d'erreurs de traduction (TER) et des erreurs de graphèmes (GER) jusqu'à 3,28 et 6,19 points respectivement (Tableau 5.2).

De plus, le système 3 a obtenu les performances les plus élevées par rapport aux deux autres systèmes 1 et 2, avec des gains de 3,55 et 1,55 points BLEU, des réductions de -6,41 et -1,42 points TER, -12,09 et -2,69 points GER respectivement (Tableau 5.2).

---

6. `https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html`

Tableau 5.2 Évaluation des expérimentations avec les métriques d'évaluation telles que *BLEU (BiLingual Evaluation Understudy)* (Papineni *et al.*, 2002), *taux d'erreurs de traduction (TER - Translation Error Rate)* (Snover *et al.*, 2009) et *taux d'erreurs de graphèmes (GER - Grapheme Error Rate).*

Expérimentations	**BLEU ↑**	**TER ↓**	**GER ↓**
Système de base A (pbSMT)	61,30	24,08	44,20
Système de base B (modèle de séquences multi-jointes)	65,70	20,50	38,00
Système 1 (encodeur-décodeur + attention)	92,38	9,69	18,28
Système 2 (encodeur-décodeur + attention + représentation d'alignement)	94,41	4,70	8,87
Système 3 (encodeur-décodeur + attention + représentation d'alignement + plongements de sources et de cibles pré-entraînés)	**95,96**	**3,28**	**6,19**

Dans les configurations à faibles ressources linguistiques, nous souhaitons effectuer la modification de la taille de la dimension pour trois systèmes de translittération neuronaux. Par conséquent, nous observons certains impacts lorsque la dimension de l'architecture d'encodeur-décodeur augmente de 64 à 128 (Tableau 5.3).

Les différences de performance au sein du système 1 via les expérimentations de deux différentes dimensions sont significatives, avec un gain de +4,69 points BLEU et des réductions de -4,37 points TER, -8,25 points GER. En outre, les systèmes 2 et 3 ont obtenu des performances comparables dans les expérimentations de deux différentes

Tableau 5.3 Évaluation sur l'architecture d'encodeur-décodeur de LSTM bi-directionnel, avec différentes dimensions de 64 et de 128.

	64 dimensions			128 dimensions		
Expérimentations	**BLEU ↑**	**TER ↓**	**GER ↓**	**BLEU ↑**	**TER ↓**	**GER ↓**
Système 1 (encodeur-décodeur + attention)	87,69	14,06	26,53	92,38	9,69	18,28
Système 2 (encodeur-décodeur + attention + représentation d'alignement)	93,37	5,60	10,57	94,41	4,70	8,87
Système 3 (encoder-decoder + attention + représentation d'alignement + plongements pré-entraînés)	**95,44**	**3,70**	**6,98**	**95,96**	**3,28**	**6,19**

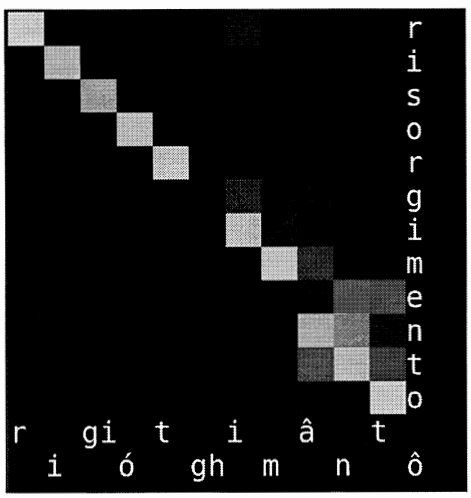

Figure 5.4 Illustration d'un alignement du mécanisme d'attention généré par le système de translittération à base de réseaux de neurones récurrents pour une paire de langues peu dotée, traitant des entités nommées bilingues, avec la source en français est « *risorgimento* » et la cible en vietnamien est « *rigiótghimântô* ».

dimensions en termes de scores de BLEU, de TER et de GER. Nous pouvons en déduire que la performance du système est améliorée lorsque la taille de la dimension augmente.

Nous remarquons également qu'un grand avantage du mécanisme de l'attention nous permet la capacité d'interpréter et de visualiser ce que fait le modèle. Par exemple, en visualisant la matrice de pondération de l'attention lorsqu'une entité nommée est traduite, nous pouvons comprendre comment le modèle traduit (Figure 5.4).

En général, l'approche proposée a très bien réalisé la tâche de translittération avec des gains significatifs et a pu réduire les taux d'erreurs de graphèmes. Nous pouvons observer que la qualité de sortie de notre système, qui est basée sur les réseaux de neurones récurrents, est plus fluide, plus cohérente, présentant moins d'erreurs que d'autres systèmes, qui sont basés sur l'approche à base de statistiques.

Tous les résultats expérimentaux ont montré que l'utilisation de la représentation d'alignement et de plongements de source et de cible pré-entraînés ont des impacts significatifs par rapport à d'autres méthodes. D'un autre côté, nous observons que l'architecture à base de réseaux de neurones est plus simple que l'architecture à base de statistiques, qui est composée de plusieurs modules individuels, de sorte que celle-ci se doit d'améliorer chaque module individuel indépendamment.

Par la suite, nous effectuons une analyse des erreurs probables, dans la section suivante, afin de décortiquer différents types d'erreurs générés par nos divers systèmes d'expérimentation.

5.4.2 Analyse des erreurs

Afin de mieux comprendre différents types d'erreurs concernant les graphèmes prédits de la source en français à la cible en vietnamien, nous avons effectué une analyse des erreurs possibles produites par les cinq systèmes d'évaluation. Par conséquent, nous

avons réalisé une comparaison des résultats de prédiction de la translittération des entités nommées entre les cinq systèmes d'évaluation avec des noms propres qui n'ont jamais été vus au cours de la phase d'apprentissage (Tableau 5.4).

Tableau 5.4 Exemples de résultats de prédiction de translittération par tous les systèmes d'évaluation, hypothèses et références pour six noms propres, tels que *PARIS, TIGRANE, TOULOUSE, TOURS, TRUFFAUT et ZURICH.*

Références	**Système de base A**	**Système de base B**	**Système 1**	**Système 2**	**Système 3**
p a r i	p a r í t	p a r í t	p a r i	p a r í : t x : ɑ	p a r i _
t i g ờ r a n n ɑ	t i g ờ r a n n ɑ	t i g ờ r a n n ɑ	t i g ờ r a n n ɑ	t i g : ờ r a n n : ɑ	t i g : ờ r a n n : ɑ
t u l u g i ɑ	t u l u x ɑ	t u l u x ɑ	t u l u g i ɑ	t u l u y s : ờ	t u l u y s : ɑ
t u a	t u ố c t x ɑ	t u ố c t x ɑ	t u ý t x ɑ	t : ɑ u y _ x : ɑ	t u a _ x : ɑ
t r u y p h ô	t r u y p h ố t	t r u y p h ố t	t r ú p p h ô	t r u y p h ô	t r u p h ô
g i u y r í c h	g i u y r i	g i u y r i	g i u y r í c h	g i u y r í : c h	g i u y r í : t

Les résultats ont montré que les systèmes de base A (*pbSMT*) et B (*modèle de séquences multi-jointes*) ont produit certaines erreurs de prédiction de la translittération des noms propres, alors que les trois autres systèmes 1, 2 et 3 (*l'approche proposée*) en ont produit de meilleurs résultats. Il est à noter que les systèmes de base ont rencontré des difficultés pour prédire de façon optimale toutes les hypothèses de translittération dues à la variété d'origine des noms propres (*français, anglais, italien, espagnol, portugais, russe, etc.*) ainsi que la prononciation de différentes syllabes finales (*coda*), muette ou prononcée, telles que :

- « -s » ou « -ce » (xɑ or ϕ) : *nice* → *nít-xɑ* ou *paris* → *pa-ri*
- « -x » ou « -ch » (ích ou ít ou ϕ) : *zurich* → *giuy-rích* ou *zurich* → *giuy-rít*
- et d'autres graphèmes comme « -u- » (/y/ = uy ou u) : *truffaut* → *tru-phô* ou *truffaut* → *truy-phô*

Tableau 5.5 Exemples de résultats de prédiction de translittération par deux systèmes d'évaluation, hypothèses et références pour six noms propres tels que *PARIS, NICE, JACQUES, VINCENTE, YUKON et ZIMBABWE.*

Évaluation	Hypothèses	Références
Système de base A (pbSMT)	p a r í t	p a r i
	n í t	n í t x ờ
	gi ắ c	gi ắ c c ơ
	v a nh s â n t ơ	v a nh x â n t ơ
	iu k ô ng	d uy c ô ng
	gi anh b a b u ê	gi i m b a b u ê
Notre approche (Système 3)	p a r i	p a r i
	n í t x ờ	n í t x ờ
	gi ắ c c ơ	gi ắ c c ơ
	v a nh x â n t ơ	v a nh x â n t ơ
	d uy c ô ng	d uy c ô ng
	gi i m b a b ơ	gi i m b a b u ê

Dans une autre analyse (Tableau 5.5), nous avons observé que parmi les deux systèmes de base, en particulier le système de base A, à base de statistiques, des erreurs ont été commises telles que 'x' à 's' (pour VINCENTE) ou 'd uy' à 'iu' (pour YUKON) ou 'im' à 'anh' (pour ZIMBABWE), alors que notre approche (Système 3) a translittéré avec succès ces noms propres. Nous avons également constaté que le système de base A a omis le dernier graphème dans les hypothèses concernant NICE et JACQUES, alors que notre système a correctement produit ces prédictions. Par contre, nous avons remarqué que notre approche a prédit une autre hypothèse de translittération dans 'b ơ' au lieu de 'b u ê' (pour ZIMBABWE).

5.5 Conclusion

Dans ce chapitre, nous avons présenté l'approche neuronale, impliquant plusieurs techniques, notamment l'architecture d'encodeur-décodeur, le mécanisme d'attention, la représentation d'alignement pour les séquences d'entrée et les plongements de source-cible pré-entraînés, dans la tâche de translittération des entités nommées bilingues. Cette approche a été expérimentée dans l'étude de cas d'une paire de langues peu dotée en français-vietnamien. Évidemment, cette approche peut être étendue à n'importe quelle paire de langues peu dotée pourvu qu'un dictionnaire de prononciation bilingue soit disponible.

Nous avons observé que combiner la représentation d'alignement et les plongements de source-cible pré-entraînés a produit des améliorations significatives en termes de taux d'erreurs par rapport aux autres systèmes de base, notamment à base de statistiques et à base de séquences multi-jointes.

Dans le chapitre suivant, nous allons concevoir un système de traduction et de translittération automatique des entités nommées pour une paire de langues peu dotées, avec une étude sur la paire de langues français-vietnamien.

CHAPITRE VI

SYSTÈME DE TRADUCTION ET DE TRANSLITTÉRATION AUTOMATIQUE DES ENTITÉS NOMMÉES (TAEN)

6.1 Introduction

Depuis des décennies, la traduction automatique (TA) a fait l'objet de nombreuses recherches dans le domaine du traitement automatique des langues naturelles (TALN). Le but principal de la TA consiste à traduire un texte d'une langue à une autre langue. Cette tâche complexe est traitée par un humain expert en traduction, mais pas pour un ordinateur. C'est la raison pour laquelle la recherche dans le domaine de la TA a récemment proposé de nombreuses méthodes basées sur l'apprentissage automatique. À cet égard, la traduction automatique statistique (TAS) était l'une des approches les plus populaires, en particulier celle basée sur les segments. Cette approche s'appuie sur les corpus monolingues et parallèles afin d'entraîner les systèmes de traduction automatique.

De nos jours, avec l'émergence de l'approche basée sur l'apprentissage profond, un des développements majeurs de la TAS consiste en l'adoption de réseaux de neurones, d'où la traduction automatique neuronale (TAN). D'après Koehn (2017), les modèles de réseaux de neurones promettent un meilleur partage des évidences statistiques entre les mots similaires et l'inclusion des contextes riches avec la transformation en vecteurs numériques de mots.

Par ailleurs, dans le contexte de la TA pour les paires de langues peu dotées, en particulier pour la tâche de traduction automatique des entités nommées, la TA doit faire face à de nombreux défis, à savoir (1) les particularités linguistiques de chaque langue et (2) les ressources linguistiques bilingues ainsi que les outils de TALN.

(1) En ce qui concerne les particularités linguistiques, entre le français et le vietnamien, la construction des mots n'est pas similaire. La frontière des mots n'est pas identifiée par les espaces.

(2) En ce qui concerne les ressources linguistiques bilingues et les outils de TALN, nous ne possédons pas encore de données d'apprentissage bilingues assez grandes, ni de dictionnaires des noms propres bilingues, ni d'outils linguistiques appropriés pour pouvoir construire un tel système de TA performant.

D'où, dans ce chapitre, nous souhaitons appliquer les deux approches de TAS et de TAN, dans une étude sur la paire de langues peu dotée du français vers le vietnamien, dans le but de mieux comprendre les impacts sur la qualité des systèmes de traduction automatique, spécialement en relation avec la traduction automatique des entités nommées bilingues.

Le reste de ce chapitre est organisé comme suit : La section 6.2 décrit l'approche proposée. Puis, la section 6.3 aborde différentes expérimentations. Ensuite, la section 6.4 présente les évaluations de performances des systèmes, avec une analyse des erreurs constatées. Enfin, la section 6.5 conclut le chapitre.

6.2 Approches proposées

Dans cette section, nous présentons la méthodologie concernant les deux approches, à savoir à base de statistiques et à base de réseaux de neurones, pour la tâche de traduction automatique d'une paire de langues peu dotée.

Pour faire face au problème de la pénurie des ressources linguistiques bilingues, nous allons appliquer différentes stratégies d'adaptation, notamment avec un pré-traitement de segmentation de phrases au niveau des mots et au niveau des caractères, un large modèle de langue monolingue vietnamien, avec notamment l'utilisation d'un ensemble de données de développement en ce domaine des entités nommées bilingues pour la phase de réglage (en anglais : *tuning*). Aussi, nous faisons appel aux deux modules de reconnaissance et de translittération automatique des entités nommées afin de mieux identifier les entités nommées au sein de notre corpus d'apprentissage.

6.2.1 Approche à base de statistiques

Notre système de traduction automatique statistique des entités nommées bilingues se compose de différents modules. Ce système est construit en se basant sur la combinaison de deux modèles de traduction au niveau des mots et au niveau des caractères. Il se compose de deux phases principales, à savoir *(i)* la phase d'apprentissage et *(ii)* la phase de décodage. La figure 6.1 illustre l'architecture générale de notre système de traduction automatique statistique des entités nommées bilingues.

Dans la première phase *(i)*, le corpus parallèle français-vietnamien est divisé en deux corpus monolingues. Ensuite, en ce qui concerne le corpus en français, nous appliquons, comme pré-traitement, une segmentation de mots et une segmentation de caractères afin de construire deux corpus segmentés au niveau des mots et au niveau des caractères. En ce qui concerne le corpus en vietnamien, nous appliquons également, comme pré-traitement, une segmentation de mots et une segmentation de syllabes afin de construire deux corpus segmentés au niveau des mots et au niveau des syllabes.

Puis, nous entraînons deux modèles de traduction au niveau des mots et au niveau des caractères en utilisant les corpus appropriés annotés. Les modèles de langue sont entraînés avec un large corpus monolingue cible, en vietnamien, qui doit être segmenté au

niveau des mots et au niveau des syllabes. Les sorties de la première phase d'apprentissage servent à la phase suivante de décodage.

Dans la deuxième phase *(ii)*, tout d'abord, un texte en français est segmenté et traduit au niveau des mots en appliquant le premier modèle de traduction au niveau des mots. Donc nous obtenons le premier résultat (1) des hypothèses de traduction en vietnamien. Nous constatons qu'il y a encore des mots non traduits, soit de nouveaux mots soit de mots hors vocabulaire.

Puis, en utilisant nos deux modules de reconnaissance et de translittération des entités nommées bilingues, les candidats de traduction dans le résultat (1) sont filtrés en terme des entités nommées. Par la suite, une nouvelle liste d'entités nommées est projetée dans le décodage de TAS (2). Le deuxième modèle de traduction au niveau des caractères permet de trouver de meilleures hypothèses de traduction non seulement pour les entités nommées, mais aussi pour la plupart des mots non traduits qui demeurent dans le premier décodage de TAS (1). Enfin, nous obtenons le deuxième résultat final (2) des hypothèses de traduction en vietnamien.

6.2.2 Approche à base de réseaux de neurones

Notre système de traduction automatique neuronale des entités nommées bilingues est construit en se basant sur une architecture d'encodeur-décodeur de réseaux de neurones récurrents bidirectionnel avec le mécanisme d'attention. De même, il se compose de deux phases principales, à savoir *(i)* la phase d'apprentissage et *(ii)* la phase de décodage. La figure 6.2 illustre l'architecture générale de notre système de traduction automatique neuronale des entités nommées bilingues.

Dans la première phase *(i)*, le corpus parallèle est pré-traité avec un encodage en vecteurs numériques de mots, appelés plongements de mots. Ensuite, les plongements de

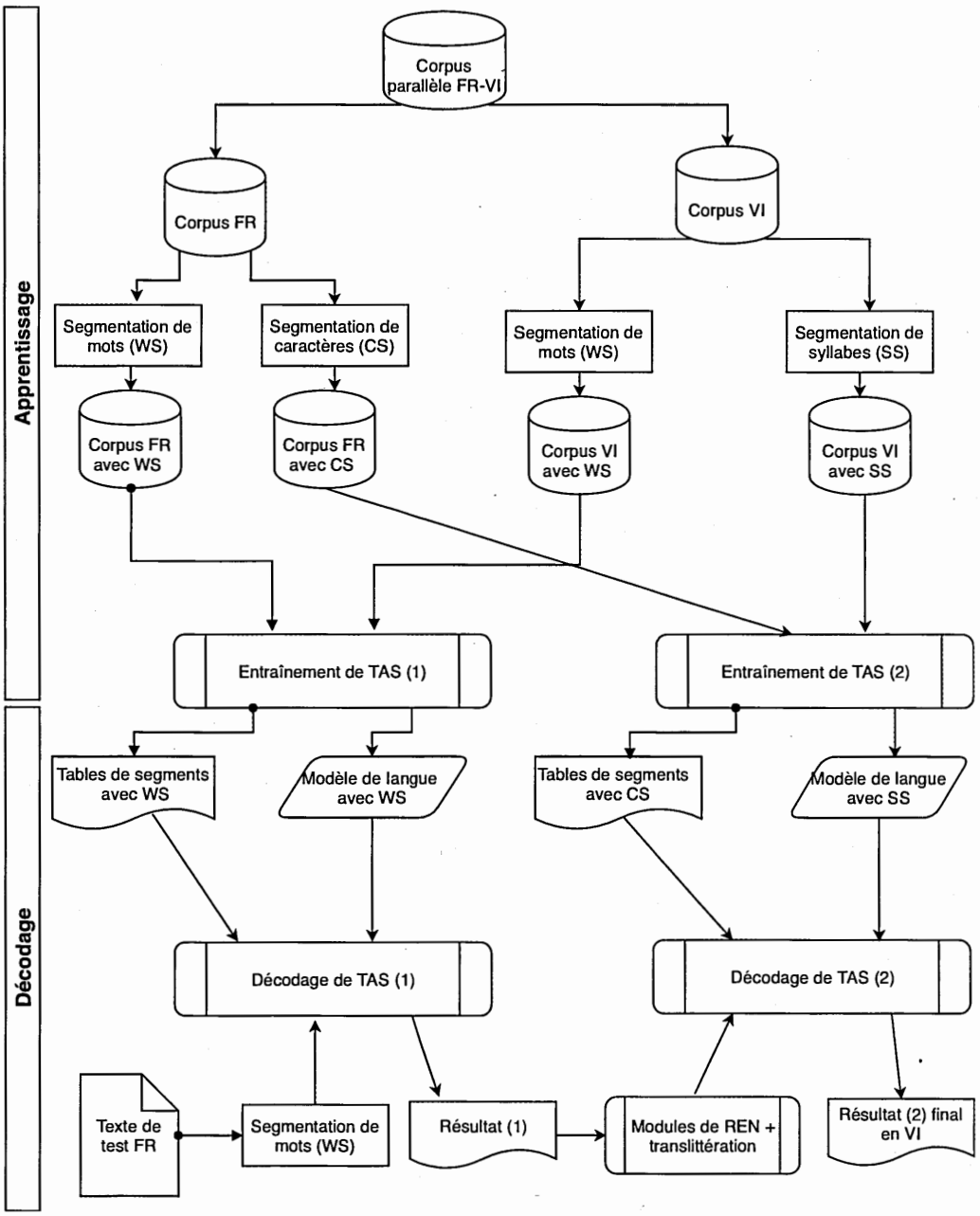

Figure 6.1 Architecture générale de notre système de traduction automatique statistique des entités nommées bilingues du français vers le vietnamien

mots, qui constituent la première couche d'entrée, sont projetés dans le réseau de neurones récurrent bidirectionnel.

Le modèle d'apprentissage incorpore le mécanisme d'attention (Bahdanau *et al.*, 2014). Ce mécanisme permet de capter différentes informations contextuelles pour de longues phrases. En plus, il permet au décodeur, dans la deuxième phase, de mieux focaliser sur différentes parties des séquences d'entrée. Donc les relations complexes dans les phrases longues ne sont pas perdues.

Dans la deuxième phase *(ii)*, nous procédons à la prédiction de mots de sortie, étant donnée une phrase test en français. À noter que la fonction de coût (erreur) est individuellement calculée pour chaque mot de sortie et totalisée à la phrase. À travers les états cachés du décodeur, les mots de sortie corrects dans l'état précédent sont utilisés comme contexte de conditionnement. Ainsi, la fonction objective d'apprentissage est basée sur la masse de probabilités donnée au mot correct, étant donné un contexte parfait. Enfin, nous obtenons, comme résultat final, un plongement de mots de sortie, en vietnamien, correspondant à la phrase test.

6.3 Expérimentations

Cette section présente la configuration des expérimentations, les résultats obtenus avec les évaluations en comparant nos deux méthodes dans la tâche de traduction automatique des entités nommées pour la paire de langues peu dotée français-vietnamien. Nous présentons également une analyse des erreurs produites par nos modèles de traduction.

6.3.1 Préparation des données

En ce qui concerne les données d'apprentissage, nous collectons un corpus bilingue français-vietnamien, à partir des sites Web d'actualité multilingues, tel que présenté

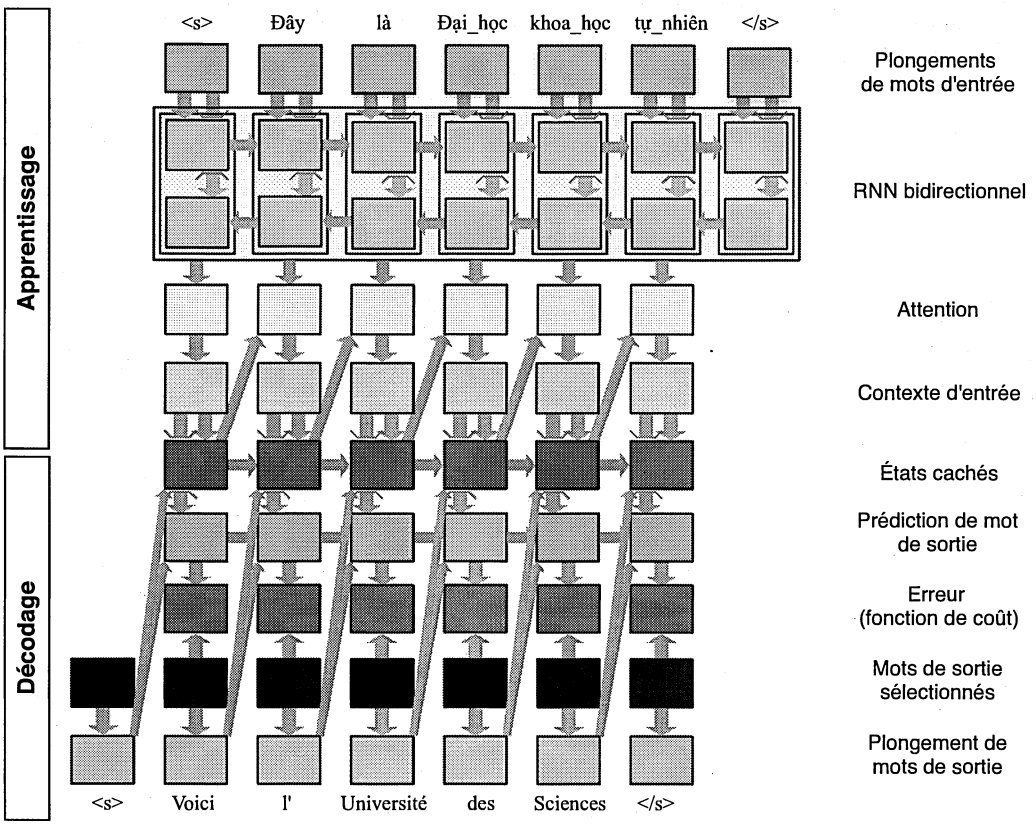

Figure 6.2 Architecture générale de notre système de traduction automatique neuronale des entités nommées bilingues. Illustration inspirée de (Koehn, 2017), avec un graphe de computation entièrement déroulé pour un exemple d'apprentissage ayant 7 *tokens* d'entrée *<s> Đây là Đại_học khoa_học tự_nhiên </s>* et 7 *tokens* de sortie *<s> Voici l'Université des Sciences </s>*.

dans Diep *et al.* (2010). Les données d'apprentissage se compose de 70 000 paires de phrases bilingues (Tableau 6.1).

Dans nos expérimentations, nous divisons tout le corpus avec un ratio de 80% - 10% - 10% pour les corpus d'apprentissage, de développement et de test, respectivement.

Tableau 6.1 Statistiques des données d'apprentissage pour la tâche de traduction automatique des entités nommées pour la paire de langues peu dotée en français-vietnamien.

Données	**Description**	**#paires de phrases**	**#tokens fr - vi**
Apprentissage	conférence Tam Dao 2009	5k	**105 009 - 143 255**
	conversations, livres bilingues, Web	15k	**234 422 - 287 178**
	corpus utilisé dans la thèse de Diep (2010)	50k	**1 783 349 - 2 337 997**

6.3.2 Configuration

Pour surmonter le problème de la dispersion des données d'apprentissage, nous effectuons, tout d'abord, une première étape de pré-traitement avec une normalisation en minuscule et une segmentation au niveau des mots / *tokens* et au niveau des syllabes / caractères.

Le processus de la TAS se déroule en plusieurs étapes suivantes :

(1) Pré-traitement :

En ce qui concerne le corpus vietnamien, toutes les phrases sont segmentées au niveau des mots et au niveau des syllabes en raison de la particularité linguistique de la langue vietnamienne. La boîte à outils CLC_VN_WS du Centre de linguistique computationnel [1] (CLC) à l'Université des Sciences, au Vietnam est utilisée pour cette segmentation de phrases en mots. En ce qui concerne le corpus

1. `http://www.clc.hcmus.edu.vn/?lang=fr`

français, toutes les phrases sont segmentées au niveau des *tokens* et au niveau des caractères.

(2) Construction de modèle de langue :

Pour entraîner un modèle de langue, nous utilisons un large corpus monolingue non étiqueté, contenant 17 095 994 de phrases en vietnamien (équivalent à 372 720 876 de *tokens*). Nous utilisons l'outil de *SRI Language Modeling (SRILM)* (Stolcke, 2002) pour extraire tous les traits correspondants au modèle de langue.

(3) Entraînement :

Le modèle d'alignement de mots est généré par l'outil *mgiza*[2], une autre version de multi-tâche de *GIZA++* (Och et Ney, 2003). La longueur maximale de phrase est configurée à 20.

Les expérimentations de notre système de TAS utilisent la boîte à outils de *Moses*[3], qui se base sur l'approche de traduction automatique statistique à base de segments (Koehn *et al.*, 2007). Nous utilisons les scripts, fournis par *Moses*, pour différentes tâches telles que la mise en minuscule (*lowercase.pl*), la normalisation (*normalize-punctuation.pl, tokenize.pl*) et le calcul de score de BLEU (*multi-bleu.pl*).

En outre, nous conservons également les valeurs de configuration par défaut lors de l'exécution de *Moses* : modèle log-linéaire avec 15 scores de composantes pondérées, dont 6 scores de ré-ordonnancement lexical, 1 score de distorsion, 1 score de modèle de langue, 1 score de pénalité de mot, 1 score de pénalité de phrase, 4 scores de modèle de traduction et 1 score de pénalité mot inconnu (Potet *et al.*, 2010). Par la suite, les poids de ces traits sont réglés en utilisant

2. `https://github.com/moses-smt/mgiza`

3. `https://github.com/moses-smt`

la procédure d'apprentissage du taux d'erreur minimum (*Minimum Error Rate Training - MERT*) (Och et Ney, 2003), intégré dans *Moses*.

Dans la phase de décodage, nous utilisons les options suivantes pour générer les informations des langues source et cible, telles que les meilleures hypothèses *n* du système de TAS et les informations d'alignement des mots dans la liste de *n-best* (*n-meilleures hypothèses*) :

- *-include-segmentation-in-n-best* et *-print-alignment-info-in-n-best* : à extraire l'information des alignements de mot-à-mot dans la liste de *n-meilleures*. Il est à noter que les alignements de mot-à-mot sont exclus de la table de segments.
- *-n-best-list NOM_FICHIER SIZE [distinct]* : à extraire une *n-meilleures* hypothèses de taille *SIZE* au chemin du fichier NOM_FICHIER.

En ce qui concerne le processus de la TAN, nous appliquons un pré-traitement tel que cité ci-dessus. Pour entraîner les plongements de source et de cible pré-entraînés, nous appliquons l'outil de *word2vec*[4] avec une dimension de 100, une taille de la fenêtre contextuelle de 5 grammes et l'option de *skip-gram*.

Afin d'entraîner notre modèle de traduction, nous appliquons la boîte à outils de *nmt-keras*[5] (Peris, 2017) dans un cas d'étude d'une paire de langues peu dotée, le français-vietnamien. Le choix de cette boîte à outils permet d'implémenter l'architecture d'encodeur-décodeur, le mécanisme d'attention, l'intégration des plongements, et surtout de construire rapidement un système de TAN. Dans la configuration de notre système de TAN, nous avons utilisé deux couches de l'architecture d'encodeur-décodeur LSTM bidirectionnel, avec une couche d'entrée à 64 dimensions pour encoder les séquences d'entrée et 128 nœuds dans la couche

4. `https://code.google.com/archive/p/word2vec/`

5. `https://github.com/lvapeab/nmt-keras/`

cachée. Nous avons utilisé l'optimiseur « *Adam* » pour apprendre les poids du réseau avec un taux d'apprentissage par défaut de 0,001. Pour le décodage, la recherche en faisceau est configurée à la taille de fenêtre contextuelle de 6.

Nous avons effectué quatre expérimentations de traduction automatique des entités nommées, pour la paire de langues peu dotée en français-vietnamien, comme suit :

- Expérimentation 1 (Système 1 - TAS) : Dans cette première expérimentation pour un système de traduction automatique statistique, qui est considérée comme système de base, nous utilisons un corpus d'apprentissage et un corpus de test composés de 50k et 359 entités nommées extraites de 1 060 paires de phrases bilingues, respectivement. Il est à noter que, dans ce scénario, le modèle de langue est entraîné sur un petit corpus cible, en vietnamien, qui constitue l'ensemble des données d'apprentissage.
- Expérimentation 2 (Système 2 - TAS) : Dans cette deuxième expérimentation pour un système de traduction automatique statistique, nous utilisons le même corpus d'apprentissage ayant 50k paires de phrases bilingues et un corpus de test ayant 359 entités nommées extraites de 1 060 paires de phrases bilingues. Il est à noter que, dans ce scénario, le modèle de langue est entraîné sur un large corpus ayant 17 095 994 phrases en vietnamien.
- Expérimentation 3 (Système 3 - TAS) : Dans cette troisième expérimentation pour un système de traduction automatique statistique, nous utilisons la totalité du corpus d'apprentissage de 70k paires de phrases bilingues et un corpus de test de 359 entités nommées extraites de 1 060 paires de phrases bilingues. Le modèle de langue est le même que dans l'expérimentation 2.
- Expérimentation 4 (Système 4 - TAN) : Dans cette quatrième expérimentation, pour un système de traduction automatique neuronale, nous utilisons la même configuration que dans l'expérimentation 3.

6.4 Évaluations

6.4.1 Résultats obtenus

Afin d'évaluer nos deux méthodes, nous appliquons différentes métriques d'évaluation :

- score de *BLEU (BiLingual Evaluation Understudy)* (Papineni *et al.*, 2002),
- métrique pour l'évaluation de la traduction avec ordonnancement explicite (*METEOR - Metric for Evaluation of Translation with Explicit ORdering*) (Banerjee et Lavie, 2005b),
- taux d'erreur de traduction (*TER - Translation Error Rate*) (Snover *et al.*, 2009),
- taux de mots hors-vocabulaire (MHV).

Nous constatons que la performance des trois premiers systèmes de TAEN, qui se basent sur l'approche de statistiques, a été améliorée, avec les scores de BLEU de 31,40%, 40,00%, 51,80% pour le système de base 1, le système 2 et le système 3, respectivement (Tableau 6.2). En comparant avec un travail de recherche similaire, Do *et al.* (2009) a rapporté leur système de TAS du français vers le vietnamien ayant un score de BLEU de 40,09% avec le même corpus d'apprentissage. Nous observons des améliorations significatives dans notre approche, à travers les scores d'autres métriques d'évaluation.

En comparant avec le système de base 1, nous observons des augmentations de +22,90% et de +27,90% de score de METEOR pour le système 2 et le système 3, respectivement. D'ailleurs, nous observons des diminutions du taux d'erreur de traduction (TER) et du taux de mots hors vocabulaire (MHV), avec les scores de TER de -17,90% et -31,40%, les scores de taux MHV de -11,20% et -22,30% pour le système 2 et le système 3, respectivement.

En ce qui concerne le système 4, qui se base sur l'approche des réseaux de neurones, nous constatons une performance médiocre, avec seulement 12,04% de BLEU. Ce score

de BLEU est de -19,36% par rapport au système de base 1. En plus, nous remarquons que toutes les métriques d'évaluation, pour le système 4, sont très basses, avec 28,30%, 91,94% et 69,72% pour les scores de METEOR, de TER et de taux de MHV, respectivement. Nous trouvons une diminution de -12,20% de METEOR, des augmentations du taux d'erreur de traduction (TER) de +24,34% et du taux de MHV de +19,52% par rapport au système de base 1.

Tableau 6.2 Évaluation des performances de différents systèmes de traduction automatique des entités nommées pour la paire de langues peu dotée français-vietnamien, en appliquant nos deux méthodes.

Expérimentation	**BLEU**	**METEOR**	**TER**	**Taux de MHV**
Système 1 - TAS (*base*)	31,40	40,50	67,6	49,80
Système 2 - TAS	40,00	63,40	49,70	38,60
Système 3 - TAS	**51,80**	**68,40**	**36,20**	**27,50**
Système 4 - TAN	12,04	28,30	91,94	69,72

Pour le système de base 1, nous avons utilisé seulement un petit corpus d'apprentissage et un petit corpus monolingue cible pour entraîner le modèle de langue. Alors que le système 2 et le système 3 ont bénéficié d'un plus grand corpus d'apprentissage et d'un large corpus monolingue cible pour entraîner le modèle de langue. D'où nous constatons des améliorations de performance de ces deux derniers par rapport au système de base 1. Les systèmes 2 et 3 sont capables de prédire la traduction des entités nommées, qui ont été considérées comme hors vocabulaire pour le système de base 1.

D'ailleurs, nous constatons qu'avec le module de translittération des entités nommées, qui se base sur l'approche des réseaux de neurones, qui convertit une séquence de caractères en un vecteur de mot, nous évitons le problème d'avoir de nombreuses variantes morphologiques apparaissant comme des entités distinctes dans un vocabulaire. La traduction au niveau des caractères est beaucoup plus avantageuse que la traduction au

niveau des mots. Les modèles au niveau des caractères ne souffrent pas de problèmes hors-vocabulaire. Et ces modèles ne nécessitent pas d'informations segmentées.

Par contre, le système 4, qui se base sur l'approche des réseaux de neurones, n'est pas capable de traduire correctement les entités nommées. Une des causes est probablement due à la petite taille du corpus d'apprentissage et à la petite taille des plongements de mots de source et de cible. Normalement, la construction d'un système de TAN nécessite un corpus d'apprentissage de l'ordre de millions de paires de phrases bilingues (Koehn et Knowles, 2017). Alors que, dans notre étude pour la paire de langues peu dotée en français-vietnamien, le corpus d'apprentissage est constitué seulement de 70k de paires de phrases bilingues.

6.4.2 Analyse des erreurs

Dans cette section, nous effectuons une analyse des erreurs probables, afin de décortiquer différents types d'erreurs générées par nos divers systèmes d'expérimentation. Par conséquent, nous avons réalisé une comparaison des résultats de prédiction de la traduction des entités nommées entre les quatre systèmes d'évaluation avec des noms propres qui n'ont jamais été vus durant la phase d'apprentissage (Tableau 6.3).

Les résultats ont montré que le système de base 1 a relevé des erreurs de prédiction de traduction des entités nommées, alors que les systèmes 2 et 3 ont produit de meilleurs résultats. L'entité nommée « *Asie du Sud-Est* » a été bien traduite dans les systèmes 2 et 3, mais le mot « *Est* » (*en vietnamien, đông*), devant être traduit dans le système de base 1, a échappé . De même, les entités nommées « *Mékong* » et « *khmer* » ont été très bien traduites et translittérées par les systèmes 2 et 3, alors que le système de base 1 n'a pas réussi à proposer une hypothèse de traduction complète.

Tableau 6.3 Quelques illustrations de prédiction de la traduction automatique des entités nommées produites par les quatre systèmes d'expérimentation, avec un ensemble d'entités nommées en français {*Indonésie, Cambodge, Asie du Sud-Est, Mékong, khmer*}.

Référence	**Système 1**	**Système 2**	**Système 3**	**Système 4**
indonesia	inđônêxia	inđônêxia	inđônêxia	in-đô-nê-xia ta lực
cambodia	cam-pu-chia	campuchia	campuchia	tìm thấy miến
đông nam á	nam á	đông nam á	đông nam á	nam á thế á thế
mêkông	sông	sông mê công	mê công	sông cửu long
khmer	dân	khơme	khơme	liệu họ .

D'ailleurs, nous avons remarqué que le système de base 1 et surtout le système 4 ont rencontré beaucoup de difficultés pour prédire, de façon optimale, toutes les hypothèses de traduction à cause de la faible quantité du corpus d'apprentissage. Néanmoins, le système 4 a pu prédire une autre hypothèse synonyme pour l'entité nommée «*Mékong*» par « *sông cửu long* » (*en français, littéralement, fleuve des neuf dragons*).

Tableau 6.4 Évaluation de la performance de notre modèle, le système 3 - TAS, pour la tâche de traduction automatique de 100 noms propres de la Bible bilingue française-vietnamienne, ainsi que 100 termes techniques de chimie en général.

Expérimentation	**BLEU**	**METEOR**	**TER**	**Taux de MHV**
Bible	**42,10 (+8.90)**	**34,30 (-3,30)**	**38,80 (-15,40)**	**58,30 (-20,30)**
Termes techniques	33,20	37,60	54,20	78,60

Pour aller plus loin, nous avons effectué deux expérimentations, en utilisant le meilleur système que nous proposons, le système 3 - TAS, (1) avec la traduction de 100 noms

propres, extraits depuis la Bible bilingue française-vietnamienne [6] et (2) avec la traduction de 100 termes de chimie, extraits d'un dictionnaire bilingue de mots d'emprunt du français dans la langue vietnamienne (Minh, 2011). Le tableau 6.4 montre la performance de notre approche.

Les métriques d'évaluation montrent que notre approche, à base de statistiques, performe mieux pour l'expérimentation des noms de la Bible que dans l'expérimentation des termes techniques de chimie. Le score de BLEU de la TAEN de la Bible est meilleur que celui de la TAEN des termes techniques, avec un gain de 8,9%. Les taux d'erreurs, pour traduire les noms propres de la Bible, sont beaucoup inférieurs à celui de la TAEN pour les termes techniques, avec des diminutions de -15,40% et -20,30% pour TER et taux de MHV, respectivement (Tableau 6.4),

Les tableaux 6.5 et 6.6 illustrent quelques exemples de prédiction de traduction automatique des noms propres de la Bible bilingue et des termes techniques de chimie en utilisant notre approche à base de statistiques.

Tableau 6.5 Quelques exemples de prédiction de traduction automatique des noms propres, extraits de la Bible bilingue française-vietnamienne, produits par notre approche (système 3 - TAS).

Source FR	**Référence VI**	**TAS**	**Source FR**	**Référence VI**	**TAS**
israel	ysơraên	isờraen	zabulon	sabulôn	giabulông
ruben	rubên	cuben	benjamin	bêngiamin	bendamin
simeon	simêôn	simêôn	jacob	giacốp	dacốp
levi	lêvi	lêvi	joseph	giôsep	puydốtxơ
juda	giuđa	duđa	ramses	ramse	ramxơ
issacar	ysaca	ixờtaca	pharaon	pharaôn	baraông

6. http://www.transcripture.com/francais-vietnamese-psaumes-1.html

Tableau 6.6 Quelques exemples de prédiction de la traduction automatique de termes techniques de chimie, extraits d'un dictionnaire bilingue de mots d'emprunt du français dans la langue vietnamienne, produits par notre approche (système 3 - TAS).

Source FR	**Référence VI**	**TAS**
adrenaline	ađrênalin	ađờrênalinnơ
aleurone	alơron	alêuronnơ
amygdale	amiđan	amycđanlơ
aminoacide	aminôaxít	aminôaxiđơ
amine	amin	aminnơ
amibe	amíp	amibê
ammoniac	amôniắc	ammôniác
acetylene	axêtylen	asờtilânơ
blouson	bờludông	bờlôusông
hemoglobine	hêmôglôbin	hêmôgờlôbinnơ
paludisme	baluyđít	baluđisờme
vaseline	vadơlin	vagienlinnơ
virus	virút	virútxơ
vitamine	vitamin	vitaminnơ

En ce qui concerne l'expérimentation (1), nous constatons beaucoup d'irrégularités linguistiques au sein de certaines hypothèses de traduction des noms propres avec différentes lettres (Tableau 6.5), telles que :

- « i » → « y » ou « i » :
  israel → **y**sơraên ou **i**sờraen
  issacar → **y**saca ou **i**xờtaca

- « j » → « gi » ou « d » :

  juda → **gi**uđa ou **d**uđa

  benjamin → bên**gi**amin ou ben**d**amin

En ce qui concerne l'expérimentation (2), nous constatons que les consonnes finales ont été découpées ou abrégées (Tableau 6.6) :

- « nơ », « lơ » ou « xơ » :

  adrenaline → ađrênalin ou ađờrênalin**nơ**

  amygdale → amiđan ou amycđan**lơ**

  virus → virút ou virút**xơ**

Nous observons ici que les voyelles accentuées, qui ont été rajoutées, permettent de former une variété de tons lexicaux. Ainsi la prononciation de termes techniques est plus conforme à celle du mot d'origine.

En somme, nous avons essayé de montrer que notre approche est capable de prédire la traduction automatique non seulement des noms propres, mais aussi de termes techniques spécifiques.

6.5 Conclusion

Dans ce chapitre, nous avons présenté les modèles de traduction automatique, notamment à base de statistiques et à base de réseaux de neurones, en incorporant les modules de reconnaissance et de translittération des entités nommées, dans une étude de cas d'une paire de langues peu dotée en français-vietnamien.

Nous avons pu découvrir des points forts et des points faibles dans les deux approches pour la tâche de traduction automatique des entités nommées bilingues. D'ailleurs, nous avons obtenu de meilleures prédictions de traduction des entités nommées avec l'approche de TAS qu'avec l'approche de TAN, en raison de plusieurs facteurs quantitatifs et qualitatifs.

Par conséquent, nous avons détecté multiples régularités et irrégularités linguistiques pour cette tâche de traduction automatique des entités nommées. Nous avons également remarqué que la traduction des entités nommées dépendait fortement de la qualité des systèmes de REN et de translittération.

Par ailleurs, en raison de la taille petite du corpus parallèle développé et utilisé, la TAN a montré une faible performance par rapport à la TAS. Avec un plus grand corpus d'apprentissage ou une méthode basée sur l'apprentissage par transfert, la TAN pourrait surpasser la TAS.

Dans le chapitre suivant, nous allons récapituler les principales contributions du point de vue cognitif et informatique, les obstacles rencontrés tout au long de la thèse, ainsi que les recherches à venir dans le domaine de traduction automatique en particulier et dans celui du traitement automatique de langues naturelles en général.

CONCLUSION

En guise de conclusion, nous allons récapituler, en regard de la présente thèse, les principales contributions du point de vue cognitif et informatique, les obstacles rencontrés ainsi que les recherches à venir dans le domaine de la traduction automatique en particulier et dans celui du traitement automatique de langues naturelles en général.

1. Contributions de la thèse

1.1. Composante cognitive

Le but principal de notre choix, l'étude de cas de la paire de langues français-vietnamien, consiste à contribuer aux défis de la recherche sur la traduction automatique en particulier et sur le traitement automatique de langues naturelles (TALN) en général, tout en considérant la langue vietnamienne, une langue peu dotée, ainsi que cette paire de langues comme peu dotée en ressources linguistiques informatisées.

En effectuant diverses expérimentations concernant les entités nommées, monolingues et bilingues, nous constatons que les extractions des traits linguistiques additionnels non tonaux ne sont pas suffisants pour faire face à toutes les problématiques possibles. Concrètement, nous remarquons que les règles de conversion graphémique ne suivent pas les mêmes règles de conversion classique. La diversité des transcriptions phonétiques du français dans la langue vietnamienne rajoute encore la complexité de la tâche de conversion. C'est une des raisons pour lesquelles nous envisageons, en perspective, de traiter d'autres traits phonétiques tonaux (Bharadwaj *et al.*, 2016) dans le but d'améliorer la performance de notre système de translittération.

Par ailleurs, d'après la revue de littérature dans le domaine du TALN, une grande partie des mots qui restent inconnus (ou non traduits) dans les textes bilingues français-vietnamiens, sont liés aux entités nommées. D'où la reconnaissance, la translittération et la traduction des entités nommées dans la langue vietnamienne demeurent toujours des défis majeurs à résoudre. Nous remarquons qu'en ce qui concerne un des défis majeurs de la translittération, nous ne pourrions pas simplement compter sur les corpus parallèles bilingues, mais sur des modèles construits séparément, notamment un modèle à base de phonétique et/ou un modèle à base d'orthographe.

En somme, la contribution à la translittération des entités nommées permet non seulement de réduire des mots hors vocabulaires, non traduits et/ou incorrectement traduits, mais aussi d'améliorer la qualité du système de la traduction automatique.

1.2. Composante informatique

En ce qui concerne la composante informatique de la thèse, nous avons effectué de multiples contributions.

- Premièrement, nous avons développé une méthode d'extraction des entités nommées d'une langue source (français) vers une langue cible (vietnamien) en appliquant la projection trans-linguistique. Par la suite, nous avons construit des ressources [7] linguistiques informatisées pour une paire de langues peu dotée, dont un corpus bilingue français-vietnamien.
- Deuxièmement, nous avons modélisé un module de reconnaissance des entités nommées pour une langue peu dotée, dans une étude de cas du vietnamien. Ce module est réalisé en utilisant l'approche à base de réseaux de neurones récurrents, avec différents traits linguistiques additionnels, notamment surface de

7. https://github.com/NgocTanLE

mots, lemme, grappe de mots (*word cluster*), étiquettes morpho-syntaxiques, segment syntaxique (*chunk*) et entités nommées.

- Troisièmement, nous avons également modélisé un module de translittération des entités nommées bilingues pour une paire de langues peu dotée, dans une étude de cas du français-vietnamien. Ce module est réalisé en utilisant l'approche à base de réseaux de neurones récurrents, avec les informations d'alignement entre les séquences d'entrée et de sortie, l'architecture d'encodeur-décodeur et le mécanisme d'attention.
- Quatrièmement, nous avons adapté l'approche d'apprentissage profond dans un système de traduction et de translittération automatique des entités nommées, avec l'application sur une paire de langues peu dotée en français-vietnamien. Cette approche permet d'améliorer la performance du système de traduction automatique des entités nommées bilingues, en incorporant les modules de la reconnaissance des entités nommées et de la translittération des entités nommées dans le processus de décodage.

2. Obstacles rencontrés

Le premier obstacle rencontré pour réaliser notre projet de thèse demeure la construction d'un large corpus bilingue français-vietnamien. Jusqu'à présent, il n'existe aucune ressource linguistique mise à la disposition de la communauté de recherche internationale du TALN concernant cette paire de langues. La langue vietnamienne est catégorisée comme une langue peu dotée parmi d'autres langues asiatiques ou africaines, telles que les langues *thaïe, khmère, malaise, mongole, zoulou, xhosa, etc.* De plus, les ressources linguistiques en ligne pour cette paire de langues sont assez restreintes, avec un certain nombre de sites Web bilingues ou multilingues. Il faudrait donc envisager de collecter un corpus de textes bilingues français-vietnamiens à partir de l'Internet

(par exemple actualités, nouvelles, journaux informatisés) ou de livres bilingues, de colloques bilingues, etc.

Le deuxième obstacle est lié à l'extraction de lexiques bilingues des entités nommées pour des fins de construction d'un système de traduction automatique statistique. Une solution serait de trouver des traits linguistiques similaires entre les deux langues vietnamienne et française, ainsi que leurs régularités linguistiques réciproques.

Le troisième obstacle concerne la tâche de translittération des entités nommées. Il n'existe pas encore de ressources linguistiques de translittération pour la paire de langues français-vietnamien. Il est crucial de créer un dictionnaire de transcription phonétique pour le français-vietnamien. En raison des particularités linguistiques du vietnamien, notamment la composition d'un mot dans cette langue, nous envisageons d'étudier ce problème en décomposant un mot en sous-mots, c'est-à-dire en syllabes et en caractères. Nous envisageons la construction de modèles de translittération des entités nommées, notamment un modèle à base de phonétique et/ou un modèle à base orthographique.

Le quatrième obstacle est lié à l'adaptation de l'approche d'apprentissage profond dans la traduction automatique, afin d'améliorer la qualité et l'efficacité de notre système. Surtout, nous souhaitons examiner les impacts possibles de l'apprentissage profond dans la traduction automatique des entités nommées, tout en comparant entre la traduction automatique statistique et la traduction automatique neuronale.

3. Perspectives

En ce qui concerne le corpus d'apprentissage, nous allons nous concentrer sur la collecte d'un corpus bilingue à grande échelle, en incorporant d'autres ressources lin-

guistiques, notamment des dictionnaires bilingues, des lexiques des entités nommées bilingues.

Les tâches de reconnaissance et de translittération automatique des entités nommées demeurent toujours des défis majeurs à résoudre dans la communauté de recherche du TALN en particulier, et dans la fouille de données en général. En ce qui concerne la tâche de reconnaissance des entités nommées, nous avons l'intention d'appliquer d'autres architectures au sein de notre modèle telles que l'architecture de *Convolutional Neural Network* (*en français : réseau neuronal convolutif*) (Chiu et Nichols, 2015; Ma et Hovy, 2016) ou l'adaptation de domaine cross-linguistique (Cotterell et Duh, 2017; Feng *et al.*, 2018).

En ce qui concerne la tâche de translittération des entités nommées, nous avons l'intention d'expérimenter notre approche avec un corpus d'apprentissage plus large. Une simple altération utilisant les informations d'alignement pour la représentation des séquences d'entrée et de sortie dans les modèles de translittération nécessite des recherches plus approfondies concernant la conception de représentations afin d'améliorer davantage les performances de ces modèles. En plus, nous envisageons de traiter non seulement les noms propres, mais aussi les noms communs, qui font partie de l'ensemble de mots d'emprunts du français dans la langue vietnamienne, comme étudié dans les travaux de recherche de Huynh (2010) et de Minh (2011).

En ce qui concerne la traduction automatique des entités nommées pour une paire de langue peu dotée, nous avons l'intention de rajouter de nouvelles composantes neuronales dans la traduction automatique statistique. Aussi, nous envisageons d'incorporer les unités lexicales à base de segments dans le modèle de traduction afin de compenser la traduction du mot-à-mot dans la traduction automatique neuronale (Wang *et al.*, 2017).

La construction d'un corpus parallèle synthétique en utilisant le système de TAS avec une traduction inverse (en anglais, *back-translation*) (Hoang *et al.*, 2018) afin d'augmenter la taille des données d'apprentissage (en anglais, *data augmentation*) (Fadaee *et al.*, 2017) et ainsi appliquer la TAN. Ceci pourrait être considéré comme une hybridation des deux systèmes de TAS et TAN (Huang *et al.*, 2017; Dahlmann *et al.*, 2017).

En outre, nous avons l'intention de combiner avec d'autres approches semi-supervisées ou non-supervisées afin de tirer le plus d'avantages possibles d'une grande masse de données textuelles non annotées, ainsi que d'explorer d'autres architectures de la traduction automatique neuronale dans l'étude de cas d'une paire de langues peu dotée de ressources linguistiques informatisées (Johnson *et al.*, 2016; Britz *et al.*, 2017; Lample *et al.*, 2018).

ANNEXE A

PUBLICATIONS

1. Articles publiés

a. Articles de revues

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard, Dien Dinh. Low Resource Machine Transliteration Using Recurrent Neural Network. *In Proceedings of ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), ISSN 2375-4699, Inf. Process. 18(2) : 13 :1-13 :14 (2019).*

Phuoc Tran, Dien Dinh, **Ngoc Tan Le**, Long H. B. Nguyen (2017). Linguistic-Relationships-Based Approach for Improving Word Alignment. *In Proceedings of ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), ISSN 2375-4699, doi 10.1145/3133323, Volume 17 Issue 1, Article No. 5, 5 :1–5 :16, November 2017.*

b. Chapitres de livres

Ngoc Tan Le, Fatiha Sadat (2018). Cross-linguistic Projection for French-Vietnamese Named Entity Recognition. *Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, Springer LNAI (Lecture Notes in Artificial Intelligence) Series, Lecture Notes in Computer Science, DOI : 10.1007/978-3-319-93782-3_29. In book : Human Language Technology. Challenges for Computer Science and Linguistics, 407-419, June 2018.*

Fatma Mallek, **Ngoc Tan Le**, Fatiha Sadat (2017). Automatic Machine Translation for Arabic Tweets. *Intelligent Natural Language Processing : Trends and Applications. Scientific Research Group in Egypt (SRGE) editor Springer, 2017, book chapter.*

c. Articles de conférences et d'ateliers

Ngoc Tan Le, Fatiha Sadat. Transfer Learning to Enhance Korean Word Segmentation on Social Media Texts. *In proceedings of the 2019 16th International Conference of the Pacific Association for Computational Linguistics (PACLING), October 11-13, 2019, Hanoi, Vietnam.*

Ghaith Dekhili, **Tan Ngoc Le** and Fatiha Sadat. Improving Named Entity Recognition with Commonsense Knowledge Pre-training. *In the proceedings of the 2019 Pacific Rim*

Knowledge Acquisition Workshop (PKAW 2019), Quan Bai and Kouzou Ohara Volume Edition, colocated in the 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019), Fiji, USA, August 26-27, 2019.

Rakhee Sharma, **Tan Ngoc Le**, Fatiha Sadat (2018). Multimodal Sentiment Analysis Using Deep Learning. *Dans Proceedings of The 17th IEEE International Conference on Machine Learning and Applications (ICMLA 2018), Orlando, Florida, 17-20 December 2018.*

Tan Ngoc Le, Fatiha Sadat (2018). Deep Neural Networks for Social Media Word Segmentation of Asian Languages. *Dans Proceedings of The IEEE International Conference on Big Data (IEEE Big Data 2018), Seattle, USA, 10-13 December 2018.*

Tan Ngoc Le, Fatiha Sadat (2018). Improving Low-Resource Neural machine Transliteration of Asian Languages. *Dans Proceedings of The 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32), The Hong Kong Polytechnic University, Hong Kong SAR, 1-3 December 2018.*

Ngoc Tan Le, Fatiha Sadat (2018). Low-Resource Machine Transliteration Using Recurrent Neural Networks of Asian Languages. *In Proceedings of the Seventh Named Entities Workshop, Association for Computational Linguistics (ACL), Melbourne, Australia, July 20, 2018, 1-6.*

Ngoc Tan Le, Fatiha Sadat (2018). Data Adaptation for Named Entity Recognition in Twitter with Features-Rich CRF. *In Proceedings of NAACL HLT 2018, Widening Natural Language Processing workshop, New Orleans, Louisiana, USA, 1 to 6 June 2018.*

Ngoc Tan Le, Long Nguyen, Fatiha Sadat (2018). Neural Network-based Model for Named Entity Recognition in Low-Resource Settings. *In Proceedings of NAACL HLT 2018, Widening Natural Language Processing workshop, New Orleans, Louisiana, USA, 1 to 6 June 2018.*

Ngoc Tan Le, Fatiha Sadat (2017). A neural network architecture for low-resource machine transliteration. *In proceeding of the 8th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, in the 5th LRL Workshop (a Joint LTC-ELRA-FLaReNet-META-NET Workshop on Less-Resourced Languages) : « Language Technology for Less Resourced Languages », November 17-19, 2017, Poznań, Poland.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2017). A Neural Network Transliteration Model in Low Resource Settings. *In Proceedings of the 16th International Conference*

of Machine Translation Summit, September 18-22 2017, Nagoya, Japan, volume 1, Research Track. pp.337–345.

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2017). Translittération automatique pour une paire de langues peu dotée. *24ème Traitement Automatique des Langues Naturelles, Orléans, France, 26-30 Juin 2017.*

Ngoc Tan Le, Fatma Mallek, Alexandro Fonseca, Billal Bilainine, Fatiha Sadat (2017). Reconnaissance des entités nommées dans les messages Twitter en français. *In Proceedings of 19ème Conférence sur l'apprentissage automatique, Grenoble, France, 28-30 Juin 2017.*

Mallek F, **Le Ngoc T**, Sadat F. (2017). Improved Machine Translation for Arabic Tweets on Scarce-resource settings. *In proceedings of The Workshop on Social Media and User Generated Content Machine Translation, May 31st 2017, EAMT, Prague, Czech Republic.*

Ngoc Tan Le, Fatiha Sadat (2017). Information Extraction in Real-World Business Documents. *CICLing : International Conference on Computational Linguistics and Intelligent Text Processing, April 17-23, 2017 Budapest, Hungary.*

Ngoc Tan Le, Fatiha Sadat (2016). UQAM-NTL : Named entity recognition in Twitter messages. *COLING 2016, the 26th International Conference on Computational Linguistics, Association for Natural Language Processing (ANLP), the second Workshop on Noisy User-generated Text Osaka, Japan, 11-16 December 2016.*

Ngoc Tan Le, Fatiha Sadat (2015). Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection. *22ème Traitement Automatique des Langues Naturelles, Caen, France, 2015.*

Ngoc Tan Le, Fatiha Sadat (2015). Error Analysis of Named Entity Translation output for Poor-Resourced Bilingual Vietnamese-French Pair. *7th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, November 27-29, 2015, Poznań, Poland, pp.304-308.*

2. Présentations

Ngoc Tan Le (2019). Traduction et translittération des entités nommées pour une paire de langues peu dotée. *Séminaire du Doctorat en Informatique Cognitive, Université du Québec à Montréal, 26 Septembre 2019.*

Ngoc Tan Le (2018). Reconnaissance des entités nommées pour une langue peu dotée de ressources linguistiques. *Séminaire du Doctorat en Informatique Cognitive, Université du Québec à Montréal, 06 Décembre 2018.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2018). Un modèle de translittération neuronal pour une paire de langues peu dotée. *Colloque annuel du Doctorat en Informatique Cognitive, Université du Québec à Montréal, TÉLUQ, 20 Juin 2018.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2018). Nouvelle méthode de construction d'un système de translittération pour une paire de langues peu dotée en utilisant l'approche à base des réseaux de neurones. *86^{me} Congrès de l'ACFAS – Association francophone pour le savoir, Université du Québec à Chicoutimi, Québec, Canada, 7-11 Mai 2018, communication libre.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2017). Translittération automatique des entités nommées pour une paire de langues peu dotées. *Colloque annuel du Doctorat en Informatique Cognitive, Université du Québec à Montréal, 26 Mai 2017.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2016). Construction de ressources linguistiques pour un corpus annoté bilingue franco-vietnamien et traduction automatique des entités nommées pour Français-Vietnamien. *84ème Congrès de l'ACFAS – Association francophone pour le savoir, Université du Québec à Montréal, Québec, Canada, 9-13 Mai 2016.*

Ngoc Tan Le, Fatiha Sadat, Lucie Ménard (2015). Building a Bilingual Vietnamese-French Named Entities Annotated Corpus. *Colloque annuel du Doctorat en Informatique Cognitive, Université du Québec à Montréal, 10 Juin 2015.*

RÉFÉRENCES

Alonso, J. A. et Bocsák, A. (2005). Machine translation for Catalan-Spanish. The real case for productive MT. Dans *Proceedings of the tenth Conference on European Association of Machine Translation (EAMT 2005), Budapest, Hungary.*

Artetxe, M., Labaka, G., Agirre, E. et Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv :1710.11041.*

Azpeitia, A., Cuadros, M., Gaines, S. et Rigau, G. (2014). Nerc-fr : supervised named entity recognition for French. Dans *International Conference on Text, Speech, and Dialogue*, 158–165. Springer.

Babych, B., Hartley, A. et Sharoff, S. (2007). Translating from under-resourced languages : comparing direct transfer against pivot translation. Dans *Proceedings of the Machine Translation Summit XI*, 412–418. Citeseer.

Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473.*

Banerjee, S. et Lavie, A. (2005a). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. Dans *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the forty-third Annual Meeting of the Association of Computational Linguistics.*

Banerjee, S. et Lavie, A. (2005b). METEOR : An automatic metric for mt evaluation with improved correlation with human judgments. Dans *Proceedings of the Association for Computational Linguistics workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Barros, M. J. et Weiss, C. (2006). Maximum entropy motivated grapheme-to-phoneme, stress and syllable boundary prediction for portuguese text-to-speech. *IV Jornadas en Tecnologías del Habla. Zaragoza, Spain*, 177–182.

Bautista, M. L. S. et Bolton, K. (2008). *Philippine English : Linguistic and Literary.* Hong Kong University Press.

Bengio, Y., Simard, P. et Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *Transactions on Neural Networks*, *5*(2), 157–166.

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues « peu dotées »*. (Thèse de doctorat). Université Joseph-Fourier-Grenoble I.

Bhalla, D., Joshi, N. et Mathur, I. (2013). Rule based transliteration scheme for english to punjabi. *arXiv preprint arXiv :1307.4300*.

Bharadwaj, A., Mortensen, D., Dyer, C. et Carbonell, J. (2016). Phonologically aware neural model for named entity recognition in low resource transfer settings. Dans *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1462–1472.

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

Britz, D., Goldie, A., Luong, T. et Le, Q. (2017). Massive exploration of neural machine translation architectures. *arXiv preprint arXiv :1703.03906*.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. et Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Cao, N. X., Pham, N. M. et Vu, Q. H. (2010). Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model. Dans *Proceedings of the 2010 Symposium on Information and Communication Technology*, 59–63. Association for Computing Machinery.

Chen, L., Li, M., Zhang, J., Zhu, Z. et Yang, Z. (2014). A statistical method for translating chinese into under-resourced minority languages. Dans *China Workshop on Machine Translation*, 49–60. Springer.

Chen, S. F. (2003). Conditional and joint models for grapheme-to-phoneme conversion. Dans *Proceedings of the Interspeech 2003, Geneva, Switzerland*, 2033–2036.

Chen, Y., Liu, Y. et Li, V. O. (2018). Zero-resource neural machine translation with multi-agent communication game. Dans *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cheng, Y., Liu, Y., Yang, Q., Sun, M. et Xu, W. (2016). Neural machine translation with pivot languages. *arXiv preprint arXiv :1611.04928*.

Cheung, H. et Chen, H.-C. (1998). Lexical and conceptual processing in Chinese-English bilinguals : Further evidence for asymmetry. *Memory and Cognition*, *26*(5), 1002–1013.

Chinchor, N., Robinson, P. et Brown, E. (1998). Hub-4 named entity task definition version 4.8. *National Institute of Standards and Technology (NIST)*.

Chinchor, N. A. (1998). *Overview of MUC-7/MET-2*. Rapport technique, Science Applications International Corporation, San Diego, CA.

Chiu, J. P. C. et Nichols, E. (2015). Named entity recognition with bidirectional LSTM-CNNs. *CoRR - Computing Research Repository - arXiv, abs/1511.08308*.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cotterell, R. et Duh, K. (2017). Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. Dans *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, 91–96.

Crystal, D. (2011). *A dictionary of linguistics and phonetics*, volume 30. John Wiley and Sons.

Dahlmann, L., Matusov, E., Petrushkov, P. et Khadivi, S. (2017). Neural machine translation leveraging phrase-based models in a hybrid search. *arXiv preprint arXiv :1708.03271*.

DeFrancis, J. (1977). *Colonialism and language policy in Viet Nam*. Numéro 19. Mouton De Gruyter.

Della Pietra, S., Della Pietra, V. et Lafferty, J. (1997). Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, *19*(4), 380–393.

Diep, D. T. N., Besacier, L. et Castelli, E. (2010). Improved Vietnamese-French parallel corpus mining using English language. Dans *International Workshop on Spoken Language Translation (IWSLT) 2010*.

Do, T. N. D. (2011). *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. (Thèse de doctorat). Université de Grenoble.

Do, T.-N.-D., Le, V.-B., Bigi, B., Besacier, L. et Castelli, E. (2009). Mining a comparable text corpus for a Vietnamese-French statistical machine translation system. Dans *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 165–172. Association for Computational Linguistics.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. et Weischedel, R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. Dans *The fourth of International Conference on Language Resources and Evaluation, Lisbon, Portugal*, 1–4.

Duan, X., Banchs, R. E., Zhang, M., Li, H. et Kumaran, A. (2016). Report of NEWS 2016 machine transliteration shared task. *Association for Computational Linguistics 2016*, 58–72.

Ehrmann, M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation.* (Thèse de doctorat). Paris Diderot University.

Fadaee, M., Bisazza, A. et Monz, C. (2017). Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv :1705.00440.*

Feng, X., Feng, X., Qin, B., Feng, Z. et Liu, T. (2018). Improving low resource named entity recognition using cross-lingual knowledge transfer. Dans *The twenty-seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 4071–4077.

Ferlus, M. (1999). Les dishannonies tonales en viet-muong et leurs implications historiques. *Cahiers de Linguistique Asie Orientale*, *28*(1), 83–100.

Féry, C. (2003). Markedness, faithfulness, vowel quality and syllable structure in french. *Journal of French Language Studies*, *13*(2), 247–280.

Finch, A., Liu, L., Wang, X. et Sumita, E. (2016). Target-bidirectional neural models for machine transliteration. *Association for Computational Linguistics 2016*, 78–82.

Firat, O., Cho, K. et Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv :1601.01073.*

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis.*

Friburger, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques.* (Thèse de doctorat). Université de Tours.

Gao, Q. et Vogel, S. (2008). Parallel implementations of word alignment tool. Dans *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 49–57. Association for Computational Linguistics.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40(3 and 4)*, 237–264.

Grass, T. (2010). A quoi sert encore la traduction automatique ? *Les Cahiers du GEPE. Autres exploitations des outils électroniques, Strasbourg : Presses universitaires de Strasbourg.*

Gu, J., Hassan, H., Devlin, J. et Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv :1802.05368*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H. et Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv :1503.03535*.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y. et Ma, W.-Y. (2016). Dual learning for machine translation. Dans *Advances in Neural Information Processing Systems*, 820–828.

Hirschman, L., Reeder, F., Burger, J. et Miller, K. (2000). Name translation as a machine translation evaluation task. Dans *Proceedings of The International Conference on Language Resources and Evaluation, Athens, Greece*.

Hirst, D. et Di Cristo, A. (1998). *Intonation systems : a survey of twenty languages*. Cambridge University Press.

Hoang, V. C. D., Koehn, P., Haffari, G. et Cohn, T. (2018). Iterative back-translation for neural machine translation. Dans *Proceedings of the second Workshop on Neural Machine Translation and Generation*, 18–24.

Hochreiter, S. et Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, P.-S., Wang, C., Huang, S., Zhou, D. et Deng, L. (2017). Towards neural phrase-based machine translation. *arXiv preprint arXiv :1706.05565*.

Huynh, S. (2010). *Les mécanismes d'intégration des mots d'emprunt français en vietnamien*. L'Harmattan.

Jelinek, F. et Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. Dans *Proceedings Workshop on Pattern Recognition in Practice, 1980*, 381–397.

Jiampojamarn, S., Kondrak, G. et Sherif, T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. Dans *Human Language Technologies - North America Chapter of the Association for Computational Linguistics (NAACL)*, 372–379.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M. et Corrado, G. (2016). Google's multilingual neural machine translation system : enabling zero-shot translation. *arXiv preprint arXiv :1611.04558*.

Jurafsky, D. et Martin, J. H. (2000). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (1st éd.). Upper Saddle River, NJ, USA : Prentice Hall PTR.

Kang, Y., Phạm, A. H. et Storme, B. (2016). French loanwords in Vietnamese : the role of input language phonotactics and contrast in loanword adaptation. Dans *Proceedings of the Annual Meetings on Phonology*, 3–15.

Karimi, S., Scholer, F. et Turpin, A. (2011). Machine transliteration survey. *Association for Computing Machinery (ACM) Computing Surveys (CSUR)*, *43*(3), 17.

Kashani, M. M., Popowich, F. et Sarkar, A. (2007). Automatic transliteration of proper nouns from Arabic to English. Dans *Proceedings of the Second Workshop on Computational Approaches to Arabic Script-based Languages*, 275–282.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. Dans *IEEE Transactions on Acoustics, Speech and Signal Processing*, 400–401.

Kingma, D. P. et Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980.*

Klementiev, A. et Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 817–824. Association for Computational Linguistics.

Kneser, R. et Ney, H. (1995). Improved backing-off for m-gram language modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 181–184.

Knight, K. et Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, *24*(4), 599–612.

Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

Koehn, P. (2017). Neural machine translation. *arXiv preprint arXiv :1709.07809.*

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C. et Zens, R. (2007). Moses : Open source toolkit for statistical machine translation. Dans *Proceedings of the forty-fifth annual meeting of the Association for Computational Linguistics on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics.

Koehn, P. et Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv :1706.03872*.

Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical phrase-based translation. Dans *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48–54. Association for Computational Linguistics.

Kumano, T., Kashioka, H., Tanaka, H. et Fukusima, T. (2004). Acquiring bilingual named entity translations from content-aligned corpora. Dans *International Conference on Natural Language Processing*, 177–186. Springer.

Kumaran, A., Khapra, M. M. et Li, H. (2010). Report of the named entities workshop 2010 transliteration mining shared task. Dans *Proceedings of the 2010 Named Entities Workshop*, 21–28. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A. et Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. Morgan Kaufmann Publishers Incorporation.

Lagrée, S., Cling, J.-P., Razafindrakoto, M. et Roubaud, F. (2009). *Stratégies de réduction de la pauvreté : approches méthodologiques et transversales*. Editions The Gioi, Hanoi.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. et Dyer, C. (2016). Neural architectures for named entity recognition. *CoRR - Computing Research Repository - arXiv, abs/1603.01360*.

Lample, G., Ott, M., Conneau, A., Denoyer, L. et Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv :1804.07755*.

Laurent, A., Deléglise, P., Meignier, S. et Spécinov-Trélazé, F. (2009). Grapheme to phoneme conversion using an SMT system. Dans *Proceedings of the Interspeech 2009, ISCA, Brighton, United Kingdom*, 708–711.

Le, N. T., Mallek, F. et Sadat, F. (2016). UQAM-NTL : Named entity recognition in twitter messages. Dans *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, Association for Natural Language Processing (ANLP), the 2nd Workshop on Noisy User-generated Text, Osaka, Japan, 11-16 December*, 197–202.

Le, N. T. et Sadat, F. (2017). A neural network transliteration model in low resource settings. Dans *Proceedings of the 16th International Conference of Machine Translation Summit, 18-22 September 2017, Nagoya, Japan, Research Track*, 337–345.

Lewis, W. D. et Yang, P. (2012). Building MT for a severely under-resourced language : White Hmong. *Association for Machine Translation in the Americas, October*.

Luong, M.-T., Nakov, P. et Kan, M.-Y. (2010). A hybrid morpheme-word representation for machine translation of morphologically rich languages. Dans *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 148–157. Association for Computational Linguistics.

Ma, X. et Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv :1603.01354*.

Mangeot, M. et Sadat, F. (2014). TALN-RECITAL 2014 workshop TALAf 2014 : Traitement automatique des langues africaines (TALAf 2014 : African language processing). Dans *TALN-RECITAL 2014 Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014 : African Language Processing)*.

Martinet, A. (1970). *Economie des changements phonétiques*. Editions A. Francke.

Mastropavlos, N. et Papavassiliou, V. (2011). Automatic acquisition of bilingual language resources. Dans *Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Dans *Advances in neural information processing systems*, 3111–3119.

Miller, G. (1998). *WordNet : An electronic lexical database*. MIT press.

Minh, D. T. (2011). Pour une réflexion sociolinguistique sur les emprunts directs faits au français par la langue vietnamienne. *Synergies pays riverains du Mékong numéro spécial. Dictionnaire vietnamien - français : Les mots vietnamiens d'origine française*, 243–251.

Modard, D. et Vignes, L. (2006). *Les enjeux sociaux du langage. Hommage à Bernard Gardin*.

Moore, R. C. (2003). Learning translations of named-entity phrases from parallel corpora. Dans *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 259–266. Association for Computational Linguistics.

Nadeau, D. et Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Ngo, H. G., Chen, N. F., Nguyen, B. M., Ma, B. et Li, H. (2015). Phonology-augmented statistical transliteration for low-resource languages. Dans *Proceedings of the Interspeech 2015, Dresden, Germany*, 3670–3674.

Nguyen, D. B., Hoang, S. H., Pham, S. B. et Nguyen, T. P. (2010). Named entity recognition for Vietnamese. Dans *Asian Conference on Intelligent Information and Database Systems*, 205–214. Springer.

Nguyen, K. A. et Dinh, D. (2011). Tích hợp thông tin từ loại vào hệ dịch máy thống kê. Dans *Proceedings of the National Conference, Can Tho, Viet Nam*, 150–157.

Nicolai, G., Hauer, B., Salameh, M., St Arnaud, A., Xu, Y., Yao, L. et Kondrak, G. (2015). Multiple system combination for transliteration. Dans *Proceedings of the Fifth Named Entities Workshop*, 72–79.

Nouvel, D., Antoine, J.-Y. et Friburger, N. (2011). Pattern mining for named entity recognition. Dans *Proceedings of the Language and Technology Conference*, 226–237. Springer.

Nouvel, D., Antoine, J.-Y., Friburger, N. et Soulet, A. (2013). Fouille de règles d'annotation pour la reconnaissance d'entités nommées. *Traitement Automatique des Langues*, *54*(2), 13–41.

Nuessel, F. H. (1992). *The study of names : a guide to the principles and topics*. Greenwood Publishing Group.

Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51.

Oh, J.-H., Choi, K.-S. et Isahara, H. (2006). A machine transliteration model based on correspondence between graphemes and phonemes. *Association for Computing Machinery Transactions on Asian Language Information Processing (TALIP)*, *5*(3), 185–208.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N. et Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Dans *Proceedings of the 51th annual meeting of the North America Chapter of the Association for Computational Linguistics*, 380–390. Association for Computational Linguistics.

Palmer, D. D. et Day, D. S. (1997). A statistical profile of the named entity task. Dans *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLC-97)*, 190–193. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. Dans *Proceedings of the fortieth annual meeting on Association for Computational Linguistics*, 311–318. Association for Computational Linguistics.

Peris, Á. (2017). NMT-Keras. `https://github.com/lvapeab/nmt-keras`. GitHub repository.

Pham, T. T. X., Tran, T., Dinh, D. et Nigel, C. (2007). Named entity recognition in Vietnamese using classifier voting. *Association for Computing Machinery Transaction on Asian and Low-Resource Language Information Processing*, *6*(4), 1–18.

Phe, H. (1997). Vietnamese dictionary. *Vietnam Lexicography Centre, Da Nang Publishing House*.

Phuong, C. X. (1973). *L'étude de la vietnamisation du vocabulaire français depuis l'intervention française au Vietnam jusqu'à nos jours*. (Thèse de doctorat).

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M. et Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. Dans *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, 20–21.

Poibeau, T. (2003). The multilingual named entity recognition framework. Dans *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 155–158. Association for Computational Linguistics.

Potet, M., Besacier, L. et Blanchon, H. (2010). The LIG machine translation system for WMT 2010. Dans *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, 161–166. Association for Computational Linguistics.

Ramsay-Brijball, M. (1999). Understanding Zulu-English code-switching : A psychosocial perspective. *South African Journal of Linguistics*, *17*(2-3), 161–172.

Rao, K., Peng, F., Sak, H. et Beaufays, F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. Dans *Proceedings of the 2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4225–4229. Institute of Electrical and Electronics Engineers.

Rijsbergen, C. J. V. (1979). *Information Retrieval* (second éd.). Newton, MA, USA : Butterworth-Heinemann.

Ritter, A., Clark, S., Mausam et Etzioni, O. (2011). Named entity recognition in tweets : An experimental study. Dans *Proceedings of the Conference on Empirical Methods*

in Natural Language Processing (EMNLP-11), 1524–1534. Association for Computational Linguistics.

Rumellhart, D. (1986). Learning internal representations by error propagation. *Parallel distributed processing*, 318–362.

Sajjad, H., Schmid, H., Fraser, A. et Schütze, H. (2017). Statistical models for unsupervised, semi-supervised and supervised transliteration mining. *Computational Linguistics*.

Sánchez, L. (2003). *Quechua-Spanish bilingualism : Interference and convergence in functional categories*. John Benjamins Publishing.

Sekine, S. et Isahara, H. (2000). Irex : IR and IE evaluation project in Japanese. Dans *Proceedings of the International Conference on Language Resources and Evaluation*, 1977–1980. Citeseer.

Sennrich, R., Haddow, B. et Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv :1511.06709*.

Shao, Y. et Nivre, J. (2016). Applying neural networks to English-Chinese named entity transliteration. Dans *Proceedings of the Sixth Named Entity Workshop, joint with the fifty-fourth Association for Computational Linguistics*.

Sharwood, B. (2013). Machine translation of under-resourced languages. *University of Cape Town, Honours project report*.

Snover, M. G., Madnani, N., Dorr, B. et Schwartz, R. (2009). TER-plus : paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, *23*(3), 117–127.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. Dans *Proceedings of the Interspeech 2002, Denver, Colorado, USA*, 2002–2005.

Storme, B. (2015). The loi de position in french. *Unpublished manuscript, MIT*.

Sundheim, B. M. (1996). Overview of results of the MUC-6 evaluation. Dans *Proceedings of a workshop on held at Vienna, Virginia, May 6-8, 1996*, 423–442. Association for Computational Linguistics.

Thi, L. N., My, L. H., Viet, H. N., Minh, H. N. T. et Le Hong, P. (2013). Building a treebank for vietnamese dependency parsing. Dans *The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*, 147–151. IEEE.

Thompson, L. C. (1987). *A Vietnamese reference grammar*, volume 13. University of Hawaii Press.

Thu, Y. K., Pa, W. P., Sagisaka, Y. et Iwahashi, N. (2016). Comparison of grapheme–to–phoneme conversion methods on a Myanmar pronunciation dictionary. Dans *Proceedings of the sixth Workshop on South and Southeast Asian Natural Language Processing 2016*, 11–22.

Tjong Kim Sang, E. F. et De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. Dans *Proceedings of the seventh conference on Natural language learning at Human Language Technologies - North America Chapter of the Association for Computational Linguistics*, 142–147. Association for Computational Linguistics.

Toan, V. (1992). Từ gốc pháp trong tiếng việt. *Nxb KHXH, Hanoi, Vietnam.*

Tran, P., Dinh, D. et Nguyen, H. T. (2016). A character level based and word level based approach for Chinese-Vietnamese machine translation. *Computational intelligence and neuroscience, 2016.*

Tran, Q. T., Pham, T. T., Ngo, Q. H., Dinh, D. et Collier, N. (2007). Named entity recognition in Vietnamese documents. *Progress in Informatics Journal, 5*, 14–17.

Tsuji, K., Daille, B. et Kageura, K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. Dans *Language Resources and Evaluation Conference (LREC)*. Citeseer.

Tu, N. C., Oanh, T. T., Hieu, P. X. et Thuy, H. Q. (2005). Named entity recognition in Vietnamese free-text and web documents using conditional random fields. Dans *Proceedings of the 8th Conference on some selection problems of Information Technology and Telecommunication*. Citeseer.

Turian, J., Ratinov, L. et Bengio, Y. (2010). Word representations : a simple and general method for semi-supervised learning. Dans *Proceedings of the forty-eighth annual meeting of the association for computational linguistics*, 384–394. Association for Computational Linguistics.

Udupa, R., Saravanan, K., Kumaran, A. et Jagarlamudi, J. (2009). Mint : A method for effective and scalable mining of named entity transliterations from large comparable corpora. Dans *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 799–807. Association for Computational Linguistics.

Vu, D. H. (2011). Phân đoạn từ tiếng Việt ngữ dụng. Dans *Proceedings of the National Conference, Can Tho, Viet Nam*, 140–147.

Wang, X., Tu, Z., Xiong, D. et Zhang, M. (2017). Translating phrases in neural machine translation. *arXiv preprint arXiv :1708.01980*.

Waxmonsky, S. et Reddy, S. (2012). G2P conversion of proper names using word origin information. Dans *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, 367–371. Association for Computational Linguistics.

Wen, L., Lei, C., Wudabala, H. et Miao, L. (2010). Chained machine translation using morphemes as pivot language. Dans *Proceedings of the Eighth Workshop on Asian Language Resources*, 169–177.

Witten, I. H. et Bell, T. C. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inf. Theor.*, *37*(4), 1085–1094. `http://dx.doi.org/10.1109/18.87000`. Récupéré de `http://dx.doi.org/10.1109/18.87000`

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q. et Macherey, K. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*.

Xuan, X. V. (2002). Le phonème français et le phonème vietnamien. *Études de linguistique appliquée*, (2), 225–236.

Yang, Z., Chen, W., Wang, F. et Xu, B. (2018). Unsupervised neural machine translation with weight sharing. *arXiv preprint arXiv :1804.09057*.

Yao, K. et Zweig, G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv :1506.00196*.

Yao, Y. et Huang, Z. (2016). Bi-directional LSTM recurrent neural network for Chinese word segmentation. *CoRR - Computing Research Repository - arXiv, abs/1602.04874*.

Zhang, J. et Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. Dans *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1535–1545.

Zuercher, K. (2009). *Azerbaijani-Russian code-switching and code-mixing : Form, function, and identity*. The University of Texas at Arlington.