UNIVERSITÉ DE FRANCHE-COMTÉ
ÉCOLE DOCTORALE «LANGAGES, ESPACES, TEMPS, SOCIÉTÉS»

Thèse en vue de l'obtention du titre de docteur en

SCIENCES DU LANGAGE
SPÉCIALITÉ : TRAITEMENT AUTOMATIQUE DES LANGUES

MACHINE TRANSLATION OF PROPER NAMES FROM ENGLISH AND FRENCH INTO VIETNAMESE: AN ERROR ANALYSIS AND SOME PROPOSED SOLUTIONS

Traduction automatique des noms propres de l'anglais et du français vers le vietnamien : analyse des erreurs et quelques solutions

Présentée et soutenue publiquement par

Thao PHAN THI THANH

Le 11 mars 2014

Sous la direction de Mme le Professeur Sylviane CARDEY-GREENFIELD
et la co-direction de M. Dr. Ha LE AN et Mme. Dr. Izabella THOMAS

Membres de Jury:

Sylviane CARDEY-GREENFIELD, Directrice de recherche, université de Franche-Comté, France
Ha LE AN, Docteur, HDR, Co-directeur de recherche, université de Wolverhampton, Royaume-Uni
Denis MAUREL, Professeur, université François Rabelais de Tours, France, Rapporteur
Ruslan MITKOV, Professeur, université de Wolverhampton, Royaume-Uni, Rapporteur
Izabella THOMAS, Docteur, Co-directrice de recherche, université de Franche-Comté, France.

RÉSUMÉ

Dans l'ère de l'information et de la connaissance, la traduction automatique (TA) devient progressivement un outil indispensable pour transposer la signification d'un texte d'une langue source vers une langue cible. La TA des noms propres (NP), en particulier, joue un rôle crucial dans ce processus, puisqu'elle permet une identification précise des personnes, des lieux, des organisations et des artefacts à travers les langues. Malgré un grand nombre d'études et des résultats significatifs concernant la reconnaissance d'entités nommées (dont le nom propre fait partie) dans la communauté de TAL dans le monde, il n'existe presque aucune recherche sur la traduction automatiques des noms propres (TANP) pour le vietnamien.

En raison des caractéristiques particulières d'écriture, de translittération/ transcription et de traduction des NP dans une variété des langues telles que l'anglais, le français, le russe, le chinois, etc., la TANP depuis ces langues vers le vietnamien constitue un défi complexe. Notre recherche se focalise sur les erreurs de TANP d'anglais vers le vietnamien et de français vers le vietnamien, résultant des systèmes courants de TA. Tout d'abord, elle propose une classification des NP basée sur le corpus, ensuite une analyse des erreurs de la TANP, pour conclure par une proposition de solution de prétraitement pour améliorer la qualité de la TA.

A travers l'analyse et la classification d'erreurs de la TANP faites sur deux corpus parallèles de textes avec PN (anglais-vietnamien et français-vietnamien), nous proposons les solutions concernant deux problématiques importantes : (1) l'annotation de corpus, afin de préparer des bases de données pour le prétraitement et (2) la création d'un programme pour prétraiter automatiquement les corpus annotés, afin de réduire les erreurs de la TANP et d'améliorer la qualité de traduction des systèmes de TA, tels que Google, Vietgle, Bing et EVTran.

L'efficacité de différentes méthodes d'annotation des corpus avec des NP ainsi que les taux d'erreurs de la TANP avant et après l'application du programme de prétraitement sur les deux corpus annotés sont comparés et discutés dans cette thèse. Ils prouvent que le prétraitement réduit significativement le taux d'erreurs de la TANP et, par la même, contribue à l'amélioration de traduction automatique vers la langue vietnamienne.

Mots-clés : nom propre, prétraitement, corpus parallèle, qualité de traduction automatique, anglais-vietnamien, français-vietnamien, erreur de traduction

TABLE DES MATIÈRES

RÉSUMÉ I
CHAPITRE 1 : INTRODUCTION 1
1.1 RAISONNEMENT 1
1.2 OBJECTIFS D'ÉTUDE 3
1.3 MÉTHODOLOGIE 4
1.4 STRUCTURE DE THÈSE 4
CHAPITRE 2 : ÉTAT DE L'ART 7
2.1 TRADUCTION AUTOMATIQUE 7
2.2 NOMS PROPRES 8
2.3 TRADUCTION AUTOMATIQUE DE NOMS PROPRES 9
CHAPITRE 3 : ANALYSE D'ERREURS DE LA TANP À BASE DE CORPUS 11
3.1 PRÉPARATION D'UNE BASE DE DONNÉES : CONSTRUCTION DU CORPUS PARALLÈLE DE TEXTES AVEC DES NOMS PROPRES 11
3.2 ANALYSE D'ERREURS DE LA TANP À BASE DE CORPUS 12
3.2.1 Classification de NPs 12
3.2.2 Classification d'erreurs de TANP 12
3.2.3 Analyse qualitative des erreurs de la TANP avec des exemples 15
CHAPITRE 4 : ANNOTATION DE CORPUS 17
4.1 CADRE D'ANNOTATION 17
4.2 MÉTHODES D'ANNOTATION 18
4.2.1 Annotation manuelle 18
4.2.2 Annotation automatique et annotation semi-automatique 19
4.3 RÉSULTATS D'ANNOTATION DE CORPUS 20
4.4 COMPARAISON DES RÉSULTATS DES DIFFÉRENTES MÉTHODES D'ANNOTATION 21
CHAPITRE 5 : AMÉLIORATION DE LA TRADUCTION AUTOMATIQUE PAR PRÉTRAITEMENT 22
5.1 PRÉ-TRAITEMENT 22
5.1.1 Définition et objectif du prétraitement 22
5.1.2 Tâches spécifiques du prétraitement pour la TA anglais-vietnamienne et français-vietnamienne 22
5.2 DESCRIPTION DU PROGRAMME DE PRÉTRAITEMENT POUR LA RÉDUCTION D'ERREURS DE LA TANP 24
5.2.1 Tâches principales 24
5.2.3 Interface et démos du programme de prétraitement 24
5.2.4 Avantages et limites du programme de prétraitement 28
5.3 RÉSULTATS DES TESTES AVREC ET SANS LE PRÉTRAITEMENT 29
5.3.1 Résultats des testes sur le corpus anglais-vietnamien avant et après le prétraitement 29
5.3.2. Résultats des testes sur le corpus franco-vietnamien de PNs avant et après le prétraitement 31
5.4 COMPARAISON GÉNERALE ET ÉVALUATION DES SYSTÈMES DE TA ANGLAIS-VIETNAMIEN ET FRANÇAIS-VIETNAMIEN 32
5.4.1 Systèmes de TA anglais-vietnamien 32
5.4.2 Systèmes de TA français-vietnamien 34
CHAPTER 6 : CONCLUSION 36
6.1 CONTRIBUTION DE LA THÈSE 36
6.2 LIMITATIONS DE LA THÈSE ET RECHERCHE A VENIR 38
BIBLIOGRAPHIE SELECTIVE 40

CHAPITRE 1 : INTRODUCTION

1.1 RAISONNEMENT

L'importance croissante de la traduction automatique (TA) n'est plus à souligner, particulièrement dans notre ère d'information et de la connaissance, dans laquelle les besoins d'accès, d'échange et de transmission d'une grande quantité d'informations augmentent rapidement dans le monde entier. La TA est devenue indispensable pour plusieurs raisons que l'on peut résumer ainsi.

Premièrement, la TA est un outil simple d'utilisation, rapide et bon marché ; elle peut être utilisée pour remplacer des humains pour traduire facilement une grande quantité de documents, en peu de temps et avec un modeste investissement financier.

Deuxièmement, la meilleure qualité des traductions effectuées par les systèmes de TA a changé l'attitude des utilisateurs potentiels envers ces outils. Somers (2003:514) affirme que les bonnes performances d'un certain nombre de systèmes de TA, gratuitement disponibles sur le World Wide Web, ont amplifié l'intérêt du public pour la TA.

Troisièmement, le nombre d'internautes, qui s'intéressent à la TA, augmente constamment. Grâce à sa facilité d'utilisation et à sa vitesse, de plus en plus de personnes se servent de la TA pour leurs communications quotidiennes (Wilks 2008:10)

Quatrièmement, la mauvaise qualité de la TA provient des erreurs de la traduction parmi lesquelles *la mauvaise traduction d'*entités nommées (ENs) est particulièrement importante. La bonne reconnaissance et traduction des ENs (Hermjakob et al. 2008) peut avoir un grand impact sur la qualité des résultats de traduction.

Wilks (2008) estime que la TA est une technologie vive et essentielle. Son importance dans le monde de l'information multilingue ne peut qu'augmenter, aussi bien intellectuellement que commercialement. Malgré le développement constant et les achèvements significatifs dans le domaine du TAL, les systèmes de TA rencontrent encore de nombreux défis ; en particulier concernant la traduction de noms propres (Aone et 1998 Maloney, Hirschman et al. 2000, Somers 2003 :523). Selon Aone (1998), il y a deux cas où un système de TA échoue à traduire correctement les noms propres (NPs).

D'abord, les systèmes de TA échouent à déterminer les frontières d'un nom propre; cela cause un important problème pour les langues utilisant les logogrammes telles que le japonais (kanji), le chinois (hanzi), le coréen (hanja), etc. Les systèmes de TA souvent "découpent" des NPs en mots et traduisent chaque mot individuellement, d'autant plus que dans ces langues les NPs ne commencent pas par une majuscule. Par exemple, un nom propre japonais d'une personne "Mori Hanae", écrit avec des caractères kanji, est segmenté en trois mots incluant "mori" (la forêt), "hana" (l'Angleterre) et "e" (la bénédiction) (Aone 1998). Dans le fait, ce nom de personne ne devrait être ni segmenté, ni traduit.

Ensuite, les systèmes de TA échouent parfois à distinguer des NPs des noms communs. Hirschman et al. (2000) identifient cette erreur de traduction de NPs comme s'ils étaient des mots "normaux" noms communs significatifs, comme typique de la TA. Ce type d'erreurs, que commettent également les systèmes de TA anglais-vietnamien, arrive souvent avec des noms de personnes, des abréviations ou des acronymes de noms géographiques et d'organisations. Par exemple, quelques patronymes anglais comme "*Brown, Rice, Greenfield, Mark*" sont automatiquement traduits vers le vietnamien comme s'ils étaient des noms communs ayant une signification : "*marron, riz, champ vert, marque*". Dans le fait, ces noms de personnes devraient rester inchangés.

Concernant la traduction automatique de NPs de l'anglais et du français vers le vietnamien, les systèmes de TA font souvent deux types majeurs d'erreurs, que nous classifions en erreurs de non-traduction et erreurs de mauvaise traduction. Les erreurs de non-traduction arrivent fréquemment avec les abréviations, les titres professionnels, les noms d'organisation, de jours, etc. Les erreurs de mauvaise traduction de NPs sont dues à leurs différentes caractéristiques telles que des caractéristiques graphiques, lexicales, syntaxiques et des problèmes de translittération ou de transcription. Par exemple, les jours et les mois sont graphiquement écrits avec les initiales en majuscule en anglais, mais avec une minuscule en vietnamien, d'où les erreurs de traduction (par exemple, *Sunday* est mal traduit en vietnamien par *Chủ nhật* ou *Chủ Nhật*, alors que la bonne traduction serait *chủ nhật*). En français, les noms des nationalités ne commencent pas en majuscule, contrairement au vietnamien. Par exemple, *russe* est mal traduit en vietnamien par *tiếng nga*, ce qui devrait être corrigé par *tiếng Nga*.

Les erreurs de mauvaise traduction résultent de l'absence de distinction entre des NPs et des noms communs. Certains noms géographiques en anglais sont

traduits en vietnamien; par exemple, le *West Ham* (nom d'une place à Londres) est traduit par EVTran comme *Đùi Phương tây* (c'est-à-dire "jambon à l'Ouest*"*). La même problématique se pose avec les noms d'organisations, lesquelles ne devraient pas être traduits, par exemple, *Marks et Spencer* sont mal rendus par le système de TA EVTran comme *Những sự đánh dấu và Spence* (c'est-à-dire: "marquer et Spencer*"*).

1.2 OBJECTIFS D'ÉTUDE

Cette thèse se propose d'atteindre les neuf objectifs suivants :

Objectif 1 : Revoir les différentes classifications des NPs, aussi bien linguistiques que computationnelles, en anglais et en français pour établir une classification des NPs la plus appropriée pour une traduction automatique de NPs de l'anglais et du français vers le vietnamien.

Objectif 2 : Créer deux corpus bilingues annotés de NPs : les corpus anglais-vietnamiens et français-vietnamiens de textes avec NPs. Premièrement, ces corpus serviront aux objectifs de cette thèse. Deuxièmement, ils seront mis à la disposition de la communauté universitaire pour mener des recherches plus approfondies sur la traduction automatique de noms propres (TANP).

Objectif 3 : Proposer une classification de noms propres basée sur l'analyse de corpus et satisfaisant à la fois les domaines du TAL et de la linguistique.

Objectif 4 : Effectuer l'analyse et la classification d'erreurs de TANP faites par les systèmes actuels de TA en utilisant à la fois la classification et les deux corpus de noms propres établis auparavant: corpus parallèle anglais-vietnamien (EVC) et corpus parallèle français-vietnamien (FVC).

Objectif 5 : Etablir la meilleure méthode d'annotation de corpus en NPs; nous proposerons et comparerons les résultats de trois méthodes d'annotation : l'annotation manuelle, l'annotation automatique et l'annotation semi-automatique.

Objectif 6 : Construire le programme de prétraitement des NPs pour l'anglais et le français afin de réduire les erreurs de TANP et d'améliorer la qualité de la TA en général, ainsi que la qualité de traduction de quatre systèmes de la TA (Vietgle, Google Translate, Bing Translator et EVTran) en particulier.

Objectif 7: Tester le programme de prétraitement sur les deux corpus de NPs et évaluer les résultats de traduction de NP fournis par les différents systèmes avec et

sans utilisation de ce programme. La comparaison des systèmes de TA sera effectuée pour révéler les avantages et les limites de chaque système.

Objectif 8 : Indiquer les avantages et les limites du programme de prétraitement afin de fournir des pistes d'amélioration pour la TANP.

Objectif 9 : Créer deux glossaires bilingues des NPs pouvant servir à améliorer la traduction de NPs de l'anglais et le français vers le vietnamien.

1.3 MÉTHODOLOGIE

La méthodologie de notre thèse comporte cinq étapes majeures :

i/ Construction des corpus parallèles de textes avec NPs afin de préparer la base de données pour l'analyse d'erreurs de TANP et d'évaluer la performance de systèmes de TA ;

ii/ Analyse d'erreurs de TANP, à base de corpus et leur classification, selon les deux méthodes : linguistique et computationnelle ;

iii/ Annotation de corpus et évaluation de méthodes d'annotation ;

iv/ Conception et testes du programme de prétraitement sur deux corpus annotés de NPs ;

v/ Évaluation d'amélioration de qualité de TA après utilisation du programme de prétraitement.

1.4 STRUCTURE DE THÈSE

Le Chapitre 1 présente les motivations principales de notre étude, les objectifs de recherche et la méthodologie utilisée pour réaliser nos objectifs.

Dans **le Chapitre 2**, nous décrivons l'état de l'art dans les deux domaines liés à notre recherche : la traduction automatique et les différents traitements de noms propres. La première partie de ce chapitre décrit le développement des systèmes de TA au Vietnam et, particulièrement, des moteurs de traduction utilisés les plus couramment, avec leurs avantages et limites. La seconde partie du chapitre est consacrée à la description des classifications des NPs, aussi bien du point de vue de la linguistique que du TAL; elle traite aussi les problématiques de la TANP de l'anglais et du français vers le vietnamien, constatant une absence de travaux de recherche dans ce domaine.

Dans **le Chapitre 3**, nous décrivons les deux corpus parallèles bilingues de textes avec NPs que nous avons créés, à savoir le corpus parallèle anglais-vietnamien (EVC) et le corpus parallèle français-vietnamien (FVC) ; nous proposons aussi notre propre classification de NPs utilisée ultérieurement pour l'analyse d'erreurs de TANP. Nous décrivons les deux classifications des erreurs de la TANP de E-V et F-V à partir d'une série d'exemples extraits de nos corpus.

Dans **le Chapitre 4**, nous abordons les questions relatives à l'annotation de corpus anglais et français de NPs, et nous focalisons sur la meilleure méthode d'annotation pour permettre au programme de prétraitement d'améliorer la qualité de la TA. Nous présentons des différentes méthodes d'annotation incluant la méthode manuelle, la méthode automatique et la méthode semi-automatique. Nous analysons, classifions et comparons les résultats d'annotation de deux corpus anglais et français de NPs pour chacune de ces méthodes.

Le chapitre 5 décrit la problématique de prétraitement et la création du programme de prétraitement pour la réduction d'erreurs de la TANP. Ce programme est établi pour effectuer les tâches spécifiques, identifiées auparavant comme améliorant la TA de NP : étiquetage en "Ne-pas-traduit" (DNT, DO-NOT-TRANSLATE) de NPs du corpus annoté, changement des structures possessives avec NPs en anglais et l'omission des groupes français "de + déterminants" précédant certains noms géographiques et des noms d'organisations. La description du programme est détaillée en tâches principales et spécifications, interface et demos, et avantages et limitations de l'utilisation.

Le Chapitre 5 propose les résultats quantitatifs de notre recherche et la comparaison générale et l'évaluation des systèmes de TA anglais-vietnamien et français-vietnamien. D'abord, nous montrons les résultats de tests d'erreurs de TANP sur les deux corpus de NPs, avec et sans utilisation du programme de prétraitement. Ensuite, nous évaluons les quatre moteurs de TA anglais-vietnamien et les deux moteurs de TA français-vietnamien utilisés dans nos expériences, nous les comparons selon des différents critères tels que la vitesse de traduction, la taille des documents source, le nombre d'options de langue traitées, la vitesse de traitement des documents source, le nombre d'erreurs de TANP avant et après l'utilisation du programme de prétraitement, etc.

Le chapitre 6 récapitule la thèse avec une discussion portant sur les contributions et limites de l'étude. Ce chapitre propose différentes voies d'investigation en vue de recherches ultérieures.

CHAPITRE 2 : ÉTAT DE L'ART

Ce chapitre introduit le contexte théorique ayant trait à trois questions majeures de notre étude : 1) traduction automatique, 2) noms propres et 3) traduction automatique de noms propres.

2.1 TRADUCTION AUTOMATIQUE

La première partie de ce chapitre concerne le développement de la TA au Vietnam. Bien que les systèmes de TA aient été développés depuis les années 1940 dans le monde, on a commencé à les étudier au Vietnam seulement depuis les années 1960. L'intérêt pour la TA s'est plus particulièrement manifesté à partir des années 2000. Le développement de la TA au Vietnam peut être divisé en 4 périodes :

1. 1960 à 1970 : début des projets de recherche sur la TA du vietnamien ;

2. De 1970 aux années 1990 : une période appelée "le temps de fermeture" puisque la TA a été complètement négligée ; il n'existe aucun progrès significatif dans cette période ;

3. Des années 1990 aux années 2000 : apparition du premier système de TA au Vietnam – EVTran ;

4. Des années 2000 jusqu'à présent : période florissante de la TA au Vietnam grâce au fort développement de la technologie de l'information ; apparition de quatre systèmes de TA pour "anglais-vietnamien" : Vietgle, Google Translate, Bing Translator et EVTran.

La seconde partie de ce chapitre étudie l'impact des méthodes de prétraitement sur la qualité de la TA, particulièrement en ce qui concerne les deux paires de langues : anglais-vietnamien et français-vietnamien. Le prétraitement est défini comme une tâche spécifique, permettant d'anticiper certains problèmes de traduction concernant le lexique, la sémantique et la syntaxe d'un texte source. L'objectif principal de la tâche de prétraitement consiste à détecter et à "corriger" les erreurs anticipées dans le texte source avant de le soumettre à un système de TA. Cette étude se propose d'élaborer un programme de prétraitement, pouvant être appliqué aux différents systèmes de TA, afin d'améliorer la TA de NPs de l'anglais et du français vers le vietnamien.

2.2 NOMS PROPRES

La deuxième partie de ce chapitre présente les différentes classifications de NPs, aussi bien du point de vue de la linguistique que du TAL. Puisque les NPs jouent un rôle important dans toutes sortes de textes, il existe beaucoup de recherches les concernant. Leroy (2004) énumère plusieurs études linguistiques concernant les NPs, parmi lesquelles Algeo J. (1973), Kleiber (1981), Molino (1982), Siblot (1987), Le Bot (1989), Gary-Prieur (2001), Jonasson (1994), Maurel et Geuthner (2000), Van de Velde et Flaux (2000), Montecot, Osipov, Vassilaki (2001), Vaxelaire (2005), etc. Nous présentons et analysons certaines classifications de NPs, allant de classifications simples aux plus complexes dans l'ordre chronologique, proposées par les linguistes suivants : Zabeeh (1968), Molino (1982), Bauer (1985) et Grass (1999) (cité par Daille et al. 2000), Leroy (2004) et Vaxelaire (2005).

À l'instar les linguistes, de nombreux chercheurs en TAL ont enquêté sur les NPs, spécifiquement dans le domaine de la reconnaissance d'entités nommées (NER), qui a un grand impact sur les diverses applications du TAL, telles que la traduction automatique (TA), l'extraction de l'information interlangue (CLIR), extraction de l'information (IE), systèmes de question-réponse (QA), moteurs de recherche sur l'Internet, peuplement d'ontologies (Kozareva et al. 2008), filtrage de contenu du Web (Hidalgo, Garcia et Sanz, 2005), désambiguïsation de mots en majuscule pour identifier les NPs (Mikheev 2002) etc. Plusieurs événements importants se concentrent sur les questions de NER : MUC-7 (Chinchor 1997), IREX (Sekine & Isahara 2000), CoNLL-2002 et CoNLL-2003 (Sang 2002, Sang & De Meulder 2003), ACE (Doddington et al. 2002), et HAREM (Santos et al. 2006) ; lors de ces manifestations, on propose plusieurs classifications et traitements différents de NPs dans le domaine du TAL.

Les problèmes les plus souvent rencontrés dans ces deux types de classifications de NPs (linguistiques et TAL) sont l'absence de certains types et les sous-types de NPs ainsi que l'inconsistance de classification de NPs. Par conséquent, nous proposons d'établir notre propre classification de NPs destinée à l'analyse de NPs dans des domaines de la linguistique et du TAL, à la fois et orientée vers des recherches menées en TA.

2.3 TRADUCTION AUTOMATIQUE DE NOMS PROPRES

La troisième partie de ce chapitre fait un état de l'art de la TANP d'une langue étrangère au vietnamien. Bien qu'il existe des études portant sur la TANP pour différentes paires de langues et pour certaines questions de NER relatives au vietnamien, la TANP concernant l'anglais-vietnamien et le français-vietnamien n'a jamais été abordée jusqu'à ce jour. Ainsi, notre étude est la première contribution à l'étude de la TANP et nous espérons qu'elle sera utile pour le développement de la TA.

Il existe plusieurs études sur la TA de NPs pour les diverses paires de langues telles que : anglais-chinois (Chen et al. 1998), espagnol-anglais (Hirschman et al. 2000), français-anglais (Noir 1995, Moore et Robert 2003, Moshop 2007), anglais-arabe (Izwaini 2006, Kashani et al. 2007), anglais-japonais (Kumano et al. 2004). Nous signalons notamment l'étude de la TANP utilisant Prolexbase (Maurel et al. 2006) concernant la traduction vers l'allemand, l'anglais, l'italien, le hollandais, le polonais, le portugais, et le serbe. Nous n'avons trouvé aucune recherche concernant la TANP pour l'anglais-vietnamien et le français-vietnamien. Les systèmes de TA doivent faire face à de nombreux problèmes concernant ces deux paires de langues. Ces problèmes sont dérivés des caractéristiques des NPs de l'inconsistance de leur écriture et de la transcription ou la translittération en vietnamien.

En ce qui concerne la traduction de NPs d'une langue étrangère vers le vietnamien, nous discutons trois principales questions :

1) les principes d'écriture des noms propres en vietnamien ;

2) la transcription/translittération et la traduction de NPs d'une langue étrangère vers le vietnamien ;

3) les défis de la TANP de l'anglais et du français vers le vietnamien.

Afin de proposer les principes d'écriture de NPs en vietnamien et de normaliser le processus de translittération/transcription et de la traduction de NPs d'une langue étrangère vers le vietnamien, nous divisons l'ensemble des NPs vietnamiens en deux types : NPs d'origine vietnamienne et NPs d'origine étrangère. En ce qui concerne les NPs d'origine vietnamienne, nous présentons quelques principes d'écriture liées à la mise en majuscule. Quant aux NPs d'origine étrangère,

nous traitons la question de la transcription ou de la translittération des NPs, dont l'écriture est basée sur les caractères latins ou sur d'autres caractères. La seconde question concerne la traduction des NPs d'une langue étrangère vers le vietnamien en général et de l'anglais et du français vers le vietnamien en particulier.

CHAPITRE 3 : ANALYSE D'ERREURS DE LA TANP À BASE DE CORPUS

3.1 PRÉPARATION D'UNE BASE DE DONNÉES : CONSTRUCTION DU CORPUS PARALLÈLE DE TEXTES AVEC DES NOMS PROPRES

Les corpus parallèles sont des ressources précieuses dans les études sur les langues, dans l'enseignant et dans plusieurs domaines du TAL, tels que TA, CLIR, IE, QA, désambiguïsation du lexique (WSD), extraction de terminologie bilingue, etc. Malheureusement, il n'existe pas suffisamment de corpus parallèles dans les langues minoritaires.

La procédure de construction des corpus parallèles pour des paires de langues minoritaires est loin d'être insignifiante. Dans cette section, nous décrivons la méthode que nous avons utilisée afin d'établir les deux corpus bilingues de textes avec NPs pour l'anglais-vietnamien et le français-vietnamien. Nous décrivons en détail les deux corpus parallèles de NPs (EVC et FVC).

L'EVC est une collection de 1500 textes (incluant 101289 mots et 575166 caractères) extraits de BBC News en ligne et liés à différents domaines tels que la politique, la santé, science-environnementale, éducation, divertissement et arts, technologie, etc. concernant des actualités et des événements de l'Afrique, l'Amérique, l'Asie-Pacifique, l'Europe, les États-Unis, le Canada, l'Amérique Latine et les Caraïbes, le Moyen-Orient, l'Asie du Sud, le Japon, le Royaume-Uni, etc. Tous ces textes ont été traduits en vietnamien par les systèmes de traduction les plus populaires au Vietnam à savoir : Vietgle, Google Translate, Bing Translator et EVTran.

Le corpus parallèle français-vietnamien (FVC) compte 1500 textes (incluant 109584 mots et 781347 caractères) extraits des articles en ligne du jourrnal Le Monde. Ce corpus inclut les textes concernant les actualités, l'économie, le sport, la culture, l'éducation, etc., appartenant aux catégories diverses comme International, l'Afrique, l'Amérique, l'Europe, l'Asie-Pacifique et le Moyen-Orient. Nous rassemblons aléatoirement ces différents articles et les mettons dans des systèmes de TA français-vietnamien.

3.2 ANALYSE D'ERREURS DE LA TANP À BASE DE CORPUS

3.2.1 Classification de NPs

Sur la base d'études menées sur les deux corpus, nous proposons une classification de NPs pour identifier et extraire des NPs d'un corpus. En comparaison des classifications de NPs précédemment réalisées par des linguistes et des chercheurs en TAL, notre classification a la caractéristique de ne pas être trop complexe, mais suffisamment spécifique pour décrire les caractéristiques typiques de NPs tant du point de vue de la linguistique que de la perspectives du TAL (voir la Figure 1)

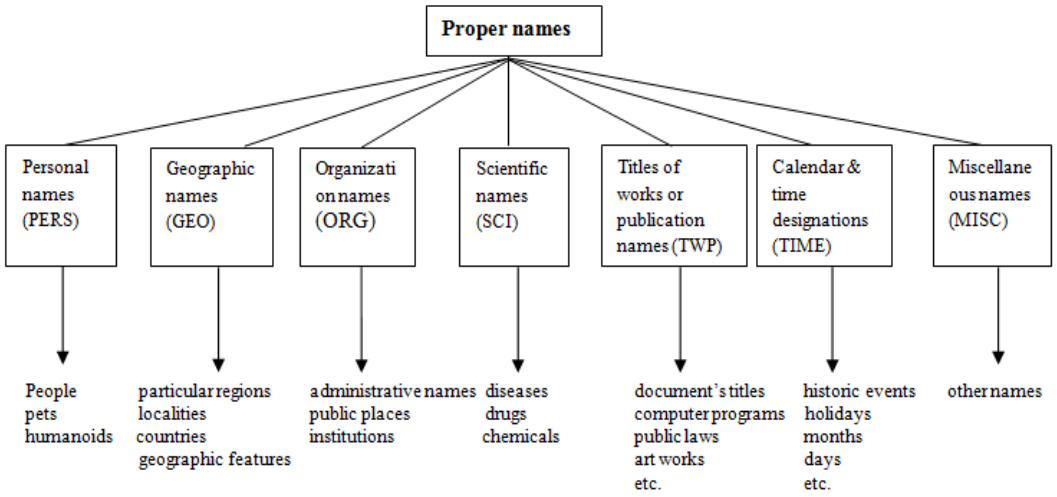


Figure 1 : Classification de PNs

3.2.2 Classification d'erreurs de TANP

La classification d'erreurs de la TANP est basée sur la classification des NP que nous avons établie auparavant. Comme nous l'avons mentionné précédemment, nous nous concentrons essentiellement sur deux types d'erreurs rencontrés en corpus, à savoir les erreurs de non-traduction et les erreurs de la mauvaise traduction de NPs. Nous divisons les erreurs de NT en six sous-types incluant des erreurs de la non-traduction d'abréviations ou d'acronymes, des titres professionnels et humains, des noms géographiques, des noms d'organisations, des titres d'œuvres, des jours et des mois. Les erreurs de mauvaise traduction sont divisées selon quatre critères concernant le graphisme, le lexique, la syntaxe et la transcription/translittération. En conséquence, nous classifions ces erreurs en quatre types principaux, à savoir des

erreurs graphiques (GE), des erreurs lexicales (LE), des erreurs syntaxiques (SY) et des erreurs de transcription (TE).

Les erreurs graphiques résultent d'une mise en majuscule erronée de certains NPs lors de leur traduction automatique. Elles peuvent être subdivisées en 2 types. Le premier type concerne des noms géographiques, des titres professionnels, des noms d'organisation, des titres d'œuvres, des noms d'événements, des nationalités, etc., qui devraient être mis en majuscule lors de la traduction, mais ne le sont pas. Le second type concerne les mois, les jours, les titres humains, qui sont souvent écrits avec des lettres initiales majuscules en anglais ; cependant, en vietnamien, ils ne devraient pas être écrits en majuscule.

Les erreurs lexicales concernent la mauvaise traduction des mots et peuvent être subdivisés en erreurs causées par des mots incorrects, des mots manquants, des mots superflus et des mots inconnus (Vilar et al. 2006). Des erreurs de "mots incorrects*"* ont lieu lorsque les systèmes de TA traduisent incorrectement un mot ou plusieurs mots constituant un nom propre. En raison de la nature polysémique des mots, les systèmes de TA ne proposent pas toujours les traductions correctes pour des contextes donnés. Par exemple, ils traduisent un nom commun écrit en majuscule comme s'il s'agissait d'un nom propre. Inversement, ils traduisent un nom propre ayant une signification dans un dictionnaire comme s'il s'agissait d'un nom commun. Par exemple, la phrase anglaise "*Turkey is my favourite food"* (La dinde est ma nourriture favorite) est automatiquement traduit en vietnamien comme "*Thổ Nhĩ Kỳ là thực phẩm mà tôi thích nhất*" (La Turquie est ma nourriture favorite).

Des erreurs de "mots manquants" se produisent lorsqu'il manque un mot dans la traduction (Vilar et al. 2006). Par exemple, les quantificateurs devant précéder des noms pluriels en vietnamien comme "*các, những, nhiều*, etc.", manquent parfois dans la TA de syntagmes nominaux anglais et français.

Des erreurs de "mots superflus" surviennent lorsqu'on constate la présence de mots dupliqués, inutiles ou superflus dans les traductions (Vilar et al. 2006). Par exemple, certains quantificateurs superflus "*các, những, nhiều,* etc.*"* précédent les syntagmes nominaux au singulier en vietnamien lorsqu'ils sont traduits de l'anglais.

Des erreurs de "mots inconnus*"* sont des erreurs dues à la présence de mots non-identifiés, ou de mots écrits dans une autre langue dans les traductions. Par exemple, lors de la traduction automatique des noms géographiques du français vers

le vietnamien, on obtient des noms géographiques écrits en anglais, mais pas en vietnamien.

Les erreurs syntaxiques concernent la mauvaise traduction de structures de NPs, par exemple, la mauvaise traduction de structures possessives, d'ordre des mots dans des syntagmes nominaux avec NPs, d'ordre de noms de personnes et d'ordre de dates.

Les erreurs de transcription ou translittération (TE) concernent les NPs, qui sont mal transcrits ou translittérés par les systèmes de TA. Ces types d'erreurs concernent quelquefois des noms géographiques et autres noms divers incluant des unités monétaires, des noms de nationalités et des langues.

Nous proposons 2 classifications d'erreurs de TANP, selon la paire de langue traitée : i/Classification d'erreurs de TANP pour l'anglais-vietnamien ; ii/Classification d'erreurs de TANP pour le français-vietnamien. Pour distinguer les erreurs de TANP d'anglais-vietnamien et les erreurs de TANP de français-vietnamien, nous utilisons les abréviations comme NTE1 (les erreurs de non-traduction en anglais), NTF1 (les erreurs de non-traduction en français), GEE1 (les erreurs graphiques en anglais), SYF1 (les erreurs syntaxiques en français), etc (voir la Figure 2 et la Figure 3).

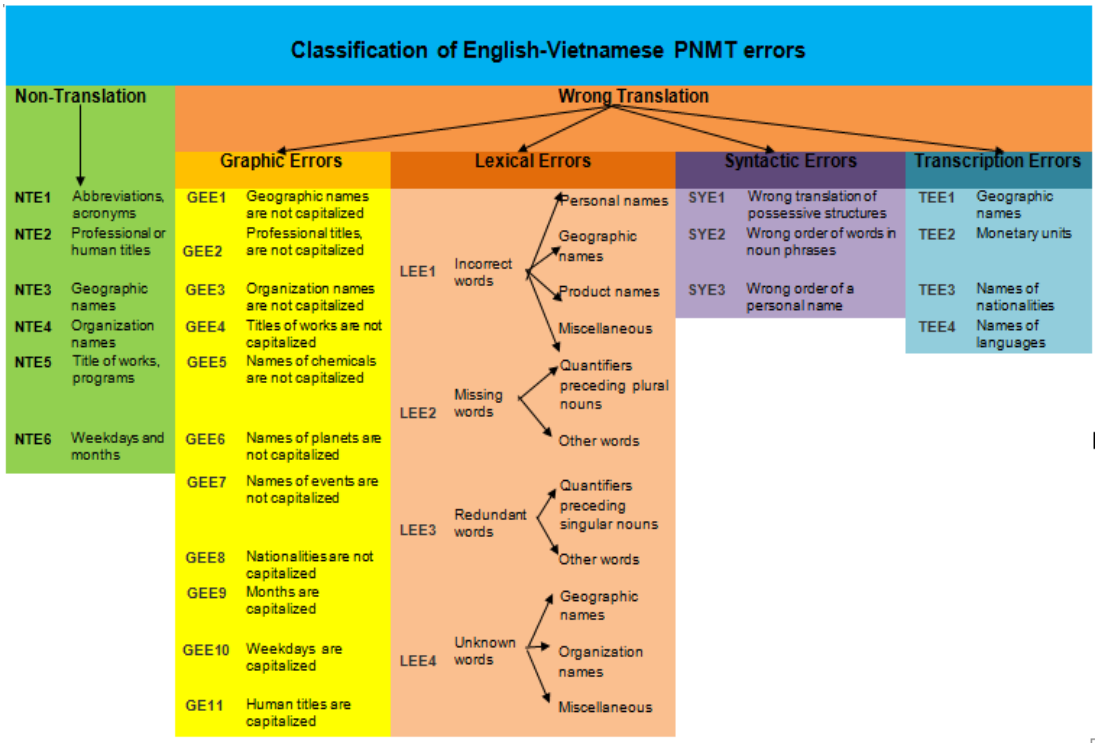


Figure 2 : Classification des erreurs de la TANP d'anglais vers le vietnamien

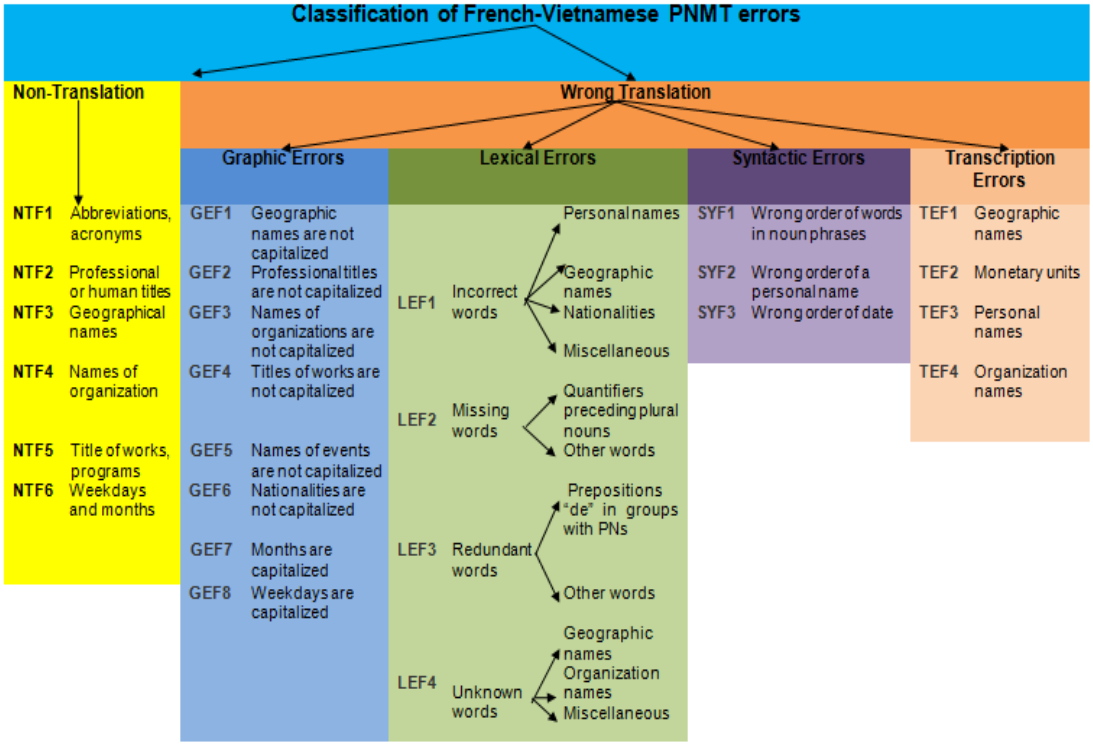


Figure 3 : Classification des erreurs de la TANP de français vers le vietnamien

3.2.3 Analyse qualitative des erreurs de la TANP avec des exemples

Cette section illustre la classification des erreurs de la TANP anglais-vietnamien et français-vietnamien avec une série d'exemples extraits des deux corpus parallèles de NPs mentionnés dans le Chapitre 2. Ce sont des exemples authentiques tirés du corpus et utiles à l'analyse et à la classification d'erreurs de TANP (Tableau 1). Les résultats de notre étude constituent des premières ressources à l'étude de la traduction de NPs de l'anglais et le français vers le vietnamien.

Type d'erreurs		**EVC**	**FVC**
Non-traduction		**NTE1:** RSPB (The Royal Society for the Protection of Birds)	**NTF1:** JDD (Journal du dimanche), PS (Parti socialiste)
		NTE2: Gen Mladic (Tướng Mladic), Mr. Shapiro (ông Shapiro)	**NTF2:** M. Le Graët (ông Le Graët), Comtesse de Besarn (nữ bá tước de Besarn)
		NTE3: Buckingham Palace (Cung điện Buckingham), Blue Cypress Lake (Hồ Bách xanh)	**NTF3:** Palais de Tokyp (Cung điện Tokyo), Quai d'Orsay (Bến tàu Orsay)
		NTE4: Museum of Innocence (Bảo tàng Ngây thơ)	**NTF4:** Théâtre de l'Aquarium (Nhà hát Thủy cung)
		NTE5: No Country for Old Men (Không chốn dung thân)	**NTF5:** Le Monde (tờ báo Thế giới)
		NTE6: May (tháng năm)	**NTF6:** mecredi (thứ tư)
Mauvaise traduction	GE	**GEE1:** Red Square => quảng trường đỏ (Quảng trường Đỏ)	**GEF1:** Côte d'Ivoire => bờ biển ngà (Bờ biển Ngà)
		GEE2: CIA director Leon Panetta=>Leon Panetta giám đốc Cục tình báo trung ương Hoa Kỳ (Leon	**GEF2**: le premier ministre Georges Papandreou=> thủ tướng Georges Papandreou (Thủ tướng

		Panetta, Giám đốc Cục tính báo trung ương Hoa Kỳ)	Georges Papandreou)
		GEE3: parliament=> quốc hội (Quốc hội)	**GEF3:** la Cour pénale internationale=> tòa án hình sự quốc tế (Tòa án Hình sự Quốc tế)
		GEE4: Financial Times => thời báo tài chình (Thời báo Tài chình)	**GEF4:** la Constitution afghane=> hiến pháp Afghanistan (Hiến pháp Afghanistan)
		GEE5: elements Thorium and Hafnium => các yếu tố thori và hafini (các nguyên tố Thori và Hafini)	**GEF5:** la Renaissance=> phục hưng (Phục hưng)
		GEE6: Solar System=> thái dương hệ (Thái dương Hệ)	**GEF6:** Arabes=> người ả Rập (người Ả Rập)
		GEE7: Toronto International Film Festival=> liên hoan phim quốc tế Toronto (Liên hoan Phim Quốc tế Toronto)	**GEF7:** avril=> tháng Tư (tháng tư)
		GEE8: Italian extraction => những người ý (Ý) khai thác	**GEF8:** mardi=> thứ Ba (thứ ba)
		GEE9: January=> tháng Giếng (tháng giêng)	
		GEE10: Monday=> thứ Hai (thứ hai)	
		GEE11: Mr. Obama=> Ông Obama (ông Obama)	
	LE	**LEE1:** Vince Cable => Cáp Vince (le cable Vince) (Vince Cable)	**LEF1:** Marine Le Pen=> biển Le Pen (mer Le Pen) (Marine Le Pen)
		LEE2: Russian TV channels=> (các) kênh truyền hình Nga	**LEF2:** municipales partielles italiennes=> (những) vùng thành phố Ý
		LEE3: Planet Jupiter=> các hành tinh sao Mộc (Hành tinh sao Mộc)	**LEF3:** la prairie de Pont-Aven=> đồng cỏ ~~của~~ Pont-Aven
	SY	**SYE1:** Mr. Brown's success => Ông Nâu là thành công (M. Brown est succès) (thành công của ông Brown)	**SYF1:** L'Estat de Kaduna=> Kaduna Nhà nước (Nhà nước Kaduna)
		SYE2: Chinese flags=> Trung quốc cờ xanh (cờ xanh Trung Quốc)	**SYF2:** Moussa Ibrahim=>Ibrahim Moussa (Moussa Ibrahim)
		SYE3: Husain Haqqani=> Haqqani Husain (Husain Haqqani)	**SYF3:** lundi 30 mai 2011=> Thứ 2 Tháng 5 30 2011 (lundi mai 30 2011)
	TE	**TEE1:** the Irish Republic => Cộng hòa Ailen (Cộng hòa Ireland)	**TEF1:** Turin=> Torino (Turin)
		TEE2: $2bn => 2 tỷ USD (2 tỷ đô la Mỹ)	**TEF2:** 294 milliards de dollars=> 294 tỷ USD (294 tỷ đô-la Mỹ)
		TEE3: Syrians=> Syrians (người Syria)	**TEF3:** Hu Chunhua=> Hu thuan (Hồ Xuân Hoa)
		TEE4: Turkish=> tiếng Turk (tiếng Thổ Nhĩ Kỳ)	**TEF4:** parti Baas=> Đảng Baath (Đảng Baas)

Tableau 1: Exemples des erreurs de la TANP de deux corpus anglais et français

CHAPITRE 4 : ANNOTATION DE CORPUS

4.1 CADRE D'ANNOTATION

L'annotation en noms propres des corpus parallèles est basée sur les principes généraux d'annotation de corpus. Leech (2004) définit l'annotation de corpus comme pratique consistant à ajouter des informations linguistiques interprétatives à un corpus. Il s'agit généralement d'ajouter des tags ou des étiquettes indiquant qu'un mot ou un groupe de mots appartiennent à une certaine classe comportant des caractéristiques semblables.

Notre cadre d'annotation de corpus parallèles est constitué de cinq grandes étapes:

1. Collecte de textes originaux en anglais et en français pour construire les corpus ;

2. Analyse linguistique de NPs : identification de certains NPs spécifiques, par exemple :

   a. Noms propres "simples": par exemple, *Bill Gates, South Korea, EU*, etc.;

   b. Noms propres incluant des prépositions de coordination, par exemple : *University of California, Hotel de Glace* (noms d'organisation), *Good Night and Good Luck, Gone with the Wind* (les titres des livres, des films, des chansons, des programmes, etc.);

   c. Certains syntagmes nominaux dépendant d'un NP, par exemple. *CBS's "60 Minutes" program, Britain's first black Conservative peer*, etc.

3. Création de tags multilingues pour l'annotation de NPs : cette étape dépend de l'analyse linguistique de chaque langue ;

4. Mise en œuvre d'annotation de corpus en utilisant la méthode manuelle, la méthode automatique et la méthode semi-automatique ;

5. La comparaison et l'évaluation de toutes les méthodes d'annotation.

4.2 MÉTHODES D'ANNOTATION

Pour les besoins de notre recherche, nous avons mis en œuvre et comparé trois méthodes d'annotation, à savoir la méthode manuelle, la méthode automatique et la méthode semi-automatique.

4.2.1 Annotation manuelle

Nous avons crée un ensemble d'étiquettes qui servent à reconnaître et, si possible, à faciliter la correction des NPs de l'anglais et du français. En se référent à la classification des NPs (voir la Figure 1) et aux catégories des erreurs de TANP d'EVC et de FVC (voir la Figure 2 et la Figure 3), nous avons créé 16 étiquettes d'annotation, que nous présentons avec une définition simple et une exemplification dans le Tableau 2.

No	Tags d'annotation	Signification
1	ACRO	Acronymes ou abréviations
2	GEO1	Noms géographiques anglais/français qui doivent être traduits
3	GEO2	Noms géographiques anglais/français qui ne doivent pas être traduits
4	Human Title	Par exemple, *Dr., Mrs., Miss, M., Mme, Mlle*, etc.
5	MISC	Noms divers
6	NE	Frontières d'entités nommées
7	NP	Syntagme nominal anglais dans les structures possessives avec NPs
8	ORG	Noms d'organisation anglais/français
9	ORG1	Noms français d'organisations exigeant l'omission du groupe "de + déterminant" les précédant
10	PER	Noms personnels (noms de personnes précédés par des titres humains ou titres professionnels), par exemple, Président Barack Obama
11	PERS	Noms personnels (contenant trois cas : prénoms ou noms de famille ou tous les deux sans titres humains/ titres professionnels), par exemple, *Obama, Barack Obama*.
12	PROD	Noms de produits
13	Professional Title	Par exemple, *Pilot/Pilote, Secretary/Secrétaire, Doctor/Docteur, Professor/Professeur*, etc.
14	SCI	Noms scientifiques
15	TIME	Désignations de temps
16	TWP	Titres des œuvres et des publications

Tableau 2: Etiquettes pour annotation manuelle des corpus anglais et français de NPs

4.2.2 Annotation automatique et annotation semi-automatique

Pour annoter nos corpus automatiquement et semi-automatiquement, nous utilisons la version 7.1 de logiciel GATE développé par Cunningham et al. De (2012) l'Université de Sheffield; il permet aux utilisateurs d'annoter automatiquement des entités nommées tels que *Person*, *Location*, *Organization*, *Date*, *Percents*, *Money* et *Address*. En raison de l'architecture ouverte de GATE, les modules de NER sont flexibles et facilement adaptables parce qu'ils consistent en des ensembles de règles d'appariement de formes, créés manuellement et pouvant être étendues pour ajouter et modifier des nouveaux types d'entités (Bontcheva et al. 2002). Nous pouvons alors éditer, modifier ou ajouter des nouveaux tags d'annotation à la liste d'annotation d'ANNIE dans l'environnement visuel de GATE.

GATE est un outil d'annotation très utile et très efficace pour effectuer l'annotation semi-automatique des corpus multilingues de NPs. En fait, GATE a été utilisé pour annoter des documents non seulement en anglais, mais aussi dans d'autres langues, comme le français, l'allemand, l'italien, l'arabe, le chinois, le roumain, le hindi et le cebuano. Pour le rendre approprié à notre travail, nous ajoutons à la liste d'annotations déjà présentes les étiquettes spécifiques incluant Science, Product, TitleOfWorks, NP, et Misc (voir Tableau 3).

No	**Le type des tags d'annotation**		**Signification**
	Ensemble d'annotation de GATE	Etiquettes supplémentaires	
1	<Address>		Liens de site Web
2	<Date>		Désignations de temps
3	<FirstPerson>		Prénoms
4	<JobTitle>		Titres professionnels (par exemple, Pilot/Pilote, Secretary/Secrétaire, Doctor/Docteur, Professor/Professeur, etc.)
5	<Location>		Noms géographiques
6		<Misc>	Noms divers
7	<Money>		Unités monétaires avec chiffres
8		<NP>	Syntagmes nominaux anglais dans des structures possessives avec NPs
9	<Organization>		Noms d'organisations
10	<Percent>		Pourcentage
11	<Person>		Noms personnels (incluant les noms, les prénoms, et les titres humains ou titres professionnels)
12		<Product>	Noms de produits
13		<Science>	Noms scientifiques
14	<Title>		Par exemple, Dr., Mrs., Miss, M., Mme, Mlle, etc.

15		<TitleOfWorks>	Titres des œuvres et des publications

Tableau 3: Les étiquettes pour l'annotation semi-automatique des corpus anglais et français de NPs

En résumé, cette section présente les avantages de l'annotation de corpus et se concentre sur trois méthodes d'annotation que nous utilisons pour nos corpus de NPs. La méthode d'annotation manuelle demande énormément de temps et de travail, mais fournit des résultats avec une très haute precision. La méthode d'annotation automatique, effectuée par GATE, permet d'épargner du temps, mais fournit des résultats avec beaucoup moins de précision. La méthode d'annotation semi-automatique combine l'annotation automatique avec la correction manuelle de ses résultats et réduit environ de 90 % le temps d'annotation manuelle, et fournit une très bonne précision.

4.3 RÉSULTATS D'ANNOTATION DE CORPUS

Dans cette section, nous analysons les erreurs d'annotation automatiques des corpus anglais et français de PNs donné par GATE et comparons ces résultats avec des résultats d'annotation manuelle pour calculer combien d'erreurs d'annotation nous devons corriger et ajuster dans l'annotation semi-automatique. Le Tableau 4 présente les résultats d'annotations automatiques effectuée par GATE.

Nom du corpus		**Corpus anglais**		**Corpus français**	
Ressources		BBC News		Le Monde	
Taille	Mots	101289		109584	
	Caractères	575166		781347	
	Paragraphes	1512		1501	
Résultats d'annotation		Type de tags	Total d'annotations	Type de tags	Total d'annotations
		Location	2743	Location	2369
		Title	1245	Title	581
		JobTitle	1219	JobTitle	603
		Organization	1443	Organization	449
		Person	2290	Person	1530
		FirstPerson	1782	FirstPerson	1483
		Date	1697	Date	1869
Total de NPs annotés			**12419**		**8884**

Tableau 4 : Les résultats d'annotation du corpus anglais et du corpus français obtenus par l'annotation automatique

Parmi ces annotations automatiques, il existe une série d'erreurs d'annotation. Le tableau 5 indique le total d'annotations automatiques pour chaque type de tags et montre le nombre et le pourcentage d'erreurs.

Type de tags	Corpus anglais			Corpus français		
	Nombre d'annota-tions automa-tiques	Nombre d'erreurs d'annota-tions automa-tiques	Pour-centage	Nombre d'annota-tions automa-tiques	Nombre d'erreurs d'annota-tions automa-tiques	Pour-centage
Date	1697	815	48%	1869	1042	55.75%
FirstPerson	1782	333	18.68%	1483	784	52.86%
JobTitle	1219	354	28.30%	603	144	23.88%
Location	2743	543	19.79%	2369	809	31.14%
Organization	1443	424	29.38%	449	1258	63.67%
Person	2290	635	27.72%	1530	875	57,18%
Title	1245	635	51.0%	581	342	58.86%
Total	**12419**	**3739**	**30.10%**	**8884**	**5254**	**49.04%**

Tableau 5 : Statistiques d'erreurs pour l'annotation automatique des corpus anglais et français de NPs

4.4 COMPARAISON DES RÉSULTATS DES DIFFÉRENTES MÉTHODES D'ANNOTATION

Pour comparer les résultats d'annotation obtenus par les trois méthodes (la méthode manuelle avec la méthode automatique et la méthode automatique avec la méthode semi-automatique et la méthode manuelle avec la méthode semi-automatique), nous devons comparer les étiquettes correspondantes. Les différences les plus importantes, issues de la comparaison des résultats portent sur le temps d'annotation et le nombre total d'annotations pour chaque type d'étiquettes correspondant entre les différentes paires de méthodes d'annotation. L'annotation manuelle fournit un taux élevé de précision et de rappel, mais exige un temps de travail conséquent.

L'annotation semi-automatique offre les taux de précision et de rappel aussi élevés que ceux de l'annotation manuelle. L'avantage d'annotation semi-automatique consiste à minorer considérablement le temps d'exécution des annotations. Par conséquent, elle constitue la meilleure méthode d'annotation puisqu'elle combine à la fois l'avantage de l'annotation manuelle de bonne qualité et l'avantage de l'annotation automatique, économe en temps d'annotation. Les corpus semi-annotés constituent les données de départ pour le programme de prétraitement pour l'améliorer la TANP.

CHAPITRE 5 : AMÉLIORATION DE LA TRADUCTION AUTOMATIQUE PAR PRÉTRAITEMENT

5.1 PRÉ-TRAITEMENT

5.1.1 Définition et objectif du prétraitement

Dans le domaine du TAL, le *prétraitement* est une tâche informatique effectuée sur les données de départ et destinée à produire de meilleures données en sortie. En ce qui concerne la TA, le *prétraitement* se situe dans le processus de la traduction automatique aidée par l'humain (HAMT), afin que les systèmes de TA génèrent des traductions de bonne qualité.

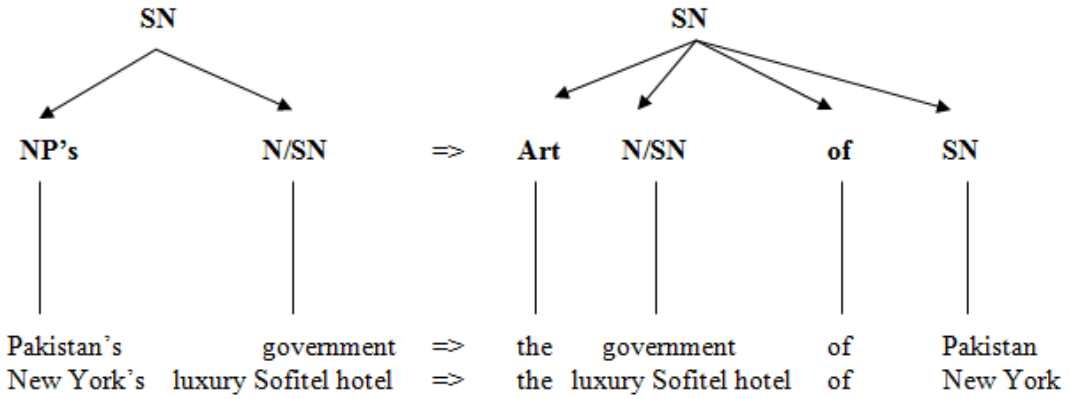
Il existe plusieurs sortes de prétraitements efficaces pour améliorer les résultats de systèmes de TA. Ils peuvent être fonction des systèmes de TA et des types de données de sorties espérés. Hutchins et Somers (1992:151) disent que le prétraitement typique implique la recherche de problèmes de traduction prévisibles dans des textes originaux et leurs suppressions. Le prétraitement typique peut inclure l'identification de noms (noms propres), le marquage des catégories grammaticales homographes, l'indication de propositions enchâssées, la mise en parenthèses de structures coordonnées, le marquage ou la substitution de mots inconnus, etc. Un grand nombre de tâches de prétraitement peuvent être effectuées manuellement et/ou automatiquement, par exemple, le changement de l'ordre des mots dans les syntagmes et les phrases; le réagencement des adverbes ou des expressions adverbiales de temps, de lieu, de fréquence, de manière, etc. ; le transfert des structures passives aux structures actives ; la correction de la ponctuation pour éviter des erreurs graphiques concernant l'utilisation des majuscules et des minuscules); la suppression de mots pour éviter des erreurs de traduction causées par des mots superflus.

5.1.2 Tâches spécifiques du prétraitement pour la TA anglais-vietnamienne et français-vietnamienne

Parmi les diverses erreurs de la TANP, que nous avons recensées dans notre étude, nous nous concentrons sur les types d'erreurs suivants : i/les erreurs qui résultent de la mauvaise traduction de structures possessives anglaises (les erreurs

syntaxiques); ii/les erreurs causées par la mauvaise traduction de NPs anglais et français, qui ne devraient pas être traduits; iii/les erreurs qui résultent de la mauvaise traduction de mots superflus dans les expressions françaises avec NPs. Puisque ces erreurs sont statistiquement significatives, relativement faciles à reconnaître et à corriger à l'aide du prétraitement, nous définissons trois tâches de prétraitement :

a/ Restructurer les génitifs anglais avec NPs (les structures possessives de l'anglais) ;



(SN= Syntagme nominal ; NP= nom propre ; Art= article ; N= nom)

Figure 4 : Restructuration de génitifs avec NPs en anglais

b/ Omettre des groupes français "de + déterminant" précédant certains noms géographiques et noms d'organisation ;

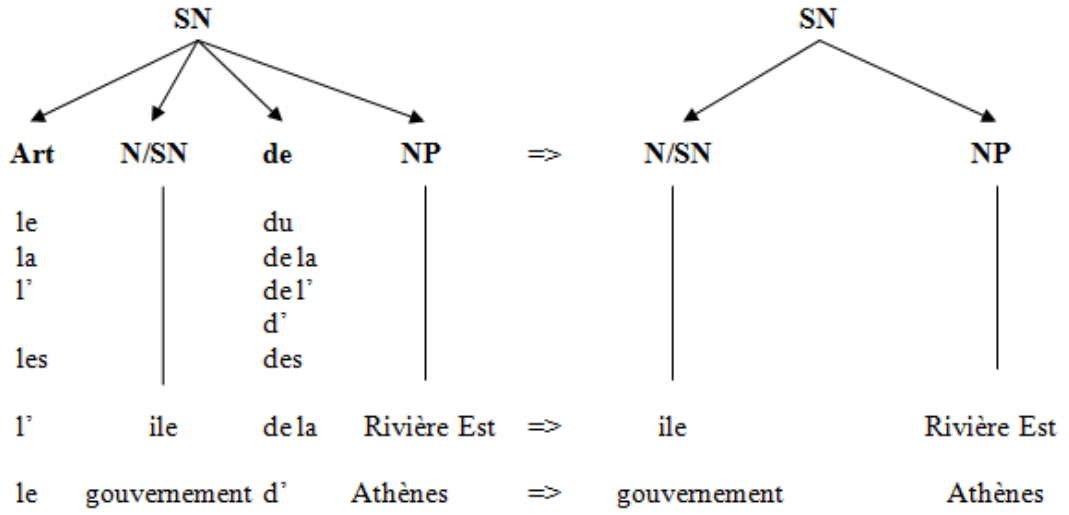


Figure 5 : Omission des groupes français "de + déterminant" dans les syntagmes nominaux avec NPs

c/ Attribuer l'étiquette DNT (Do-Not-Translate) à certains NPs anglais et français (noms personnels, certains noms géographiques, titres d'œuvres, noms de produits en anglais et en français)

5.2 DESCRIPTION DU PROGRAMME DE PRÉTRAITEMENT POUR LA RÉDUCTION D'ERREURS DE LA TANP

5.2.1 Tâches principales

Ce programme de prétraitement est conçu pour exécuter les tâches suivantes :

1. Identifier les NPs et les extraire, tout d'abord de nos corpus anglais et français et ensuite, des autres corpus annotés avec NPs ;
2. Grouper les NPs en différentes catégories basées sur la classification que nous avons proposée, afin de créer des listes de NPs par catégorie : liste de noms personnels, liste de noms géographiques, liste de noms d'organisations, etc. ;
3. Compter le nombre total de NPs par corpus annoté, en général et par catégorie ;
4. Attribuer l'étiquette Do-Not-Translate à tous les NPs, qui ne devraient pas être traduits ;
5. Détecter et simplifier les structures possessives (génetif) anglaises avec NPs ;
6. Détecter et supprimer les groupes français "de + déterminant" précédant certains noms géographiques et certains noms d'organisations pour simplifier les structures françaises avec NPs.

5.2.3 Interface et démos du programme de prétraitement

Le programme de prétraitement comporte trois menus principaux : *File, Options* et *Help*. Le menu *File* inclut les sous-options suivantes :
1/ *New task :* permet aux utilisateurs de créer un nouveau projet ;
2/ *Open input document :* permet d'ouvrir un fichier (fichiers au format *.txt, .doc, .rtf*). Le taille maximum d'un fichier est d'environ 80,000 mots ou 800,000 caractères ;

3/ *Save input text* : permet de sauvegarder le corpus d'entrée ou le fichier ;

4/ *Save final results* : permet de sauvegarder les résultats finaux ;

5/*Exit* : permet de quitter le programme.

Le menu *Options* inclut les sous-options *Define Tags* et *Reset all tags* (voir la Figure 6). *Define tags* ouvre le gestionnaire d'étiquettes d'annotation ; il fournit la liste de toutes les étiquettes (*Tag names)* et leur *Description*, ainsi que l'option de marquage en DNT et l'option *Check French determiner*. L'option *Define tags*, permet d'ajouter/éditer/enlever *(add/edit/remove*) les étiquettes d'annotation. Nous pouvons définir de nouvelles étiquettes avec l'option *Tag name* et les colorier en utilisant l'option *Color*.

L'option DNT propose de sélectionner *Yes* pour les NPs, qui ne devraient pas être traduits, ou de sélectionner *No* pour les NPs, qui devraient être traduits en langue cible.

L'option *Check French Determiner* permet de choisir entre deux valeurs : *Yes* (pour supprimer les groupes "de + déterminant" de certains NPs) ou *No* (pour laisser ces groupes inchangés). Pour notre corpus, la valeur Yes sera choisie les NPs étiquetés noms géographiques, noms d'organisations et dates.

Si nous cliquons sur la commande *Update (*Mise à jour), tous les ajustements effectués seront mis à jour. Au contraire, *Undo* permet de ne pas enregistrer les ajustements. L'option *Reset all tags* est utilisée pour revenir aux valeurs initiales du programme. Finalement, nous cliquons sur la boîte *OK* pour exécuter le programme ajusté. Si nous cliquons sur la boîte *Cancel*, les rajustements ne seront pas acceptés.

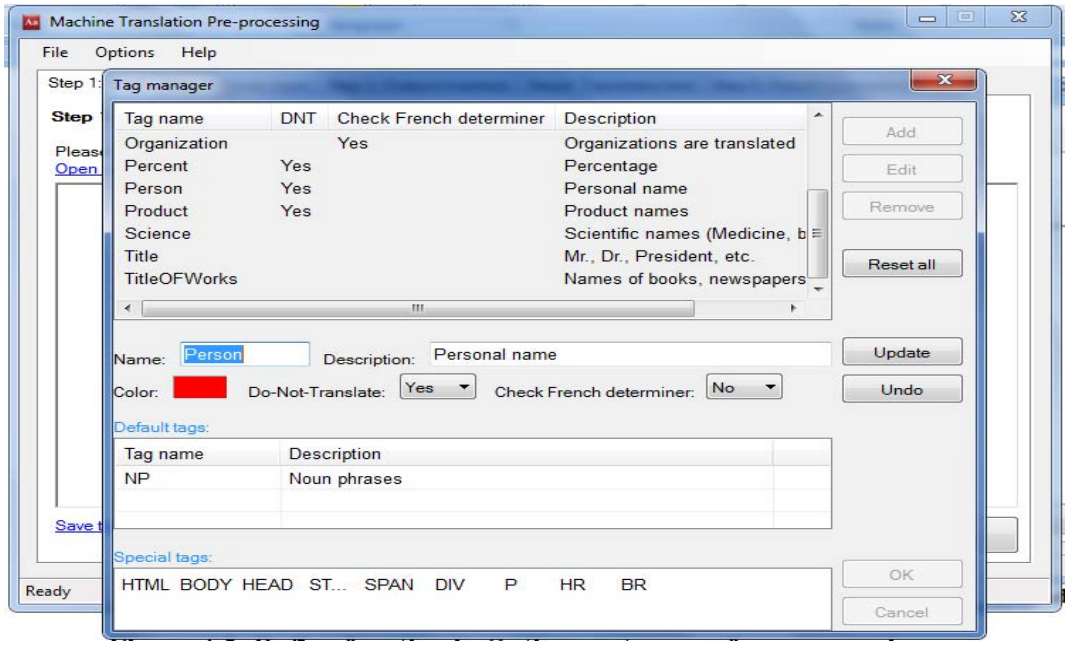


Figure 6 : Définition des nouvelles étiquettes dans le Gestionnaires d'étiquettes

Le programme du prétraitement fonctionne en cinq étapes.

Étape 1 (Fenêtre Step 1) : Choix et chargement des documents à traiter (voir la Figure 7). Il y a deux façons de sélectionner des documents et de les charger dans le programme de prétraitement : 1/copier et coller un texte dans l'espace central destiné à des données d'entrée ; 2/ouvrir un fichier avec l'option Open input document. Pour stocker un document pour l'usage futur, il faut choisir *Saving Input* (Sauvegarder le fichier d'entrée) et créer un lien dans le répertoire de bases de données. Dans cette étape, il faut aussi choisir la langue des textes sources (*Input Language*).

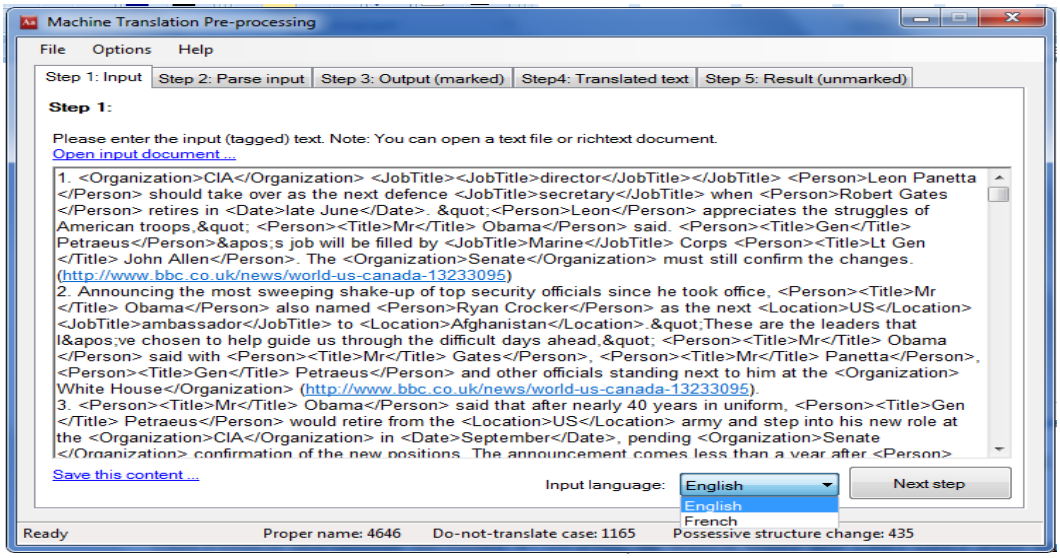


Figure 7 : Saisie des textes annotés dans le programme de prétraitement

Étape 2 (Fenêtre Step 2) **:** Traitement du document

Étape 3 (Fenêtre Step 3) **:** Présentation du document prétraité (voir la Figure 8)

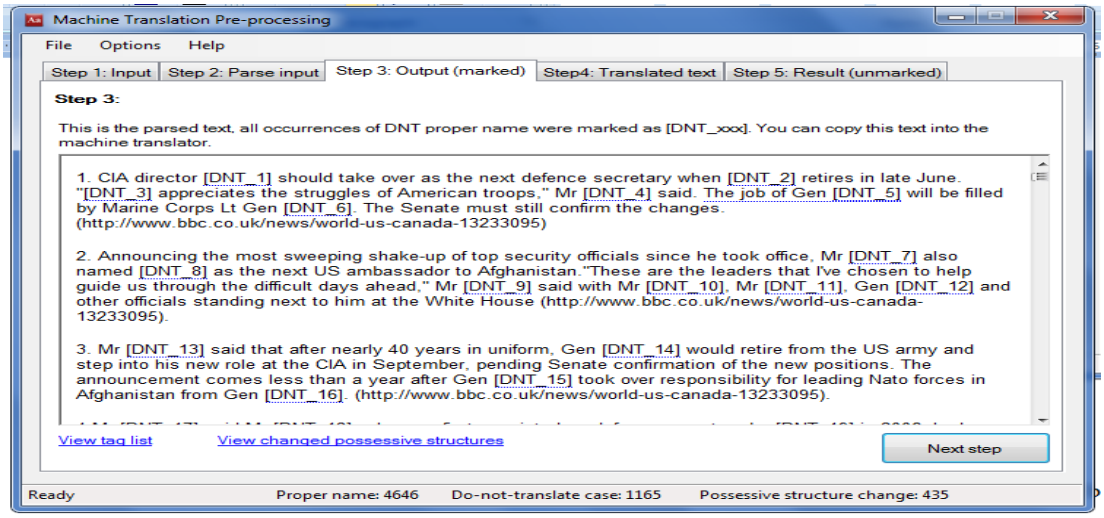


Figure 8 : Exemple d'un document anglais prétraité (Étape 3)

Dans l'Étape 3, nous pouvons voir le texte résultant du traitement réalisé dans l'Étape 2 et avec les résultats quantificatifs affichés dans la barre d'outils (dans l'exemple de la figure 8, 4646 NPs incluant 1165 DNT NPs et 435 cas de structures possessives). Ce texte doit être copié manuellement et soumis à de différents systèmes de TA.

Étape 4 (Fenêtre Step 4) : Récupération des textes traduits par des systèmes de TA

Cette étape sert à récupérer les textes prétraités et traduits par différents systèmes de TA (indiqué dans l'Étape 3). Ces documents doivent être collés manuellement dans l'espace central de l'étape 4 pour procéder à des réajustements finaux.

Étape 5 (Fenêtre Step 5) : Résultat final.

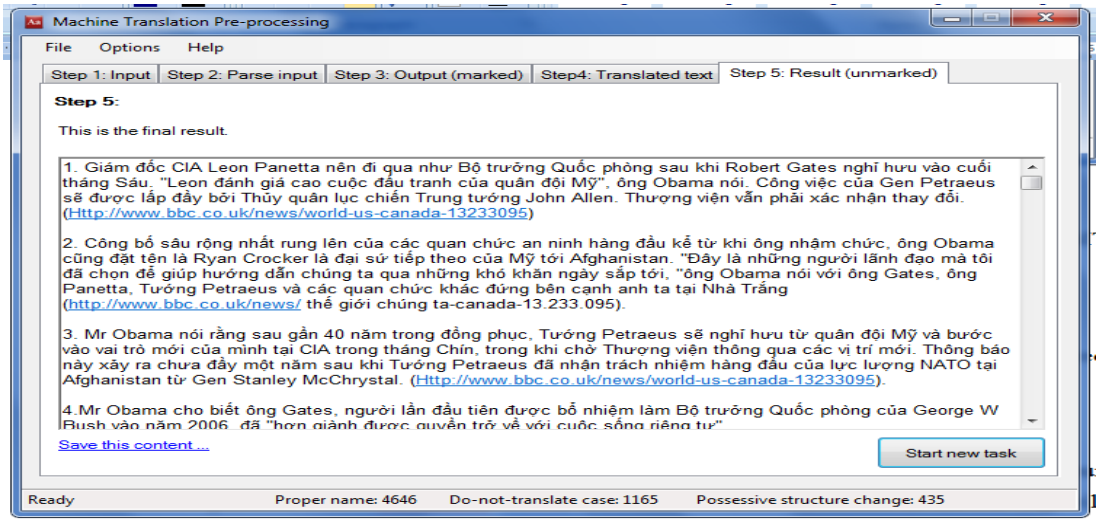


Figure 9 : Le résultat final dans l'Étape 5

5.2.4 Avantages et limites du programme de prétraitement

Ce programme a été construit pour effectuer des tâches spécifiques de prétraitement en vue de réduction des erreurs de la TANP de l'anglais et du français vers le vietnamien.

Il permet :

i/ d'améliorer la qualité de traductions effectuées par des systèmes de TA (voir la section 5.3) ;

ii/ réduire le taux d'erreurs de transcription faites par ces systèmes ;

iii/ simplifier les structures françaises avec des NPs ;

iv/ créer diverses listes de NPs dans différentes catégories ;

v/ établir les deux lexiques bilingues de NPs ;

vi/ fournir des statistiques des NPs en total et par catégories.

Ce programme est applicable pour deux langues, anglais et français. Son plus grand avantage consiste en la diminution du temps de traitement par rapport à la solution manuelle. Il dispose aussi d'une interface conviviale et est facile à prendre en main et à manipuler.

Il possède aussi quelques limitations.

La première limite concerne la préparation des documents d'entrée. Le programme ne prend en entrée que des documents annotés, ce qui nécessite un temps de préparation de ces documents. Bien que l'annotation d'un document en anglais ou en français puisse être effectuée automatiquement, elle doit être corrigée pour obtenir des meilleurs résultats de traduction.

Deuxièmement, ce programme ne peut pas corriger tous les types d'erreurs de la TANP faites par les différents systèmes de TA, par exemple des erreurs de non-traduction, des erreurs graphiques, quelques sous-types d'erreurs lexicales, syntaxiques et des erreurs de transcription.

Troisièmement, le passage par les systèmes de TA n'est pas directement intégrés dans le programme de prétraitement ; on doit copier et coller les documents dans différents systèmes, les remettre ensuite dans le programme de prétraitement pour enfin obtenir le résultat final. Il serait idéal de connecter directement le programme de prétraitement aux systèmes de TA, pour qu'il n'y ait aucune autre manipulation nécessaire pour obtenir la traduction finale.

5.3 RÉSULTATS DES TESTES AVREC ET SANS LE PRÉTRAITEMENT

Dans cette section, nous présentons les statistiques concernant le taux d'erreurs de la TANP, pour les deux corpus parallèles anglais-vietnamien et français-vietnamien. Ces statistiques résultent des testes effectués sur les corpus avant et après l'utilisation du programme de prétraitement et mettent en évidence l'avantage du prétraitement. Nous nous en servons aussi pour illustrer les avantages et les limites de chaque système de TA utilisé pour notre étude.

5.3.1 Résultats des testes sur le corpus anglais-vietnamien avant et après le prétraitement

Le Tableau 6 détaille le nombre et le taux d'erreurs de chaque catégorie décrite dans la classification d'erreurs de TANP de l'anglais vers le vietnamien (voir la Figure 1). Les résultats présentés dans ce tableau concernent quatre différents traducteurs automatiques et ont été recueillis sur les corpus non - prétraités (donc, avant utilisation du programme de prétraitement).

Type d'erreurs de TANP			**Erreurs de Vietgle**		**Erreursde Google**		**Erreursde Bing**		**Erreursde EVTran**	
			(1)	**%**	**(2)**	**%**	**(3)**	**%**	**(4)**	**%**
Non-traduction		NTE1	319	9.89	418	15.85	391	10.92	193	11.18
		NTE2	116	3.59	20	0.75	79	2.20	2	0.11
		NTE3	42	1.30	25	0.94	180	5.02	10	0.57
		NTE4	39	1.20	17	0.64	38	1.06	7	0.40
		NTE5	4	0.12		0.00	4	0.11	7	0.40
		NTE6	3	0.09	15	0.56	15	0.41		0.00
		Total	**523**	**16.22**	**495**	**18.77**	**707**	**19.75**	**219**	**12.68**
Mauvaise traduction	***GEE***	GEE1	162	5.02	140	5.31	197	5.50	72	4.17
		GEE2	218	6.76	125	4.74	264	7.37	44	2.54
		GEE3	211	6.54	94	3.56	265	7.40	51	2.95
		GEE4		0.00	1	0.03	4	0.11		0.00
		GEE5	15	0.46	8	0.30	25	0.69	7	0.40
		GEE6	5	0.15	6	0.22	12	0.33	4	0.23
		GEE7	8	0.24	17	0.64	28	0.78	8	0.46
		GEE8	6	0.18	7	0.26	23	0.64	2	0.11
		GEE9	19	0.58	40	1.51	27	0.75	24	1.39
		GEE10	96	2.97	111	4.21	24	0.67	41	2.37
		GEE11	120	3.72	60	2.22	6	0.16	31	1.79
		Total	**860**	**26.67**	**609**	**23.10**	**875**	**24.44**	**284**	**16.45**
	LEE	LEE1	805	24.95	402	15.23	556	15.53	395	22.86
		LEE2	122	3.78	271	10.27	229	6.39	62	3.58
		LEE3	65	2.01	86	6.23	70	1.95	22	1.27
		LEE4	100	3.10	189	7.17	440	12.29	31	1.79
		Total	**1,092**	**33.87**	**948**	**35.96**	**1,295**	**36.18**	**510**	**29.54**
	SYE	SYE1	248	7.69	264	10.01	364	10.17	287	16.62
		SYE2	290	8.99	117	4.43	147	4.10	247	14.31
		SYE3	105	3.25	81	3.07	85	2.37	102	5.90
		Total	**643**	**19.94**	**462**	**17.52**	**596**	**16.65**	**636**	**36.84**

	TEE	TEE1	55	1.70	79	2.99	74	2.06	54	3.12
		TEE2	11	0.34	19	0.72	20	0.55	10	0.57
		TEE3	37	1.14	21	0.79	10	0.27	13	0.75
		TEE4	3	0.09	3	0.11	2	0.05		0.00
		Total	**106**	**3.28**	**122**	**4.62**	**106**	**2.96**	**77**	**4.46**
Total d'erreurs/ Total de textes traduits			**3224/ 1500**	**21.49**	**2636/ 1500**	**17.57**	**3579/ 1413**	**25.32**	**1726 /609**	**28.34**

Tableau 6 : Taux d'erreurs de la TANP de l'anglais vers le vietnamien avant le prétraitement

Le Tableau 7 propose les statistiques d'erreurs corrigées par les quatre systèmes de TA E-V après l'utilisation du programme de prétraitement. L'utilisation de ce programme a permis de réduire significativement certains types d'erreurs de la TANP notamment certains sous-types d'erreurs lexicales et syntaxiques. En général, la proportion d'erreurs corrigées est relativement élevée. Par ailleurs, les différents systèmes de TA de part leur différent fonctionnement, ne corrigent pas les mêmes nombres d'erreurs. Par rapport aux autres systèmes, EVTran a corrigé le plus grand nombre d'erreurs, et Google s'avère moins performant que Vietgle et EVTRAN. Parmi les quatre systèmes de TA, Bing a corrigé le plus petit nombre d'erreurs de la TANP.

Type d'erreurs de TANP		**Erreurs de Vietgle**		**Erreurs de Google**		**Erreurs de Bing**		**Erreurs de EVTran**	
		(1)	% (1')	(2)	% (2')	(3)	% (3')	(4)	% (4')
Erreurs lexicales (LEE)	LEEs corrigées	756/ 1092	23.44	354/ 948	13.42	413/ 1295	11.53	367/ 510	21.26
Erreurs syntaxiques (SYE)	SYE1 corrigées	244/ 248	7.63	261/ 264	9.90	361/ 364	10.08	284/ 287	16.42
	SYE3 corrigées	105/ 349		81/81		85/85		102/ 102	
	Total of SYE corrigées	349 /643	10.82	339/ 462	12.86	438/ 596	12.23	386/ 636	22.19
Total de (LEE+ SYE) corrigées/ Total de TANP erreurs		**1105/ 3224**	**34.27**	**696/ 2636**	**26.40**	**859/ 3570**	**24.06**	**753/ 1726**	**43.62**

Table 7 : Statistiques d'erreurs de TANP E-V corrigées après l'utilisation du programme prétraitement

5.3.2. Résultats des testes sur le corpus franco-vietnamien de PNs avant et après le prétraitement

Dans cette section, nous présentons les statistiques d'erreurs de TANP faites par deux traducteurs automatiques français-vietnamien avant et après l'utilisation du programme de prétraitement. L'analyse des erreurs, montrée dans le Tableau 8, est effectuée sur la base de la classification d'erreurs de la TANP du français vers le vietnamien.

Type d'erreurs			**Erreurs de Google**		**Erreurs de Bing**	
			(1)	**% (1')**	**(2)**	**% (2')**
Non-Traduction		NTF1	377	11.46	339	10.00
		NTF2	45	1.36	36	1.06
		NTF3	12	0.36	2	0.05
		NTF4	15	0.45	15	0.44
		NTF5	7	0.21	2	0.05
		NTF6	4	0.12	4	0.11
		Total	**460**	**13.99**	**398**	**11.75**
Mauvaise traduction	***Erreurs graphi-ques***	GEF1	159	4.83	282	8.32
		GEF2	148	4.50	330	9.74
		GEF3	269	8.18	560	16.53
		GEF4	**14**	0.42	31	0.91
		GEF5	15	0.45	22	0.64
		GEF6	0	0.00	2	0.05
		GEF7	121	3.68	14	0.41
		GEF8	205	6.23	27	0.79
		Total	**931**	**28.32**	**1268**	**37.43**
	Erreurs lexicales	LEF1	600	18.25	448	13.22
		LEF2	52	1.58	368	10.86
		LEF3	486	14.78	184	5.43
		LEF4	239	7.27	301	8.88
		Total	**1377**	**41.89**	**1301**	**38.41**
	Erreurs synta-xiques	SYF1	257	7.88	234	6.90
		SYF2	49	1.49	20	0.59
		SYF3	107	3.25	27	0.79
		Total	**413**	**12.56**	**281**	**8.29**
	Erreurs de transcrip-tion	TEF1	67	2.03	88	2.59
		TEF2	29	0.88	37	1.09
		TEF3	5	0.15	10	0.29
		TEF4	5	0.15	4	0.11
		Total	**106**	**3.22**	**139**	**4.10**
Total d'erreurs/ Total de textes traduits			**3287/ 1500**	**29.72%**	**3387/ 1500**	**30.71%**

Tableau 8: Statistiques d'erreurs de TANP du français vers le vietnamien avant le prétraitement

Selon ces statistiques, le taux d'erreurs lexicales est le plus élevé en comparaison avec d'autres types d'erreurs. Google et Bing possèdent les taux les plus élevés de LEFs. Le taux d'erreurs graphiques (GEFs) est moins bas que celui de

LEFs, mais plus haut que le taux de SYFs et TEFs. Le taux de GEFs et TEFs de Bing est plus élevé que celui de Google ; néanmoins, le taux de SYFs de Google dépasse celui de Bing. Le tableau 9 présente les statistiques d'erreurs de TANP français-vietnamien, qui ont été corrigées après que nous ayons utilisé le programme de prétraitement.

Type d'erreurs		**Erreurs de Google**		**Erreurs de Bing**	
		(1)	**(1')**	**(2)**	**(2')**
Erreurs lexicales	LEF1 corrigées	512	15.57	398	11.74
	LEF3 corrigées	443	13.47	167	4.93
	Total (LEF1+LEF3) corrigées	955/1377	29.05	565/1301	16.68
Erreurs syntaxiques	SYF2 corrigées	49/413	1.49	20/281	0.59
Total de (LEF+SYF) corrigées		**1004/3287**	**30.54**	**585/3387**	**17.27**

Tableau 9 : Statistiques d'erreurs de la TANP F-V corrigées après le prétraitement

5.4 COMPARAISON GÉNERALE ET ÉVALUATION DES SYSTÈMES DE TA ANGLAIS-VIETNAMIEN ET FRANÇAIS-VIETNAMIEN

5.4.1 Systèmes de TA anglais-vietnamien

Dans cette section, nous présentons des points forts et des points faibles de chaque système de TA anglais-vietnamien que nous comparons aussi sur plusieurs critères : 1) taux erreurs de TANP ; 2) vitesse de traduction et 3) taille maximale du corpus d'entrée. Le Tableau 10 compare la vitesse de traduction avec la taille maximale des documents pouvant être traduits par les quatre systèmes de la traduction automatique de l'anglais vers le vietnamien.

Les systems de TA	**Taille du corpus**		**Temps de traduction**	
	Taille maximale d'un corpus		Taille moyenne d'un corpus (6,500 mots/100 textes)	
	Taille maximale d'un corpus d'entrée	Durée de traduction de toutes les parties	Subdivision du corpus en parties pouvant être traduites	Durée de traduction de tout le corpus
Vietgle	1200 mots (~7500 caractères)	240 secondes	5	1,200 secondes
Google	12000 mots (~72000 caractères)	2 secondes	1	1 seconde
Bing	700 mots (~4200 caractères)	2 secondes	10	20 secondes
EVTran	40 mots (~200 caractères)	600 secondes	165	99,000 secondes

Tableau 10 : Comparaison entre la vitesse de traduction et la taille d'un corpus pouvant être traduit par les quatre systèmes de TA E-V

Selon nos statistiques d'erreurs de La TANP, parmi les quatre systèmes de TA de l'anglais vers le vietnamien, Google Translate exécute le mieux la traduction de textes avec les NPs. En fait, Google fait le plus petit nombre d'erreurs de TANP (17.57 %), tandis que le ratio de Vietgle est de 21.49 % et celui de Bing 25.23 %. EVTran possède le taux le plus élevé d'erreurs (28.26 %). Le deuxième système de TA d'anglais vers le vietnamien le plus performant est Vietgle, avec le plus petit nombre de NTE1, lié au fait qu'il traduit bien les acronymes et les abréviations. En général, Vietgle fait moins de NTEs que Google et Bing.

De plus, concernant les erreurs syntaxiques, en comparaison avec Google, Bing et EVTran, Vietgle fait le plus petit nombre de SYE1. Cependant, il fait le plus grand nombre d'erreurs graphiques et d'erreurs lexicales causées par des mots incorrects. Bing Translator est le troisième système de TA E-V, lequel fait le plus bas taux d'erreurs syntaxiques parmi les quatre moteurs de TA E-V. Néanmoins, Bing effectue le nombre le plus élevé d'erreurs de non-traduction et d'erreurs graphiques. Particulièrement il fait un grand nombre d'erreurs de non-traduction pour les titres humains comme Mr. Ms. Mrs. Dr., etc. (par exemple, *Mr Fayyad, Mr Balls, Ms Giffords, Dr Hyacinth Orikara*) et les noms géographiques. Le quatrième système de TA de l'anglais vers le vietnamien est EVTran, qui totalise le plus grand nombre d'erreurs résultant de la traduction des NPs ayant des significations dans un dictionnaire. EVTran possède le plus haut taux d'erreurs syntaxiques et d'erreurs de transcription parmi les quatre systèmes de TA E-V. Néanmoins, EVTran réalise le plus petit nombre d'erreurs de non-traduction grâce à sa capacité de traduire tous les mots ayant des significations, tels que des noms personnels, des noms géographiques et des noms de produits. En fait, EVTran traduit beaucoup d'acronymes et d'abréviations tandis que d'autres systèmes ne le font pas.

Sur la base des évaluations présentées dans cette section, nous pouvons conclure qu'actuellement, les meilleurs systèmes de TA anglais-vietnamien sont Google Translate et Bing Translator en raison de : la qualité de leurs produits, du nombre d'options de langue, du taux élevé de textes d'entrée traduits et du nombre d'erreurs. Vietgle constitue un bon choix pour des utilisateurs souhaitant traduire des documents lié à des sujets spécifiques. Finalement, puisque EVTran traduit peu de

textes d'entrée (609/1500) à une vitesse moindre et offre des traductions de moindre qualité, il peut être considéré comme le système de TA le moins efficace.

5.4.2 Systèmes de TA français-vietnamien

Nous évaluons la qualité des deux traducteurs automatiques français-vietnamien (Google Translate et Bing Translator) en comparant la qualité de leurs sorties, c'est-à-dire le nombre d'erreurs de TANP faites par chaque moteur de TA, la vitesse de traduction et la taille maximale du texte d'entrée qu'ils peuvent traiter.

En réalité, il n'y a pas beaucoup de différences entre Google et Bing concernant les taux d'erreurs de la TANP. Les taux d'erreur pour chaque sous-type de NTF faites par Google et Bing sont semblables. Cela signifie que Google et Bing font face aux mêmes difficultés dans la traduction de NPs du français vers le vietnamien. Aussi bien Google que Bing font un nombre semblable d'erreurs lexicales et de transcription. Néanmoins, il y a une grande différence entre les sous-types d'erreurs graphiques faites par ces deux systèmes de TA F-V. Par exemple, le nombre d'erreurs graphiques causées par des noms géographiques en minuscule, des noms professionnels, des noms d'organisation et les titres d'œuvres de Bing est deux fois plus élevé que celui de Google; par contre, Google fait plus d'erreurs syntaxiques que Bing.

En ce qui concerne la vitesse de traduction et les capacités de traitement des textes d'entrée, les deux systèmes offrent une grande vitesse de traduction. Cependant, Google fournit les sorties non seulement à grande vitesse, mais aussi pour la plus grande quantité de textes d'entrée. Bing peut traduire un texte à la même vitesse que Google, mais la taille du texte d'entrée est limitée. Le tableau 11 présente la comparaison entre la vitesse de traduction et la taille de textes traduits par les deux systèmes de TA et le Tableau 12 récapitule tous les critères d'évaluation utilisés pour comparer les deux systèmes de TA du français vers le vietnamien.

MT system	Vitesse de traduction et taille du corpus d'entrée			
	Taille moyenne d'un corpus (incluant 7000 mots dans 100 textes)		Taille maximale d'un corpus	
	Les parties divisées du corpus	Durée	Taille maximale d'un corpus	Durée
Google	1	1 seconde	15000 mots (~ 100 000 caractères sans espace)	120 secondes
Bing	13	13 secondes	600 mots (~4200 caractères sans espace)	2 secondes

Tableau 11 : Comparaison entre la vitesse de traduction et la taille des documents pour les deux systèmes de TA F-V

No	Les critères d'évaluation	Google	Bing
1	Le nombre d'erreurs graphiques et d'erreurs de transcription (voir le Tableau 8)	Plus petit	Plus grand
2	Le nombre d'erreurs de non-traduction, d'erreurs lexicales et d'erreurs syntaxiques (voir le Tableau 8)	Plus grand	Plus petit
3	Nombre d'options de langues	58 langues	37 langues
4	Taille des documents d'entrée	Plus grand	Plus petit
5	Vitesse de traitement des documents d'entrée	Plus rapide	Plus lente
6	Capacité de traduire un corpus contenant le même nombre de textes	1500 textes/ 1500 textes	1500 textes/ 1500 textes

Tableau 12 : Comparaison de deux systèmes de TA F-V

En résumé, nous avons évalué divers systèmes de TA anglais-vietnamien et deux moteurs de TA français-vietnamien en ce qui concerne la traduction de NPs, et nous avons indiqué les différents taux d'erreurs de la TANP faites par ces moteurs de TA. Sur la base de l'analyse d'erreurs de la TANP causées par chaque système de TA, nous avons exploré les avantages et les limites de différents moteurs selon le nombre d'erreurs de TANP, la vitesse de traduction, la capacité à traiter les textes d'entrée ainsi que selon le nombre d'options de langues. Nous constatons aussi que le prétraitement semble réduire significativement les erreurs de TANP pour les systèmes de TA de l'anglais et du français vers le vietnamien.

CHAPTER 6 : CONCLUSION

6.1 CONTRIBUTION DE LA THÈSE

Nous avons entrepris ce travail afin de trouver des solutions pour réduire les erreurs de TANP. Pour atteindre cet objectif, nous l'avons divisé en plusieurs sous-tâches, qui chacune, contribue significativement à l'objectif final d'amélioration de la qualité de la TANP. Nous énumérons différents apports de notre thèse dans l'ordre chronologique de notre travail.

Contribution 1: **Construction de deux corpus parallèles de NPs pour deux paires de langues : anglais-vietnamien et français-vietnamien**

Cette étude a fourni les premiers corpus parallèles annotés anglais-vietnamien et français-vietnamien pour la détection et la traduction de NPs. Ils peuvent être réutilisés par d'autres chercheurs de TAL. Simultanément ils peuvent constituer de bonnes bases de données pour des utilisateurs de TA, et servir à comparer les résultats de TANP offerts par différents systèmes de TA.

Contribution 2: **Classification de NPs combinant deux optiques linguistique et TAL**

Dans cette étude, nous présentons les classifications de NPs de plusieurs linguistes et discutons leurs avantages et limites; certaines de ces classifications sont trop complexes, incluant trop de types et de sous-types de NPs. D'autres sont trop simples pour indiquer les caractéristiques propres à chaque type de NPs. Les classifications de TAL de NPs sont trop générales pour inclure toutes les catégories de NPs. Pour répondre aux besoins de la linguistique et du TAL en ce qui concerne le traitement des NPs, nous avons proposé notre propre classification de NPs composée de sept types majeurs de NPs. Cette classification de NPs constitue une base d'analyse et de classification d'erreurs de TANP faites par les systèmes courants de TA E-V et de TA F-V.

Contribution 3: **Analyse et classification d'erreurs de TANP faites par les systèmes de TA anglais-vietnamien et français-vietnamien**

Sur la base des deux corpus parallèles de NPs, nous avons analysé et classifié les erreurs de TANP faites par quatre moteurs courants de TA E-V et deux moteurs

de TA F-V. La classification des erreurs de TANP est subdivisée en deux types majeurs : les erreurs de non-traduction et les erreurs de mauvaise traduction. Les erreurs de mauvaise traduction sont divisées en quatre types principaux basés sur des critères linguistiques relatifs à l'écriture, au lexique, à la syntaxe et à la transcription. Chaque catégorie d'erreurs de TANP est illustrée par la série d'exemples extraits des deux corpus parallèles de NPs. L'analyse quantitative d'erreurs de TANP fournit des informations sur les types d'erreurs les plus fréquents; par ailleurs, elle souligne aussi la nécessité de trouver des solutions appropriées afin de traiter les erreurs de TANP.

Contribution 4: **Comparaison de méthode manuelle, automatique et semi-automatique d'annotation des corpus anglais et français de NPs**

Puisque notre solution pour l'amélioration de TANP exige des corpus annotés, nous comparons trois méthodes d'annotation de corpus afin d'établir la méthode la plus appropriée à notre objectif. D'abord, nous proposons l'annotation manuelle pour créer un corpus de référence pour d'autres méthodes d'annotation. Cette méthode nécessite beaucoup de travail et de temps, mais elle produit les meilleurs résultats en termes de rappel et de précision d'annotation. L'annotation automatique est effectuée à l'aide de l'outil GATE. C'est la méthode la plus rapide et la moins exigeante, cependant, ses résultats ne sont pas suffisamment précis pour être directement utilisés par le programme de prétraitement. La méthode optimale est la méthode semi-automatique. C'est une combinaison des deux méthodes précédentes : nous appliquons d'abord la méthode automatique et corrigeons ensuite manuellement les résultats. De cette façon, le temps nécessaire à l'annotation est diminué par rapport à la méthode manuelle, et le taux de précision augmente par rapport à la méthode automatique.

Contribution 5: **Création du programme de prétraitement pour réduire les erreurs de TANP**

Pour faciliter le processus de prétraitement, nous avons créé un programme de prétraitement automatique, qui peut être utilisé efficacement pour corriger certains erreurs de TANP et par conséquent améliorer la qualité des systèmes de TA. Le programme de prétraitement effectue automatiquement certaines tâches prédéfinies de prétraitement : changement de structures possessives anglaises avec NPs, marquage en DNT pour certains types de noms propres et suppression de

groupes français "de + déterminant" précédant certains noms géographiques et noms d'organisation. Dans l'avenir, il serait intéressant de développer ce programme dans deux directions : d'abord, élargir les types d'erreurs de TANP pouvant être corrigés, et ensuite, l'adapter à d'autres paires de langues.

Contribution 6: **Amélioration de la qualité de TA par le prétraitement**

L'utilisation du programme de prétraitement permet de réduire un grand nombre d'erreurs lors de la traduction automatique des textes avec NPs de l'anglais et du français vers le vietnamien. Le taux moyen de réduction d'erreurs est de 32.08 % pour les systèmes anglais-vietnamien et de 23.90 % pour les systèmes français-vietnamien. Par conséquent, la qualité de traduction s'en trouve améliorée. Puisque le prétraitement est fait automatiquement, le temps de pré-édition est aussi significativement réduit. Il en est de même pour la post-édition, grâce à la meilleure qualité de traduction. Par conséquent, l'application du programme de prétraitement à la TA peut réduire le temps et le travail de pré-édition et de post-édition pour des traducteurs.

Contribution 7: **Établissement de deux glossaires de NPs pour la TA anglais-vietnamien et français-vietnamien**

Un autre avantage du programme de prétraitement consiste en la possibilité de créer des glossaires des NPs pour chaque langue. Ces glossaires énumérant un ensemble de NPs dans la langue source et leur traduction en langue cible, peuvent être réutilisés à d'autres fins, aussi bien par les utilisateurs humains que d'autres systèmes automatiques.

Contribution 8: **Évaluation de différents systèmes de TA utilisés dans la communauté du TAL au Vietnam**

Finalement, nous avons proposé une évaluation d'actuels systèmes de TA anglais-vietnamien et français-vietnamien. Cette évaluation porte plus particulièrement sur les taux d'erreurs de TANP. Elle permettra d'explorer de nouvelles solutions de réduction de ces erreurs.

6.2 LIMITATIONS DE LA THÈSE ET RECHERCHE A VENIR

Les limitations de notre travail constituent aussi les orientations futures pour notre recherche. En voici une liste non-exhaustive :

1. Certains types d'erreurs de TANP ne sont pas corrigées.

Bien que le programme de prétraitement ait été construit pour prévenir les erreurs de traduction automatique de textes avec NPs certains types d'erreurs de TANP ne peuvent toujours pas être corrigées, telles que les erreurs graphiques, les erreurs de transcription, plusieurs sous-types d'erreurs sémantiques (LEE2, LEE3, LEE4 et LEF2) et les erreurs syntaxiques (SYE2, SYF1 et SYF3). Nous continuerons notre recherche pour trouver des solutions susceptibles de corriger ces types d'erreurs.

2. Le programme de prétraitement exige des textes annotés.

Une des limites de la thèse est la préparation des documents d'entrée, et notamment l'annotation qui nécessite du temps et du travail. Pour obtenir un document correctement annoté, nous devons vérifier et corriger toutes les erreurs d'annotation, parce que le fonctionnement du programme est limité par des erreurs d'annotation.

3. Limitation du programme d'annotation de corpus

La méthode d'annotation semi-automatique est basée sur le système GATE, qui obtient de très bons résultats en anglais. Néanmoins, les résultats pour le français ne sont pas aussi bons, puisque ce système n'est pas encore suffisamment entrainé pour annoter cette langue.

4. Le programme de prétraitement ne peut pas être lié directement aux systèmes de TA

Les utilisateurs doivent copier les textes prétraités pour les soumettre à différents systèmes de TA ; ils doivent recopier les textes traduits dans le programme de prétraitement pour récupérer les résultats finaux. Il serait idéal de pouvoir connecter directement le programme de prétraitement aux différents systèmes de TA afin d'éviter ces manipulations superflues.

5. Limitation des langues pouvant être prétraitées

Le programme de prétraitement peut être utilisé pour les documents en deux langues : l’anglais et le français. Il ne peut pas être appliqué à d'autres langues telles que le vietnamien, le chinois, l'espagnol, etc. Le développement de ce programme pour d'autres langues constitue l’une de nos orientations de future recherche.

BIBLIOGRAPHIE SELECTIVE

Aone, C. and Maloney, J. (1997). Reuse of Proper Noun Recognition System in Commercial and Operational NLP Applications. In Proceedings of ACL'97 Workshop on From Research to Commercial Applications: Making NLP Technology Work in Practice.

Babych, B. and Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, Centre for Translation Studies, University of Leeds, UK., pp. 1-8.

Balabantaray, R., and Sahoo, D. (2013). An Experiment to Create Parallel Corpora for Odia. *International Journal of Computer Applications*, *67*(19), pp. 18-20.

Cunningham, H. et al. (2012). Developing Language Processing Components with GATE Version 7 (a User Guide), the University of Sheffield, Department of Computer Science 2001-2012. Retrieved on the 12th March 2013 from http://gate.ac.uk/userguide.

Dien, D. (2005). Building an annotated English-Vietnamese parallel corpus, *MKS: A Journal of Southeast Asian Linguistics and Languages*, Volume 35, pp. 21-36.

Dinh, D., Hoang, K. & Hovy, E. (2004). BTL: A Hybrid Model for English-Vietnamese Machine Translation

Hassan, A., Fahmy, H., & Hassan, H. (2007). Improving named entity translation by exploiting comparable and parallel corpora. *AMML07*

Héja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In Proceedings of the LREC2010 Conference, La Valletta, Malta.

Hermjako, U., Knight K. & Daumé III, H. (2008). Name Translation in Statistical Machine Translation: Learning When to Transliterate. In Proceedings of the 46th Annual Meeting on Association for Computational Linguistics, Columbus, Ohio, pp. 389-397.

Hidalgo, J. M. G., Garcia, F.C., and Sanz, E.P. (2005). Named Entity Recognition for Web Content Filtering. In Montoyo, A., Munoz, R., Métais, E. (Eds.) *Natural*

Language Processing and Information Systems, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, *NLDB 2005*, Springer publisher, pp.286-297.

Krstev, C., Vitas, D., Maurel, D. & Tran, M. (2005). Multilingual Ontology of Proper Names. In Proceedings of 2nd Language & Technology Conference, Poznań, Poland, ed. Zygmunt Vetulani, pp. 116-119.

Leech, G. (2004). Adding Linguistic Annotation. In Wynne M. (Ed.) Developing Linguistic Corpora: a Guide to Good Practice. Retrieved on the 15th September 2012 from http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm.

Leroy, S. (2004). *Le Nom propre en français*. (Coll. *L'Essentiel français*). Paris, Gap: Ophrys.

Maurel, D., et al. (2008). Prolexbase, ProLMF version 1.2, Université François-Rabelais de Tours, retrieved at : http://www.cnrtl.fr/lexiques/prolex/

Maurel, D., and Bouchou-Markhoff, B. (2013). Prolmf: A Multilingual Dictionary of Proper Names and their Relations. In *LMF Lexical Markup Framework*, pp. 67-82

Munteanu, D.S. and Marcu, D. (2006). Improving Machine Translation Performance by Exploiting Non-parallel Corpora, Association for Computational Linguistics, 31(4), pp. 476-504.

Phan, T.T.T. and Thomas , I. (2012). English-Vietnamese Machine Translation of Proper Names: Error Analysis and Some Proposed Solutions", Proceedings of the 15th internaltional conference TSD 2012 (Text, Speech and Dialogue), September 3-7, 2012, Brno, Czech Republic, Springer Edition, pp.386-393.

Phan, T.T.T. and Thomas, I. (2013). Pre-processing as a Solution to Improve French-Vietnamese Named Entity Machine Translation. In Vetulani Z. and Uszkoveit H. (Eds.) Proceedings of Human Language Technologies as a Challenge for Computer Sciences and Linguistics, the 6th Language and Technology Conference (LTC 2013), December 7-9, 2013, Poznan, Poland, pp. 142-146.

Sang, T.K.E.F., and De Meulder, F.(2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Association for Computational Linguistics, Volume 4, pp. 142-147.

Somers, H. (2003). Machine Translation: Latest Developments. In Mitkov R. (Ed.) *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 511-529.

Vilar, D., Xu, X., D'Haro, L.F., Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genova, Italy.

Virga, P. & Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Lingual Information Retrieval. In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed- Language named Entity Recognition, pp. 57-64.

Wilks, Y. (2008). *Machine Translation:Its Scope and Limits*, Publisher: Springer,p.5

UNIVERSITÉ DE FRANCHE-COMTÉ
ÉCOLE DOCTORALE «LANGAGES, ESPACES, TEMPS, SOCIÉTÉS»

Thèse en vue de l'obtention du titre de docteur en

SCIENCES DU LANGAGE
SPÉCIALITÉ : TRAITEMENT AUTOMATIQUE DES LANGUES

MACHINE TRANSLATION OF PROPER NAMES FROM ENGLISH AND FRENCH INTO VIETNAMESE: AN ERROR ANALYSIS AND SOME PROPOSED SOLUTIONS

Traduction automatique des noms propres de l'anglais et du français vers le vietnamien : analyse des erreurs et quelques solutions

Présentée et soutenue publiquement par

Thao PHAN THI THANH

Le 11 mars 2014

Sous la direction de Mme le Professeur Sylviane CARDEY-GREENFIELD
et la co-direction de M. Dr. Ha LE AN et Mme. Dr. Izabella THOMAS

Membres de Jury:

Sylviane CARDEY-GREENFIELD, Directrice de recherche, université de Franche-Comté, France
Ha LE AN, Docteur, HDR, Co-directeur de recherche, université de Wolverhampton, Royaume-Uni
Denis MAUREL, Professeur, université François Rabelais de Tours, France, Rapporteur
Ruslan MITKOV, Professeur, université de Wolverhampton, Royaume-Uni, Rapporteur
Izabella THOMAS, Docteur, Co-directrice de recherche, université de Franche-Comté, France.

ABSTRACT

In the present era, in respect of information and knowledge, machine translation (MT) has increasingly become an indispensable tool for decoding the meaning of a text written in a source language into a target language. In particular, MT of proper names (PN) plays a crucial role in providing the specific and precise identification of persons, places, organizations and artefacts across languages. Despite a large number of studies and significant achievements in named entity recognition in the NLP community around the world, there has been almost no research on PNMT for Vietnamese.

Due to the different features of PN writing, transliteration or transcription and translation from a variety of languages including English, French, Russian, Chinese, etc. into Vietnamese, the PNMT from these languages into Vietnamese is still a challenging and problematic issue. This study focuses on the problems of English-Vietnamese and French-Vietnamese PNMT arising with current MT systems. First, the study proposes a corpus-based PN classification, and follows with a detailed PNMT error analysis, concluding with some pre-processing solutions in order to improve MT quality.

By means of the analysis and classification of PNMT errors from two English-Vietnamese and French-Vietnamese parallel corpora of texts with PNs, we propose solutions concerning two major issues: (1) corpus annotation for preparing the pre-processing databases, and (2) design of the pre-processing program to be used on annotated corpora to reduce PNMT errors and enhance the quality of MT systems, these including Google, Vietgle, Bing and EVTRAN.

The efficacy of different annotation methods for English and French corpora of PNs and the results of PNMT errors before and after using the pre-processing program on the two annotated corpora are compared and discussed in this study. It is shown that the pre-processing solution reduces significantly PNMT errors and contributes to the improvement of the MT systems for Vietnamese.

Key words: proper name, pre-processing, parallel corpus, machine translation quality, English-Vietnamese, French-Vietnamese, translation error.

RÉSUMÉ

Dans l'ère de l'information et de la connaissance, la traduction automatique (TA) devient progressivement un outil indispensable pour transposer la signification d'un texte d'une langue source vers une langue cible. La TA des noms propres (NP), en particulier, joue un rôle crucial dans ce processus, puisqu'elle permet une identification précise des personnes, des lieux, des organisations et des artefacts à travers les langues. Malgré un grand nombre d'études et des résultats significatifs concernant la reconnaissance d'entités nommées (dont le nom propre fait partie) dans la communauté de TAL dans le monde, il n'existe presque aucune recherche sur la traduction automatiques des noms propres (TANP) pour le vietnamien.

En raison des caractéristiques particulières d'écriture, de translittération/ transcription et de traduction des NP dans une variété des langues telles que l'anglais, le français, le russe, le chinois, etc., la TANP depuis ces langues vers le vietnamien constitue un défi complexe. Notre recherche se focalise sur les erreurs de TANP d'anglais vers le vietnamien et de français vers le vietnamien, résultant des systèmes courants de TA. Tout d'abord, elle propose une classification des NP basée sur le corpus, ensuite une analyse des erreurs de la TANP, pour conclure par une proposition de solution de prétraitement pour améliorer la qualité de la TA.

A travers l'analyse et la classification d'erreurs de la TANP faites sur deux corpus parallèles de textes avec PN (anglais-vietnamien et français-vietnamien), nous proposons les solutions concernant deux problématiques importantes : (1) l'annotation de corpus, afin de préparer des bases de données pour le prétraitement et (2) la création d'un programme pour prétraiter automatiquement les corpus annotés, afin de réduire les erreurs de la TANP et d'améliorer la qualité de traduction des systèmes de TA, tels que Google, Vietgle, Bing et EVTran. L'efficacité de différentes méthodes d'annotation des corpus avec des NP ainsi que les taux d'erreurs de la TANP avant et après l'application du programme de prétraitement sur les deux corpus annotés sont comparés et discutés dans cette thèse. Ils prouvent que le prétraitement réduit significativement le taux d'erreurs de la TANP et, par la même, contribue à l'amélioration de traduction automatique vers la langue vietnamienne.

Mots-clés : nom propre, prétraitement, corpus parallèle, qualité de traduction automatique, anglais-vietnamien, français-vietnamien, erreur de traduction

TABLE OF CONTENTS

ABSTRACT I
RÉSUMÉ III
TABLE OF CONTENTS IV
CHAPTER 1: INTRODUCTION 1
1.1 RATIONALE 1
1.2 STUDY OBJECTIVES 3
1.3 METHODOLOGY 4
1.4 STRUCTURE OF THESIS 4
CHAPTER 2: STATE OF ART 6
2.1 MACHINE TRANSLATION 6
2.2 PROPER NAMES 7
2.3 PROPER NAME MACHINE TRANSLATION 8
CHAPTER 3: CORPUS-BASED PNMT ERROR ANALYSIS 10
3.1 DATABASE PREPARATION: BUILDING THE PARALLEL CORPUS OF TEXTS WITH PROPER NAMES 10
3.2 CORPUS-BASED ANALYSIS OF PNMT ERRORS 11
3.2.1 Classification of PNs 11
3.2.2 Classification of PNMT errors 11
3.2.3 Qualitative PNMT error analysis with examples 14
CHAPTER 4: CORPUS ANNOTATION 16
4.1 ANNOTATION FRAMEWORK 16
4.2 ANNOTATION METHODS 16
4.2.1 Manual annotation 17
4.2.2 Automatic and semi-automatic annotation 17
4.3 CORPUS ANNOTATION RESULTS 19
4.4 COMPARISON OF ANNOTATION RESULTS USING DIFFERENT METHODS 20
CHAPTER 5: IMPROVING MT BY PRE-PROCESSING 21
5.1 PRE-PROCESSING 21
5.1.1 Definition and general pre-processing tasks 21
5.1.2 Specific pre-processing tasks for English-Vietnamese and French-Vietnamese MT 21
5.2 DESCRIPTION OF THE PRE-PROCESSING PROGRAM FOR PNMT ERROR REDUCTION 23
5.2.1 Principal tasks 23
5.2.3 The pre-processing program's interface and demos 23
5.2.4 Advantages and limitations of the pre-processing program 26
5.3 TESTING RESULTS OF PRE-PROCESSING 27
5.3.1 Testing results of English-Vietnamese corpus of proper names before and after pre-processing 27
5.3.2 Testing results of FVC of proper names before and after pre-processing 29
5.4 OVERALL COMPARISON AND EVALUATION OF ENGLISH-VIETNAMESE AND FRENCH-VIETNAMESE MT SYSTEMS 30
5.4.1 Issues of English-Vietnamese MT systems 30
5.4.2 Issues of French-Vietnamese MT systems 32
CHAPTER 6: CONCLUSION 34
6.1 CONTRIBUTIONS OF THE THESIS 34
6.2 LIMITATIONS OF THE THESIS AND FUTURE RESEARCH 36
SELECTIVE BIBLIOGRAPHY 38

CHAPTER 1: INTRODUCTION

1.1 RATIONALE

The growing importance of machine translation (MT) is an obvious fact, especially in our current era of information and knowledge, in which the need of accessing, exchanging and transmitting a huge amount of information sources is rapidly increasing all over the world. The role of MT has become emerging and indispensible for the following major reasons:

Firstly, MT is highly agreed to possess the advantage of simple, high-speed and low-cost usage in the NLP and other information technology applied domains. In fact, first MT is an easy, fast and cost-effective tool, which can be used to replace human beings to translate easily a large quantity of documents in a short time with a low amount of money.

Second, the increasing amount of good MT systems has changed people's thinking and raised their knowledge of MT. Somers (2003:514) states that the success of a number of MT systems, which are available on the World Wide Web, usually free, has heightened awareness of MT for the public.

Third, the number of computer users, who take MT into consideration, is growing constantly. More and more people are using MT for their daily communication because of its beneficial and universal features (Wilks (2008:10).

Fourth, MT quality has been mostly influenced by its errors in the translation of source texts into target texts, in which name entity (NE) mistranslation and missing content words should be notably noticed. NEs are particularly challenging MT systems to translate documents correctly (Hermjakob et al. 2008), which can have a major impact on the research of result translation.

As Wilks (2008) concludes, MT is a lively and essential technology, and its importance in a multi-lingual and information driven world can only increase intellectually and commercially. Despite the constant development and significant achievements in the NLP domain, MT systems are still dealing with numerous challenges; in particular, MT systems have problems in translating proper names (Aone and Maloney 1998, Hirschman et al. 2000, Somers 2003:523). According to Aone (1998), there are two cases where an MT system fails to translate names.

First, MT systems fail to recognize "where a name starts and ends in a text string"; this causes a non-trivial problem in languages with logogram characters

such as Japanese (kanji), Chinese (hanzi), Korean (hanja), etc. The MT systems often "cut" names into words and translate each word individually in those languages in which proper names are not capitalized. For instance, a Japanese person name "Mori Hanae" in kanji characters is segmented into three words including "mori" (forest), "hana" (England) and "e" (blessing) by MT systems (Aone 1998). In reality, this person name should be neither segmented nor translated.

Then MT systems sometimes fail to distinguish names and non-names. Hirschman et al. (2000) identify this typical error of translation of PNs as if they were normal meaningful words. This type of errors, which is also made by English – Vietnamese MT systems, happens to the person names and abbreviations or acronyms of geographic names and organizations. For example, some English persons names as "*Brown, Rice, Greenfield, Mark*, etc." are automatically translated into Vietnamese as if they were the common nouns having the meaning of "*brown colour, grains used as food, a green field, a distinguishing symbol*". Actually, those personal names should be kept unchanged.

Concerning machine translation of PNs from English and French into Vietnamese, the MT systems often make two major types of mistakes, that we classify as the non-translation and wrong translation errors. Non-translation errors frequently happen to abbreviations, professional titles, organization names, weekdays, etc. Wrong translation errors of PNs are due to different linguistic features such as graphical, lexical, syntactic and transcription or transliteration characteristics. For instance, all the weekdays and months are graphically written in upper case with the initial letters in English, but they are not in Vietnamese (e.g. *Sunday* is incorrectly translated into Vietnamese by *Chủ nhật* or *Chủ Nhật*, which should be corrected to *chủ nhật*). On the contrary, the majority of languages, nationalities are not capitalized in French, but they are often written in upper case in Vietnamese, e.g., *russe* is wrongly translated into Vietnamese by *tiếng nga*, which should be corrected to *tiếng Nga*.

Wrong translation errors are also caused by failing to distinguish names and non-names, which should or should not be translated. Some English geographic names are translated into Vietnamese, but others are not; for example, *West Ham* (i.e. a place in the London Borough of Newham in England) is translated by EVTran as *Đùi Phương tây* (i.e. meat cut from the thigh of a hog in the West); in fact, it should be kept untranslated. The same problems occur to some organization names, which should not be translated, e.g., *Marks and Spencer* is wrongly translated by

EVTran MT system as *Những sự đánh dấu và Spencer* (i.e. a marking that consists of lines that cross each other and Spencer).

1.2 STUDY OBJECTIVES

This thesis is expected to achieve the nine following targets:

Target 1 is to review the linguistics and NLP treatment of proper names in English and French in order to establish the proper name classification, which is the most suitable for the machine translation of PNs from English to Vietnamese and French to Vietnamese.

Target 2 is to create two annotated bilingual corpora of proper names: English-Vietnamese and French-Vietnamese corpora of texts with PNs. Firstly, these corpora will serve to the proper objectives of this thesis. Secondly, they can be made widely available for further research on PNMT.

Target 3 is to propose an appropriate classification of proper names based on corpus analysis and satisfying both linguistics and NLP domains; it will be further used as a criterion to idendify and analyze the PNMT errors made by the current English-Vietnamese and French-Vietnamese MT systems.

Target 4 is to perform the analysis and classification of PNMT errors made by the current MT systems using the established classification of PNs and the two corpora of proper names: English-Vietnamese parallel corpus (EVC) and French-Vietnamese parallel corpus (FVC).

Target 5 is to propose the best method of corpus annotation for preparing the database for the pre-processing program; we will compare the annotation results given by three annotation methods, namely manual, automatic and semi-automatic annotation.

Target 6 is to build the pre-processing program for English, French used to reduce the PNMT errors and improve the quality of MT in general, and the four MT systems (Vietgle, Google Translate, Bing Translator and EVTran) in particular.

Target 7 is to test the pre-processing program on the two corpora of PNs and to evaluate the results given by the MT systems with and without using this program. The comparison of the MT systems will be carried out to reveal the strengths and weaknesses of each system.

Target 8 is to indicate the advantages and limitations of the pre-processing program in the enhancement of PNMT from English to Vietnamese and French to Vietnamese for future directions of research.

Target 9 is to create two bilingual glossaries of proper names that can enhance further studies on PN translation from English and French into Vietnamese, as well as linguists and computational linguists.

1.3 METHODOLOGY

The methodology of our thesis is divided into five major steps:

i/ Building parallel corpora of texts with PNs is order to prepare the database for the PNMT error analysis and to evaluate the MT systems performance;

ii/ Corpus-based analysis and classification on PNMT errors, using linguistics and NLP approaches;

iii/ Corpus annotation and evaluation of annotation methods;

iv/ Design and test of the pre-processing program on two annotated corpora of PNs;

V/ Evaluation of MT quality improvement after using the pre-processing program

1.4 STRUCTURE OF THESIS

This thesis is structured in six chapters. **Chapter 1** presents the main motivations of our study, the research objectives and the general methodology to achieve our targets.

In **Chapter 2,** we address the questions of the state of art in two domains related to our research: machine translation and different treatment of proper names. The first part of this chapter describes the MT systems' development in Vietnam, and in particular, the current widely used translation engines with their strengths and limitations. The second part of the chapter is dedicated to the linguistic, NLP classifications of proper names, and focuses on the problems of PNMT from English, and French into Vietnamese, which have not been much studied so far.

In **Chapter 3,** we describe the two bilingual parallel corpora of texts with PNs that we have created, namely English-Vietnamese parallel corpus (EVC) and French-Vietnamese parallel corpus (FVC) and propose our proper classification of

PNs used for the corpus-based PNMT error analysis. We also show the two classifications of E-V PNMT and F-V PNMT errors illustrate them with a series of examples extracted from our corpora.

In **Chapter 4**, we address questions of the corpus annotation for the both English and French corpora of PNs and concentrate on the annotation method, which would be the best to support the pre-processing program for MT quality improvement. We present different methods including manual, automatic and semi-automatic annotation methods, analyze, classify and compare the corpus annotation results achieved for the two English and French corpora of PNs.

Chapter 5 concerns the pre-processing issue and the creation of the pre-processing program for PNMT errors reduction. This pre-processing program is established on the basis of specific pre-processing tasks such as the parsing and "Do-not-translate" (DNT) marking of PNs from the annotated corpus, the change of the possessive structures with PNs in English, and the omission of French groups of "de + determiners" preceding some geographic names and organization names. The program's description is detailed with its principal tasks, specifications, interface and demos and its advantages and limitations in use.

Chapter 5 gives the quantitative results of our research and the overall comparison and evaluation of the English-Vietnamese and French-Vietnamese MT systems. First, we show the testing results of PNMT errors on the two corpora of PNs with and without using the pre-processing program. Second, we evaluate the four English-Vietnamese and the two French-Vietnamese MT engines used in our experiments, comparing different criteria such as translation speed, size of input document, number of language options, processing speed of the input documents, number of PNMT errors before and after using the pre-processing program, and so forth.

Chapter 6 summarizes the whole thesis with a discussion about some contributions and limitations of the study. It proposes also the future directions of the research.

CHAPTER 2: STATE OF ART

This chapter summarizes the theoretical background relating to three major issues of our study: 1) machine translation, 2) proper names and 3) proper name machine translation.

2.1 MACHINE TRANSLATION

The first part of this chapter concerns the MT development in Vietnam. Although MT has been developed since 1940s in the world, it has started to be studied since 1960s and has particularly grown from the year 2000 to present for Vietnamese language. The Vietnam's MT development is divided into 4 periods as follows:

1. From 1960s to 1970: the beginning of ideas and projects of MT research for Vietnamese;

2. From 1970 to 1990s: the period called the "Closing time" since MT has almost been neglected without any significant achievements;

3. From 1990s to 2000s: the appearance of the first MT system in Vietnam - EVTran;

4. From 2000s to present: the flourishing period of MT in Vietnam due to the strong development of information technology with a series of MT systems' establishment, specifically, four MT systems for English-Vietnamese: Vietgle, Google Translate, Bing Translator and EVTran.

In addition, this thesis studies the impact of pre-procesing methods on the MT outputs, in particular, concentrates on how the pre-processing improves the quality of MT systems for English-Vietnamese and French-Vietnamese language pairs. Since the pre-processing is defined as a specific task to resolve a large number of linguistic problems concerning lexicon, semantics and syntax that exist in a source text, the main objectives of the pre-processing task are to correct the anticipated mistakes of the source texts before being put into a MT system. This study aims to build the effective pre-processing program, which can be applied in different MT systems to ameliorate the automatic translation of PNs from English and French into Vietnamese.

2.2 PROPER NAMES

The second part of this chapter presents the linguistic and NLP treatment and classifications of PNs. Since PNs play a vital role in all kinds of texts in European language, especially English and French, there was a particular tendency of doing research on proper names. According to Leroy (2004), a variety of linguists have paid great attention to linguistic features of proper names such as Algeo J. (1973), Kleiber (1981), Molino (1982), Siblot (1987), Le Bot (1989), Gary-Prieur (2001), Jonasson (1994), Maurel et Geuthner (2000), Van de Velde et Flaux (2000), Montecot, Osipov, Vassilaki (2001), Vaxelaire (2005), etc. Based on the linguistic features including phonological and graphic, morphological and lexical or semantic, syntactic and pragmatic features, PNs are divided into numerous categories. We present and analyze some proper name classifications from the basic to the complex ones in chronological order proposed by the following linguists: Zabeeh (1968), Molino (1982), Bauer (1985) and Grass (1999) (cited in Daille et al. 2000), Leroy (2004) and Vaxelaire (2005).

Along with linguists, numerous NLP researchers began to investigate the PNs, specifically in the domain of Named-Entity Recognition (NER), which has a great impact on various NLP's applications such as machine translation (MT), cross linguistics information retrieval (CLIR), information extraction (IE), Question-Answering (QA), Internet search engines or ontology population (Kozareva et al. 2008), web content filtering (Hidalgo, Garcia and Sanz, 2005), disambiguation of capitalized words to identify PNs (Mikheev 2002) and so on. Many important events focusing on the NER issues are: MUC-7 (Chinchor 1997), IREX (Sekine & Isahara 2000), CoNLL-2002 and CoNLL-2003 (Sang 2002, Sang & De Meulder 2003), ACE (Doddington et al. 2002), and HAREM (Santos et al. 2006), which mention different treatments and classificitaions of PNs in NLP domain.

To some extent, the PN classifications are still dealing with some problems related to both linguistics and NLP domain such as the existence of many different classifications, the lack of certain types and sub-types of PNs in some classifications, and the inconsistence of PN classification in some linguistic and NLP viewpoints. Consequently, our great concern on the establishment of a PN classification, which can be used for PN analysis in both linguistic and NLP domains, and particularly, can be useful for MT purpose.

2.3 PROPER NAME MACHINE TRANSLATION

The third part of this chapter mentions the overview of previous works on PNMT and the focus of PNMT from a foreign language into Vietnamese. In spite of certain related studies on PNMT for different language pairs such as English-Chinese, Spanish-English, etc. and some NER issues concerning Vietnamese, PNMT for English-Vietnamese and French-Vietnamese has never been addressed before. Thus, our study is the first contribution to the study of PNMT, and we hope it will be useful for MT development.

There exist many studies on MT of PNs in various pairs of languages such as English-Chinese (Chen et al. 1998), Spanish-English (Hirschman et al. 2000), French-English (Noir 1995, Moore and Robert 2003, Moshop 2007), Arabic-English (Izwaini 2006, Kashani et al. 2007), Japanese-English (Kumano et al. 2004) and in particular, the study on MT of multi-lingual PNs in the *Prolexbase* (Maurel et al. 2006) introduces French PNs translated to other languages including German, English, Intalian, Dutch, Polish, Portuguese, Serbic. Despite some previous studies addressing some issues on PNMT for different language pairs on NER in Vietnamese, there has been still no research on PNMT for the English-Vietnamese and French-Vietnamese language pairs. Actually, MT systems are coping with numerous problems for PNMT in these two language pairs. These problems derive as well from the PNs' characteristics as from the inconsistence of their writing and transcription or transliteration in Vietnamese.

Concerning PN translation from a foreign language into Vietnamese, we discuss three major questions:

1) the principles of writing proper names in Vietnamese;

2) the transcription/transliteration and the translation of PNs from a foreign language into Vietnamese;

3) challenges of PNMT from English and French into Vietnamese.

In order to propose the principles of writing PNs in Vietnamese and to regulate the process of transliteration/transcription and translation of PNs from a foreign language into Vietnamese, we divide PNs into two main types: Vietnamese PNs and foreign PNs. As regards the Vietnamese PNs, we present some writing principles related to capitalization. As to the foreign PNs, they are involved in two

different issues. The first one deals with the transcription or transliteration of PNs, of which writing is based on either Latin characters or other characters. The second concerns the PN translation from a foreign language into Vietnamese in general, and from English and French into Vietnamese in particular.

CHAPTER 3: CORPUS-BASED PNMT ERROR ANALYSIS

3.1 DATABASE PREPARATION: BUILDING THE PARALLEL CORPUS OF TEXTS WITH PROPER NAMES

Parallel corpora are valuable resources in language studies, teaching and many NLP domains such as MT, CLIR, IE, QA, Word Sense Disambiguation (WSD), bilingual terminology extraction, and so on. Unfortunately, the sources of parallel texts are very limited and the number of parallel corpora is restricted for certain language pairs.

The fact is that the needs of using parallel corpora are constantly increasing due to their benefits and the potential for retrieving bilingual texts. Computational linguists thus are attempting to construct parallel corpora for more pairs of languages. In this section, we focus on the procedures and objectives of building parallel corpora in general, and on the method, we used to establish the two bilingual corpora of texts with PNs for English-Vietnamese and French-Vietnamese. We focus on describing the two parallel corpora of PNs (EVC and FVC).

EVC is a collection of 1,500 texts (including 101,289 words or 575,166 characters) extracted from BBC News online related to different topics like Politics, Business, Health, Science-Environment, Education, Entertainment and Arts, Technology, etc. concerning news from Africa, America, Asia-Pacific, Europe, US, Canada, Latin America & Caribbean, Middle East, South Asia, Japan, UK, etc. All of these texts have been translated into Vietnamese with the most popular translation engines used in Vietnam at present: Vietgle, Google Translate, Bing Translator and EVTran.

The French-Vietnamese parallel corpus (FVC) consists of 1,500 texts (including 109,584 words or 781,347 characters) extracted from online articles of Le Monde. This corpus includes texts of all topics such as News, Economy, Sport, Culture, Education, etc. belonging to various categories such as International, Africa, America, Europe, Asia-Pacific, and Middle East. We randomly collect those texts from different articles and put them into French-Vietnamese MT systems.

3.2 CORPUS-BASED ANALYSIS OF PNMT ERRORS

3.2.1 Classification of PNs

Our study objective is to find the most appropriate classification of PN used as the robust training data for identifying and extracting named entities from a corpus. In comparison to the previous PN classifications by linguists and NLP researchers, our classification requires being not too complex but sufficient to describe the major typical features of PNs from both linguistics and NLP perspectives (see Figure 1)

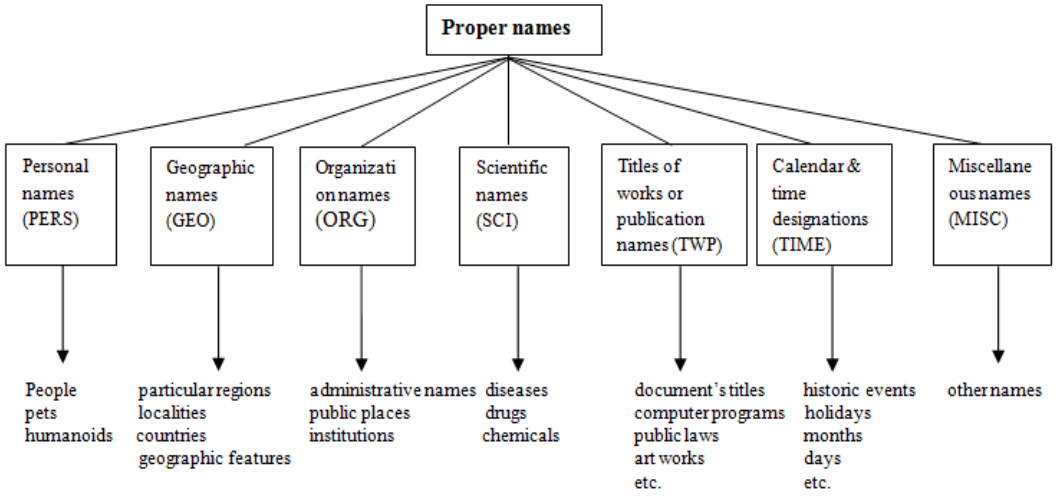


Figure 1: Our proposed classification of PNs

3.2.2 Classification of PNMT errors

Automatic translation of PNs is not free of mistakes for any MT systems, in particular, English-Vietnamese and French-Vietnamese translation engines. In our study, we essentially focus on two corpora - driven types of errors, namely non-translation and wrong translation of PNs . Non-translation (NT) errors made by MT systems often happen to most of abbreviations or acronyms, professional titles, organization names, weekdays, etc. Based on our proposed PN classification, we divide NT errors into six sub-types including non-translation of abbreviations or acronyms, professional and human titles, geographic names, organizations names, titles of works, weekdays and months. Wrong translation errors are based on four criteria concerning graphics, lexis, syntax, and transcription/transliteration. Hence, we classify the wrong translation errors of PNMT into four major types of errors

including graphic errors (GE), lexical errors (LE), syntactic errors (SY) and transcription errors (TE).

Graphic errors result from wrong capitalization of PNs translated by MT systems. There are two cases of wrong capitalization of PNs. The first relates to PNs such as geographic names, professional titles, organization names, titles of works, names of events, nationalities, etc., which should be capitalized in translations, but they are not. The second occurs to some PNs like months, weekdays, human titles, which are often capitalized with initial letters in English; however, they should not be written in upper case in Vietnamese.

Lexical errors refer to wrong translation at word level and include errors concerning incorrect words, missing words, redundant words and unknown words (Vilar et al. 2006). Incorrect word errors occur when MT systems do not provide a correct sense of a word being a constituent of a proper name. Due to the polysemous nature of words, MT systems cannot easily select the correct translations for a given contexts. MT systems translate a common noun written in upper case as if it were a proper name. In contrast, they translate a proper name having a meaning in a dictionary as if it were a common noun. For example, the word *"turkey"* which has different meanings can be translated incorrectly and inappropriately for the given circumstance e.g., the sentence "*Turkey is my favourite food*" in English is automatically translated as "*Thổ Nhĩ Kỳ là thực phẩm mà tôi thích nhất*" in Vietnamese (i.e. *Turkey country is my favourite food*). In this sentence, "turkey" should be translated as *a large gallinaceous bird with fan-shaped tail, widely domesticated for food*, but not a location name - *Turkey, a Eurasian republic in Asia Minor and the Balkans.*

Missing word errors are produced when a word in the generated sentence is missing (Vilar et al. 2006). For instance, the quantifiers preceding plural nouns in Vietnamese such as "*các, những, nhiều*, etc.", are sometimes missing in the MT of English and French noun phrases with PNs.

Redundant word errors are derived from the use of duplicative, unnecessary or superfluous words (Vilar et al. 2006). For example, this type of errors arises from the superfluousness of quantifiers "*các, những, nhiều,* etc.*"* preceding Vietnamese singular noun phrases with PNs when translated from English.

Unknown word errors are mistakes made by the use of unidentified, unrecognized, unfamiliar words, and words written in another language in PNMT. For example, some geographic names should be translated automatically from French into Vietnamese; however, the obtained outputs are geographic names written in English, but not in Vietnamese.

Syntactic errors are the errors concerning the wrong translation of PN structures, for instance, wrong translation of possessive structures, wrong order of words in noun phrases with PNs, wrong order of personal name, and wrong order of dates.

Transcription or transliteration errors (TE) occur to the PNs, which are incorrectly transcribed or transliterated by the MT systems. This type of errors occasionally happens to geographic names and miscellaneous names including monetary units, names of nationalities and languages.

Dealing with PNMT problems for two pairs of languages, we classify the PNMT errors into two categories: i/ Classification of PNMT errors for English-Vietnamese; ii/ Classification of PNMT errors for French-Vietnamese. To distinguish the English-Vietnamese PNMT errors from French-Vietnamese PNMT errors, we use the abbreviations such as NTE1 (non-translation errors of English), NTF1 (non-translation errors of French), GEE1 (graphic errors of English), SYF1 (syntactic errors of French), etc. (see Figure 2 and Figure 3).

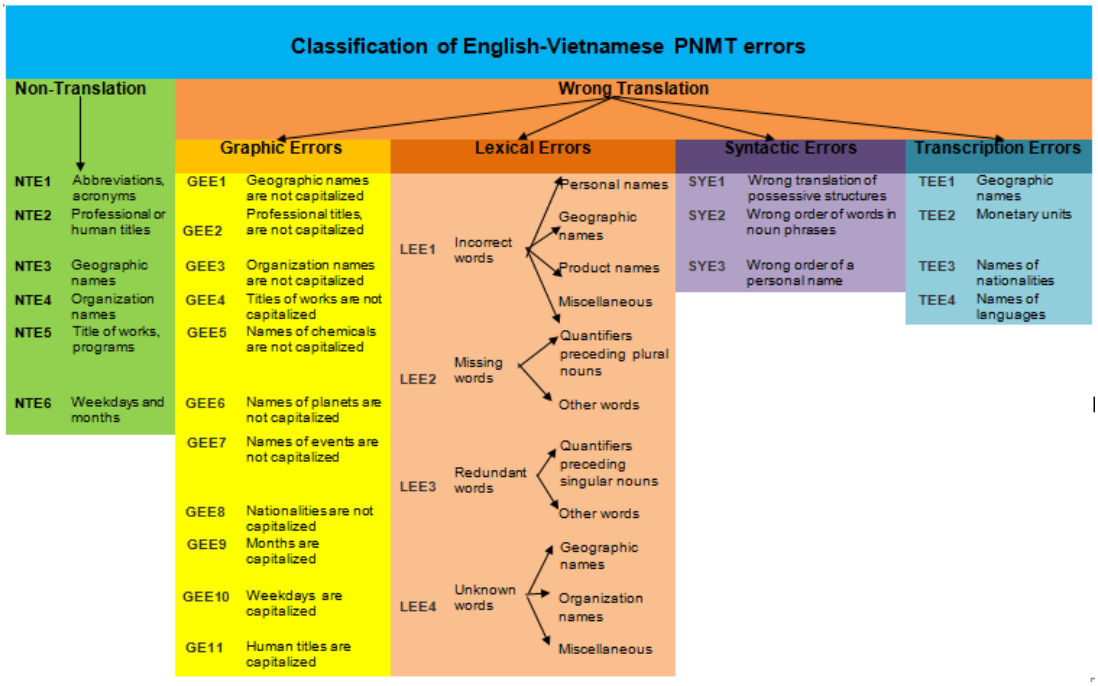


Figure 2: Classification of English-Vietnamese PNMT errors

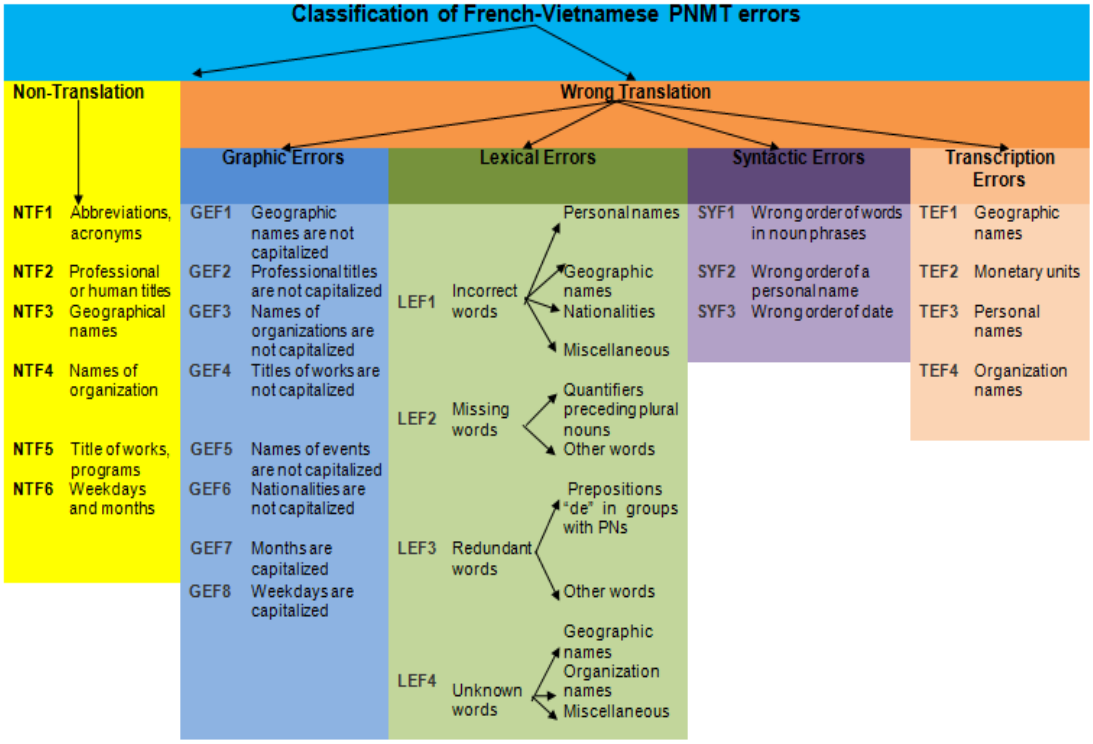


Figure 3: Classification of French-Vietnamese PNMT errors

3.2.3 Qualitative PNMT error analysis with examples

This section illustrates the classification of English-Vietnamese and French-Vietnamese PNMT errors with a series of examples extracted from the two parallel corpora of PNs mentioned in Chapter 2. Since they are authentic and useful examples for the analysis and classification of PNMT errors (Table 1) on both linguistic and NLP criteria, the analysis results that we have made, bring the linguists and NLP researchers the very first essential resources for PN translation study from English and French into Vietnamese

Type of errors		EVC	FVC
Non-translation		**NTE1:** RSPB (The Royal Society for the Protection of Birds)	**NTF1:** JDD (Journal du dimanche), PS (Parti socialiste)
		NTE2: Gen Mladic (Tướng Mladic), Mr. Shapiro (ông Shapiro)	**NTF2:** M. Le Graët (ông Le Graët), Comtesse de Besarn (nữ bá tước de Besarn)
		NTE3: Buckingham Palace (Cung điện Buckingham), Blue Cypress Lake (Hồ Bách xanh)	**NTF3:** Palais de Tokyp (Cung điện Tokyo), Quai d'Orsay (Bến tàu Orsay)
		NTE4: Museum of Innocence (Bảo tàng Ngây thơ)	**NTF4:** Théâtre de l'Aquarium (Nhà hát Thủy cung)
		NTE5: No Country for Old Men (Không chốn dung thân)	**NTF5:** Le Monde (tờ báo Thế giới)
		NTE6: May (tháng năm)	**NTF6:** mecredi (thứ tư)
Wrong translation	GE	**GEE1:** Red Square => quảng trường đỏ (Quảng trường Đỏ)	**GEF1:** Côte d'Ivoire => bờ biển ngà (Bờ biển Ngà)
		GEE2: CIA director Leon Panetta=>Leon Panetta giám đốc Cục	**GEF2**: le premier ministre Georges Papandreou=> thủ

		tính báo trung ương Hoa Kỳ (Leon Panetta, Giám đốc Cục tính báo trung ương Hoa Kỳ)	tướng Georges Papandreou (Thủ tướng Georges Papandreou)
		GEE3: parliament=> quốc hội (Quốc hội)	**GEF3:** la Cour pénale internationale=> tòa án hình sự quốc tế (Tòa án Hình sự Quốc tế)
		GEE4: Financial Times => thời báo tài chình (Thời báo Tài chình)	**GEF4:** la Constitution afghane=> hiến pháp Afghanistan (Hiến pháp Afghanistan)
		GEE5: elements Thorium and Hafnium => các yếu tố thori và hafini (các nguyên tố Thori và Hafini)	**GEF5:** la Renaissance=> phục hưng (Phục hưng)
		GEE6: Solar System=> thái dương hệ (Thái dương Hệ)	**GEF6:** Arabes=> người ả Rập (người Ả Rập)
		GEE7: Toronto International Film Festival=> liên hoan phim quốc tế Toronto (Liên hoan Phim Quốc tế Toronto)	**GEF7:** avril=> tháng Tư (tháng tư)
		GEE8: Italian extraction => những người ý (Ý) khai thác	**GEF8:** mardi=> thứ Ba (thứ ba)
		GEE9: January=> tháng Giếng (tháng giêng)	
		GEE10: Monday=> thứ Hai (thứ hai)	
		GEE11: Mr. Obama=> Ông Obama (ông Obama)	
	LE	**LEE1:** Vince Cable => Cáp Vince (le cable Vince) (Vince Cable)	**LEF1:** Marine Le Pen=> biển Le Pen (mer Le Pen) (Marine Le Pen)
		LEE2: Russian TV channels=> (các) kênh truyền hình Nga	**LEF2:** municipales partielles italiennes=> (những) vùng thành phố Ý
		LEE3: Planet Jupiter=> các hành tinh sao Mộc (Hành tinh sao Mộc)	**LEF3:** la prairie de Pont-Aven=> ~~đồ~~ng cỏ của Pont-Aven
	SY	**SYE1:** Mr. Brown's success => Ông Nâu là thành công (M. Brown est succès) (thành công của ông Brown)	**SYF1:** L'Estat de Kaduna=> Kaduna Nhà nước (Nhà nước Kaduna)
		SYE2: Chinese flags=> Trung quốc cờ xanh (cờ xanh Trung Quốc)	**SYF2:** Moussa Ibrahim=>Ibrahim Moussa (Moussa Ibrahim)
		SYE3: Husain Haqqani=> Haqqani Husain (Husain Haqqani)	**SYF3:** lundi 30 mai 2011=> Thứ 2 Tháng 5 30 2011 (lundi mai 30 2011)
	TE	**TEE1:** the Irish Republic => Cộng hòa Ailen (Cộng hòa Ireland)	**TEF1:** Turin=> Torino (Turin)
		TEE2: $2bn => 2 tỷ USD (2 tỷ đô la Mỹ)	**TEF2:** 294 milliards de dollars=> 294 tỷ USD (294 tỷ đô-la Mỹ)
		TEE3: Syrians=> Syrians (người Syria)	**TEF3:** Hu Chunhua=> Hu thuan (Hồ Xuân Hoa)
		TEE4: Turkish=> tiếng Turk (tiếng Thổ Nhĩ Kỳ)	**TEF4:** parti Baas=> Đảng Baath (Đảng Baas)

Table 1: Examples of PNMT errors from EVC and FVC corpora of PNs

CHAPTER 4: CORPUS ANNOTATION

4.1 ANNOTATION FRAMEWORK

The annotation of proper names from parallel corpora relies on principles of corpus annotation. According to Leech G. (2004), corpus annotation is the practice of adding interpretative linguistic information to a corpus. A very common type of annotation is the addition of tags or labels indicating a word or a group of words belonging to a certain class with similar features. Our framework for multi-language corpus annotation, consists of five major steps:

1. Collecting the original texts in English and French to build the corpora;

2. Linguistic analysis of PNs: this step concerns the identification of some specific PNs, namely:

   a. Names: e.g. *Bill Gates, South Korea, EU*, etc.;

   b. Proper names including prepositions and coordination, e.g. *University of California, Hotel de Sofitel* (organization names), *Good Night and Good Luck, Gone with the Wind* (titles of books, films, songs, programs, etc.);

   c. Some noun phrases following a PN, e.g. *CBS's "60 Minutes" program, Britain's first black Conservative peer*, etc.

3. Creation of multi-language tags for PN annotation: this step depends on the linguistic analysis of each language.

4. Implementation of corpus annotation using manual, automatic and semi-automatic annotation methods;

5. Comparison and evaluation of those annotation methods

4.2 ANNOTATION METHODS

In this study, we have implemented and compared the three methods of corpus annotation including manual, automatic and semi-automatic annotation methods.

4.2.1 Manual annotation

Since our study concentrates on issues of PNs in English and French languages, we set up the tags serving to recognize and, if possible, to correct these sorts of named entities. In reference to the classification of PNs (see Figure 1) and the categories of PNMT errors of EVC and FVC (see Figure 2 and Figure 3), we have created 16 tags, which are listed together with a simple definition and exemplification in Table 2.

No	Annotation tags	Meaning
1	ACRO	Acronyms or abbreviations
2	GEO1	English/French geographic names to be translated
3	GEO2	English/French geographic names not to be translated
4	Human Title	e.g., Dr., Mrs., Miss, M., Mme, Mlle, etc.
5	MISC	Miscellaneous names
6	NE	Named Entity boundary
7	NP	English noun phrase in possessive structures with PNs
8	ORG	English/French organization names
9	ORG1	French organization names requiring the omission of "de + determiner" groups preceding them
10	PER	Personal names (including full names and human titles or professional titles), e.g., President Barack Obama
11	PERS	Personal names (containing three cases: either first names or last names or both of them without human titles/ professional titles), e.g. Obama, Barack Obama.
12	PROD	Product names
13	Professional Title	e.g. Pilot/Pilote, Secretary/Secrétaire, Doctor/Docteur, Professor/Professeur, etc.
14	SCI	Scientific names
15	TIME	Time designations
16	TWP	English/French titles of works and publications

Table 2: Manual annotation tags for English and French corpora of PNs

4.2.2 Automatic and semi-automatic annotation

To automatically and semi-automatically annotate our corpora, we use GATE software version 7.1 by Cunningham et al. at the University of Sheffield, U.K. (2012), which allows users to annotate automatically entities such as *Person*, *Location*, *Organization*, *Date*, *Percents*, *Money* and *Address*. Due to GATE's open architecture, NER modules are easily customizable and flexible because they consist of manually created sets of pattern-matching rules that can be extended to add new

entity types or modified for new domains (Bontcheva et al. 2002a). We may edit or modify and even add new annotation tags to the list of ANNIE default annotation tags in GATE's visual environment.

GATE can be seen as the most useful and efficient annotation tool to carry out the semi-automatic annotation for multi-language corpora of PNs. In fact, GATE has been used for annotating documents not only in English but also in other languages such as French, German, Italian, Arabic, Chinese, Romanian, Hindi and Cebuano. Moreover, we add some annotation tags including Science, Product, TitleOfWorks, NP, and Misc to the list of default annotations to create the annotation sets for semi-automatic annotation of our English and French corpora of PNs (see Table 3).

No	**Type of annotation tags**		**Meaning**
	GATE annotation set	Added annotation tags	
1	<Address>		Website links
2	<Date>		Time designations
3	<FirstPerson>		First names
4	<JobTitle>		Professional titles (e.g. Pilot/Pilote, Secretary/Secrétaire, Doctor/Docteur,Professor/Professeur, etc.)
5	<Location>		Geographic names
6		<Misc>	Miscellaneous names
7	<Money>		Monetary units with digits
8		<NP>	English noun phrases in possessive structures with PNs
9	<Organization>		Organization names
10	<Percent>		Percentage
11	<Person>		Personal names (including full name and human titles or professional titles)
12		<Product>	Product names
13		<Science>	Scientific names
14	<Title>		e.g., Dr., Mrs., Miss, M., Mme, Mlle, etc.
15		<TitleOfWorks>	Titles of works and publications

Table 3: Annotation tags for semi-automatic annotation of English and French corpora

In summary, this section introduces the benefits of corpus annotation and concentrate on three annotation methods that we use for annotating our corpora of PNs. The manual annotation method takes a huge amount of time and labour, but

provides annotation results with high rate of accuracy. The automatic annotation method implemented with the GATE tool is less time consuming, but provides the results with less accuracy. The semi-automatic annotation method combines the automatic annotation with the manual correction of its results and reduces about 90% of manual annotation time, and provides such very good results in terms of accuracy.

4.3 CORPUS ANNOTATION RESULTS

In this section, we analyze the automatic annotation errors of English and French corpora of PNs given by GATE system and compare those automatic results with manual annotation results in order to calculate how many annotation errors we have to corrected and adjusted in the semi-automatic annotation. Table 3 shows the number of automatic annotations attributed by GATE system.

Name of corpus		**English corpus**		**French corpus**	
Resources		BBC News		Le Monde	
Size	Words	101,289		109,584	
	Characters	575,166		781,347	
	Paragraphs	1,512		1,501	
Annotation Results		Type of tags	Total of annotations	Type of tags	Total of annotations
		Location	2,743	Location	2,369
		Title	1,245	Title	581
		JobTitle	1,219	JobTitle	603
		Organization	1,443	Organization	449
		Person	2,290	Person	1,530
		FirstPerson	1,782	FirstPerson	1,483
		Date	1,697	Date	1,869
Total of annotated PNs			**12,419**		**8,884**

Table 4: Annotation results of English and French corpora given by automatic annotation

Among these automatic annotations there exists a variety of errors. Table 5 shows the total of automatic annotations for each tag type and indicates the number and percentage of automatic annotation errors per type.

Type of tags	English corpus			French corpus		
	Number of automatic annota-tions	Number of automatic annotation errors	Percen-tage	Number of automatic annota-tions	Number of automatic annotation errors	Percen-tage
Date	1,697	815	48%	1,869	1,042	55.75%
FirstPerson	1,782	333	18.68%	1,483	784	52.86%
JobTitle	1,219	354	28.30%	603	144	23.88%
Location	2,743	543	19.79%	2,369	809	31.14%
Organization	1,443	424	29.38%	449	1,258	63.67%
Person	2,290	635	27.72%	1,530	875	57,18%
Title	1,245	635	51.0%	581	342	58.86%
Total number	**12,419**	**3,739**	**30.10%**	**8,884**	**5,254**	**49.04%**

Table 5: Statistics of automatic annotation errors in English and French corpora of PNs

4.4 COMPARISON OF ANNOTATION RESULTS USING DIFFERENT METHODS

To compare the annotation results given by each pair of annotation methods (manual vs. automatic, automatic vs. semi-automatic and manual vs. semi-automatic), we describe their corresponding annotation tags. The most important differences in the compared results concern the time of annotation and the total number of annotations for each corresponding annotation type between the annotation method pairs. The manual annotation provides the high rate of accuracy and precision, but it requires a huge time and work.

It can be said that the semi-automatic annotation offers the same high rate of precision and recall as the manual annotation. The advantage of the semi-automatic annotation lies in the amount of time needed to perform the annotation. Therefore, it is considered the best annotation method since it combines the advantage of the manful annotation in terms of quality of annotated data and the advantage of automatic annotation in terms of time needed for annotation process. The semi-annotated corpora constitute the input for the pre-processing program aiming at amelioration of PNMT.

CHAPTER 5: IMPROVING MT BY PRE-PROCESSING

5.1 PRE-PROCESSING

5.1.1 Definition and general pre-processing tasks

In the NLP domain, *pre-processing* is a programming task that processes input data to produce better outputs. In particular, in MT aspect, *pre-processing* is notably defined as a human-aided machine translation (HAMT) process, which supports MT systems to produce high-qualified outputs.

There exist several kinds of pre-processing tasks, which are significantly effective for improving results of MT systems. Coping with a variety of texts, different MT systems yielding diverse outputs require special pre-processing tasks to generate the best and the most appropriate products. Hutchins and Somers (1992:151) state that typically pre-processing involves checking source texts for foreseeable problems for the systems and trying to eradicate them. Furthermore, it can include identification of names (proper nouns), marking of grammatical categories of homographs, indication of embedded clauses, bracketing of coordinate structures, flagging or substitution of unknown words, etc. A large number of pre-processing tasks can be carried out both manually and/or automatically for many MT systems: changing the word order in phrases and clauses; reordering the adverbs or adverbial phrases of time, place, frequency, manner, and so on; transferring the passive structures into active ones; correcting the punctuation to avoid graphic errors (e.g. upper-case and lower case errors); deleting words to avoid translation errors caused by redundant words; and so forth.

5.1.2 Specific pre-processing tasks for English-Vietnamese and French-Vietnamese MT

There exist various PNMT errors, among which it should be noticed the following types of errors: i/ errors derived from wrong translation of English possessive structure (syntactic errors); ii/ errors caused by wrong translation of English and French PNs, which should not be translated; iii/ errors resulting from wrong translation of superfluous words in French phrases with PNs. Since those errors are statistically significant, easy to recognize and correct with pre-processing

solutions, we focus on three distinctive pre-processing tasks to reduce or limit them as follows:

a/ restructuring English possessive phrases with PNs;

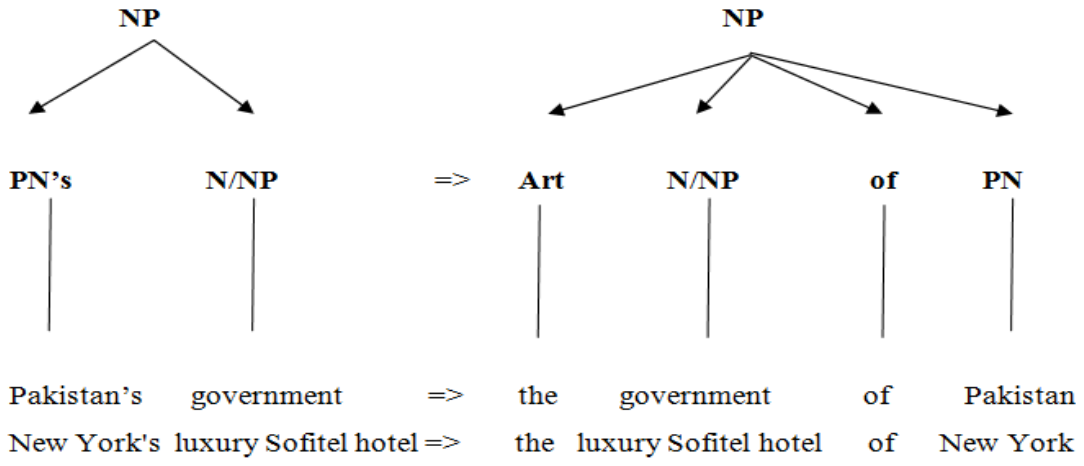


Figure 4: Restructuring possessive noun phrases with PNs in English

b/ omitting French groups of "de + determiner" preceding some geographic names and organization names;

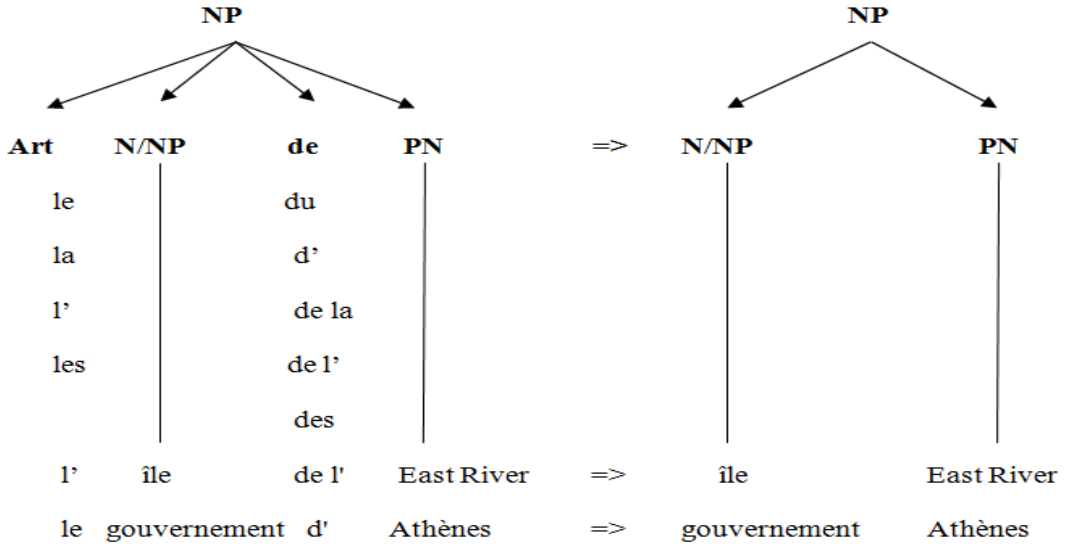


Figure 5: Omission of French groups "de + determiner" in NPs with PNs

c/ giving Do-Not-Translate (DNT) marks to PNs of personal names, certain geographic names, titles of works, and names of products in both English and French languages.

5.2 DESCRIPTION OF THE PRE-PROCESSING PROGRAM FOR PNMT ERROR REDUCTION

5.2.1 Principal tasks

This pre-processing program is designed to execute the following tasks:

1. To identify all kinds of PNs and extract them from our English and French corpora and other annotated corpora of texts with PNs;
2. To classify PNs into different categories based on our proposed NLP classification of PNs, so that we can create various lists of PNs such as list of personal names, list of geographic names, list of organization names, etc.;
3. To count the total number of PNs in general and the total number of each PN category from an annotated corpus of PNs;
4. To give Do-Not-Translate marks to all the PNs, which should not be translated;
5. To search for and change the English possessive structures with PNs into the simple ones;
6. To search for and delete the French groups of "de + determiner" preceding certain geographic names and some organization names in French annotated corpus in order to simplify the French structures with PNs.

5.2.3 The pre-processing program's interface and demos

The pre-processing program includes three main functions: *File, Options* and *Help*. The ***File*** category includes the following sub-options:

1/ *New task:* This allows users to create a new task for the pre-processing program.

2/ *Open input document*: This option permits the users to open a file (.txt, .doc, .rtf file). The file put into the program can have a size of about 80,000 words or 800,000 characters.

3/ *Save input text*: It is used to save the input text or file

4/ *Save final results*: This is used to save the final results after we put the texts processed with this program into different MT systems.

5/*Exit*: This allows users to leave the program.

Options category includes sub-options: *Define Tags* and *Reset all tags* (see Figure 6). *Define tags* open the tag manager, which provides the list of all the *Tag names* with their *Description*, the *DNT* option of marking PNs which should not be translated, *Check French determiner* to verify the French PNs that go after French groups of "de + determiner". The *Define tags* option permits us to *add/edit/remove* the tags to/from the list of tag names. We can define the new tags in the *Tag name* option and associate them with various colours in the *Color* box.

The *DNT* (Do-not-translate marks for PNs*)* option provides either the selection of *Yes* for those PNs, which should not be translated, or the selection of *No* for those PNs, which should be translated into the target language.

Check French Determiner option allows selecting one from two sub-options: *Yes* or *No*. For the types of PNs including geographic names, organization names and date, we choose *Yes*; that is to say, the program will delete the French groups "de + determiner" preceding those PNs. In contrast, when we choose *No*, the French groups "de + determiner" will not be omitted.

If we click on *Update* box, all the adjustments including addition, edition and removal of tag names will be updated. On the contrary, we click on *Undo* box when we do not want any adjustments. The *Reset all tags* option is used to reset all the define tags that had been added, edited or removed to return to the default setting of the program. Finally, we click on OK box to implement all the adjustments to our program. If we click on Cancel box, the adjustments will not be accepted.

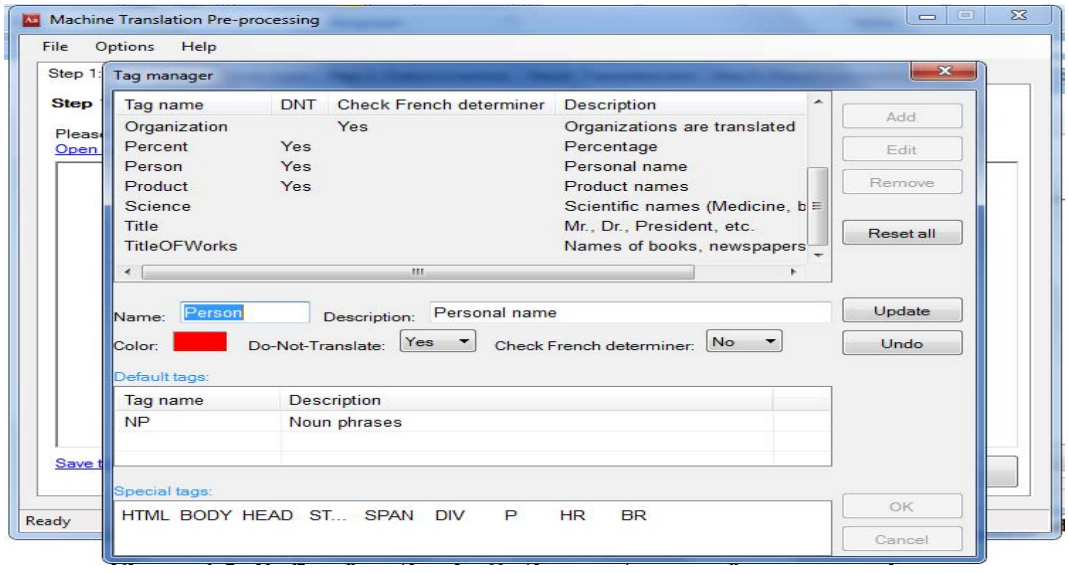


Figure 6: Define tags in Options category of the pre-processing program

The pre-processing program includes five major steps.

Step 1: Input the document into the program (see Figure 7). There are two ways to input the document into the pre-processing program: *1/* We can copy and paste a document file or a text to the blank box used for input data; *2/* We can open a document file or a text and link it to the pre-processing program. In order to store the document for the future use, we select the function of *Saving Input* text and create a link to the database directory. In the first step, depending on the language of input texts, we select English or French in the *Input Language* box.

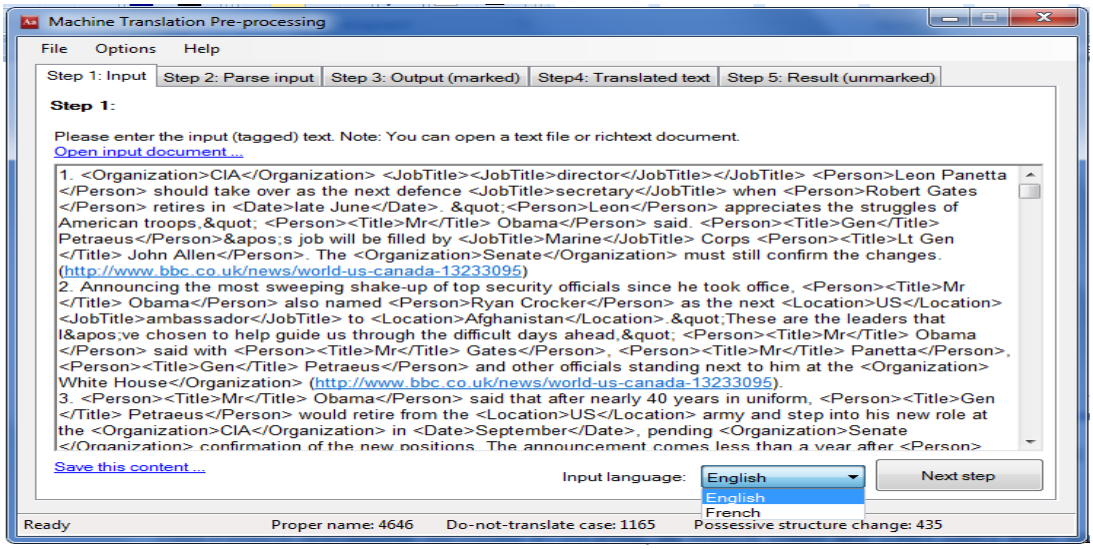


Figure 7: Input the annotated texts into the pre-processing program

Step 2: Processing the document.

Step 3: Presenting the pre-processed output (see Figure 8).

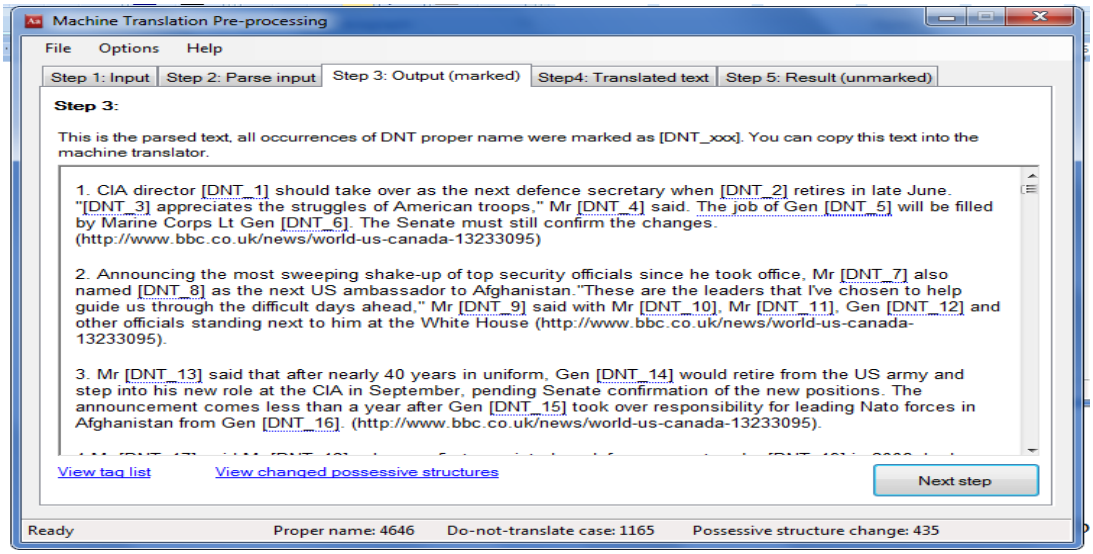


Figure 8: A pre-processed English document in the Step 3

In the Step 3, we can see the result of the processing done in the Step 2 shown in the tool bar (e.g. there are 4,646 PNs including 1,165 DNT PNs and 435 cases of possessive structures). Moreover, we copy all those processed texts and put them into different MT systems. Therefore, we obtain different translations from various MT systems for one document processed by our pre-processing program.

Step 4: Recovering the processed document translated by MT systems.

This step serves to recover the pre-processed document translated by different MT systems (indicated in the Step 3). We copy each document translated by each MT system and put it back to the pre-processing program.

Step 5: Achieving the final result

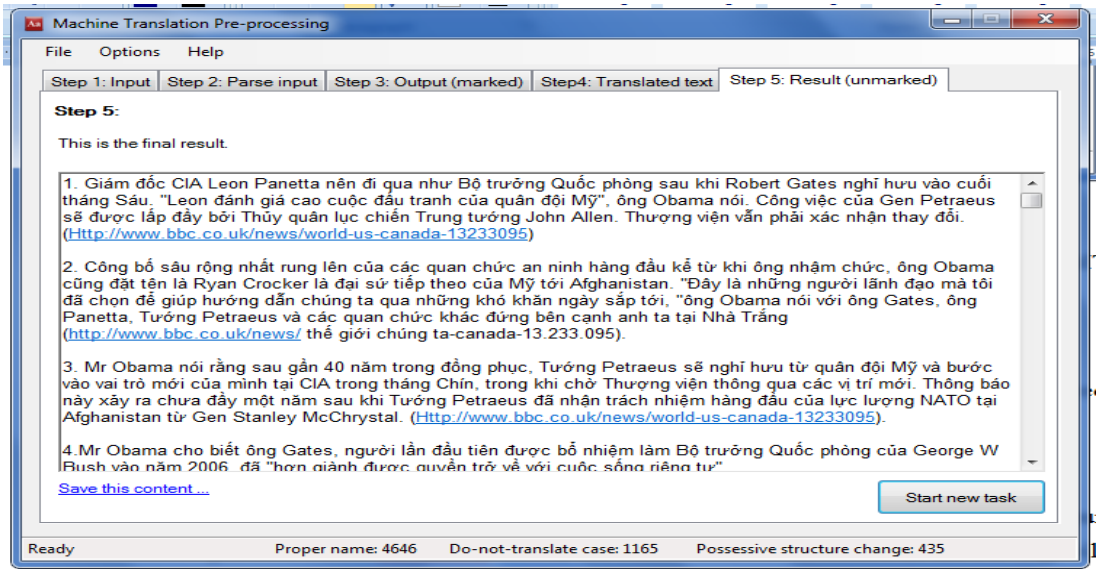


Figure 9: The final result in the Step 5

5.2.4 Advantages and limitations of the pre-processing program

This program has been built to perform the specific pre-processing solutions on reducing the PNMT errors made by English-Vietnamese and French-Vietnamese MT systems, which is evaluated with its benefits and limitations in the NLP's application tasks.

It permits us to:

i/ Correct some PNMT errors and improve the quality of MT systems

ii/ Reduce of transcription errors made by MT systems

iii/ Simplify French structures with PNs

iv/ Create various lists of PNs in different categories

v/ Establish two bilingual lexicons of PNs

vi/ Provide the statistics of PNs in different categories

This program can be applicable to both English and French. The greatest advantage concerns the saving of the pre-processing time in comparison with the manual solution. It also has a user-friendly interface, which can be easy to run.

It also possesses some limitations.

The first limitation concerns the preparation of a database for the program. The pre-processing program requires the input document to be annotated, which takes a great deal of time and labour. Although the annotation of an English or French document can be implemented semi-automatically, it requires the annotators much time and work to adjust or amend and correct the annotation errors made by automatic annotation system.

Second, this program cannot resolve all the problems relating to all types of PNMT errors made by different MT systems, for example, non-translation errors, graphic errors, some sub-types of lexical errors, syntactic and transcription errors.

Third, the machine translation systems are not directly integrated in the pre-processing program; one has to copy and paste the documents into different systems, then put them back to the pre-processing program to obtain the final result. It would be ideal to connect directly the pre-processing program to MT systems, so that there is no other manipulation needed in order to get the final translation.

5.3 TESTING RESULTS OF PRE-PROCESSING

In this section, we present the statistics of PNMT errors from the two English-Vietnamese and French-Vietnamese parallel corpora , which we carry out the testing before and after using our pre-processing system. By testing those corpora on different MT engines, we have achieved distinguished results showing various rates of PNMT errors. The results also offer the proofs and illustrations of certain strengths and weaknesses of each MT system.

5.3.1 Testing results of English-Vietnamese corpus of proper names before and after pre-processing

Based on the classification of English-Vietnamese PNMT errors (see Figure 1), Table 6 details the number and the types of errors. The results presented in this

table have been offered by the four different English-Vietnamese MT systems in the non-preprocessing corpus of PNs (i.e. before we use the pre-processing program).

Type of PNMT Errors			Errors by Vietgle		Errors by Google		Errors by Bing		Errors by EVtran	
			(1)	%	(2)	%	(3)	%	(4)	%
Non-Translation		NTE1	319	9.89	418	15.85	391	10.92	193	11.18
		NTE2	116	3.59	20	0.75	79	2.20	2	0.11
		NTE3	42	1.30	25	0.94	180	5.02	10	0.57
		NTE4	39	1.20	17	0.64	38	1.06	7	0.40
		NTE5	4	0.12		0.00	4	0.11	7	0.40
		NTE6	3	0.09	15	0.56	15	0.41		0.00
		Total	**523**	**16.22**	**495**	**18.77**	**707**	**19.75**	**219**	**12.68**
Wrong translation	*GEE*	GEE1	162	5.02	140	5.31	197	5.50	72	4.17
		GEE2	218	6.76	125	4.74	264	7.37	44	2.54
		GEE3	211	6.54	94	3.56	265	7.40	51	2.95
		GEE4		0.00	1	0.03	4	0.11		0.00
		GEE5	15	0.46	8	0.30	25	0.69	7	0.40
		GEE6	5	0.15	6	0.22	12	0.33	4	0.23
		GEE7	8	0.24	17	0.64	28	0.78	8	0.46
		GEE8	6	0.18	7	0.26	23	0.64	2	0.11
		GEE9	19	0.58	40	1.51	27	0.75	24	1.39
		GEE10	96	2.97	111	4.21	24	0.67	41	2.37
		GEE11	120	3.72	60	2.22	6	0.16	31	1.79
		Total	**860**	**26.67**	**609**	**23.10**	**875**	**24.44**	**284**	**16.45**
	LEE	LEE1	805	24.95	402	15.23	556	15.53	395	22.86
		LEE2	122	3.78	271	10.27	229	6.39	62	3.58
		LEE3	65	2.01	86	6.23	70	1.95	22	1.27
		LEE4	100	3.10	189	7.17	440	12.29	31	1.79
		Total	**1,092**	**33.87**	**948**	**35.96**	**1,295**	**36.18**	**510**	**29.54**
	SYE	SYE1	248	7.69	264	10.01	364	10.17	287	16.62
		SYE2	290	8.99	117	4.43	147	4.10	247	14.31
		SYE3	105	3.25	81	3.07	85	2.37	102	5.90
		Total	**643**	**19.94**	**462**	**17.52**	**596**	**16.65**	**636**	**36.84**
	TEE	TEE1	55	1.70	79	2.99	74	2.06	54	3.12
		TEE2	11	0.34	19	0.72	20	0.55	10	0.57
		TEE3	37	1.14	21	0.79	10	0.27	13	0.75
		TEE4	3	0.09	3	0.11	2	0.05		0.00
		Total	**106**	**3.28**	**122**	**4.62**	**106**	**2.96**	**77**	**4.46**
Total of Errors/ Total of texts translated			**3224/ 1500**	**21.49**	**2636/ 1500**	**17.57**	**3579/ 1413**	**25.32**	**1726/ 609**	**28.34**

Table 6: Results of English-Vietnamese PNMT errors before pre-processing

Table 7 sums up the statistics of errors corrected by the four E-V MT systems after we use the pre-processing program. The use of the pre-processing program has reduced significantly certain PNMT errors including some sub-types of lexical and syntactic errors. In general, the proportion of corrected errors is relatively high. Furthermore, due to different MT systems built with distinguished structures,

the number of PNMT errors corrected by those four systems is dissimilar to one another. For instance, EVTran has corrected the biggest number of PNMT errors, Google achieves the lower rate of errors corrected than Vietgle and EVTran. Among the four MT systems, Bing has corrected the lowest number of PNMT errors.

Type of PNMT Errors		Errors by Vietgle		Errors by Google		Errors by Bing		Errors by EVTran	
		(1)	%	(2)	%	(3)	%	(4)	%
Lexical errors (LEE)	LEEs corrected	756/ 1,092	23.44	354/ 948	13.42	413/ 1,295	11.53	367/ 510	21.26
Syntactic errors (SYE)	SYE1 corrected	244/ 248	7.63	261/ 264	9.90	361/ 364	10.08	284/ 287	16.42
	SYE3 corrected	105/ 349		81/ 81		85/ 85		102/ 102	
	Total of SYE corrected	349 /643	10.82	339/ 462	12.86	438/ 596	12.23	386/ 636	22.19
Total of LE +SY errors corrected/ Total of PNMT errors		**1,105 / 3,224**	**34.27**	**696/ 2,636**	**26.40**	**859/ 3,570**	**24.06**	**753/ 1,726**	**43.62**

Table 7: Statistics of the E-V PNMT errors corrected by the pre-processing program

5.3.2 Testing results of FVC of proper names before and after pre-processing

In this section, we present the statistics of PNMT errors made by two current French-Vietnamese MT engines before and after using the pre-processing program. Based on the *Classification of French-Vietnamese PNMT errors* we conduct the statistical analysis of the F-V PNMT errors shown in Table 8.

Type of Errors			Errors by Google		Errors by Bing	
			(1)	% (1')	(2)	% (2')
Non-Translation		NTF1	377	11.46	339	10.00
		NTF2	45	1.36	36	1.06
		NTF3	12	0.36	2	0.05
		NTF4	15	0.45	15	0.44
		NTF5	7	0.21	2	0.05
		NTF6	4	0.12	4	0.11
		Total	**460**	**13.99**	**398**	**11.75**
Wrong translation	***Graphic Errors***	GEF1	159	4.83	282	8.32
		GEF2	148	4.50	330	9.74
		GEF3	269	8.18	560	16.53
		GEF4	**14**	0.42	31	0.91
		GEF5	15	0.45	22	0.64
		GEF6	0	0.00	2	0.05
		GEF7	121	3.68	14	0.41
		GEF8	205	6.23	27	0.79
		Total	**931**	**28.32**	**1,268**	**37.43**

	Lexical Errors	LEF1	600	18.25	448	13.22
		LEF2	52	1.58	368	10.86
		LEF3	486	14.78	184	5.43
		LEF4	239	7.27	301	8.88
		Total	**1,377**	**41.89**	**1,301**	**38.41**
	Syntactic errors	SYF1	257	7.88	234	6.90
		SYF2	49	1.49	20	0.59
		SYF3	107	3.25	27	0.79
		Total	**413**	**12.56**	**281**	**8.29**
	Transcri p-tion errors	TEF1	67	2.03	88	2.59
		TEF2	29	0.88	37	1.09
		TEF3	5	0.15	10	0.29
		TEF4	5	0.15	4	0.11
		Total	**106**	**3.22**	**139**	**4.10**
Total of Errors/ Total of texts translated			**3,287/ 1,500**	**29.72%**	**3,387/ 1,500**	**30.71%**

Table 8: Statistics of French-Vietnamese PNMT errors before pre-processing

According to our statistics, lexical error rate is the highest one in comparison with other types of errors. Both Google and Bing have the high rate of LEFs. The rate of graphic errors (GEFs) is lower than that of LEFs, but much higher than the rate of SYFs and TEFs. Bing has the higher rate of GEFs and TEFs than Google does; nevertheless, Google has made much higher rate of SYFs than Bing has. Table 9 shows the statistics of French-Vietnamese PNMT errors, which have been corrected after we use the pre-processing program.

Type of Errors		**Errors by Google**		**Errors by Bing**	
		(1)	**(1')**	**(2)**	**(21')**
Lexical errors	LEF1 corrected	512	15.57	398	11.74
	LEF3 corrected	443	13.47	167	4.93
	Total (LEF1+LEF3) corrected	955/1,377	29.05	565/1,301	16.68
Syntactic errors	SYF2 corrected	49/413	1.49	20/281	0.59
Total (LEF+SYF) corrected		**1,004/3,287**	**30.54**	**585/3,387**	**17.27**

Table 9: Statistics of F-V PNMT errors corrected after pre-processing

5.4 OVERALL COMPARISON AND EVALUATION OF ENGLISH-VIETNAMESE AND FRENCH-VIETNAMESE MT SYSTEMS

5.4.1 Issues of English-Vietnamese MT systems

In this section, we present strengths and weaknesses of each MT system and make a comparison of the four E-V MT systems based on the following criteria:1)

the number of PNMT errors; 2) the translation speed and 3) the maximum size of input corpus. Table 10 shows the comparison of the translation speed and the maximum size of documents translated by the four E-V MT systems.

MT systems	Size of a corpus		Translation time	
	Maximum size of a corpus		Average size of a corpus (6,500 words/100 texts)	
	Maximum size of the text	Translation duration of all parts	Subdivisions of the corpus into parts translated	Translation duration of all the parts in corpus
Vietgle	1,200 words (~7,500 characters)	240 seconds	5	1,200 seconds
Google	12,000 words (~72,000 characters)	2 seconds	1	1 second
Bing	700 words (~4,200 characters)	2 seconds	10	20 seconds
EVTran	40 words (~200 characters)	600 seconds	165	99,000 seconds

Table 10: Comparison of translation speed and the size of a corpus translated by four E-V MT systems

According to our statistics of PNMT errors, Google Translate engine performs the best among the four English-Vietnamese MT systems in translating texts with PNs from English into Vietnamese. In fact, Google has made the lowest number of PNMT errors (17.57%), while the ratio for Vietgle is 21.49 % and for Bing 25.23%. EVTran is the worst working system with the highest rate of errors (28.26%). The second English-Vietnamese MT system compared in this section is Vietgle, which makes the lowest number of NTE1, i.e., it is good at translating acronyms, and abbreviations. In general, Vietgle makes fewer NTEs than Google and Bing.

Besides, concerning the syntactic errors, compared with Google, Bing and EVTran, Vietgle makes the lowest number of SYE1. However, it makes the highest number of graphic errors and lexical errors caused by incorrect words. Bing Translator is the third E-V MT system, which has the lowest rate of syntactic errors among the four E-V MT engines. Nonetheless, Bing makes the highest number of non-translation errors and graphic errors. In particular, it makes a large number of non-translation errors for human titles such as Mr. Ms. Mrs. Dr., etc.(e.g. *Mr Fayyad, Mr Balls, Ms Giffords, Dr Hyacinth Orikara*) and geographic names. The fourth English-Vietnamese MT system is EVTran, which makes the highest number of typical PNMT errors arising from the translation of the PNs containing meanings

in a dictionary. EVTran has the highest rate of syntactic errors and transcription errors among the four E-V MT systems. Nevertheless, due to EVTran's typical feature of translating all the words having meanings, which include personal names, geographic names and product names, EVTran makes the lowest number of non-translation errors. In fact, EVTran translate many acronyms and abbreviations while other systems do not.

Based on the evaluation in this section, we can conclude that, nowadays, the best English-Vietnamese MT system are Google Translate and Bing Translator due to the quality of their outputs, the number of language options, the high rate of input texts translated and the lowest rate of some error types. Vietgle is a good choice for those who would like to translate the document related to specific topics. In fact, it has been verified that choosing the right option for the input document can ameliorate the quality of translation. Finally, since EVTran translates a few input texts (609/1,500) at a very slow speed and offers low quality translations, it may be considered as the least performing MT system.

5.4.2 Issues of French-Vietnamese MT systems

We assess the quality of the two French-Vietnamese MT engines (Google Translate and Bing Translator) by comparing the quality of their output, i.e. the number of PNMT errors made by each MT engine, the translation speed and the maximum size of input text they can process.

Actually, there are not many differences between Google and Bing concerning the rates of PNMT errors.The rates of each sub-type of NTF made by Google and Bing are also similar to each other. It means that Google and Bing face the same difficulties in translating PNs from French into Vietnamese. Moreover, both Google and Bing make the similar number of lexical errors and transcription errors. Nonetheless, there is a big difference between the sub-types of graphic errors made by these two F-V MT systems. For instance, the graphic errors caused by uncapitalized geographic names, professional names, organization names and titles of works made by Bing are twice as high as those made by Google; as to syntactic errors, Google performs worse than Bing does.

Concerning the translation speed and capacities of input processing, both Google and Bing offer the high speed of translation. However, Google system provides the outputs not only at the high speed, but also for larger amount of input

texts. Even if Bing can translate a text at the same speed as Google, the size of the input text is limited. Table 11 shows the comparison of the translation speed and size of texts translated by the two French-Vietnamese MT systems, and Table 12 summarizes all the criteria of evaluation to compare the two French-Vietnamese MT systems.

MT system	Translation speed and size of an input corpus			
	An average size of a corpus (including 7,000 words in 100 texts)		Maximum size of a corpus	
	Divided parts of the corpus	Duration	Maximum size of a corpus	Duration
Google	1	1 second	15,000 words (~ 100, 000 characters without spaces)	120 seconds
Bing	13	13 seconds	600 words (~4,200 characters without spaces)	2 seconds

Table 11: Comparison of the translation speed and size of texts of the two F-V MT systems

No	Criteria of evaluation	Google	Bing
1	Number of graphic errors and transcription errors (*see Table 36*)	lower	higher
2	Number of non-translation errors, lexical errors and syntactic errors (*see Table 36*)	higher	lower
3	Number of language options	58 languages	37 languages
4	Size of an input document	bigger	smaller
5	Speed of processing the input document	faster	slower
6	Capacity of translating a corpus containing the same number of texts	1,500 texts/ 1,500 texts	1,500 texts/ 1,500 texts

Table 12: Comparison the two F-V MT systems

In summary, we have evaluated various English-to-Vietnamese machine translation systems and two French-to-Vietnamese MT engines with respect to proper name translation, and indicate the different rates of PNMT errors made by those MT engines. Based on result analysis and discussion about PNMT errors caused by each MT system, we explore the strengths and weaknesses of different engines relating to the number of PNMT errors, translation speeds, capacities of processing the input texts, and the number of language options as well. We also recommend that pre-processing is one of the best methods to reduce the PNMT errors for not only English-Vietnamese MT systems, but also French-Vietnamese engines.

CHAPTER 6: CONCLUSION

6.1 CONTRIBUTIONS OF THE THESIS

We have conducted this thesis in order to find a solution to reduce the PNMT errors for the systems translating from English and French into Vietnamese. To meet this objective, we divided it into several sub-tasks, each one making a significant contribution to the final goal of PN machine translation improvement. We enumerate all of them in the chronological order of our work.

Contribution 1: **Building two parallel corpora of PNs for two pairs of languages, English-Vietnamese and French Vietnamese**

Parallel corpora are necessary resources in many NLP applications such as MT, CLIR, IE, QA, WSD, bilingual terminology extraction, and so on, but some restricted language pairs, such as English-Vietnamese and French-Vietnamese, do not dispose of valuable parallel corpora. In consequence, the building of these real databases is of great necessity in developing the NLP applications and MT systems in particular. This study has provided the first English-Vietnamese and French-Vietnamese parallel corpora annotated for PN detection and translation; other NLP researchers can reuse them, simultaneously they can constitute helpful resources and good databases for MT users to compare the PNMT results offered by different MT systems.

Contribution 2: **Classification of PNs based on the viewpoints of both linguistics and NLP domains**.

In this study, we present the PN classifications of many linguists, and discuss their strengths and weaknesses; for example, some classification are too complex with too many types and sub-types of PNs, others are too simple to indicate the explicit features of each type of PNs. The NLP classifications of PNs are too general to include all the typical kinds of PNs. To meet both linguistics and NLP criteria of PNs, we have proposed our classification of PNs containing seven major types of PNs. This classification of PNs has set up the base for the analysis and classification of PNMT errors made by current E-V MT and F-V systems.

Contribution 3: **An analysis and a classification of PNMT errors made by current English-to-Vietnamese and French-to-Vietnamese MT engines**

On the basis of the two parallel corpora of PNs in English-Vietnamese and French- Vietnamese, we have analyzed and classified the PNMT errors made by the four current E-V MT and two F-V MT engines. The classification of PNMT errors is divided into two major types: non-translation and wrong translation errors. The wrong translation errors are further divided into four main types based on linguistic criteria concerning graphics, lexis, syntax and transcription. Each category of PNMT errors is also illustrated by series of examples extracted from the two parallel corpora of PNs. The quantitative analysis of PNMT errors provides the information about the most frequent and important types of translation errors; furthermore, it also highlights the necessity to find proper solutions to deal with the PNMT errors.

Contribution 4: **Comparison of manual, automatic and semi-automatic annotation methods for English and French corpora of PNs**

Since our solution on the PNMT improvement demands annotated corpora, we compare three methods of corpus annotation in order to establish the most appropriate one for our purpose. First, we propose the manual annotation to create a point of reference for other annotation methods. It takes a lot of work and time, but provides the best results for the pre-processing task. Then, we evaluate the automatic annotation with GATE tool. It is the most rapid and less demanding method, but its results are not good enough in terms of precision and recall to be directly used by the pre-processing program. The optimal method, called semi-automatic, is a combination of the two previous ones, i.e. we first apply the automatic method, and then correct manually the results. In this way, we save consistently the work and time needed for annotation process, and simultaneously, we increase significantly the rate of precision and recall.

Contribution 5: **Creation of the pre-processing program for reducing the PNMT errors**

To facilitate the pre-processing process, we have created an automatic pre-processing program, which can be used effectively to correct some PNMT errors and consequently to improve MT systems' quality. The pre-processing program takes an annotated corpus as an input, and then carry out automatically some specific predefined tasks: changing English possessive structures with PNs, marking DNT for certain types of proper names, and deleting French groups "de + determiner" preceding some geographic and organization names. It will be interesting to expand

this program in two directions in the future: first, to prevent other types of PNMT errors, and second, to develop it for other language pairs.

Contribution 6: **Application of the pre-processing program to the MT quality enhancement**

The pre-processing program applied on an annotated corpus, reduces a large number of PNMT errors when translating the texts with PNs from English and French into Vietnamese. Specifically, it reduces on average 32.08 % of PNMT errors made by English-Vietnamese MT systems and 23.90 % of PNMT errors made by French-Vietnamese MT engines. Consequently, it enhances the MT systems' quality. Since the pre-processing is done automatically, it reduces significantly the time of pre-edition. Moreover, it reduces significantly the post-edition too, which is considered a time-consuming and expensive work. Therefore, the application of the pre-processing program to MT can reduce the time and work of pre-edition (in comparison to manual one) and post-edition for translators.

Contribution 7: **Establishment of two glossaries of PNs for English-Vietnamese and French-Vietnamese languages**

The other advantage of the pre-processing program lies in the possibility to create the glossaries of PNs for each language. These glossaries list a set of PNs in both source language and their translation in the target language, which can be used later on to better support the MT systems to give out the correct outputs.

Contribution 8: **A real evaluation of different MT systems used in Vietnam's NLP community**

Finally, with our study's results, we have proposed a real evaluation on the current MT systems for the two language pairs including English-Vietnamese and French-Vietnamse. In particular, the results of PNMT errors permit to think of and to find other new solutions on avoiding or reducing those errors.

6.2 LIMITATIONS OF THE THESIS AND FUTURE RESEARCH

Based on the thesis limitations analyzed in the previous section, we discuss some possible directions for future research.

1. Some kinds of PNMT errors are still incorrected.

Although the pre-processing program has been built to deal with the machine translation of texts with PNs from English and French into Vietnamese, it cannot resolve all the PNMT problems. In fact, there are still some types of PNMT errors, which cannot be corrected such as graphic errors, transcription errors, and plenty of sub-types of semantic errors (LEE2, LEE3, LEE4, and LEF2) and syntactic errors (SYE2, SYF1, and SYF3). We will continue our research to find solutions to correct these types of errors.

2. Pre-processing program can be only used for annotated texts.

One of the limitations of the thesis is the preparation of the input document, which requires the annotation before being put into pre-processing program. This not only takes much time, but also a lot of work. In order to obtain a precisely annotated document, we have to check and correct all the annotation errors, because the pre-processing program cannot run well if there are any errors appearing in the input texts.

3. Corpus annotation limitations

The semi-automatic annotation method is based on the GATE's system, which is highly appreciated for its results for the English language. Nonetheless, the results for the French language are not so good, since this system has just started applying for certain languages including French, Arabic, Chinese, etc. Furthermore, it requires adjusting a large number of annotations, which is time-consuming for annotators.

4. Pre-processing program cannot be linked directly to MT systems

It requires the users to copy the pre-processed texts to put into different MT systems to achieve the results (shown in Step 4, Figure 33), and should be copied again to the pre-processing program to retrieve the final results. It will be ideal if we can connect directly this pre-processing program to any current MT systems to translate the document better.

5. Limited languages used for the pre-processing program

Pre-processing program can be used for documents in only two languages, English and French. It cannot be applied to other languages such as Vietnamese, Chinese, Spanish, etc. Developing this program used for other languages is one of our research directions in the future.

SELECTIVE BIBLIOGRAPHY

Aone, C. and Maloney, J. (1997). Reuse of Proper Noun Recognition System in Commercial and Operational NLP Applications. In Proceedings of ACL'97 Workshop on From Research to Commercial Applications: Making NLP Technology Work in Practice.

Babych, B. and Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, Centre for Translation Studies, University of Leeds, UK., pp. 1-8.

Balabantaray, R., and Sahoo, D. (2013). An Experiment to Create Parallel Corpora for Odia. *International Journal of Computer Applications*, *67*(19), pp. 18-20.

Cunningham, H. et al. (2012). Developing Language Processing Components with GATE Version 7 (a User Guide), the University of Sheffield, Department of Computer Science 2001-2012. Retrieved on the 12th March 2013 from http://gate.ac.uk/userguide.

Dien, D. (2005). Building an annotated English-Vietnamese parallel corpus, *MKS: A Journal of Southeast Asian Linguistics and Languages*, Volume 35, pp. 21-36.

Dinh, D., Hoang, K. & Hovy, E. (2004). BTL: A Hybrid Model for English-Vietnamese Machine Translation

Hassan, A., Fahmy, H., & Hassan, H. (2007). Improving named entity translation by exploiting comparable and parallel corpora. *AMML07*

Héja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In Proceedings of the LREC2010 Conference, La Valletta, Malta.

Hermjako, U., Knight K. & Daumé III, H. (2008). Name Translation in Statistical Machine Translation: Learning When to Transliterate. In Proceedings of the 46th Annual Meeting on Association for Computational Linguistics, Columbus, Ohio, pp. 389-397.

Hidalgo, J. M. G., Garcia, F.C., and Sanz, E.P. (2005). Named Entity Recognition for Web Content Filtering. In Montoyo, A., Munoz, R., Métais, E. (Eds.) *Natural Language Processing and Information Systems*, *Proceedings of the 10th*

International Conference on Applications of Natural Language to Information Systems, *NLDB 2005*, Springer publisher, pp.286-297.

Krstev, C., Vitas, D., Maurel, D. & Tran, M. (2005). Multilingual Ontology of Proper Names. In Proceedings of 2nd Language & Technology Conference, Poznań, Poland, ed. Zygmunt Vetulani, pp. 116-119.

Leech, G. (2004). Adding Linguistic Annotation. In Wynne M. (Ed.) Developing Linguistic Corpora: a Guide to Good Practice. Retrieved on the 15th September 2012 from http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm.

Leroy, S. (2004). *Le Nom propre en français*. (Coll. *L'Essentiel français*). Paris, Gap: Ophrys.

Maurel, D., et al. (2008). Prolexbase, ProLMF version 1.2, Université François-Rabelais de Tours, retrieved at : http://www.cnrtl.fr/lexiques/prolex/

Maurel, D., and Bouchou-Markhoff, B. (2013). Prolmf: A Multilingual Dictionary of Proper Names and their Relations. In *LMF Lexical Markup Framework*, pp. 67-82

Munteanu, D.S. and Marcu, D. (2006). Improving Machine Translation Performance by Exploiting Non-parallel Corpora, Association for Computational Linguistics, 31(4), pp. 476-504.

Phan, T.T.T. and Thomas, I. (2012). English-Vietnamese Machine Translation of Proper Names: Error Analysis and Some Proposed Solutions", Proceedings of the 15th internaltional conference TSD 2012 (Text, Speech and Dialogue), September 3-7, 2012, Brno, Czech Republic, Springer Edition, pp.386-393.

Phan, T.T.T. and Thomas, I. (2013). Pre-processing as a Solution to Improve French-Vietnamese Named Entity Machine Translation. In Vetulani Z. and Uszkoveit H. (Eds.) Proceedings of Human Language Technologies as a Challenge for Computer Sciences and Linguistics, the 6th Language and Technology Conference (LTC 2013), December 7-9, 2013, Poznan, Poland, pp. 142-146.

Sang, T.K.E.F., and De Meulder, F.(2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Association for Computational Linguistics, Volume 4, pp. 142-147.

Somers, H. (2003). Machine Translation: Latest Developments. In Mitkov R. (Ed.) *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 511-529.

Vilar, D., Xu, X., D'Haro, L.F., Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genova, Italy.

Virga, P. & Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Lingual Information Retrieval. In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed- Language named Entity Recognition, pp. 57-64.

Wilks, Y. (2008). *Machine Translation:Its Scope and Limits*, Publisher: Springer,p.5