UNIVERSITÉ DE GRENOBLE

N° attribué par la bibliothèque
/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

DOCTEUR ÈS SCIENCES

délivré par l'UNIVERSITÉ DE GRENOBLE

Spécialité : "INFORMATIQUE"

Thèse préparée au laboratoire GETALP-LIG (CNRS-INPG-UJF) dans le cadre de

l'École Doctorale "Mathématiques, Sciences et Technologies de l'Information, l'informatique"

présentée et soutenue publiquement

par

Lingxiao WANG

Le 14 décembre 2015

Outils et environnements pour l'amélioration incrémentale, la post-édition contributive et l'évaluation continue de systèmes de TA. Application à la TA français-chinois.

JURY

Mme. Catherine BERRUT	Président
M. Xiaodong SHI (Xiamen)	Rapporteur
M. Eric Wehrli (Unige)	Rapporteur
M. François BROWN DE COLSTOUN (L&M)	Invité
M. Jean-Pierre CHEVALLET	Invité
M. Hong-Thai NGUYEN	Invité
M. Christian BOITET	Directeur de thèse
Mme Valérie BELLYNCK	Codirecteur de thèse

Résumé

Cette thèse, effectuée dans le cadre d'une bourse CIFRE, et prolongeant un des aspects du projet ANR Traouiero, aborde d'abord la production, l'extension et l'amélioration de corpus multilingues par traduction automatique (TA) et post-édition contributive (PE). Des améliorations fonctionnelles et techniques ont été apportées aux logiciels SECTRA_W et IMAG, et on a progressé vers une définition générique de la structure d'un corpus multilingue, multi-annoté et multimédia, pouvant contenir des documents classiques aussi bien que des *pseudo-documents* et des *méta-segments*. Cette partie a été validée par la création de bons corpus bilingues français-chinois, l'un d'eux résultant de la toute première application à la traduction littéraire.

Une seconde partie, initialement motivée par un besoin industriel, a consisté à construire des systèmes de TA de type Moses, spécialisés à des sous-langages, en français↔chinois, et à étudier la façon de les améliorer dans le cadre d'un usage en continu avec possibilité de PE. Dans le cadre d'un projet interne sur le site du LIG et d'un projet (TABE-FC) en coopération avec l'université de Xiamen, on a pu démontrer l'intérêt de l'apprentissage incrémental en TA statistique, sous certaines conditions, grâce à une expérience qui s'est étalée sur toute la thèse.

La troisième partie est consacrée à des contributions et mises à disposition de supports informatiques et de ressources. Les principales se placent dans le cadre du projet COST MUMIA de l'EU et résultent de l'exploitation de la collection CLEF-IP 2011 de 1,5 M de brevets partiellement multilingues. De grosses mémoires de traductions en ont été extraites (17,5 M segments), 3 systèmes de TA en ont été tirés, et un site Web de support à la RI multilingue sur les brevets a été construit. On décrit aussi la réalisation en cours de JIANDAN-EVAL, une plate-forme de construction, déploiement et évaluation de systèmes de TA.

Abstract

This thesis, conducted as part of a CIFRE grant, and extending one of the aspects of the ANR project Traouiero, first addresses the production, extension and improvement of multilingual corpora by machine translation (MT) and contributory post-editing (PE). Functional and technical improvements have been made to the SECTra and iMAG software produced in previous PhD theses (P.C. Huynh, H.T. Nguyen), and progress has ben made toward a generic definition of the structure of a multilingual, annotated and multi-media corpus that may contain usual documents as well as *pseudo-documents* (such as Web pages) and *meta-segments*. This part has been validated by the creation of good French-Chinese bilingual corpora, one of them resulting from the first application to literary translation (a Jules Verne novel).

A second part, initially motivated by an industrial need, has consisted in building MT systems of Moses type, specialized to sub-languages, for french↔chinese, and to study how to improve them in the context of a continuous use with the possibility of PE. As part of an internal project on the LIG website and of a project (TABE-FC) in cooperation with Xiamen University, it has been possible to demonstrate the value of incremental learning in statistical MT, under certain conditions, through an experiment that spread over the whole thesis.

The third part of the thesis is devoted to contributing and making available computer tools and resources. The main ones are related to the COST project MUMIA of the EU and result from the exploitation of the CLEF-2011 collection of 1.5 million partially multilingual patents. Large translation memories have been extracted from it (17.5 million segments), 3 MT systems have been produced (de-fr, en-fr, fr-de), and a website of support for multilingual IR on patents has been constructed. One also describes the on-going implementation of JianDan-eval, a platform for building, deploying and evaluating MT systems.

Remerciements

Résumé en chinois

Table des matières

Résumé 2
Abstract 2
Remerciements 3
Résumé en chinois 3
Table des matières 4
Table des figures 7
Table des tableaux 9
Glossaire et abréviations 11
Introduction 13
Partie A Production, extension et amélioration de corpus multilingues par TA et PE contributive 16
Résumé 16
Chapitre I Amélioration d'aspects fonctionnels et techniques de SECTra et du logiciel iMAG pour les passerelles d'accès multilingue 16
I.1 SITUATION ET ÉTAT DE L'ART AU DÉBUT DE LA THÈSE 16
I.1.1 Modélisation et exploitation de corpus de traductions : SECTra_w 16
I.1.2 Accès multilingue à des sites Web : le logiciel iMAG 26
I.1.3 Travaux comparables et idées directrices pour le futur 33
I.2 AMÉLIORATIONS DE SECTRA_W DANS LE CADRE DU PROJET TRAOUIERO 37
I.2.1 Extension de fonctions existantes 37
I.2.2 Aspects de génie logiciel 39
I.2.3 Travail de spécification 41
I.3 AMÉLIORATIONS DES IMAG DANS LE CADRE DU PROJET TRAOUIERO 42
I.3.1 Paramétrisation 42
I.3.2 Travail de spécification 43
Chapitre II Travail sur des aspects plus conceptuels et définition de nouvelles fonctionnalités 45
II.1 MODÉLISATION DE CORPUS DE TRADUCTIONS VARIÉS 45
II.1.1 Variété des corpus de traductions et esquisse d'une méthode de formalisation 45
II.1.2 Esquisse d'une méthode de modélisation des corpus de traductions 51
II.1.3 Validation au niveau des métadonnées 53
II.2 CONCEPTION DE NOUVELLES FONCTIONNALITÉS ET DÉVELOPPEMENTS EN COURS 62
II.2.1 Nouvelles fonctionnalités 62
II.2.2 Programmabilité 64
II.2.3 Modularité 65
Chapitre III Variété des iMAG et de leurs usages : de l'accès multilingue à la création de bons corpus bilingues et à la traduction littéraire contributive de qualité 67
III.1 LISTE (AVEC LES MT ASSOCIÉES) 67
III.2 COMMENTAIRES SUR LES UTILISATIONS ACTUELLES 69
III.2.1 Accès à des sites Web d'organismes ou de sociétés 69
III.2.2 Aide à la traduction de documents : rapports, de parties de thèse, manuels 69
III.2.3 Accès multilingue à des documents pédagogiques : MACAU 70
III.2.4 Évaluation 71
III.3 UTILISATIONS PLUS NOVATRICES : PRODUCTION DE BONS CORPUS PARALLÈLES, ET POST-ÉDITION DE TEXTES LITTÉRAIRES POUR L'AUTO-APPRENTISSAGE OU POUR LA TRADUCTION CONTRIBUTIVE 72
III.3.1 Production de « corpus parallèles » de qualité 72
III.3.2 « Voyage au centre de la terre » de Jules Verne 74
III.3.3 « The Book of Me » de Powers 75
III.3.4 « IITB : Monastery, Sanctuary, Laboratory » de Rohit Manchanda 76
Partie B Construction de systèmes de TA spécialisés à des sous-langages en français ↔ chinois 79
Chapitre IV Revue des systèmes TA français ↔ chinois en contexte industriel 80
IV.1 DEMANDE DE GROSSES SOCIÉTÉS 80
IV.2 ÉTAT DE L'ART DE LA TA DU CHINOIS 81
IV.2.1 Historique 81
IV.2.2 Expérimentations 83
IV.3 CONSTRUCTION DE SYSTÈMES DE TA POUR LE CHINOIS BASÉS SUR MOSES EN CONTEXTE INDUSTRIEL 87
IV.3.1 Choix du sous-langage et des couples à traiter 87
IV.3.2 Recherche infructueuse de corpus parallèles adaptés 88

IV.3.3 Production de corpus par PE de résultats de Google ... *90*
IV.3.4 Construction de systèmes français→chinois ... *91*
IV.3.5 Évaluations et perspectives ... *93*
Chapitre V Construction de systèmes de TA pour le chinois avec Moses en contexte de recherche : le projet TABE-FC ... **96**
V.1 BUTS DU PROJET TABE-FC ... 96
V.1.1 Buts théoriques ... *96*
V.1.2 Buts pratiques ... *97*
V.1.3 Définition du projet ... *97*
V.2 CONSTITUTION DES CORPUS D'APPRENTISSAGE ... 98
V.2.1 Recherche de sites et collecte de pages Web monolingues et bilingues ... *98*
V.2.2 Nettoyage et filtrage ... *98*
V.2.3 TA par GT, puis PE (production d'un corpus parallèle) ... *100*
V.3 CONSTRUCTION DE SYSTÈMES DE TA ... 102
V.3.1 Construction de systèmes Moses "ligne de base" ... *102*
V.3.2 Avancement de l'expérimentation ... *103*
V.3.3 Résultats provisoires ... *103*
Chapitre VI Démonstration de l'intérêt de l'apprentissage incrémental en TA statistique ... **105**
VI.1 CONTEXTE ... 105
VI.1.1 Motivations ... *105*
VI.1.2 Expérience sur le site du LIG ... *106*
VI.2 EXPÉRIMENTATION ... 107
VI.2.1 Phase 1 (2-6/2013) ... *107*
VI.2.2 Phase 2 (7-9/2013) ... *110*
VI.2.3 Phase 3 (9-12/14 et 7-11/15) ... *113*
VI.3 ANALYSE DES RÉSULTATS ... 116
Partie C Contribution d'outils et de ressources ... **117**
Chapitre VII Construction de systèmes de TA et support à la RI multilingue pour MUMIA ... **118**
VII.1 CONTEXTE ET MOTIVATIONS ... 118
VII.1.1 Description du projet MUMIA et du WG2 ... *118*
VII.1.2 PerFedPat et Khresmoi ... *119*
VII.1.3 Objectif poursuivi ... *120*
VII.2 CONSTRUCTION DE MT ET DE STA À PARTIR DES CORPUS CLEF-IP 2011 ... 120
VII.2.1 Description du corpus CLEF-IP ... *120*
VII.2.2 Extraction de MT à partir de CLEF-IP 2011 ... *122*
VII.2.3 Construction des systèmes de TA ... *125*
VII.3 EXPÉRIMENTATION ET ÉLARGISSEMENT À D'AUTRES LANGUES ... 125
VII.3.1 Reconstruction de trois sites Web de brevets monolingues ... *125*
VII.3.2 Accès multilingue en utilisant les systèmes de TA créés ... *127*
VII.3.3 Accès multilingue utilisant d'autres systèmes et pour d'autres langues ... *127*
Chapitre VIII Mise à disposition de ressources ... **130**
VIII.1 CONTRIBUTION DE RESSOURCES STATIQUES SOUS FORME DE MT ... 130
VIII.1.1 Formats choisis ... *130*
VIII.1.2 Méthode de création ... *132*
VIII.1.3 Résultats ... *132*
VIII.2 CONTRIBUTION SOUS FORME DE SYSTÈME DE TA ... 135
VIII.2.1 Systèmes de TA téléchargeables ... *135*
VIII.2.2 Systèmes de TA utilisables comme des service Web ... *135*
VIII.3 PASSERELLES IMAG VERS DES SITES WEB STATIQUES OU DYNAMIQUES ... 135
VIII.3.1 Passerelles iMAG pour des sites statiques ... *135*
VIII.3.2 Passerelles iMAG pour des sites dynamiques ... *136*
VIII.3.3 Structure d'une contribution « dynamique » par iMAG ... *137*
VIII.3.4 Remarques sur la création de certains des sites « contribués » ... *138*
Chapitre IX Vers une plate-forme de construction, déploiement et évaluation de systèmes de TA: JianDan-eval ... **140**
IX.1 BUTS DU SYSTÈME JIANDAN-EVAL ... 140
IX.2 ARCHITECTURE LOGICIELLE ... 141
Conclusions et perspectives ... **143**
Bibliographie 145

Table des définitions **150**
Annexes **153**
ANNEXE 1 : CORPUS DE LA CAMPAGNE D'ÉVALUATION DE TA DU PROJET TRANSAT 153
ANNEXE 2 : PROTOCOLE D'ÉVALUATION POUR LE PROJET TRANSAT 155
ANNEXE 3 : UN EXEMPLE DU CORPUS B@BEL 157
ANNEXE 4 : CORPUS EOLSS 158
ANNEXE 5 : UN EXEMPLE DE LA GRAPHE UNL AVEC CORRECTION 159
ANNEXE 6 : DOCUMENT DE BREVET DU CORPUS CLEF-IP 2011 (EP-0000007-B2.XML) 161
ANNEXE 7 : STRUCTURE DES DONNÉES DE MINIDICTIONAIRES 164
ANNEXE 8 : 50 SEGMENTS EN « *VUE SECTRA/POST-ÉDITION* » 165
ANNEXE 9 : EXEMPLE DU CORPUS PARALLÈLE FRANÇAIS-CHINOIS CRÉÉ POUR L&M DANS LE DOMAINE DE L'ÉNERGIE 172
ANNEXE 10 : SCRIPT DE FILTRAGE DE CORPUS 199
ANNEXE 11 : SOURCE DE PROGRAMME POUR CALCULER $\boldsymbol{Dmix}$ 202
ANNEXE 12 : 100 BISEGMENTS ANGLAIS-FRANÇAIS EXTRAITS DE CLEF-IP 2011 205
ANNEXE 13 : 20 BISEGMENTS ANGLAIS-FRANÇAIS EN FORMAT TMX 213

Table des figures

Figure 1 : Architecture générale d'une iMAG pour un site élu....29
Figure 2 : Capture d'écran de l'iMAG LIG-LAB en chinois....30
Figure 3 : Architecture par agents SECTra_w, iMAG, PIVAX (Nguyen, 2010)....31
Figure 4 : Interface de « Translate Corpus »....38
Figure 5 : Interface de TRADOH....39
Figure 6 : Options de la fonction "export"....40
Figure 7 : Interface de sélection paramétrable dans SECTra_w....40
Figure 8 : Structure logique d'une base de données de corpus multilingues....42
Figure 9 : Fichier HTML et fichier compagnon .unl....49
Figure 10 : Document 2 traduit de l'anglais vers le français (*GROUND AND SOIL WATER CHARACTERISTICS*)....49
Figure 11 : Exemple de structure et de description d'un dialogue du corpus ERIM....50
Figure 12 : Exemple du fichier *french.wpl* et *vietnamese.wpl*....50
Figure 13 : Capture d'écran de panneau de dictionnaires ajouté à SECTra_w....63
Figure 14 : Interface de SECTra_w intégrant les boutons « *Delete* », « *Clean* » et « *Get* »....63
Figure 15 : Traduction des segments sélectionnés et ajout à la MT....64
Figure 16 : Exemple de l'API « Call Tradoh »....64
Figure 17 : Post-édition d'un document français accédé en anglais (résumé de la thèse de Lingxiao WANG)....70
Figure 18 : Extraction of a "good" TM from a TM produced by "natural" post-edition....72
Figure 19 : Export of a « good » part of a TM....73
Figure 20 : Capture d'écran de iMAG français→chinois pour « Voyage au centre de la terre »....74
Figure 21 : Exemple de post-édition d'un chapitre de « Monastery, Sanctuary, Laboratory: 50 Years of IIT-Bombay » de Rohit Manchanda....77
Figure 22 : Architecture à 3 niveaux et 7 « missions » du projet TABE-FC (Chen, Wang et al., 2014)....97
Figure 23 : Exemple de page Web économique parallèle....98
Figure 24 : Exemple d'une page Web du site de "Bourse de Hong Kong" en format html....99
Figure 25 : Exemple de segments chinois-anglais extraits à partir de pages Web....100
Figure 26 : Capture d'écran de l'iMAG "Bourse de Paris" en chinois....101
Figure 27 : Comparaison de la traduction de GT et de la post-édition humaine....102
Figure 28 : Site du LIG vu en chinois à travers une iMAG....107
Figure 29 : Diminution de temps moyen de PE (par page standard) avec AI dans la phase 1 de l'expérience....109
Figure 30 : Capture d'écran de l'iMAG « Corpus par jour »....111
Figure 31 : Capture d'écran de Chamilo affichant le lien AXiMAG....112
Figure 32 : Diminution du temps moyen de PE (par page) avec AI dans la phase 3 de l'expérience....115
Figure 33 : Architecture de PerFedPat....119
Figure 34 : KHRESMOI....120
Figure 35 : Exemple de fichier XML Dans CLEF-IP....122
Figure 36 : Exemple de champ <*claims*> contenant 6 sous-champs <*claim*> dans *EP-0260000-B1.xml*....122
Figure 37 : Exemple d'un champ <*invention-title*> avec 3 attributs de langue différents et les contenus correspondants en 3 langues différentes....123
Figure 38 : Un champ <*patent-document*> avec attribut lang = "EN"....123
Figure 39 : Exemple de fichier XML monolingue....126
Figure 40 : Exemple de revendication dans le fichier EP0203923B1.xml....126
Figure 41 : Exemple de fichier HTML décoré....127
Figure 44 : PE en mode avancé, avec pseudo-trace montrant les différences entre les sorties de TA, la post-édition (utilisée comme référence), et la MT.....128
Figure 45 : Retraduction des segments du français vers le chinois pour DOC6 avec le système de TA français→chinois MosesLIG....129
Figure 46 : Exemple des données en format TXT (MT CLEF-IP anglais-français)....131

Figure 47 : Exemple des données en format TMX (MT CLEF-IP anglais-français) 131
Figure 48 : Extraction d'une "bonne" MT de la MT produite par post-édition "naturelle" 132
Figure 49 : Segments post-édités pour la ressource énergie 134
Figure 50 : Exemple de contribution au format HTML (Chapitre 1 : Voyage au centre de la Terre) 137
Figure 51 : Capture d'écran du site Web monolingue de CLEF-IP 138
Figure 52 : Capture d'écran de l'iMAG dédiée CLEF-IP 139
Figure 53 : Architecture initiale de gestion de travaux 142

Table des tableaux

Tableau 1 : Sites Web élus des iMAG dédiées disponibles en 2010....17
Tableau 2 : Données statistiques sur les segments post-édités dans SECTra_w depuis 2010....18
Tableau 3 : Liste des iMAG à MT dédiée construites depuis 2010....28
Tableau 4 : Exemples de sites Web de partage de corpus parallèles....33
Tableau 5 : Comparaison de l'organisations logiques, physiques, et interne de quelque corpus....51
Tableau 6 : Métadonnées du corpus BTEC (les segments extraits)....53
Tableau 7 : Métadonnées des données d'évaluation à la TRANSAT....54
Tableau 8 : Métadonnées du corpus UNESCO-B@bel....55
Tableau 9 : Métadonnées du corpus EOLSS au niveau de la macrostructure....56
Tableau 10 : Métadonnées d'un fichier HTML au niveau de la microstructure....57
Tableau 11 : Métadonnées d'un fichier UNL au niveau de la microstructure....57
Tableau 12 : Métadonnées d'un corpus EOLSS au niveau de la mésostructure....58
Tableau 13 : Métadonnées d'un corpus ERIM au niveau de la macrostructure....58
Tableau 14 : Métadonnées de la séance dans le corpus ERIM a au niveau de *la microstructure*....59
Tableau 15 : Métadonnées d'un corpus ERIM au niveau de la mésostructure....59
Tableau 16 : Métadonnées du corpus CLEF-IP 2011....60
Tableau 17 : Métadonnées d'un document de brevet....61
Tableau 18 : 10 paramètres de l'API de CREATDICO....62
Tableau 19 : Exemple de fichier de configuration de CREATDICO....62
Tableau 20 : Exemple d'un lien pour l'utilisation de l'API de CREATDICO....62
Tableau 21 : iMAG pour les sites Web de laboratoires et d'universités....67
Tableau 22 : iMAG pour les sites Web d'organismes et de sociétés....68
Tableau 23 : iMAG pour des projets et des expérimentations....68
Tableau 24 : Nombre de langue du projet MACAU (06/2013)....71
Tableau 25 : Statistiques de documents dans MACAU (06/2013)....71
Tableau 26 : Statistique sur 21 chapitres de « Voyage au centre de la terre »....74
Tableau 27 : Corpus source, cible traduite et cible corrigée....75
Tableau 28 : Statistique sur les données....83
Tableau 29 : Formule d'évaluation de l'automaticité et de la qualité d'un système de TA....84
Tableau 30 : Exemple de traduction de GT....84
Tableau 31 : Paramètres de configuration de Joshua....85
Tableau 32 : Comparaison d'exemples de traductions obtenues par TA et d'une référence....86
Tableau 33 : Exemple de résultat d'évaluation....87
Tableau 34 : Corpus collectés en cherchant des corpus pour français→chinois....89
Tableau 35 : Exemples de bisegments français→chinois parmi les 9000 collectés ou produits....90
Tableau 36 : Comparaison des temps d'entraînement de Moses....91
Tableau 37 : Configuration de la machine....91
Tableau 38 : Scores BLEU pour différentes tailles du corpus d'entraînement....92
Tableau 39 : Statistiques sur le corpus MultiUN....92
Tableau 40 : Exemple de données de test....93
Tableau 41 : Statistiques des données de test....94
Tableau 42 : Score BLEU et exemples de sorties de systèmes de TA....94
Tableau 43 : Statistiques des pages Web collectées....99
Tableau 44 : Exemple de conversion des caractères chinois du traditionnel vers le simplifié....99
Tableau 45 : Statistiques sur la ressource économique et boursière....102
Tableau 46 : Statistiques sur les données d'entraînement de la phase 1....107
Tableau 47 : Évaluation du temps de post-édition (2-6/2013)....109
Tableau 48 : Évaluations basées sur des références (BLEU, NIST, TER)....110
Tableau 49 : Statistiques de post-édition sur 21 articles français 4/7-13/9/2013....111
Tableau 50 : Statistiques de post-édition sur les supports de cours....112
Tableau 51 : Résultat de l'expérimentation (en français-chinois)....113
Tableau 52 : Nombre de segments dans chaque MT....113
Tableau 53 : Statistiques sur les données pour l'AI (phase 3 de l'expérience)....114

Tableau 54 : Évaluation du temps de post-édition (9-12/2014) 115
Tableau 55 : Données de test et scores BLEU 116
Tableau 56 : Nombre de segments extraits comme source et cible après l'alignement de segments dans les champs <*title*> et <*claims*> 124
Tableau 57 : Scores BLEU des systèmes de TA tirés de CLEF-IP 125
Tableau 58 : Segments post-édités dans SECTra_w à partir de 3 langues source 133
Tableau 59 : Segments parallèles obtenus à partir des MT (mêmes remarques) 133
Tableau 61 : Statistique des données pour la ressource énergie 134
Tableau 62 : Systèmes de TA téléchargeables 135
Tableau 63 : Systèmes de TA utilisables comme des services Web 135
Tableau 64 : Passerelles iMAG pour des sites statiques 136
Tableau 65 : iMAG pour des sites dynamiques 136

Glossaire et abréviations

AI	Apprentissage incrémentale
API	Application Programming Interface
ARIANE-G5	Environnement de développement et d'exécution de systèmes (linguiciels) de traduction automatique
CLEF-IP	Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum - Intellectual Property
CLIPS	Communication Langagière et Interaction Personne Système
CNRS	Centre national de la recherche scientifique
DC	Dublin core
DSR	Digital Silk Road
EDF	Électricité de France
EOLSS	Encyclopedia of Life Support Systems
FC	Français-Chinois
GETA	Groupe d'Etude pour la Traduction Automatique
GETALP	Groupe d'Étude pour la Traduction/le Traitement Automatique des Langues et de la Parole
GI	Génie informatique
GT	Google Translate
HQ	Haute Qualité
IMAG	Passerelle interactive d'accès multilingue (interactive Multilingual Access Gateway)
ISCC	Institut des sciences de la communication
L&M	SAS Lingua et Machina
LIG	Laboratoire d'Informatique de Grenoble
LSPL	Langage Spécialisé pour la Programmation Linguistique
MACAU	Multilingual Access & Contributive Appropriation for Universities
MT	Mémoire de Traductions
MUMIA	Multilingual and multifaceted interactive information access
OMNIA	Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins
ONU	Organisation des Nations Unies
PAHO	Pan American Health Organization
PCT	Patent Cooperation Treaty
PE	Post-Édition
PIVAX	Base lexicale à pivot par acceptions (monolingues et interlingues)
QCM	Question à choix multiples
RI	Recherche d'Information
SAAS	Software as a Service
SECTRA_W	Système d'Exploitation de Corpus de Traductions sur le Web
SECTRA/EVAL	Première version de SECTra, développée pour le projet TRANSAT d'Orange Labs
SECTRA/TRAD	Deuxième version de SECTra, développée pour le projet EOLSS/UNL++
SECTRA/WEB	Troisème version de SECTra, développée pour le projet iMAG
SEGDOC	Segmentation de documents XML
TA	Traduction Automatique
TH	Traduction Humaine
TMX	Translation Memory eXchange

TRADOH	Un outil permet d'obtenir une traduction dans sa langue, par mise en œuvre automatique d'un ou plusieurs systèmes de TA disponibles en local ou à distance, avec composition éventuelle.
TRANSAT	Projet de TA de parole d'Orange 2004~2007
TXT	Texte brut
UNESCO/B@BEL	Partie du site Web de l'Unesco consacrée à la communication multilingue
UNL	Universal Networking Language (projet lancé par l'UNU (fin 1995) et langage "anglosémantique" d'hypergraphes associables aux énoncés en langue naturelle
XLIFF	XML Localisation Interchange File Format
XRCE	Xerox Research Center Europe

Introduction

Cette thèse a été effectuée dans l'équipe GETALP du LIG, et dans le cadre d'une bourse CIFRE avec Lingua et Machina, une jeune société qui vise à "prendre en charge la communication multilingue de l'entreprise". Le sujet initialement défini était centré sur l'amélioration de plusieurs aspects de génie logiciel du logiciel SECTRA_W/IMAG, réalisé par Cong Phap HUYNH dans le cadre de sa thèse (Huynh, 2010). Le point principal concernait la transformation de la partie SECTRA_W (Service web d'Exploitation de Corpus de Traductions) en un système programmable et extensible. On visait à pouvoir l'utiliser comme un "serveur corporal" gérant des suites de test ainsi que des corpus de développement pour le compte de systèmes de TA (traduction automatique) munis d'un environnement de développement complet, comme ARIANE-G5 et son successeur, en cours d'implémentation, ARIANE-Y. Un autre objectif était de pouvoir non seulement exploiter des corpus parallèles existants, pour les évaluer et/ou les améliorer par post-édition collaborative en ligne, mais aussi de pouvoir les étendre à de nouvelles langues par appel à des serveurs de TA, suivi de post-édition. Un dernier thème était la recherche d'une méthode de spécification formelle implémentable des "vrais" corpus multilingues, c'est-à-dire pas seulement des listes de "segments" multilingues, comme le BTEC (Boitet et al., 2007), qui ne sont en fait que de grandes "mémoires de traduction", même si on les appelle "corpus parallèles". On souhaitait s'attaquer à la complexité des "vrais corpus", et passer à l'échelle, de façon, par exemple, à pouvoir traiter des corpus fortement structurés et de très grande taille comme ceux des brevets. Dans un corpus complexe, un document est formé d'un document maître (en XML par exemple) ou d'une hiérarchie de tels documents, accompagné d'une collection de fichiers "satellites" (images, vidéos…) et éventuellement d'annotations contenues dans des fichiers "compagnons".

On visait aussi à résoudre un certain nombre de problèmes liés à l'interaction entre SECTRA_W et le logiciel IMAG (**i**nteractive **M**ultilingual **A**ccess **G**ateways) qui utilise SECTRA_W comme un "dorsal" et permet d'accéder à des sites Web "élus" dans un grand nombre de langues, avec possibilité d'améliorer les "prétraductions" produites par des serveurs de TA en les corrigeant ("post-éditant") directement sur la page Web, ou dans l'interface de SECTRA_W dédiée à la post-édition.

Ces objectifs ont évolué à cause des besoins de l'entreprise, qui désirait d'abord construire des systèmes de TA "maison" français↔chinois en utilisant l'environnement Moses. Pour cela, il faut disposer de grands corpus parallèles de bonne qualité, dans le bon sens, et représentatifs des sous-langages des clients potentiels, en l'occurrence EDF, Renault, etc. Cela a d'abord mené à l'étude, l'expérimentation et l'évaluation d'aligneurs divers, de segmenteurs du chinois, et de divers systèmes de TA existants. Dans un deuxième temps, le travail s'est plus orienté vers la TA proprement dite. Il s'est d'abord agi de construire un environnement de préparation et d'exploitation de systèmes MOSES, intégré aux outils de L&M (LIBELLEX, MYRIAM) ou utilisés par L&M (METRICC, XELDA…). L'obstacle majeur à surmonter était l'absence de corpus parallèles français-chinois. Nous avons alors construit un corpus de 9000 segments, d'abord par post-édition de résultats de GOOGLE TRANSLATE (GT), puis par post-édition de résultats d'une première version d'un système MOSES-L&M-FC. À peu près à la même période, L&M nous a demandé d'étudier la nouvelle possibilité offerte par MOSES de faire de l'apprentissage incrémental (AI). Nous l'avons fait, ainsi que quelques essais préliminaires, puis L&M m'a dirigé sur autre chose, jugeant l'approche peu prometteuse. Il est vrai que les améliorations constatées étaient faibles. Pourtant, les gains de temps étaient considérables (environ 1h pour l'AI sur quelques dizaines ou centaines de nouveaux segments au lieu de 20h pour un réapprentissage complet). Notre intuition était que, dans le cas de sous-langages, on devait arriver, en faisant quelques dizaines d'itérations d'AI, à produire des prétraductions meilleures que celles de GT, BING, SYSTRAN ou

NIUTRANS, au moins en ce qui concerne la qualité d'usage pour la tâche de post-édition, et peut-être aussi pour la tâche de compréhension. Encouragé par mes directeurs de thèse, j'ai alors orienté ma recherche dans cette direction, pendant plusieurs mois. Au terme d'une expérience sur le sous-langage du site Web du LIG, nous avons pu publier (à COLING 2012) des résultats encourageants : notre courbe de "temps de post-édition" descendait assez régulièrement, et, au bout d'une vingtaine d'itérations avec un réapprentissage total au milieu, n'était plus qu'un petit peu au-dessus de la "ligne de base" correspondant à la PE des résultats de GT (environ 10 mn/page). Depuis, j'ai préparé et mené une troisième expérience, en améliorant l'automatisation du processus et des mesures associées, et j'ai pu démontrer que, au moins dans le cas d'un sous-langage comme celui du site du LIG, et du français-chinois, la combinaison de l'apprentissage incrémental par périodes, et spécialisé à un sous-langage, pouvait donner des résultats nettement meilleurs que ceux des systèmes de TA généralistes.

Ma recherche s'est ensuite trouvée orientée vers le passage à l'échelle, la TA de brevets, et la construction de grandes ressources de bonne qualité, dans le cadre du projet COST "MUMIA" de l'UE. . Mon directeur de thèse était en effet VP de MUMIA et en charge du WG2. Dans le WG2 auquel j'ai participé, il s'agissait d'étudier et de prototyper des "infrastructures" matérielles et logicielles pour la recherche d'information dans un cadre multilingue, multimodal et multi-facette. J'ai été amené à traiter la collection CLEF-2012 (la même que CLEF-2011), constituée à partir de 1,5 millions de brevets partiellement traduits par des professionnels. Une bonne proportion des "segments", initialement rédigés en français, allemand ou anglais, a été traduite dans une deux des autres langues. Il ne s'agit pas de collections parallèles : il y a un fichier par brevet, en XML, où chaque segment contient sa version originale et éventuellement une ou des versions dans d'autres langues. À partir de cette collection, j'ai détecté la langue source de chaque segment, et construit 3 mémoires de (très bonnes) traductions, une pour chaque langue source. Je les ai aussi utilisées comme base pour l'apprentissage de 3 systèmes Moses (de-fr, fr-de, fr-en). Enfin, avec l'aide d'un étudiant de M1 en TER (Huanan SUN), j'ai construit 3 collections de brevets monolingues, chacun étant seulement dans sa langue source. Grâce à 3 iMAG, il est possible d'y accéder dans diverses langues. Les traductions sont évidemment très bonnes pour les langues initiales (en tout cas, sur les parties du corpus réservées pour les tests), mais il est aussi possible d'y accéder dans d'autres langues, par exemple en chinois, d'améliorer les résultats de TA par PE, et de recycler les "bonnes traductions" pour construire un système spécialisé s'améliorant au fur et à mesure de l'usage.

Durant la dernière partie de ma thèse, je suis revenu au thème de la TA français-chinois, en participant à deux projets, MACAU-OFI et TABE-FC. MACAU-OFI est un projet défini par R. Kalitvianski et Ch. Boitet en 2012, visant à mettre à disposition des étudiants étrangers des notes de cours dans leur langue, en utilisant une passerelle IMAG dédiée, et en demandant aux étudiants eux-mêmes de "post-éditer". Durant l'été 2013, j'ai ainsi participé à l'encadrement de deux stages d'été d'étudiants chinois (en master informatique à l'UJF), qui ont post-édité environ 520 pages standard (130K mots) dans le domaine des outils formels pour l'informatique. En 2013-2014, j'ai aussi participé de façon très active à la définition et au début de la réalisation du projet TABE-FC monté avec l'université de Xiamen, dans le cadre d'une année sabbatique passée à Grenoble par le Dr Yidong CHEN. Il s'agit de construire des systèmes de TA permettant à des Chinois d'avoir un accès en chinois de bonne qualité (et surtout bien plus fiable et fidèle que les systèmes généralistes) aux "brèves" des bourses francophones, « *flash reports* » en anglais, un tel système, ALTFLASH, a été déployé pour le Nikkei à partir de 2001, et plus généralement aux sites Web économiques en français, et inversement pour des Français désireux d'intervenir sur les bourses de Shanghai, Shenzhen et Hong Kong. Cela m'a fait revenir aux thèmes plus liés au génie logiciel.

Au total, mon apport se situe dans quatre domaines principaux : (1) le génie logiciel des systèmes d'exploitation et maintenant de création et de gestion de "vrais" corpus multilingues, (2) la TA, avec des contributions portant sur l'apprentissage incrémental, la TA français-chinois, ainsi que les environnements de construction et de déploiement de systèmes de TA de type Moses ou similaire, (3) la mise à disposition de ressources (mémoires de traductions, systèmes de TA associés), et (4) la spécification et l'implémentation en cours d'une infrastructure pour l'évaluation, la plate-forme JIANDAN-EVAL, qui permettra les évaluations classiques, ainsi que l'évaluation comparative et "en usage" de systèmes de TA de toutes les architectures existantes.

La première partie de ce mémoire concerne la production, l'extension et l'amélioration de corpus multilingues par traduction automatique (TA) et post-édition contributive (PE). Des améliorations fonctionnelles et techniques ont aussi été apportées aux logiciels SECTRA_W et IMAG produits lors des thèses de C.P. HUYNH et H.T. NGUYEN. Nous avons progressé vers une définition générique de la structure d'un corpus multilingue, multi-annoté et multimédia, pouvant contenir des documents classiques aussi bien que des *pseudo-documents* (comme des pages Web) et des *méta-segments*. Cette partie a été validée par la création de bons corpus bilingues français-chinois, l'un d'eux résultant de la toute première application à la traduction littéraire (un roman de Jules Verne), projet personnel mené pour progresser en français.

La seconde partie est centrée sur nos travaux en TA proprement dite. Initialement motivée par un besoin industriel, cette partie de notre recherche a consisté à étudier comment construire des systèmes de TA de type Moses, spécialisés à des sous-langages, en français↔chinois, et à étudier la façon de les améliorer dans le cadre d'un usage en continu avec possibilité de post-édition (PE) contributive en ligne. Dans le cadre d'un projet interne sur le site du LIG et d'un projet (TABE-FC[1]) en coopération avec l'université de Xiamen, nous avons pu démontrer l'intérêt de l'apprentissage incrémental en TA statistique, sous certaines conditions, grâce à une expérience qui s'est étalée sur toute la thèse.

Dans la troisième partie de ce mémoire, nous présentons nos contributions en termes de mise à disposition de supports informatiques et de ressources. Les principales se placent dans le cadre du projet COST MUMIA de l'EU et résultent de l'exploitation de la collection CLEF-2011 de 1,5 M brevets partiellement multilingues. De grosses mémoires de traductions en ont été extraites (17,5 M segments), trois systèmes de TA en ont été tirés (allemand→français, anglais→français, français→allemand), et un site Web de support à la RI multilingue sur les brevets a été construit. Avant de conclure, nous terminons en décrivant aussi la spécification et la réalisation en cours de JIANDAN-EVAL, une plate-forme de construction, déploiement et évaluation de systèmes de TA.

[1] TA pour les sites boursiers et économiques appliquée au français-chinois. En anglais: MTSE-FC.

Partie A Production, extension et amélioration de corpus multilingues par TA et PE contributive

Résumé

La partie A présente l'amélioration d'aspects fonctionnels et techniques de SECTRA_W et du logiciel IMAG pour les passerelles d'accès multilingue. Cette partie comporte aussi des aspects plus conceptuels, et la définition de nouvelles fonctionnalités. Nous montrons enfin la variété des IMAG et de leurs usages, de l'accès multilingue à la création de bons corpus bilingues et à la traduction littéraire contributive de qualité.

Chapitre I Amélioration d'aspects fonctionnels et techniques de SECTra et du logiciel iMAG pour les passerelles d'accès multilingue

Ce chapitre présente la situation et l'état de l'art au début de la thèse, puis les améliorations de SECTRA_W/IMAG étudiées et réalisées dans le cadre du projet TRAOUIERO.

I.1 Situation et état de l'art au début de la thèse

I.1.1 Modélisation et exploitation de corpus de traductions : SECTra_w

I.1.1.1 Présentation générale

I.1.1.1.1 Motivations et bref historique

En 2005-2007, notre équipe, alors le GETA du CLIPS, a participé au projet TRANSAT dans le cadre d'un contrat avec FRANCE TELECOM R&D. Dans ce projet, nous avons utilisé la toute première version de SECTra, SECTRA/EVAL, pour organiser une campagne d'évaluation de l'utilité potentielle des systèmes de traduction automatique commerciaux dans le domaine de la traduction de la parole. Le domaine privilégié dans cette étude était l'assistance à un touriste en situation difficile. Une autre partie de l'étude portait sur le domaine de la restauration.

De février 2008 à octobre 2008, la deuxième version de SECTra, SECTRA/TRAD, a été dévloppée puis utilisée pour le projet EOLSS/UNESCOL réalisé dans le cadre d'un contrat entre l'Association Champollion et la fondation UNDL[2]. Dans ce projet, SECTRA a assuré le support et la gestion de la traduction de bonne qualité de 25 articles de l'encyclopédie EOLSS (*Encyclopedia of Life Support Systems*), soit environ 220K mots (880 pages standard), ou 13676 segments. En particulier, SECTRA_W a fourni pour ce projet un environnent collaboratif en ligne pour la post-édition humaine appliquée aux résultats de systèmes de TA et de déconvertisseurs UNL.

Depuis 2009, la troisième version, SECTRA_W, SECTRA/WEB, développée fin 2008, est utilisée comme "dorsal" pour l'accès multilingue de bonne qualité à des sites Web "élus". La conception et le développement de SECTra ont fait l'objet de la thèse de Cong Phap HUYNH (Huynh, 2010). À ce moment, mi-2010, nous avions alors construit des passerelles "iMAG" pour une trentaine de sites Web, dont celui du LIG. Chaque IMAG que nous avons construite

[2] The UNDL FOUNDATION is a private Swiss law Foundation with head office in Geneva, Switzerland, legally representing the United Nations Organization in protecting its property rights pertaining to the UNL language and system, and legally representing the EOLSS Publishers and the UNESCO Joint Committee in translating the EOLSS with the use of the UNL technology.

apparaît comme un Wiki permettant l'accès multilingue à un site Web élu et la contribution à l'amélioration des traductions des segments (phrases et titres) de ses pages.

Essentiellement, une IMAG fournit une interface interactive permettant aux utilisateurs de voir et de post-éditer le site Web élu en plusieurs langues. En arrière-plan, tous les processus de gestion des segments "multilingualisés" du site Web sont réalisés par SECTRA_W, par SEGDOC (système de segmentation), et par TRADOH (intergiciel d'appel à des systèmes de TA).

I.1.1.1.2 Apports de la thèse de C.P. HUYNH

La thèse de C.P. HUYNH a apporté des réponses théoriques et pratiques à trois grands défis. Le premier défi consistait à offrir un support informatique unifié à l'évaluation des systèmes de TA. Le deuxième défi concernait le support contributif et collaboratif au travail humain sur des corpus variés en contexte multilingue. Le troisième défi était la construction d'un support informatique à l'exploitation de corpus de traductions dans des applications novatrices comme l'accès multilingue à des sites Web et la recherche d'information en contexte multilingue et multimédia (OMNIA). Plusieurs notions nouvelles ont été précisées (comme *segment multilingualisé et contextualisé*, *pseudo-document*, *métadocument*, etc.), et plusieurs principes généraux (*proactivité*, *délégation*, etc.) ont été introduits. C.P. HUYNH[3] a dégagé six problèmes associés à chacun de ces trois défis, à dominante conceptuelle (par exemple, définition étendue d'un « contexte » de segment), algorithmique (par exemple, programmabilité du traitement des corpus), et programmatoire (par exemple, traitement de masses de données).

I.1.1.1.3 Évolutions depuis début 2010

Jusqu'au milieu de l'année 2010, 30 IMAG avaient été définies. Chaque IMAG est associée à une mémoire de traductions (MT) gérée par SECTRA_W. Une IMAG de démonstration partage en général une MT avec d'autres IMAG. Les MT partagées les plus utilisées sont DEMO, DEMO1, DEMO2. Voici quelques données présentées dans la thèse de Phap (Tableau 1).

Tableau 1 : Sites Web élus des iMAG dédiées disponibles en 2010

LIG laboratory	Digital Silk Road
Danang city	Da Nang University of Technology
TOL	ISCC
Unesco/B@bel	Systran
La Métro	Forum Lyon
Mica	Campus France
GETALP	Floralis
TechniLang	Ordinaide
Homerica	aikicorenc
MT 25 Years On	Essilor
Michelin	Winsoft
LeMonde.fr	UNDL-foundation
GETALP presentation	UNESCO_Babel
XD-consulting	ARDI Rhône

Entre 2010 et 2012, avant le début de cette thèse, nous avons participé (lors d'un stage en alternance de M2P-GI puis d'un stage prédoctoral effectué dans le cadre du projet ANR TRAOUIÉRO) à l'amélioration SECTRA_W, qui en est à la version 2. Grâce au configurateur

[3] C'est son nom d'usage dans l'équipe.

d'IMAG réalisé par H. T. NGUYEN lors de son post-doctorat (projet IMAG/LAMETRO en 2010 pour l'Expo de Shanghai en octobre 2010), le GETALP a créé 80 IMAG de démonstration, partageant 3 mémoires de traductions principales. Il y a 8 langues source, et chaque site Web peut être accédé en plus de 10 langues cible (en fait, dans toutes les langues traitées par GT, système qui en traite le plus) et 5 autres MT.

Début 2015, environ 45% des segments visités via une IMAG (plus de 370 000), et donc « pré-traduits » automatiquement, avaient été post-édités par des contributeurs. La post-édition a concerné majoritairement les paires anglais→chinois, anglais→français, et français→chinois, comme le montre le Tableau 2.

Tableau 2 : Données statistiques sur les segments post-édités dans SECTra_w depuis 2010

Paire de langues (L1→L2)	Bisegments	Mots source L1 (Pages standard)	Mots cible L2 (Pages standard)	Taille de L1	Taille de L2
anglais → français	121 074	2 542 731 (10 170 p.)	2 613 351 (10 453 p.)	10,1Mo	10,4 Mo
anglais → chinois	208 106	4 370 530 (17 482 p.)	6 063 942 (151 159 p.)	19,1 Mo	17,6 Mo
français → anglais	29 079	627 661 (2510 p.)	610 098 (2 440 p.)	4 Mo	3,9 Mo
français → chinois	10 890	228 703 (914 p.)	317 322 (793 p.)	1,5 Mo	1,25 Mo
chinois → Anglais	2 013	58 656 (146 p.)	42 275 (169 p.)	240 Mo	263 Mo
chinois → français	10 062	291 192 (727 p.)	211 185 (844 p.)	874 Mo	1 Mo

I.1.1.2 Résumé des avancées et des limitations au niveau conceptuel

I.1.1.2.1 Avancées

Structure des données. Les données sont organisées autour des MT. L'administrateur peut créer une MT et la « dédier » à un corpus (comme les 25 articles de EOLSS) ou à un pseudo-corpus (comme les pages Web accessibles par un certain ensemble d'url), ou bien il la déclare comme « partageable » par plusieurs IMAG. Dans la partie SECTRA_W/IMAG, il n'y a pas de corpus bien identifié au sens classique, sauf pour les MT dédiées. En effet, dans le cas d'une MT partagée par plusieurs IMAG, ce sont toutes les pages Web (des "*pseudo-documents*" et pas des documents, d'ailleurs) visitées depuis ces IMAG qui constituent "le corpus", qui devient de fait l'ensemble des pages Web dont les segments ont été mis dans cette MT. Il y a bien une avancée, puisqu'on peut traiter des pseudo-documents et partager des MT, mais elle s'accompagne d'une confusion, sur laquelle nous reviendrons plus loin.

Au moment de la post-édition d'une page Web sous SECTRA_W, on édite en fait la partie de la MT constituée de l'ensemble (sans répétition) des segments apparaissant dans l'instance de ce pseudo-document. Cet ensemble est présenté en une suite de "pages logiques", chacune contenant un nombre maximum de segments, paramétrable. La valeur par défaut est actuellement 20, car une page standard contient en moyenne 20 phrases de 12 à 13 mots. Mais cela peut varier selon le type de document. Si un segment apparaît plusieurs fois dans une page Web, la suite des pages logiques ne correspond donc pas à la suite des segments dans la page.

C'est un point apparemment positif, puisqu'un segment répété ne sera post-édité qu'une fois. Cependant, il serait préférable de présenter la suite des segments dans l'ordre du document, sans éliminer les segments répétés, mais en présentant toutes leurs post-éditions dans leurs occurrences successives. En effet, deux occurrences d'un même segment dans une page Web,

et plus généralement dans un document, peuvent devoir être traduites différemment. Par exemple, dans le site Web du LIG, le segment « Recherche » doit être traduit par « *Search* » s'il apparaît dans un onglet de navigation, et par « *Research* » si c'est un item du plan du site.

Un point très important est que chaque segment source a un identifiant unique qui est utilisé pour le lier avec les éléments qu'on considère comme ses annotations (des traductions automatiques avec leur origine, des post-éditions avec leurs scores et leur dernier contributeur, et on pourrait avoir une liste de versions, un minidictionnaire, un graphe UNL, etc.).

Appel de systèmes de TA. SECTRA_W permet de paramétrer la liste des systèmes de TA appelés (comme GT, SYSTRAN, et REVERSO) pour produire des traductions candidates. Pour cela, il utilise l'intergiciel TRADOH (originellement créé en 2002-2003 par Hung VO-TRUNG (Vo Trung, 2004) dans le cadre de sa thèse, puis revu par nous en 2013.

Support de l'évaluation de systèmes de TA. SECTRA_W supporte l'évaluation de systèmes de TA depuis sa première version, mais cette partie est restée indépendante du reste. En particulier, la partie concernant la PE contributive ne permet pas de faire de l'évaluation subjective classique. On se contente d'un « score de qualité » défini à partir du profil du dernier contributeur à la PE d'un segment, et modifiable par les contributeurs inscrits et connectés.

La partie dédiée à l'évaluation reste très intéressante. Décrivons-la brièvement. SECTRA/EVAL intégrait dès octobre 2007 des outils d'évaluation subjective[4] et objective[5]. On peut effectuer les deux types les plus courants d'évaluation subjective (adéquation, fluidité) en utilisant plusieurs juges, pour un ou plusieurs systèmes de TA à la fois. L'administrateur d'une campagne d'évaluation peut interdire ou permettre à plusieurs juges d'effectuer l'évaluation subjective sur la même unité de données (segment). Les libellés et le nombre de réponses possibles sont paramétrables. Au total, on peut définir de nombreuses configurations de campagnes d'évaluation. Il y a bien sûr aussi les outils d'import, d'export, et de calcul de résultats.

SECTRA_W/TRAD fournit de plus des interfaces graphiques pour faciliter la sélection de données (tout ou partie d'un corpus), et le lancement des programmes de calcul des mesures d'évaluation objective sur les données sélectionnées.

« Base corporale ». SECTRA_W est une sorte de « base corporale[6] », Il fournit des services Web pour l'exploitation de corpus multilingues, l'amélioration des traductions, et l'extension « en largeur » à d'autres langues, par appel à des serveurs de TA. Depuis 2008, SECTRA_W permet la post-édition collaborative de pages Web à l'aide d'un ou plusieurs systèmes de TA, et leur évaluation par la distance de post-édition, qui est assez bien corrélée au temps passé à la post-édition, ce qui donne une bonne mesure de « qualité d'usage », pour la post-édition tout au moins.

SECTRA_W, et d'ailleurs aucun système d'évaluation de TA, ne donne de moyen d'évaluer la qualité d'usage quand l'utilisateur est une personne ne connaissant pas la langue source et cherchant à comprendre une page Web. Dans le cas où il s'agit d'un système de e-commerce,

[4] C'est-à-dire, faisant appel à des jugements humains, comme l'évaluation d'adéquation, de fluidité, de fidélité, de grammaticalité, de précision terminologique, etc.

[5] Il s'agit non seulement des mesures de type BLEU, NIST ou METEOR, fondées sur des calculs de similarité avec des traductions de référence, et donnant des résultats sur tout un corpus, mais pas sur tel ou tel bisegment, mais aussi de mesures liées à la tâche, comme la "distance mixte de post-édition" introduite par Boitet et Pineau en 2004 (Pineau, 2004).

[6] Terme proposé par Ch. Boitet pour faire pendant à celui de "base lexicale".

on devrait pouvoir le faire en mesurant le nombre d'achats par rapport au nombre et à la durée des visites, mais ce n'est qu'une hypothèse. Dans d'autres cas, par exemple l'accès à des manuels scientifiques, on pourrait évaluer la compréhension "via la traduction" par une technique de QCM[7].

Dans cette thèse, nous nous limitons à l'évaluation de la qualité d'usage quand les utilisateurs font de la post-édition.

Notions. Lors de la conception de SECTRA_W, C. P. HUYNH a clarifié et bien défini plusieurs notions, comme segment monolingue multilingualisé, pseudo-document, métasegment, super-segment, infra-segment, etc.

Nous les reprenons ici, et en introduisons quelques-unes dont la nécessité est apparue durant le projet TRAOUIERO.

1.1.1.2.2 Définitions de diverses notions

Les définitions nouvelles sont signalées par une étoile.

a. *Phrase et groupe*

Définition 1. Une *phrase* (*sentence* en anglais) est l'unité élémentaire d'un énoncé, formée de plusieurs mots ou groupes de mots, et qui présente un sens complet. (*TheFreeDictionary*[8]).

Définition 2*. Un *syntagme* ou *groupe* (*phrase* en anglais) est un constituant possible d'une phrase. En général, un *titre* est un groupe nominal.

b. *Segment, fragment, élément non textuel (hors-texte)*

Définition 3. Un *segment* est l'unité de traduction de base des traducteurs humains. Il s'agit d'une phrase, d'un titre, ou d'un terme dans une nomenclature.

Définition 4*. Un *fragment* (*chunk* en anglais) est une partie d'un segment, qui peut être un *groupe syntaxique* (*phrase* en anglais) ou un simple n-gramme.

Définition 5*. Un segment peut contenir des *éléments non textuels*, ou *hors-texte,* comme des images, des formules, ou des balises, qui ont un rôle linguistique et une valeur non linguistique.

Ainsi, une image ou une expression mathématique fonctionne usuellement comme un groupe nominal, une relation mathématique fonctionne comme un groupe nominal ou un groupe verbal, et une balise peut seulement contrôler la présentation (gras, souligné) ou avoir aussi un rôle de ponctuation (élément de liste à puces par exemple).

On appelle aussi « segment » (en TMX ou en XLIFF) un segment source accompagné d'une ou plusieurs traductions. Il faut donc raffiner la définition précédente.

[7] Question à choix multiples

[8] http://fr.thefreedictionary.com/phrase

Dans le format TMX[9] (*Translation Memory eXchange*), un segment multilingue est représenté par un élément <*tu*>.

Par exemple :

```
<tu>
   <tuv xml:lang="en"><seg>How are you?</seg><tuv>
   <tuv xml:lang="fr"><seg>Comment vas tu?</seg><tuv>
   <tuv xml:lang="vi"><seg>Chú có khỏe không?</seg><tuv>
</tu>
```

Si N = 2, 3…, on parle de segments bilingues, trilingues, etc. On appellera donc *segment monolingue* un segment multilingue réduit à un seul segment (source).

Définition 6. Un *segment monolingue* est un segment dont le contenu textuel est en une seule langue.

p.Définition 7*. Un *segment multilingue* est un segment dont le contenu textuel est dans plusieurs langues, chaque version étant considérée comme contenant exactement la même information, exprimée de façon correcte.

C'est par exemple le cas des segments contenus dans des documents de brevet (voir VII.2.1).

Définition 8. Un *segment monolingue multilingualisé (annoté)* est un objet contenant un segment « source » primaire, une ou plusieurs traductions (automatiques ou humaines ou automatiques post-éditées) pour une ou plusieurs langues, et des annotations, en général des objets, comme des arbres linguistiques, des graphes UNL, des résultats d'évaluation(s), et des références aux contributions ayant produit chaque objet non primaire.

Définition 9*. Un *segment multilingue multilingualisé (annoté)* est un objet contenant un segment multilingue, dans N langues « sources », et, dans M autres langues, une ou plusieurs traductions (automatiques ou humaines ou automatiques post-éditées), ainsi que des annotations, comme celles d'un segment monolingue multilingualisé et annoté.

Définition 10*. Le *chemin traductionnel* d'une annotation, en particulier d'une traduction ou d'une post-édition, est la suite des opérations l'ayant produite, ainsi que les intervenants humains impliqués, et les éventuels objets auxiliaires utilisés.

Par exemple, une traduction français-chinois peut avoir été produite par le chemin traductionnel :

```
TA (Systran, fr_en); PE(Zhong, Sectra); TA(Neon, en_zh); PE(Wang, Sectra);
```

c. *Métasegment, document, métadocument*

Définition 11*. Un *métasegment* est un segment comportant une ou plusieurs *variables*, éventuellement typées (nombre, date, balise faible…).

Définition 12*. Un *document* est un ensemble formé par un support et une information, celle-ci enregistrée de manière persistante. Nous nous intéressons aux documents textuels, qui contiennent des "segments" textuels.

Définition 13*. Un *métadocument* est un document pouvant contenir des métasegments.

Définition 14*. Un *pseudo-document* est défini par une référence (nom de fichier, url, uri) à un document qui peut varier au cours du temps.

Par exemple, une page Web peut changer un peu, pas du tout ou totalement d'un appel au suivant.

[9] http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm#Misc_GroupTU

- Un petit changement peut être dû au fait que le document est en fait un métadocument, et qu'au moins une valeur d'une variable d'un métasegment a changé (par exemple, la date, ou le nombre de visites).
- Un changement total peut être dû au fait que l'instance précédente a été remplacée par un document tout à fait différent. C'est par exemple le cas d'une page Web contenant chaque jour des nouvelles, ou un nouvel éditorial (comme le "World Web" de l'Unesco).

d. *Contexte*

Définition 15. *Contexte* : Le contexte *m-n* d'segment source par rapport à un document, ou plus généralement à une instance d'un pseudo-document, est défini par :

- la liste des *m* segments (de même langue) qui le précèdent.
- la liste des *n* segments (de même langue) qui le suivant.
- l'instanciation des variables, s'il y en a, dans ces *m+n+1* segments.

Le contexte *m-n* d'un segment (cible) dans une version résultat de TA ou de PE, dans un segment monolingue, ou multilingue multilingualisé, est défini par

- la liste des *m* segments (de même langue et de même version) qui le précèdent.
- la liste des *n* segments (de même langue et de même version) qui le suivant.
- l'instanciation des variables, s'il y en a, dans ces *m+n+1* segments pour la même version.

Deux résultats de TA appartiennent à la même version s'il ont été produits par la même instance du même système de TA.

Deux post-éditions appartiennent à la même version si elles apparaissent dans la même version d'un document (ou pseudo-document) post-édité.

Par conséquent, le contexte d'une PE peut dépendre du choix des PE des segments précédents et suivants. Exemple : 2 réviseurs choisissent des PE différentes dans la MT et produisent 2 versions (tout à fait correctes) du même document.

I.1.1.2.3 Limitations

Absence d'une définition claire des "corpus" et des "corpus de traductions". Cela est dû au fait que SECTRA_W d'abord été développé pour ce qu'on appelle des corpus parallèles en TA statistique, c'est-à-dire des listes de segments se correspondant dans une ou plusieurs langues, sur le modèle du BTEC (*Basic Travel Expression Corpus*) d'ATR (Boitet, Boguslavskij et al., 2007). Dans un second temps, on a traité un "vrai" corpus (monolingue), formé de 25 articles de l'encyclopédie EOLSS, présentés dans autant de fichiers *.aspx* et de fichiers compagnons *.unl,* auxquels on a attaché les traductions en français de chaque segment. Enfin, on a traité des corpus en quelque sorte virtuels, formés des pseudo-documents correspondant aux pages Web de tel ou tel site Web.

Définition 16*. Un *corpus* est un ensemble usuellement fermé de documents homogènes du point de vue de leur structure, de leur(s) langue(s), de leur genre et de leur domaine.

Par exemple, on peut parler du corpus (monolingue) des articles du Monde sur la culture de 2000 à 2002. Le caractère fermé est relatif, car on considère que certains corpus croissent par adjonction de nouveaux ensembles. Par exemple, le corpus bilingue HANSARD des débats du Parlement canadien peut être considéré comme une suite de corpus annuels, le dernier étant en construction.

Définition 17*. Un *corpus de traductions* est un corpus au sens précédent, contenant les traductions de tout ou partie de ses segments, dans une ou plusieurs langues.

Absence de traitement des corpus à proprement parler. SECTRA_W permet l'exploitation de corpus existants, l'évaluation des traductions et l'élargissement à plus de langues, mais n'est pas assez générique. Par exemple, il ne permet pas la création de nouveaux corpus (au sens classique), ni l'annotation selon des critères librement définissables par un linguiste. Ce progrès était prévu dans la thèse de Phap, mais n'a pas encore été réalisé.

Décalage de segments. SECTRA_W appelle les systèmes de TA pour produire des pré-traductions, et le fait maintenant via l'intergiciel TRADOH, mais il lui manque un sous-système pour synchroniser et gérer les segments source et les résultats de traduction. Parfois, un système de TA renvoie plusieurs résultats de traduction pour un même segment, et alors SECTRA_W les aligne avec d'autres, ce qui détruit la cohérence des bisegments, au moins jusqu'à ce que ces autres segments soient (re)traduits et que leurs résultats soient bien alignés.

Confusion de notions. En SECTRA_W, il y a plusieurs confusions entre des notions très différentes, ce qui bloque certaines avancées souhaitables. Par exemple, il y a confusion entre "mémoire de traductions" et « projet ». Quand l'utilisateur crée un nouveau « projet », en effet, le système crée une nouvelle MT. Mais un projet devrait pouvoir concerner plusieurs MT à la fois, ou une seule MT, existante, si par exemple on veut y ajouter des annotations telles que des arbres linguistiques, ou des caractérisations d'erreur. Or, cela ne justifie pas la création de nouvelles MT.

Il y a aussi dans l'actuel SECTRA_W une confusion entre « corpus » et « mémoire de traductions ». Cela est dû, historiquement, à l'évolution vers le traitement de pages Web. Auparavant, dans le projet EOLSS par exemple, on avait un vrai « corpus », c'est-à-dire une collection organisée de documents, avec leurs fichiers satellites (images, etc.) et leurs fichiers compagnons (pour EOLSS, des fichiers de graphes UNL). Après segmentation, on introduisait les segments dans la MT, et les documents n'étaient plus supposés changer. On pouvait donc facilement noter dans la base de données toutes les occurrences d'un segment[10].

Avec les pages Web, on a dû traiter des *pseudo-documents,* c'est à dire des documents qui ont le même nom (une *url*), mais qui peuvent changer d'un appel à l'autre. On a essayé d'adapter la méthode précédente, mais la solution retenue est source d'erreurs et augmente considérablement le temps de recherche d'un segment. Elle consiste en effet à associer une petite MT à chaque pseudo-document, liant ainsi corpus et MT. Quand on traite une page Web, on regarde dans sa MT s'il a déjà été vu, et si oui on prend la meilleure PE qu'on y trouve. Mais ce segment a pu être déjà post-édité dans un autre pseudo-document (une autre page Web), et du coup le système peut présenter une sortie de TA au lieu d'une PE déjà faite. De plus, si le système ne trouve pas le segment dans la MT du pseudo-document, il le cherche dans tous les autres documents du même "corpus", ce qui peut faire beaucoup, et cela d'autant plus que, si la MT est partagée, il y a des pseudo-documents liés à d'autres IMAG qui sont examinés, alors que, conceptuellement, ils sont dans un corpus différent.

La solution serait de reprendre toutes ces MT, de les fusionner avec la MT principale, et d'associer simplement à chaque pseudo-document la liste des segments déjà vus dans une des instances de ce pseudo-document, et, pour chaque instance, de construire un « document squelette » constitué du code html et de références aux segments présents dans cette instance.

[10] TM d'IBM et SDL TRADOS le font, ce qui permet de reconstituer la suite des segments des documents déjà traités, mais LIBELLEX ne le fait que depuis 2013.

I.1.1.3 Résumé des avancées et des limitations au niveau algorithmique (conception informatique)

I.1.1.3.1 Avancées

Post-édition "sans couture". SECTRA_W communique avec la partie IMAG, qui peut être considérée comme un "frontal" permettant la post-édition directement sur la page Web, ce qu'on appelle la post-édition "sans couture". Cela veut dire qu'on ne change pas de contexte entre lecture et correction d'une page Web. De plus, quand l'utilisateur fait de la post-édition dans la page Web, il peut visualiser la page Web source en parallèle. Un autre aspect de cette communication est qu'on peut passer de la post-édition d'un segment sur la page Web à sa post-édition dans la page (logique) de post-édition "avancée" de SECTRA contenant ce segment, puis revenir à la page Web, à l'endroit où on était, une fois qu'on a post-édité quelques segments en mode "avancé". En pratique, les post-éditeurs experts utilisent les deux modes. Le mode avancé leur permet de post-éditer très vite de nombreux segments, et le mode sans couture leur permet de voir l'ensemble de la page Web, et de corriger de petites erreurs qui sautent aux yeux alors qu'on les repère mal dans les petites zones de texte (*textareas*) de l'interface de PE de SECTRA_W.

Possibilité de partager des MT. Le fait qu'une MT puisse être partagée par une ou plusieurs IMAG est un avantage si les sites Web "élus" sont comparables, et plus précisément si leurs sous-langages diffèrent peu. C'est par exemple le cas de sites de tourisme, ou de sites de villes, ou de sites de cours d'informatique en ligne. En effet, des segments identiques (parfois longs) apparaissent alors, et la post-édition d'un site peut bénéficier à un autre.

Gestion des utilisateurs. SECTRA_W utilise le même type de gestion des utilisateurs que XWIKI, avec deux niveaux : individu et groupe, un individu pouvant appartenir à plusieurs groupes. Il serait préférable de séparer SECTRA_W du logiciel IMAG et de créer un service unique faisant le lien entre ces gestionnaires d'utilisateurs, par exemple fondé sur un annuaire du genre LDAP, de façon à éviter de demander à un contributeur de s'authentifier plusieurs fois alors qu'il a tous les droits nécessaires.

SECTRA_W associe à chaque utilisateur un profil général incluant des informations sur ses compétences et ses droits en général (par exemple, un utilisateur peut accéder à la MT DEMO1, mais pas aux autres MT). Il faudrait aussi associer à chaque utilisateur un raffinement de son profil, pour chaque projet. Par exemple, quelqu'un pourrait être administrateur mais pas post-éditeur dans un projet, et l'inverse dans un autre.

I.1.1.3.2 Limitations

Structure de la base de données. Quand l'administrateur crée un « projet » dans SECTRA_W, il crée une nouvelle « mémoire de traductions », qui donne lieu à la création d'une nouvelle base de données (BD) en MYSQL. Quand l'utilisateur crée une IMAG associée à cette MT, SECTRA_W crée 3 tables dans cette BD. La première table contient tous les segments source. La deuxième contient tous les segments cible traduits par un ou plusieurs système de TA. La troisième contient les post-éditions des segments post-édités en au moins une langue cible.

Par exemple, le site Web du LIG est en français. Ce site contient une centaine de pages Web, et son IMAG, donc aussi SECTRA_W, prévoit plus de 30 langues cible. Donc, à une page Web du site LIG, SECTRA_W associe une table source, une ou plusieurs tables contenant les TA, et plus de 30 tables contenant les segments post-édités. Cette structure de BD est difficile à manipuler, et augmente la complexité de la recherche de l'information. Cette limitation provient d'un mauvais choix au niveau du schéma conceptuel, et par suite du schéma logique des MT.

Passage à l'échelle. SECTRA_W, dans son état de 2011, était donc basé sur un schéma conceptuel de la structure des données, qui a mené à un schéma logique multipliant les tables, et ne permettant pas de traiter efficacement de grosses quantités de données. Nous savions donc qu'il faudrait intervenir à ce niveau pour pouvoir réellement passer à l'échelle.

I.1.1.4 Résumé des avancées et des limitations au niveau de l'implémentation

I.1.1.4.1 Avancées

Import de corpus pour l'évaluation. SECTRA_W permet d'importer divers corpus multilingues disponibles et de faciliter la sélection des corpus source pour organiser des campagnes d'évaluation. SECTRA_W accepte des fichiers d'entrée en format texte (*.txt*). Au niveau physique, chaque corpus[11] peut être constitué par un fichier source contenant des segments source, un ou plusieurs fichiers de traductions "candidates" (produites par différents systèmes de TA ou vers différentes langues) à évaluer, et un ou plusieurs fichiers de traductions de référence. Cependant, un corpus d'évaluation peut ne contenir que des segments source, car SECTRA_W permet de produire des traductions candidates par appel de TA, et des traductions de référence par post-édition en ligne. Au niveau interne, les segments dans les fichiers (source, candidats, et références) sont alignés.

Convivialité et souplesse de l'interface. L'interface de SECTRA_W permet de changer la taille des colonnes, et de cacher/montrer des colonnes en utilisant la souris ou un tableau de configuration. Les utilisateurs peuvent choisir le nombre de segments affichés à chaque fois sur l'interface. Leurs travaux sont enregistrés automatiquement et les données déjà manipulées sont distinguées des autres au niveau de la couleur du fond.

Mode de lecture et mode avancé. SECTRA_W/IMAG offre deux interfaces de post-édition. Le premier est intégré au mode de lecture. L'utilisateur peut faire la post-édition sur le document, et voir très facilement l'effet de la post-édition dans le contexte de lecture. Le deuxième est un mode « avancé ». L'utilisateur peut se concentrer sur le segment. Les segments sont bien présentés dans les pseudo-documents.

API. XWIKI permet de publier des services, et ainsi d'outiller leur utilisation par des logiciels tiers. Cela est fait dans SECTRA_W pour les fonctions de recherche exacte d'un texte dans un corpus (*TM exact search by API*) et pour mettre à jour la traduction mémorisée pour un texte dans un corpus (*Update TM by API*).

I.1.1.4.2 Limitations

XWiki et serveur de déploiement. SECTRA_W a été développé en utilisant la plate-forme XWIKI (version 1.3.1), la technologie AJAX, et les langages de script JAVASCRIPT, GROOVY et VELOCITY. Depuis 2010, SECTRA_W n'a pas suivi les mises à jour de XWIKI. Passer à une version plus récente de XWIKI comme la version 7 serait actuellement extrêmement lourd et risqué, car le langage VELOCITY (d'Apache Foundation) qui permet de traiter les BD MYSQL a lui-même changé. C'est pour cela que SECTRA_W est déployé sur TOMCAT 5, correspondant à l'ancien VELOCITY, et qu'on ne peut pas passer à TOMCAT 6 ou 7 pour l'instant. Enfin, une autre limitation est que XWIKI n'est pas très compatible avec certains navigateurs tels que IE, car il utilise des fonctions évoluées (comme AJAX) qui ne sont pas encore standardisées sur les différents navigateurs.

[11] Ce terme est utilisé par abus de langage dans le contexte des campagnes d'évaluation. Il ne s'agit de rien d'autre que d'une suite de tests formée d'un **ensemble** de paires (segment source, segment traduit), et pas d'un ensemble de **documents,** un document n'étant jamais réductible à l'ensemble de ses phrases.

Export. SECTRA_W offre une fonction capable d'exporter le résultat de la campagne TRANSAT, mais n'a pas de fonction permettant de réaliser l'export sur les MT génériques, comme la grande MT LAMETRO. Il n'y pas non plus de moyen de choisir les segments post-édités.

Absence de l'aide lexicale proactive annoncée. Dans sa thèse, C.P. HUYNH a présenté la conception d'une aide lexicale proactive intégrable à SECTRA_W, sous la forme d'un "minidictionnaire" précalculé et associé à chaque segment. De son côté, H.T. NGUYEN avait réalisé une fonction permettant de produire un minidictionnaire à partir d'une liste de mots, par lemmatisation, recherche dans une BD lexicale PIVAX, et formatage en HTML. Mais la jonction des deux n'a pas pu être réalisée avant leur départ du laboratoire.

I.1.2 Accès multilingue à des sites Web : le logiciel iMAG

I.1.2.1 Présentation générale

I.1.2.1.1 Bref historique et motivations

Le concept d'iMAG a été proposé par Ch. Boitet et V. Bellynck en 2005 (Boitet et al., 2005), étudié en 2006 et 2007 par les brais de projets d'étudiants, et a atteint l'état de prototype opérationnel grâce à C.P. HUYNH en novembre 2008 (Boitet et al., 2008), avec une première démonstration sur le site Web du laboratoire LIG. Il a été adapté au site Web DSR (Digital Silk Road) en avril 2009 par C.P. HUYNH, puis à plus de 80 autres sites Web grâce au configurateur réalisé par H.T. NGUYEN. Ces iMAG servent à des démonstrations et à des expérimentations. Citons ici (Boitet et al., 2010).

> *Une iMAG est une passerelle interactive d'accès multilingue [interactive Multilingual Access Gateway], ressemblant beaucoup à Google Translate, à première vue : on donne une URL [site Web de départ] et une langue d'accès, et on navigue ensuite dans cette langue d'accès. Lorsque le curseur passe sur un segment [le plus souvent une phrase ou un titre], une palette montre le segment source et propose de contribuer en corrigeant le segment cible, en fait en post-éditant un résultat de traduction automatique [TA]. Avec Google Translate, la page ne change pas après la contribution, et si une autre page contient le même segment, sa traduction est toujours le résultat de TA grossière, pas la version polie post-éditée. La boîte à outils de traduction Google Translation Toolkit permet de traduire par TA et ensuite de post-éditer en ligne des pages Web complètes tirées de sites tels que Wikipédia, mais, de nouveau, les segments corrigés n'apparaissent pas quand on regarde plus tard la page de Wikipédia dans la langue d'accès.*
>
> *En revanche, une instance d'iMAG [dite iMAG-S] est dédiée à un site Web élu [S], ou plutôt au sous-langage défini par une ou plusieurs URL et leur contenu textuel. C'est un bon outil pour rendre S accessible dans beaucoup de langues, immédiatement et sans responsabilité éditoriale. Les visiteurs de S ainsi que des post-éditeurs et des modérateurs payés ou non contribuent à l'amélioration continue et incrémentale des segments textuels les plus importants, et éventuellement de tous.*
>
> *Une instance iMAG-S contient une MT. Les segments sont pré-traduits non pas par un système de TA unique, mais par un ensemble [sélectionnable] de systèmes de TA gratuits. Systran et Google sont principalement utilisés aujourd'hui, mais des systèmes spécialisés, développés à partir de la MT post-éditée, vont bientôt être utilisés. Les visiteurs de S ainsi que des post-éditeurs et des modérateurs payés ou non contribuent à l'amélioration continue et incrémentale des segments textuels les plus importants, et éventuellement de tous.*
>
> *Les plates-formes contributives puissantes SECTra_w sont utilisées pour supporter les MT. Les pages traduites sont construites avec les meilleures traductions des segments disponibles au moment de l'accès. Pendant la lecture d'une page traduite, il est possible non seulement de contribuer au segment sous le curseur, mais aussi de passer de façon transparente sous l'environnement de post-édition en ligne de*

SECTra_w, muni de bonnes fonctions de filtrage et de recherche-remplacement, et ensuite de revenir dans le contexte de lecture.

I.1.2.1.2 Lancement de maquettages

Deux « maquettes » d'IMAG ont été étudiées avant novembre 2008, par Mohammad Daoud dans son M2R (Daoud, 2007), et par Carlos Ramisch dans son stage ENSIMAG (Ramisch, 2008).

Le premier n'a pas pu produire la maquette prévue, car il avait voulu écrire lui-même un segmenteur de pages Web (en "segments textuels", c'est à dire en phrases ou titres), alors qu'on l'avait prévenu qu'il s'agissait d'un problème difficile et qu'on lui avait demandé d'utiliser indirectement GT comme segmenteur.

Dans le cadre d'un projet du LIG appelé "IMAG/LIG", financé par 5 ou 6 mois de bourses CNRS du LIG, C. Ramisch a voulu lui aussi construire un prototype complet à partir de zéro. Il avait prévu d'utiliser LINGPIPE[12] pour la segmentation, mais il aurait fallu l'adapter aux pages Web. Il n'a pas eu le temps d'implémenter ce qu'il avait spécifié, car c'était un trop gros travail pour le temps disponible. Malheureusement, une fois son projet soutenu, il n'a pas pu "passer la main", étant pris à la fois par les cours de M2R et un stage à XRCE.

I.1.2.1.3 Prototype opérationnel de C. P. HUYNH

Durant l'été 2008, dans le cadre du même projet IMAG/LIG, C. P. HUYNH et H. T. NGUYEN ont travaillé sur le site Web du LIG, pour segmenter ses pages et aligner les segments déjà traduits en anglais. Le meilleur segmenteur qu'ils avaient trouvé était justement celui de GT. L'idée était qu'ils fournissent au logiciel IMAG en construction une MT initiale pour le site du LIG, en appelant GT sur les segments non traduits. Comme ce logiciel attendu n'était pas disponible, Phap a proposé de le réaliser très vite, par extension de SECTRA, qui était déjà assez mûr, ayant été utilisé par des dizaines de contributeurs dans le cadre du projet EOLSS-UNDL.

En à peine 3 semaines, juste avant de partir pour un stage prédoctoral international au Japon (NII), il a produit un premier prototype opérationnel, et une première IMAG, l'IMAG-LIG, qui a pu être démontrée à la direction du LIG (B. Plateau) vers le 20/11/2008.

Le fonctionnement est le suivant. La page consultée est accédée via la passerelle IMAG, qui réécrit les url de façon à ce que les pages accédées par la suite lors de la navigation passent aussi par elle. La page à traduire est envoyée à GT, on récupère le résultat, puis on détecte les segments marqués par GT. On les cherche alors dans la MT, et on les ajoute s'ils n'y sont pas. Dans les deux cas, on met aussi dans la MT associée à l'IMAG les (pré)traductions produites par GT. La page traduite est reconstruite à partir du code HTML de la page source, des meilleures traductions des segments (trouvées dans SECTRA_W), et de code JavaScript spécifique.

Au tout début, l'IMAG-LIG produite par le projet IMAG/LIG contenait la traduction d'environ environ 2000 segments du site Web du LIG. Elle en contient maintenant près de 15000.

I.1.2.1.4 Évolutions depuis début 2010

Début 2010, une IMAG dédiée a été créée pour la Métro de façon à permettre l'accès à son site Web en chinois et en anglais à l'occasion de l'exposition universelle de Shanghai qui eut lieu en octobre 2010. Ce fut la première expérience en grandeur réelle.

[12] http://alias-i.com/lingpipe/

Ensuite, plusieurs sites IMAG à mémoire de traductions dédiée ont été créés. La table suivante donne les URL des sites principaux.

Tableau 3 : Liste des IMAG à MT dédiée construites depuis 2010

Nom abrégé	Accès au site élu : remplacer dans http://service.aximag.fr/xwiki/bin/view/imag/home home par :	Page d'accueil de chaque site :
Expo de Shanghai	lametro	http://www.lametro.fr
Projet Traouiero	traouiero	http://getalp.imag.fr/traouiero/
Projet MACAU	macau	http://tools.aximag.fr/macau/chamilo-macau/index.php
AMIES	AMIES	http://www.agence-maths-entreprises.fr
Grenoble INP	GINP	http://www.grenoble-inp.fr/accueil/
Powers	Powers	http://www-clips.imag.fr/geod/User/laurent.besacier/TRANSLATION-EXP/The_Book_of_Me.html
Michelin	Michelin	http://www.michelin.fr
Bio Clean	Bio-Clean	http://projet-bioclean.3beesonline.com/fr/accueil/
Projet Akenou	akenou	http://getalp.imag.fr/akenou/

De nombreuses IMAG ont aussi été créées sur des MT partagées. En particulier, une a été créée dans le cadre d'un projet de l'ISCC (institut des sciences de la communication), un institut du CNRS dirigé par Dominique WOLTON.

I.1.2.2 Conception générale avec implémentation "intégrée"

Voici un peu plus de détails sur l'implémentation du logiciel IMAG, "intégré" à SECTRA. Lors du projet ANR TRAOUIERO, V. Bellynck a prévu de la rendre plus modulaire, grâce à l'introduction d'un "relais IMAG", de l'utilisation de TRADOH pour appeler les systèmes de TA, et de l'utilisation de SEGDOC pour effectuer la segmentation-normalisation, mais cette transformation n'est pas encore terminée.

Voici le schéma de l'architecture logicielle actuelle.

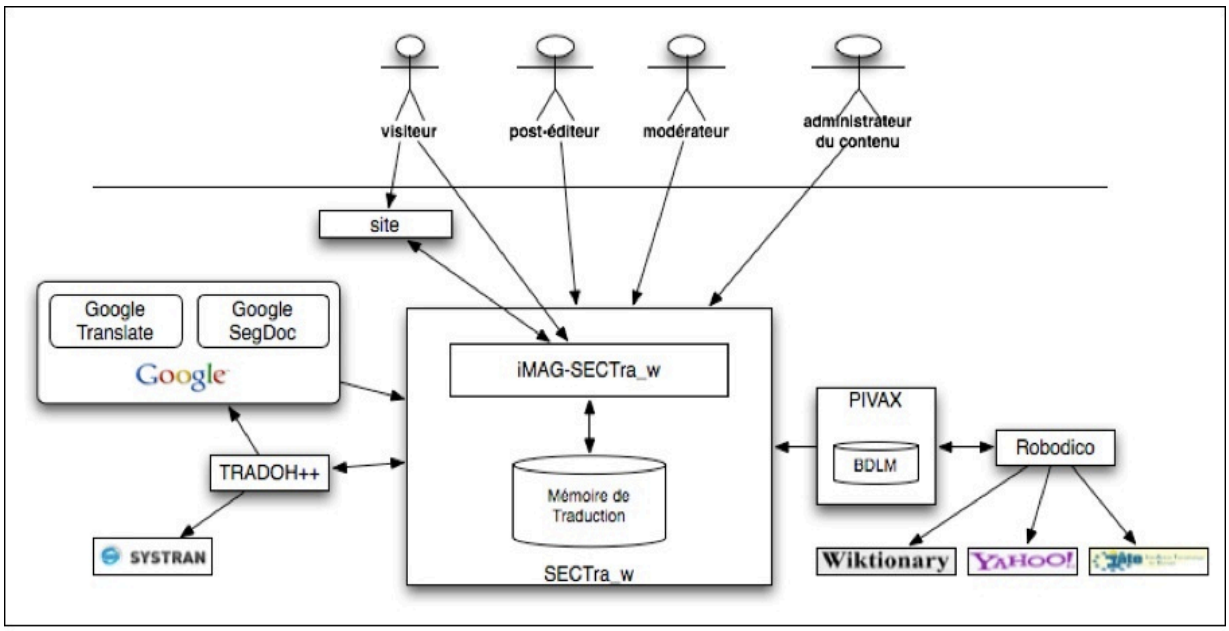


Figure 1 : Architecture générale d'une iMAG pour un site élu

Lorsque le visiteur demande à accéder au site élu S dans une autre langue que la langue source, la page d'accueil de S est envoyée à GT de façon à récupérer le découpage en segments de texte, et les traductions de chacun des segments par cet outil, s'il supporte les langues désirées en source et cible. Ultérieurement, lors de la navigation interne au site S, chaque page de S est traitée par le même procédé.

D'autres outils de traduction automatique comme REVERSO, SYSTRAN ou BING peuvent être utilisés pour prétraduire chaque segment. Les segments traduits sont mémorisés comme les correspondants (dans la langue de visite) des segments d'origine.

Dans la page Web reconstruite par l'IMAG, le texte de chaque segment en langue source est remplacé par le texte correspondant en langue cible, et le code nécessaire à l'insertion d'une bulle contextuelle (surgissant au survol par le curseur de chaque segment) est ajouté avec le texte dans la langue source, le couple (segment en langue source, segment en langue cible) constituant une correspondance de segments pour le système SECTRA_W.

Figure 2 : Capture d'écran de l'iMAG LIG-LAB en chinois

La bulle contextuelle est appelée « bulle-iMAG ». Elle montre le texte du segment en langue source, et présente un bouton dont l'action associée transforme la bulle en une bulle de contribution : elle contient alors un formulaire dont une zone de saisie de texte (« *textarea* ») permet de modifier à la volée la traduction proposée. Le visiteur peut alors contribuer à l'amélioration de la traduction de ce segment de la page Web courante selon son droit d'édition et le mode de publication du site. Une fois que la contribution est envoyée, elle est stockée dans une table associée à la page en question, identifiée par son URL comme un pseudo-document, et il y a une duplication dans une table globale des segments source-cible pour chaque site. Donc, dans une page, s'il y a un même segment (par exemple un item de menu, ou une phase type) qui est déjà traduit dans une autre page, la traduction est proposée avec la mention « *Mémoire de traductions* ».[13]

Dans l'architecture complète prévue, on a intégré l'aide lexicale proactive, ce qui explique pourquoi PIVAX apparaît dans le schéma suivant, tiré de la thèse de H.T. NGUYEN. On construit des instances spécifiques de composants SECTRA_W (pour la MT), de PIVAX (pour la base lexicale) et d'une IMAG dédiée. Il y a une connexion entre ces instances locales et une instance centrale de données afin d'initialiser les données pour un nouveau site.

Dans cette architecture, SECTRA_W, PIVAX, et les IMAG communiquent les uns avec les autres pour effectuer des tâches. SECTRA_W communique avec PIVAX pour demander des minidictionnaires, et inversement PIVAX communique avec SECTRA_W pour extraire des unités lexicales à partir des corpus.

Les IMAG communiquent avec SECTRA_W pour demander des traductions, soit à partir de MT, soit à partir de systèmes de TA, et inversement SECTRA_W communique avec les IMAG pour demander des segments source, des fichiers squelette, et des dictionnaires de hors-texte.

[13] Remarque : tous les segments d'un site élu sont actuellement supposés écrits dans une même langue.

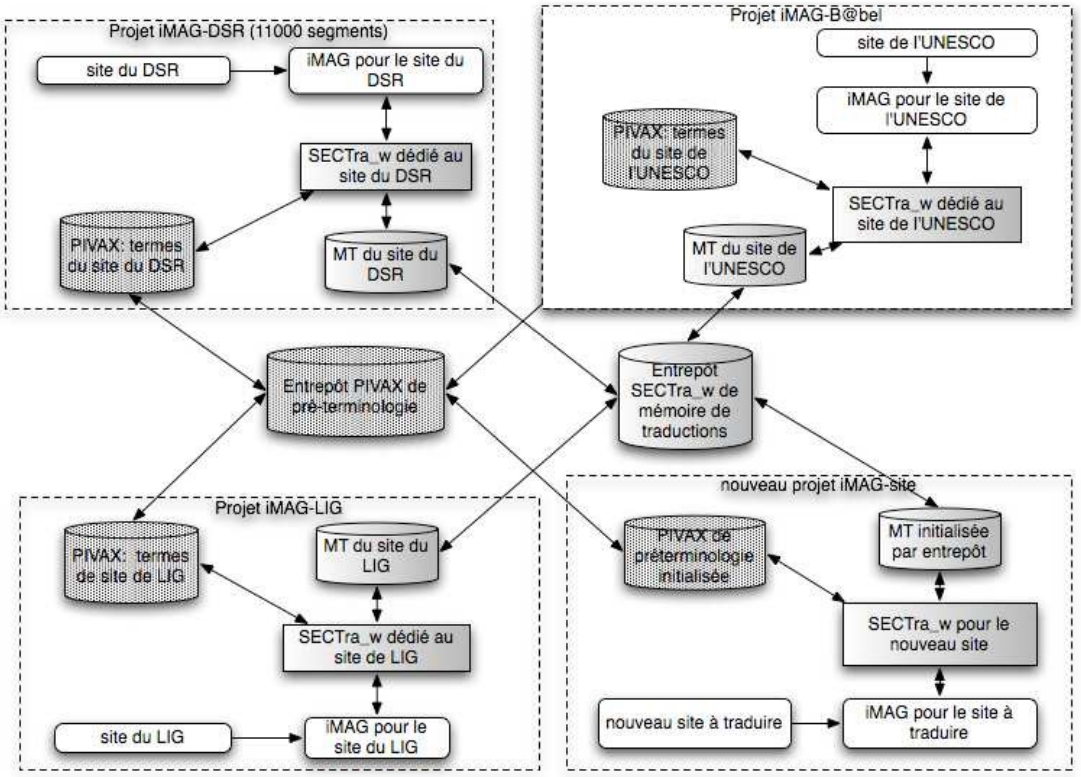


Figure 3 : Architecture par agents SECTra_w, iMAG, PIVAX (Nguyen, 2010)

I.1.2.3 Résumé des avancées et des limitations du logiciel iMAG

I.1.2.3.1 Avancées

Introduction d'une possibilité de modération. Il faut dans certains cas pouvoir éviter que des contributeurs anonymes n'améliorent pas les traductions, mais au contraire les remplacent par des textes injurieux, diffamatoires ou autres. C'est un risque non négligeable si le site élu est, par exemple, lié à des activités politiques. C'est pourquoi, à l'occasion du travail sur l'iMAG de la Métro (le "grand Grenoble"), H. T. NGUYEN a introduit une possibilité de modération, inspirée de celle de Wikipédia. En mode "modéré", un contributeur peut modifier une traduction comme il veut, et il la verra à l'écran, mais elle ne sera acceptée définitivement dans la MT qu'après avoir été validée par un modérateur ayant les droits suffisants.

Visualisation de la fiabilité supposée des segments en langue d'accès. La case à cocher « *fiabilité* » permet de voir, pour chaque segment, s'il est le résultat brut d'une TA, s'il a été produit par un contributeur anonyme (ou enregistré mais pas connecté), ou bien par un contributeur connecté. Cette visualisation ne modifie pas la couleur du fond, ou la couleur des fontes. On se borne à encadrer chaque segment par deux parenthèses spéciales, respectivement rouges, orange et vertes dans les trois cas ci-dessus.

Association d'un niveau de fiabilité supposée et d'un score de qualité à chaque segment traduit. Le terme "niveau de fiabilité" n'est pas très bon, mais nous n'en avons pas trouvé de meilleur. C'est en fait l'origine d'une traduction stockée dans la MT. Il est exprimé par 1 à 5 étoiles : ☆ pour une traduction mot-à-mot (pidgin), ☆☆ pour une traduction automatique, ☆☆☆ pour un humain connaissant les deux langues, ☆☆☆☆ pour un traducteur professionnel, et ☆☆☆☆☆ pour un traducteur certifié par le site Web associé à l'iMAG. Ce niveau est de même nature que les 3 couleurs des parenthèses spéciales, et simplement un peu plus précis. Pour les couleurs, on a voulu se limiter à 3, pour des raisons ergonomiques.

Le *score de qualité* est une note entre 0 et 20. Chaque contributeur inscrit dans SECTRA_W a un score par défaut pour chaque couple de langues sur lequel il peut intervenir, par exemple 12/20 pour "moyen", ou 14/20 pour "bon". Il peut lui-même modifier ce score pour noter ce qu'il pense de la qualité d'un segment qu'il a post-édité (ou qu'un autre a post-édité, mais on le fait plus rarement). Par exemple, on mettra 9/20 si on a conscience de ne pas avoir trouvé un équivalent convenable, et 16/20 si on est sûr d'avoir produit une très bonne traduction.

Possibilité de post-éditer directement sur la page Web. L'utilisateur peut faire de la post-édition sur une page Web en langue cible, directement dans le "contexte de lecture". Ses post-éditions sont présentées immédiatement sur la page Web[14], et le résultat est sauvegardé dans la MT associée. Beaucoup de post-éditeurs utilisent en pratique les deux modes de post-édition. Dans le mode avancé, on va très vite, mais on ne voit pas le contexte global de la page, ni la présentation. Quand on revient sur la page, on voit des coquilles, des pronoms incorrects, etc., et on les corrige directement sur la page.

Visualisation parallèle d'une page en cible et source. Quand on utilise une IMAG, on peut post-éditer sur la page Web cible, en visualisant en parallèle la page Web source.

I.1.2.3.2 Limitations

Le logiciel IMAG actuel présente trois groupes de défauts du point de vue fonctionnel et architectural.

- Les dialogues de l'interface de tous les composants devraient être multilingues. Pour l'instant, ils sont seulement en anglais.
- Des fonctionnalités critiques ne sont pas implémentées, ou pas comme il faudrait pour obtenir les résultats attendus.
    - **Gestion de la mémoire de traductions**. Il faudrait éliminer la « *sous-MT* » associée à chaque pseudo-document.
    - **Contextes linguistiques.** Il faudrait les prendre en compte.
    - **Consultations et contributions lexicales.** Il faudrait intégrer une aide lexicale non seulement au niveau de l'interface de PE de SECTRA, mais aussi dans la "palette IMAG".
    - **Visualisation de l'apport de la PE.** Il faudrait pouvoir voir non seulement la page source, mais aussi la page cible en mode « trace » (TA brute initiale et vue intuitive des modifications dues à la PE).
    - **Mesure des temps de PE.** La communication des temps de PE et des scores de qualité entre l'interface IMAG et SECTRA_W n'était pas satisfaisante.
    - **Liaison entre les deux interfaces de PE.** Il faudrait introduire une liaison directe entre un segment sur la page Web et le segment correspondant dans l'interface de SECTRA_W, et dans les deux sens. Pour l'instant, on est limité à l'accès à une « *page logique* » de SECTRA à partir d'un segment.
- Le logiciel IMAG actuel n'est toujours pas protégé contre des attaques où des robots font des requêtes qui sont autorisées pour tous les visiteurs. Cela conduit à une surcharge du système et donc à des dénis de service.

L'architecture actuelle repose essentiellement sur l'exploitation d'un seul logiciel dans lequel tout est géré, ce qui n'est pas compatible avec les besoins en modularité.

[14] Il faut rafraîchir la page pour que les parenthèses encadrant ce segment réapparaissent.

I.1.3 Travaux comparables et idées directrices pour le futur

I.1.3.1 Travaux comparables sur la création et l'exploitation des corpus multilingues

Pour améliorer notre système, nous avons essayé de étudier systématiquement les travaux comparables pour trouver les points intéressants sur la manipulation des corpus multilingues. Nous avons trouvé 2 types de systèmes, (1) les systèmes de stockage et d'accès, et (2) les systèmes de collecte directe et d'exploitation de corpus multilingues.

I.1.3.1.1 Systèmes de stockage et d'accès

Systèmes de stockage basé sur un service Web. Outre le stockage, ces systèmes permettent la navigation, le téléchargement, et la visualisation des données.

Certains fournissent aussi quelques fonctions utiles, comme la recherche dans un sous-ensemble de textes ou de phrases d'une certaine catégorie, ou la production de concordances. Par exemple, le site Web OPUS[15] (Tiedemann, 2012) contient une grande quantité de corpus multilingues. L'utilisateur peut télécharger les corpus à partir des liens du site Web. Dans le Tableau 4, nous présentons les sites Web principaux permettant de partager des corpus parallèles.

Tableau 4 : Exemples de sites Web de partage de corpus parallèles

Nom du corpus	Lien
Europarl	http://statmt.org/europarl/
JRC-Acquis	https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
Parallel corpora at PELCRA	http://pelcra.pl/new/
UM-Corpus (Tian et al., 2014)	http://nlp2ct.cis.umac.mo/um-corpus/
The Bible (Resnik et al., 1999)	http://www.umiacs.umd.edu/~resnik/parallel/bible.html

Systèmes de plus haut niveau pour les brevets. Un système de gestion de documents de brevets est un autre type de système de gestion de corpus. Généralement, un tel système contient une série de fonctionnalités permettant de manipuler les brevets, comme la recherche, la comparaison, l'extraction, etc.

Par exemple, le système PATENTSCOPE (Pouliquen and Mazenc, 2011) permet d'accéder aux demandes internationales selon le Traité de coopération en matière de brevets (PCT[16] pour PATENT COOPERATION TREATY) en texte intégral, le jour même de leur publication, ainsi qu'aux documents de brevet des offices de brevets nationaux ou régionaux participants.

Systèmes plus génériques. Ce type de système est beaucoup utilisé pour gérer et partager plusieurs types de données. Par exemple, META-SHARE (Federmann et al., 2012) est utilisé pour partager des corpus. Ce type de système a but pour gérer les publications, l'ensemble des données, des fichiers multimédia, etc. Il permet en particulier d'accéder à des corpus et de les télécharger.

I.1.3.1.2 Systèmes de collecte directe et d'exploitation

Quelques systèmes contiennent des outils ou des langages de scripts utilisables pour collecter et manipuler quelques types de corpus, et en particulier des corpus multilingues (comparables ou parallèles).

Voici trois exemples, ERIM-COLLECTE, de notre labo, BITEXTOR, et PACO2.

[15] http://opus.lingfil.uu.se

[16] http://fr.wikipedia.org/wiki/Traité_de_coopération_sur_les_brevets

ERIM-Collecte : ERIM-COLLECTE (Fafiotte, 2004) a été construit dans le cadre du projet ERIM (Environnement Réseau pour l'Interprétariat Multimodal), visant à créer un environnement en réseau pour l'aide à la communication orale multilingue, dans lequel on peut collecter des corpus de dialogues parlés spontanés bilingues et multilingues. ERIM-COLLECTE permet l'enregistrement systématique des actes et données de l'interaction, pour tous les participants (deux locuteurs ou plus, un interprète ou plus). L'enregistrement est fait localement lors de la conversation. En fin de dialogue, les descripteurs et fichiers produits localement sont transmis à un serveur de collecte, où ils sont regroupés et structurés.

Bitextor : BITEXTOR (Esplá-Gomis, 2009) est un outil d'extraction de bi-textes à partir de pages Web. Tout d'abord, il télécharge les pages Web visées. Ensuite, il normalise les fichiers téléchargés. Enfin, il compare les fichiers en utilisant divers critères pour extraire les phrases parallèles, que nous appellerons plus précisément "bisegments". Dans la terminologie actuelle, un "corpus parallèle" est une collection de bisegments, et bien plus rarement une collection de vrais documents multilingues alignés au niveau des segments (phrases et titres).

PaCo^2 : PACO^2 (*Parallel Corpora Collector*) (San Vicente and Manterola, 2012) est un outil de collecte de corpus parallèles à partir du Web. Il met en œuvre les techniques les plus modernes (à l'état de l'art) permettant de trouver des bi-textes dans un domaine Web qui en contient. De plus, il peut trouver de tels domaines Web automatiquement.

I.1.3.1.3 Systèmes d'extraction indirecte

Les corpus comparables sont considérés comme des ressources importantes pour collecter des corpus parallèles, surtout quand on n'en trouve pas directement. Ici, le terme "corpus comparable" dénote une collection de "vrais" documents, par exemple des pages Web, ou des documents en WORD ou en PDF, qui sont de typologies comparables (par exemple, des articles de WIKIPEDIA sur le même sujet, ou des recettes de cuisine, ou de la documentation technique de produits similaires).

Plusieurs méthodes ont été proposées pour extraire (ou construire) un corpus parallèle à partir de corpus comparables : elles utilisent les notions de *documents comparables* (Resnik and Smith, 2003) (Smith et al., 2010), de *phrases comparables* (Zhao and Vogel, 2002), (Munteanu and Marcu, 2005), et des ressources comme des bisegments ou des dictionnaires bilingues (Munteanu and Marcu, 2006), (Cettolo et al., 2010). Dans le cadre de sa thèse, Thi Ngoc Diep DO (Do, 2011) a aussi proposé 3 approches d'extraction de textes parallèles à partir de documents comparables.

I.1.3.2 Travaux comparables sur la post-édition de TA

I.1.3.2.1 Introduction

> *« Postediting/post-editing [PE] is by far most commonly referred to as a task related to Machine Translation [MT] and has been previously defined as the "term used for the correction of machine translation output by human linguists/editors." Another good summary statement indicates that "post-editing entails correction of a pre-translated text rather than translation 'from scratch'." In basic terms, the task of the post-editor is to edit, modify and/or correct pre-translated text that has been produced by a machine translation system from a source language into (a) target language(s). »*
>
> *— ALLEN, Jeffrey. 2003 (Allen, 2003)*

Les premières études sur la post-édition sont apparues dans les années 80, liées aux outils associés aux systèmes de TA, puis aux outils associés aux MT (Senez, 1998). Les premières parurent au Japon, et seulement en japonais.

Dès les années 1985, il y avait une trentaine de systèmes de TA commerciaux au Japon, utilisables sur miniordinateur (AS-TRANSAC de Toshiba) ou poste de travail, ou PC (comme DUET-2 de SHARP), et cette activité était enseignée et pratiquée par des professionnels de la traduction. Ailleurs, les systèmes et leur usage étaient en retard d'un quinzaine d'années à cause de l'arrêt des financements provoqué par le rapport ALPAC (12/1966—1/1967). Le centre de recherche en TA de CMU ne fut ainsi créé qu'en 1985. C'est seulement en 1999 que, pour élaborer des directives appropriées et de la formation, les membres de l'*Association for Machine Translation in the Americas* (AMTA) et *de l'European Association for Machine Translation* (EAMT) créèrent un groupe d'intérêt pour la post-édition, le *Post-editing Interest Group*, ou *Post-editing SIG* (Allen, 1999).

Depuis les années 1990, avec le développement des systèmes de TA, la post-édition de traductions automatiques s'est développée dans le monde occidental, mais nettement plus lentement qu'au Japon. En fait, les gros bureaux de traducteurs professionnels ont longtemps préféré utiliser des systèmes à mémoires de traductions comme TRANSLATION MANAGER (TM2™) d'IBM, TRADOS, TRANSIT (de STAR), ou DÉJÀ VU, même s'ils étaient couplés à des systèmes de TA, comme LMT pour TM2™. Selon (Boitet, 1996), la raison en est que les dictionnaires "main gauche" intégrés à ces systèmes et constamment mis à jour par les traducteurs suivaient l'évolution terminologique, alors que ceux des systèmes de TA évoluaient beaucoup plus lentement, de sorte que post-éditer des sorties de TA devenait plus long qu'utiliser les suggestions d'un système à mémoire de traductions (MT). Depuis que SYSTRAN a été mis sur minitel (en 1984), puis en service gratuit sur Internet (en 1994), ce sont les traducteurs indépendants qui se sont d'abord mis à utiliser des résultats de TA, en post-édition s'ils étaient assez bons, et comme "dictionnaire en contexte" sinon. Il y a quelques exceptions, dans des situations où les utilisateurs et les développeurs des systèmes de TA étaient en synergie. Ce fut le cas à la PAHO (*Pan American Health Organization*, Washington, et OMS, Genève, avec les systèmes SPANAM, ENGSPAN, puis PAHOMTS (Vasconcellos and León, 1985)), et aussi à Taipeh (Hsu and Su, 1995).

L'autre cas où la post-édition de résultats de TA a depuis très longtemps été très appréciée, et utilisée par des traducteurs professionnels, est celui où le système de TA est spécialisé à un sous-langage précis. C'est le cas du système METEO au Canada (1981-2001)[17], dédié aux *bulletins*[18] météorologiques, du système ALTFLASH de traduction des "brèves" du Nikkei (en mars 1998) (Uchino et al., 2001), et surtout des "accélérateurs de traduction" (Pouliquen et al., 2013) construits par B. POULIQUEN à partir de mémoires de traductions contenant les bisegments produits par des traducteurs professionnels pendant plusieurs années (11 ans pour les systèmes construits avec Moses pour l'ONU à Genève).

Un nouvel usage de la post-édition est apparu depuis quelques années. Il s'agit non pas d'améliorer les traductions automatiques avant qu'elles ne soient "consommées", mais d'en post-éditer une partie pour augmenter et améliorer la MT à partir de laquelle un système de TA est construit. Ce sont donc des aides au développement et à l'amélioration de systèmes.

Nous présentons maintenant les systèmes destinés uniquement à aider les traducteurs, puis les systèmes destinés à améliorer des systèmes de TA.

[17] https://en.wikipedia.org/wiki/METEO_System .

[18] et pas aux *situations* ni aux *alertes*.

I.1.3.2.2 Systèmes destinés uniquement à aider les traducteurs

Libellex. LIBELLEX[19] est l'un des deux produits-phare de la société *Lingua et Machina* (L&M). Il s'agit d'une plate-forme de traitement multilingue d'aide à la traduction qui repose sur un système de gestion de mémoires de traductions. Il vise à constituer une mémoire collective d'entreprise pour tout ce qui concerne la communication multilingue. LIBELLEX intègre un module de traitement de la terminologie (import/export à partir de ressources du client, extraction de terminologie, gestion de terminologie, etc.), et un module pour créer et utiliser des systèmes de TA de type MOSES (alignement de corpus, entraînement, appel de systèmes de TA). Il peut être utilisé en mode SAAS[20] ou installé en mode licence sur des postes de travail informatiques.

SDL Trados Studio. La société TRADOS commercialise un outil de THAM appelé SDL TRADOS STUDIO, qui a environ 200 K utilisateurs (plus de 200 K licences du produit[21]). Elle fournit des solutions innovantes en matière de logiciels de traduction à toute la chaîne logistique de traduction, y compris aux traducteurs indépendants, aux prestataires de services linguistiques, aux services internes de traduction et aux établissements universitaires. Cet outil se compose d'un gestionnaire de terminologie (SDL MULTITERM 2015), et d'un module d'alignement qui permet d'initialiser la MT utilisée à partir de documents déjà traduits.

I.1.3.2.3 Systèmes destinés à améliorer des systèmes de TA

MateCat. MATECAT (Schwenk et al., 2015) est un système visant à améliorer l'intégration d'un système de TA et de la traduction humaine (TH). L'objectif est d'améliorer la productivité des traducteurs professionnels et leur expérience du travail avec la TA. MATECAT propose une prétraduction par TA pour chaque phrase. Un système de TA basé sur MOSES est intégré dans MATECAT. Pour obtenir une bonne qualité de traduction et un gain de productivité, le système de TA est adapté au domaine (par exemple, la finance). MATECAT permet d'améliorer le système de TA en utilisant les résultats de post-édition, et il a intégré « Online model adaptation (Bertoldi et al., 2014) » de Moses pour améliorer dynamiquement le système de TA.

CASMACAT. CASMACAT (Alabau et al., 2013) (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation) est un système pour la traduction assistée par ordinateur et l'étude scientifique de la traduction humaine. Pour l'amélioration des systèmes de TA (MOSES), CASMACAT propose trois méthodes efficaces en ligne pour mettre à jour les systèmes de TA : (1) Il infère de nouvelles règles à partir des segments post-édités, et les ajoute au modèle de traduction ; (2) il utilise les segments post-édité pour mettre à jour le modèle de langue ; et (3) il met à jour les paramètres discriminants de TA dans une étape MIRA (Chiang, 2012) et (Hasler et al., 2011).

I.1.3.3 Idées directrices pour faire évoluer SECTra_w/iMAG

I.1.3.3.1 Rendre SECTra configurable au niveau des corpus

Un SECTRA[22] configurable au niveau des corpus devrait permettre non seulement de définir le descripteur du corpus (le nom du corpus, les langues source/cible, le domaine, la taille, l'auteur, etc.), mais aussi de présenter la relation entre le corpus et les fichiers compagnons, de choisir le dictionnaire, de manipuler les corpus (fusion, séparation, comparaison, filtrage, etc.), le format d'import/export, le codage, etc.

---

[19] http://libellex.fr

[20] https://fr.wikipedia.org/wiki/Logiciel_en_tant_que_service

[21] http://www.translationzone.com/fr/about

[22] Nom commun proposé dans la thèse de HUYNH Cong Phap (2010).

I.1.3.3.2 Trouver une modélisation

Nous souhaitons trouver une modélisation convenable pour les corpus multilingues, éventuellement complexes. Nous nous sommes inspiré de l'idée de structure de dictionnaire (les définitions sont présentées dans la thèse de M. MANGEOT (Mangeot, 2010)). Nous souhaitons construire une nouvelle version de SECTRA sur des corpus de traductions aussi variés et aussi gros que possible, comme JR-ACQUIS, EUROPARL, HANSARD, EOLSS augmenté par l'UNU-UNDL, ERIM, et surtout le corpus de brevets CLEF-IP (60 Go). Faire cela est nécessaire pour arriver à développer une méthode innovante et pertinente de modélisation des corpus multilingues alignés, et la partie du futur langage de programmation du système SECTRA qui permettra de décrire le type de chaque nouveau corpus (macrostructure, microstructure, comme pour les bases lexicales, et peut-être aussi "mésostructure" pour définir l'organisation des sous-documents, des fichiers satellites, et des documents ou fichiers compagnons).

I.1.3.3.3 Rendre SECTra programmable

Parmi les fonctions de SECTRA_W qu'on souhaite rendre programmables, il y a la synchronisation des données (en ligne ou hors ligne), l'import, l'export, la recherche, la sélection des données, et l'apprentissage incrémental de systèmes de TA à la MOSES.

Il faudrait que SECTra fournisse un support à la programmation, permettant de gérer et de traiter les corpus, et aussi des opérations motivées par l'utilisation dans des campagnes d'évaluation (par exemple, programmer certaines "mesures d'évaluation", comme BLEU, NIST ou TER), ou des opérations motivées par l'utilisation dans des projets de post-édition (par exemple, appel à un ou plusieurs systèmes de TA).

I.2 Améliorations de SECTra_w dans le cadre du projet Traouiero

I.2.1 Extension de fonctions existantes

Dans le cadre du projet TRAOUIÉRO, nous avons travaillé sur l'amélioration de SECTRA_W. Tout d'abord, nous avons amélioré la fonction d'appel de systèmes de TA. Elle permet de faire exécuter ou réexécuter la traduction automatique, par tel ou tel système, sur tout un corpus ou sur une sélection de documents.

Ensuite, nous avons refait la conception de la communication entre TRADOH et SECTRA_W, et réalisé la procédure d'appel des systèmes de TA via TRADOH. Enfin, nous avons participé à l'introduction de la possibilité pou un utilisateur d'ajouter à TRADOH des plugins permettant d'augmenter la collection des systèmes de TA appelables, et nous avons réalisé et intégré plusieurs plugins, dont un permettant d'appeler n'importe quel système MOSES.

I.2.1.1 Appel de systèmes de TA sur un corpus ou sur une sélection de documents

Pour ajouter de nouvelles traductions automatiques dans SECTRA_W, nous lui avons ajouté une nouvelle fonctionnalité. Dans la Figure 4, nous montrons comment choisir la langue cible et le système de TA dans l'interface de *Translate Corpus*. Une traduction via une langue pivot est aussi possible (cocher l'option *Pivot langage*).

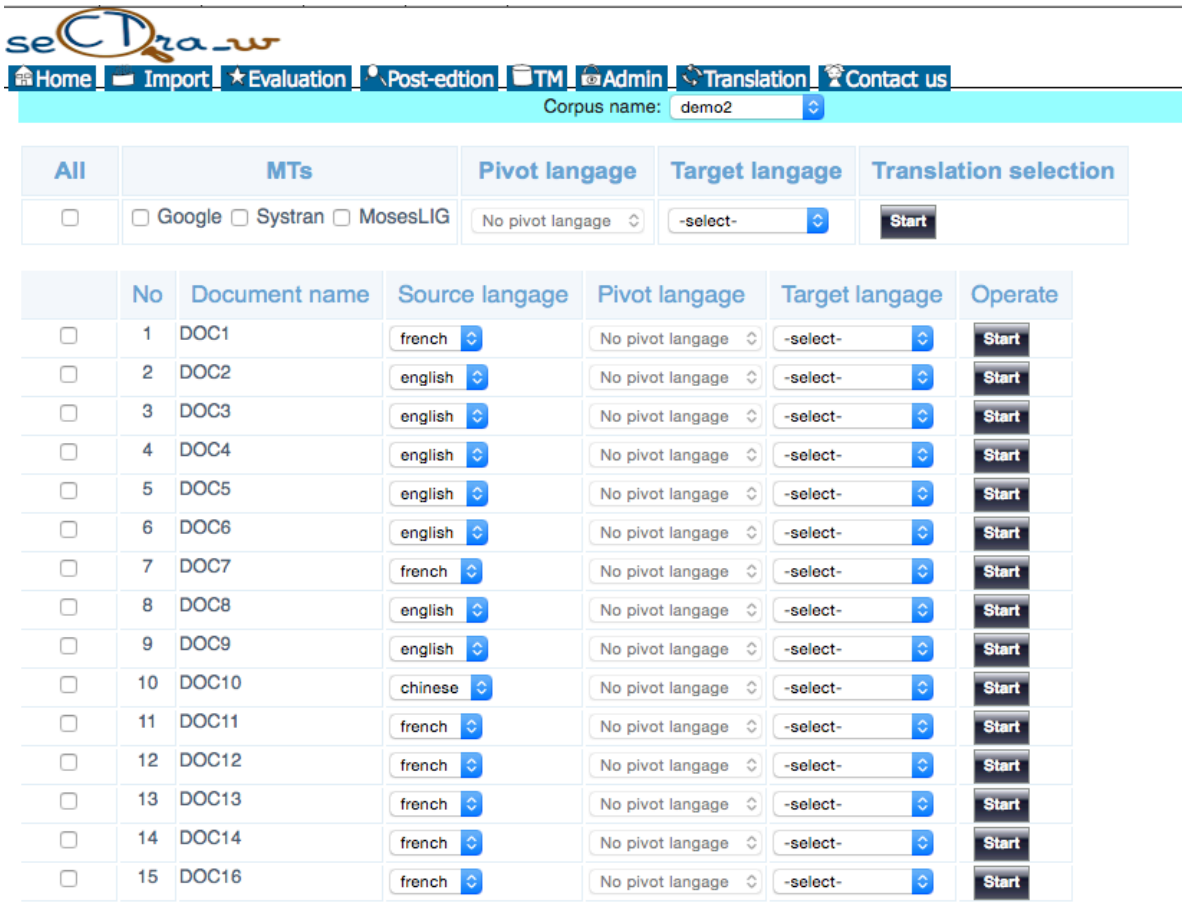


Figure 4 : Interface de « Translate Corpus »

I.2.1.2 Extension : utilisation de TRADOH et passage éventuel par une langue intermédiaire

Dans la conception de SECTRA_W, TRADOH est un important composant pour la gestion des appels à des systèmes de TA. Mais, en fait, il n'y avait pas en 2011 de vraie communication entre SECTRA_W et TRADOH. Dans le cadre du projet TRAOUIERO, nous avons reconstruit la fonctionnalité d'appel de systèmes de TA. Nous avons voulu pouvoir traiter les mêmes paires de langues avec GT et SYSTRAN, par défauts utilisés avec la même priorité pour produire les prétraductions.

Pour beaucoup de paires, GT passe en fait par un « pivot textuel » anglais. Mais d'autres systèmes, comme SYSTRAN, ne le font pas, car cela dégrade considérablement la qualité (linguistique et d'usage). Nous avons ajouté la possibilité pour un utilisateur de demander (via TRADOH) une traduction "double" passant par un texte dans une langue « pivot » qui peut être l'anglais ou toute autre langue (par exemple, le français pour faire de l'italien-espagnol). Cette possibilité est illustrée dans la Figure 4.

Dans la Figure 5, on voit une interface de TRADOH permettant de définir la langue pivot. Par exemple, le serveur SYSTRAN (V7) gracieusement prêté par SYSTRAN ne supporte pas la traduction du français vers le chinois. On prend l'anglais comme langue pivot pour obtenir la traduction en chinois (français→anglais, anglais→chinois).

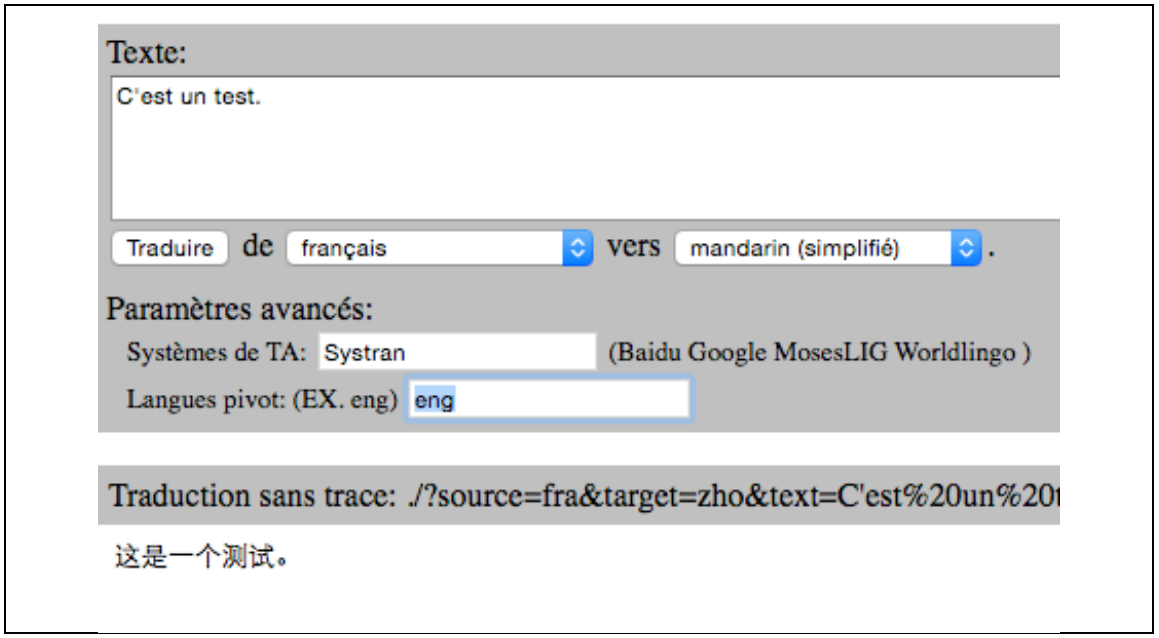


Figure 5 : Interface de TRADOH

I.2.1.3 Ajout par l'utilisateur de systèmes de TA utilisables pour les corpus associés à une MT donnée

Depuis 2004, de plus en plus d'environnements de création de systèmes de TA statistique en source ouvert (comme APERTIUM (Forcada et al., 2011), MOSES (Koehn et al., 2007), JOSHUA (Ganitkevitch et al., 2012), etc.) ont été développés par des laboratoires ou des associations. Les utilisateurs utilisent ces boîtes à outils pour entraîner leurs systèmes de TA sur leurs corpus parallèles.

Pour mettre en œuvre un système de TA qu'il a créé, un utilisateur peut utiliser le plugin de TRADOH que nous avons créé pour appeler des systèmes de TA de type MOSES. Nous avons aussi créé un plugin permettant d'appeler les systèmes de TA de type Ariane-G5 (Guilbaud, 1999) disponibles sur la plate-forme HÉLOÏSE (Boitet and Berment, 2012).

Pour l'instant, nous avons utilisé notre plugin Tradoh-Moses pour activer dans SECTRA_W le système de TA anglais→hindi MOSES-CFILT-EN_HI de l'IITB (Bombay) et nos systèmes MOSESLIG français↔chinois.

Dans la Figure 4, nous montrons comment nous avons ajouté MOSESLIG pour appeler nos systèmes de TA français↔chinois. Quand l'utilisateur veut traduire un document ou un corpus du français vers le chinois (ou l'inverse), il coche la case MOSESLIG, et le système de TA français↔chinois sera appelé.

I.2.2 Aspects de génie logiciel

I.2.2.1 Sélections paramétrables

Dans le cadre du projet TRAOUIERO, nous avons travaillé sur la "programmabilité" de SECTRA_W. Nous ainsi introduit et implémenté les *sélections paramétrables*. Le but est, par exemple, de sélectionner une « bonne » partie d'une MT et de l'exporter pour construire un système MOSES ou pour faire de l'amélioration incrémentale (voir Chapitre VI).

Nous avons ajouté la fonction « *export* » dans SECTRA_W pour exporter les segments post-édités. Nous avons implémenté un type de prédicat de sélection assez fréquent, paramétrable dans l'interface présentée dans la Figure 6.

Figure 6 : Options de la fonction "export"

Par exemple, on peut utiliser l'expression de sélection suivante :

($niveau=3 et $score>=14) ou ($niveau=4 et $score>=12) ou ($niveau=5 et $score>=11)

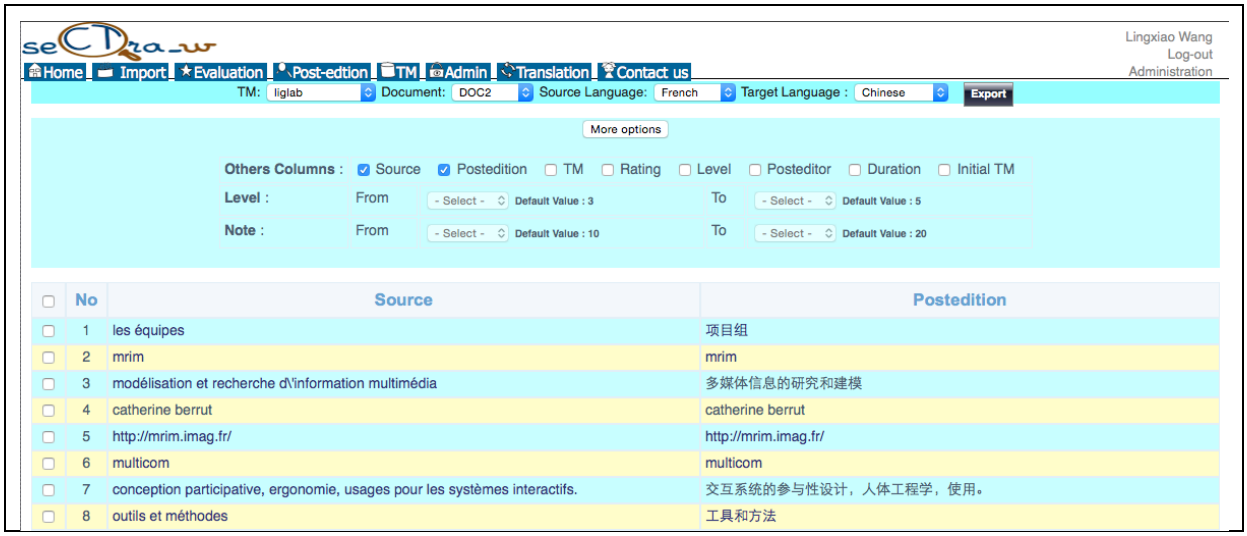


Figure 7 : Interface de sélection paramétrable dans SECTra_w

Pour aller plus loin, il faut définir ici un « petit morceau de langage » permettant de programmer n'importe quelle expression de sélection. Pour cela, nous avons enrichi l'API de SECTRA_W pour pouvoir maintenir et publier des « variables SECTRA » comme $niveau, $score, $source, $MT, $posteditor, $length, $detected_language, $date(s), etc. (Figure 7).

I.2.2.2 Clarification du traitement des "occurrences" des segments

Une page Web, qui évolue dans le temps, est représentée dans SECTRA_W par un « pseudo-document » qui contient des références à tous les « segments » déjà vus dans cette page, ainsi que le « squelette » de l'instance courante de la page.

Le problème vient surtout du fait qu'un segment peut avoir été trouvé auparavant dans une autre page Web, correspondant à un autre pseudo-document, et que, si on ne le trouve pas dans la MT du pseudo-document courant, il faut (on croit qu'il faut) aller le chercher dans les MT des autres pseudo-documents. Cela conduit à *N* recherches s'il y a *N* pseudo-documents et si le segment est nouveau… Or on devrait trouver la réponse par une seule recherche dans la MT globale.

À cause du problème algorithmique lié à ce mauvais choix de structure de données, il y a en fait une petite MT par pseudo-document (cette MT contient les segment source, les traductions obtenues par TA, et les post-éditions du document), et pas seulement un ensemble ou une liste des segments déjà vus.

Nous avons proposé une nouvelle structure de MT partagée pour résoudre ce problème, sans détruire la structure existant de SECTRA_W. Cette MT contient les segments validés par les modérateurs. Quand un nouveau segment arrive, SECTRA_W cherche d'abord la traduction dans la MT du pseudo-document en cours. S'il n'en trouve pas, il le cherche dans la MT

partagée. S'il le trouve, la traduction sera affichée sur la page Web traduite, sinon, il appellera le système de TA.

I.2.2.3 Analyse de fonctions dans XWiki, améliorant la modularité

SECTRA_W a été construit sur XWIKI. Sa programmation utilise donc les langages associés (GROOVY[23], VELOCITY[24], JAVASCRIPT, HTML). La version de XWIKI utilisée par SECTRA_W est la version 1.3, alors que la version actuelle est la 7.2. Elle nécessite l'installation de TOMCAT 5 dont on ne trouve plus d'installeur. Il a fallu procéder à l'installation complète « à la main » de TOMCAT 5. Des notes résument les manipulations effectuées, avec quelques explications, pour pouvoir les comprendre et les reproduire ou les adapter aux situations à venir.

L'installation de la version de XWIKI sur laquelle fonctionne le service a donné aussi lieu à la rédaction d'une note détaillée, en particulier pour bien mémoriser ce qui doit être fait pour définir correctement les connecteurs entre le serveur TOMCAT et les services déployés par XWIKI. Comme l'utilisation de XWIKI nous paraît en fait peu adaptée à l'implémentation de logiciels comme SECTRA_W et IMAG, même si on passait à la version 7.2, et comme un tel passage serait très lourd, nous avons préféré (et spécifié lors du projet TRAOUIERO) nous orienter vers une réimplémentation plus modulaire, et indépendante de XWIKI.

Pour faire évoluer SECTRA_W, nous avons donc commencé à travailler sur l'augmentation de sa modularité. Tout d'abord, nous avons fait une liste pour le code GROOVY qui est dans les pages XWIKI (car ce code est sauvé dans des pages Web), puis nous avons analysé les fonctions contenues dans ce code. Enfin, nous avons classé les fonctions.

I.2.3 Travail de spécification

Dans le cadre du projet TRAOUIERO, nous avons fait une spécification pour améliorer la structure de la base de données de SECTRA_W. La nouvelle base de données de SECTRA_W est centrée sur le métier « gérer des corpus multilingues de textes multilingues alignés ». Les 5 entités principales identifiées et implémentées dans le modèle conceptuel sont listées ci-dessous et on leur a rajouté 2 entités (*pseudodocument* et *segment multilingualisé multilingue*), qui manquaient dans la version courante de SECTra_w, ainsi que des précisions sur la dernière entité (Traduction). Si l'on veut avoir une représentation en « réseau corporal », comme on le fait pour des « réseaux lexicaux », il faudra donc spécialiser la notion de « segment tout court », inutilisable telle quelle, en « segment source » et « segment cible » ou « traduction segmentale ».

On pourra alors considérer qu'un segment source est un nœud d'un graphe, relié par une relation (directe ou indirecte) de traduction à des segments cibles, et que le sous-graphe connexe obtenu forme exactement un *segment multilingualisé contextualisé*. Du point de vue des traducteurs et des représentations « à la TMX », il faut cependant considérer un segment multilingualisé contextualisé comme un objet identifié comme tel, à structure complexe, et muni d'un historique propre.

> *La structure logique de la nouvelle base de données de SECTra_w illustrée dans la Figure 8 ne prend en considération l'utilisateur que dans la table provenant de l'association plusieurs à plusieurs entre les segments et leurs contextes. La stratégie de programmation choisie pour développer SECTra_w prévoit des modifications, en particulier en ce qui concerne les colonnes de ses tables[25].*

---

[23] http://www.groovy-lang.org

[24] http://velocity.apache.org

[25] Citation d'un rapport de TRAOUIERO, écrit par Valérie Bellynck.

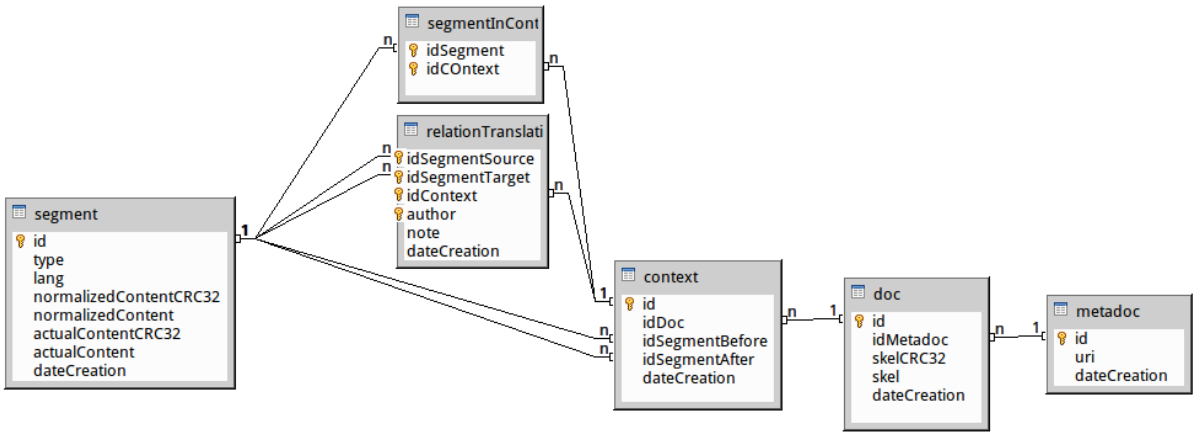


Figure 8 : Structure logique d'une base de données de corpus multilingues

I.3 Améliorations des iMAG dans le cadre du projet Traouiero

I.3.1 Paramétrisation

I.3.1.1 Réalisation d'un configurateur d'iMAG par C. P. HUYNH

Dans le cadre de la thèse de C.P. HUYNH, on a construit SECTRA_W pour permettre de mettre en œuvre des paramétrages « *externes* » au niveau du paramétrage et de la configuration par les non-informaticiens :

- Optimisation de l'interface d'évaluation subjective selon les critères d'évaluation d'une campagne, et le réglage de l'interface (nombre de segments, de colonnes, de langues, etc.).
- Filtrage de données post-éditées selon la date, l'auteur, la qualité (niveau traductionnel, note de qualité), etc.
- Recherche et remplacement interactif dans une sélection arbitraire (ensemble de segments, ensemble de documents).
- Récupération de données à partir de MT en fonction de divers paramètres.
- Lancement automatique des appels de systèmes extérieurs (TRADOH++, PIVAX, etc.).

Au niveau du paramétrage et la programmation par des contributeurs informaticiens, SECTRA_W permet d'écrire facilement des scripts pour le traitement de données, tels que des scripts pour l'import et l'export de corpus. Par exemple, A. FALAISE a écrit un script permettant l'appel depuis l'extérieur de la fonction d'import de SECTRA_W pour importer un corpus du projet SURVITRA et un autre du projet OMNIA. Cependant, nous n'avons pas encore développé les langages (langage de conditions booléennes, langage de flots de travail, langage narratif) dans la version actuelle de SECTRA_W. Nous prévoyons de les intégrer dans la prochaine version.

I.3.1.2 Améliorations prévues mais pas encore réalisées

Il faudrait améliorer le logiciel IMAG pour que, en fonction des informations sur le site S et des paramètres de l'accès multilingue à S, il appelle l'IMAG-S avec les paramètres adéquats. Ces paramètres concernent surtout les langues (lesquelles sont post-éditables, modérables, post-éditées, modérées, ou exclues...) et les systèmes de TA (lesquels sont souhaités, avec quels paramètres, avec quelles préférences entre systèmes pour chaque couple de langues, avec quels chemins traductionnels pour chaque couple, etc.).

Sur l'interface, le paramétrage d'une iMAG doit pouvoir permettre à l'administrateur d'un site Web « élu » de personnaliser la présentation (message de bienvenue, feuille de style), et de choisir les interacteurs qu'il veut proposer à ses visiteurs, selon leur rôle. Par exemple, la feuille de style pourrait permettre de générer une présentation « à la NEON », avec le texte source écrit au-dessus du texte cible, ligne par ligne, et la palette pourrait contenir les noms des contributeurs à chaque segment, et/ou son origine (niveau de ☆ à ☆☆☆☆☆), et/ou sa note (de 0 à 20), etc.

I.3.2 Travail de spécification

I.3.2.1 iMAG-Relais

Pour mieux gérer les rôles des utilisateurs entre composants différents, Valérie Bellynck a proposé un nouveau composant dit « RELAIS », qui aura pour rôle d'assurer l'identification vis-à-vis des différents services, en octroyant aux utilisateurs, de façon transparente, les droits suffisants pour assumer les tâches associées à leur rôle dans chacun d'eux, en fonction des paramétrages choisis par l'administrateur référent d'une passerelle iMAG, ou par ses administrateurs délégués. Nous reprenons ici une partie de son texte.

> *Lorsque le problème du transfert des droits a été identifié, le rôle du relais devait être de relayer les droits entre les différents services utilisés dans une architecture multiservices dans lesquels les utilisateurs peuvent avoir des rôles et des droits différents selon le service exploité. Il s'agit de transmettre ces droits sans introduire de faille de sécurité dans les services exploités.*
>
> *Non seulement les types des rôles peuvent être différents [quelqu'un peut être client dans un site commercial de souscription d'offres de services, et modérateur dans un site contributif visant à construire collaborativement de la connaissance], mais les termes utilisés pour identifier les rôles peuvent être identiques tout en ne correspondant pas à des accès identiques aux mêmes fonctions.*

I.3.2.2 Tableau blanc

Dans l'implémentation actuelle des iMAG, la plupart des composants sont des services Web, utilisés comme des serveurs. On constate deux inconvénients : (1) ces services ne sont pas autonomes (il faut les relancer à la main ou indirectement par des scripts), et (2) on n'a pas de vision sur ce qui se passe dans le système global, et qui peut être compliqué. En effet,

- il y a des appels mutuels. Par exemple, SECTRA_W appelle PIVAX en lui envoyant des segments, pour en recevoir des minidictionnaires en HTML, et PIVAX peut appeler SECTRA_W pour en obtenir la prétraduction et la post-édition des exemples d'usage contenus dans des articles de dictionnaires.
- il y a des exécutions de tâches *en boucle infinie*, qui sont pour l'instant limitées à des files d'attente de tâches considérées comme élémentaires[26]. Par exemple, un corpus peut être en train d'être traduit de français en allemand, un document EOLSS peut être en train d'être *déconverti* depuis UNL[27] vers le français ou le russe (par appel à des serveurs de TA, via TRADOH++), et plusieurs documents et mémoires de traductions peuvent être en cours de post-édition ou d'évaluation humaine.

Donc le but de cette tâche est de :

[26] D'après C. BOITET, « Il s'agit de précalculer des traductions automatiques et des minidictionnaires de façon incrémentale, et une infinité *équitable* de fois si on avait l'éternité devant soi, de façon à mettre à jour ces résultats, sachant que les systèmes de TA et les ressources lexicales évoluent constamment. »

[27] Les documents EOLSS exploités sous SECTRA_W ont des fichiers compagnons UNL qui associent un graphe UNL à chaque segment, et ces graphes sont stockés comme des annotations d'un type particulier dans SECTRA_W.

- construire, comme cela a déjà été envisagé dans le projet (ANR) OMNIA, un dispositif de contrôle du système complet par *tableau blanc* et *chef d'orchestre*[28] ;
- organiser les modules principaux comme des *agents à gros grain,* capables d'initiative, et en particulier capables de se relancer seuls, et de déterminer eux-mêmes ce qu'ils ont à faire en faisant appel au tableau blanc.

I.3.2.3 Architecture en agents à gros grains

Lorsque le système à modéliser est trop complexe, une analyse globale révèle rapidement ses limites. Il est souvent plus simple et rapide de chercher des solutions à des problèmes plus locaux, selon le principe classique en informatique du « *diviser pour régner* ». Par exemple, pour réguler le trafic aérien, de trop nombreux paramètres et contraintes influent sur le résultat de l'ensemble pour qu'une approche globale permette une résolution fiable.

En revanche, des solutions locales aident à résoudre élégamment et efficacement de telles difficultés. Les résultats obtenus sont souvent acceptables, bien qu'il soit au mieux difficile et au pire impossible de prouver cela théoriquement. L'approche multi-agents permet d'obtenir plus simplement des résultats sur des problèmes complexes, en contournant les difficultés qu'auraient des algorithmes globaux pour les gérer.

I.3.2.4 Files d'attente

Dans les cas où plusieurs clients (SECTRA_W/IMAG, TRADOH, PIVAX, etc.) veulent demander un traitement particulier, la communication par messages est souvent intéressante. En effet, elle permet de découpler les demandes des clients (comme l'appel à des systèmes de TA), et les traitements par les serveurs. Dans certains cas, les traitements à effectuer sont plus importants que le simple calcul d'une page. Il est alors beaucoup plus intéressant de passer par une liste de tâches partagées par les clients et les serveurs.

L'idée d'utiliser un système général de gestion de files d'attente sur le Web n'a finalement pas pu être réalisée lois du projet TRAOUIRO, faite de temps. Nous l'avons reprise, dans le cadre de cette thèse et de celle de Y. ZHANG. Nous en parlons plus en détaille vers la fin de ce mémoire (Voir IX.2). Disons seulement ici que nous utilisons ACTIVEMQ[29] de la fondation APACHE.

[28] Cette idée a été proposée et expérimentée il y a déjà longtemps par Ch. Boitet et M. Seligman (Boitet, Blanchon et al., 1994) Boitet, C., Blanchon, H., Seligman, M. and Bellynck, V. (1994). Evolution of MT with the Web. Proc. *Conference "Machine Translation 25 Years On"*. pp. 1-13., pour la TA de parole hétérogène, et est devenue récemment un axe de recherche reconnu, avec par exemple un atelier à LREC-2010 sur les *flux de traitements linguistiques.*

[29] http://activemq.apache.org/

Chapitre II Travail sur des aspects plus conceptuels et définition de nouvelles fonctionnalités

Résumé

Nous présentons dans ce chapitre quelques réflexions sur la modélisation des corpus de traductions, dont la structure est bien plus complexe que celle des "corpus parallèles" simples (comme le BTEC) utilisés en TA statique pour l'apprentissage, et qui sont en fait des MT. Le but poursuivi est, à terme, de séparer les concepts de MT et de corpus. À partir de cette analyse et des travaux plus pratiques décrits plus haut, nous avons aussi travaillé sur la conception de nouvelles fonctionnalités.

II.1 Modélisation de corpus de traductions variés

II.1.1 Variété des corpus de traductions et esquisse d'une méthode de formalisation

Les notions de corpus et de corpus de traductions ont été définies plus haut (Voir I.1.1.2.2). Dans un corpus de traduction, si l'on peut faire correspondre chaque unité du texte en langue source (l_1) avec chaque unité de texte en langue cible (l_2) au niveau des paragraphes, phrases ou mots, ce corpus de traduction est considéré comme un *corpus parallèle*.

On a aussi introduit plus haut la notion de *corpus comparable* (Déjean and Gaussier, 2002). Deux corpus comparables, dans deux langues l_1 et l_2, "parlent de la même chose", mais ne sont pas des corpus de traductions. Au mieux, on peut y trouver des phrases en relation de traduction, et déterminer le sens de traduction. Comme l'a montré E. DELPECH dans sa thèse (Delpech, 2013), l'espoir selon lequel on pourrait les utiliser pour trouver efficacement de la terminologie bilingue a été largement déçu. C'est pourquoi nous ne nous intéressons qu'aux corpus de traductions.

II.1.1.1 Variété des corpus de traductions

La première chose à prendre en compte si l'on veut traiter les corpus de traductions est leur grande variété, tant au niveau des contenus que des formats, des structures et des usages.

II.1.1.1.1 Variété des contenus

Texte. Un *corpus écrit* est un corpus de textes, s'il contient essentiellement des textes. Il peut être constitué de documents différents (tableau, extraits de textes, etc.). Il peut être présenté comme un livre, une page Web, ou un document (en WORD, ODS, LATEX, INTERLEAF, etc.).

Pour chaque document source, les textes et les traductions des segments peuvent être sauvegardés dans des fichiers parallèles séparés, un par langue, ou bien à l'intérieur d'un même fichier, la représentation de chaque segment contenant alors le texte source et sa ou ses traductions. Les corpus BTEC, EUROPARL, JR-ACQUIS, MULTIUN sont formés de fichiers parallèles, alors que le corpus de brevets CLEF-IP est formé de fichiers multilingues au niveau de chaque segment.

Parole. Un *corpus de parole* est un corpus constitué d'enregistrements de données orales, et éventuellement de leurs transcriptions, qui peuvent être considérées comme des annotations. À la différence de l'écrit qui n'utilise qu'un seul support, l'oral associe le plus souvent la parole enregistrée à une représentation écrite et codée (transcriptions, traductions, annotations).

Un corpus de dialogues bilingues interprétés comme le corpus ERIM contient non seulement les tours de parole des participants et leurs traductions (orales) produites par un interprète,

mais aussi des tours de parole monolingues, qui forment des sous-dialogues de clarification entre l'interprète et un participant. Un système comme SECTra doit permettre de "rejouer" ce genre de corpus, et de créer les transcriptions écrites des segments. Dans SECTRA_W, cette possibilité existe, à un niveau très « rustique », et il est possible de "déléguer" cette activité à des environnements plus adaptés (par exemple, contenant des reconnaisseurs de parole et des aligneurs texte-signal).

Visuel. Un corpus visuel est un corpus contenant des vidéos ou images, et ce type de corpus peut contenir les textes pour compléter, décrire, ou présenter les objets. Un film en vidéo contient d'ailleurs la plupart du temps, sur une ou plusieurs pistes, des sous-titres en plusieurs langues.

II.1.1.1.2 Variété des formats

À cause de la variété des usages, il existe plusieurs formats de corpus. Le format le plus utilisé est le texte brut (*.txt*). Pour un corpus monolingue, les textes sont sauvegardés dans un ou plusieurs fichiers *.txt* monolingues.

Pour un corpus bilingue/multilingue, ce format manque de balises indiquant les relations entre les phrases, et leurs langues. Donc, d'habitude, les phrases sont sauvegardées dans des fichiers séparés parallèles, un par langue. Les corpus de traductions au format .txt sont très utilisés pour l'entraînement de systèmes de TA comme MOSES. Le texte source et le texte cible sont alignés (un paragraphe par segment), et sauvegardés dans *N+1* fichiers s'il y a *N* langages cibles.

L'autre type de format le plus utilisé est « le format XML », ou plus exactement des instances de XML, comme TMX[30], TEI[31], DOCBOOK[32], NITF[33], etc., correspondant à des DTD ou à des schémas XML définis par les professionnels du traitement des documents, de la traduction, et de la localisation.

Dans le cas d'un corpus de dialogues oraux, les enregistrements sonores sont sauvegardés dans des fichiers de format audio comme *.wav*, *.mp3*, *.flac*, etc., et chaque tour de parole est représenté par un fichier XML contenant les métadonnées associées, parmi lesquelles l'url du fichier audio.

De même, un corpus "visuel", dont les documents sont des vidéos ou des textes contenant des vidéos, contient les vidéos dans des fichiers *.avi*, *.mkv*, *.mp4*, etc, et des images dans des fichiers *.svg*, *.jpeg*, *.png*, etc. Dans un film en vidéo, un sous-titre constituant un segments et ses traductions peut être considéré comme d'un segment multilingualisé.

II.1.1.1.3 Variété des structures

Structure simple. La structure la plus simple est celle où un document est une suite de segments, en format texte, avec uniquement un identificateur par segments par segment. Les traductions dans N langues sont alors contenues dans *N* fichiers de même structure.

Corpus avec annotations dans des fichiers compagnons. Dans ce type de structure, chaque fichier (source ou traduction) peut avoir un ou plusieurs fichiers « compagnons » contenant des annotations des « externes ». Cette structure peut conserver le corpus original, et ajouter l'information pour la description du corpus. Par exemple, le corpus ERIM contient dialogues

[30] https://en.wikipedia.org/wiki/Translation_Memory_eXchange
[31] http://www.tei-c.org/index.xml
[32] http://www.docbook.org
[33] http://www.gwg.nga.mil/ntb/index.html

bilingues oraux interprétés. Les dialogues sont sauvegardés dans les fichiers *.wav*, et chaque fichier *.wav* est accompagné par un fichier .xml qui décrit le contenu du fichier *.wav*.

Corpus avec des annotations internes. Les mots et/ou les segments sont étiquetés par des balises XML, ou par des chaines spéciales (ex : `Nous_Pr1pl`, `avions_VbIndImp1pl`, etc.). Les étiquettes peuvent fournir des informations de divers ordres, comme catégories syntaxiques, lemmes (forme canonique du mot fléchi), âge et sexe du locuteur, niveau d'études, etc.

Corpus arboré. Il s'agit de corpus parsés, contenant des informations sur la structure des phrases (corpus EOLSS, par exemple, voir II.1.3.3).

II.1.1.1.4 Différents usages

Les corpus de traductions différents aussi selon ce qui on veut en faire.

Corpus de traductions pour la TA

- Les corpus d'entraînement sont utilisés pour entraîner un système de TA MOSES (EUROPARL, MULTIUN, etc).
- Les corpus de développement sont utilisés pour modifier des poids dans l'étape de réglage (*Tuning*) en TA statique
- Les corpus de test sont utilisés pour évaluer les systèmes de TA et calculer des scores comme BLEU, NIST.
- D'autres corpus sont utilisés pour l'alignement, l'extraction de terminologie bilingue, etc.

Corpus de traductions pour apprenants. Un tel corpus contient des productions écrites et/ou orales faites par des apprenants d'une langue seconde servent à décrire l'interlangue et donc les difficultés des apprenants servent aussi à élaborer une typologie des erreurs pour l'utilisation dans un système de vérification grammaticale, ou production d'éditions bilingues annotées d'ouvrages littéraires.

II.1.1.2 Exemples

II.1.1.2.1 Corpus d'évaluation du projet TRANSAT

En 2005-2007, le GETA a participé au projet TRANSAT dans le cadre d'un contrat avec France Telecom R&D. À cette occasion, C. P. HUYNH a créé la première version de SECTRA_W, et nous l'avons utilisée comme support d'une campagne d'évaluation ayant pour but d'évaluer l'utilité des systèmes de traduction automatique commerciaux dans le domaine de la traduction de la parole.

Dans cette campagne, nous avons utilisé des données issues du corpus BTEC pour construire deux corpus sur deux types de tâches relatives au tourisme, ainsi que des dialogues collectés par l'équipe MULTICOM du CLIPS (Blanchon et al., 1999). Ce premier corpus source (anglais) a été extrait du corpus BTEC, et contenait un ensemble de 2224 tours de parole tous différents, en accordant une priorité aux tours de parole relatifs aux problèmes de santé. Le second corpus source (anglais), qui a aussi été extrait du corpus BTEC, était constitué de 2000 tours de parole dans le domaine de la restauration.

Les traductions candidates étaient les traductions de REVERSO TRANSLATOR 10 et de SYSTRAN (version 4). Les traductions de référence ont été produites par la post-édition de ces traductions candidates. La composition du corpus est détaillée dans l'Annexe 1.

Dans cette campagne d'évaluation, le GETA avait proposé un protocole d'évaluation pour le projet TRANSAT, donc ce corpus contient le résultat de mesures, comme la distance d'édition (Damerau, 1964, Levenshtein, 1966, Wagner and Fischer, 1974) en mots/caractères/pondérée,

BLEU, NIST, et aussi des mesures subjectives (de fluidité et d'adéquation). (Voir la définition des mesures dans l'Annexe 2).

II.1.1.2.2 Corpus B@BEL UNESCO

Le corpus UNESCO/B@BEL (Boitet, Boguslavskij et al., 2007) a été créé dans le cadre d'un contrat de recherche avec l'UNESCO. Nous avons traduit le site Web B@BEL de l'anglais vers 4 langues cible (français, russe, espagnol, et chinois). Les prétraductions ont été produites par SYSTRAN V5 PREMIUM (anglais→chinois/français/espagnol) et ETAP3[34] (anglais→russe). De plus des graphes UNL ont été construits semi-automatiquement.

Ce corpus contient environ 43200 mots (173 pages standard), et se présente comme un ensemble de multisegments, qui sont sauvegardés dans les fichiers texte, un par langue. Chaque segment source est aligné avec les 4 autres segments cible. Le corpus B@BEL contient 906 graphes UNL, et un échantillon contient 50 exemples (environs 50 pages standard). Un exemple tiré du corpus B@BEL est donné dans l'Annexe 3.

II.1.1.2.3 Corpus EOLSS

Le projet EOLSS/UNESCOL a été réalisé de février 2008 à octobre 2008 dans le cadre d'un contrat entre L'ASSOCIATION CHAMPOLLION et la fondation UNDL[35,] visant à la traduction de l'anglais vers les 5 autres langues officielles de l'UNESCO (*français, espagnol, russe, chinois, arabe*) de 25 articles de l'EOLSS (*Encyclopedia of Life Support Systems*).

Le corpus EOLSS contient 25 documents sur l'eau et l'écologie de l'encyclopédie EOLSS, représentant environ 220K mots (13676 segments) ou 880 pages standard. Chaque document est constitué d'un fichier HTML (*.aspx*), d'un fichier compagnon *.unl*, et de fichiers satellites (images, icônes, et autres hors-texte) (Figure 9). Le fichier *.unl* contient des graphes UNL représentant des segments découpés à partir du fichier *.aspx* correspondant. Le lien de téléchargement du corpus EOLSS est dans l'Annexe 4.

[34] Institute For Information Transmission Problems (Kharkevich Institute) : http://cl.iitp.ru/etap3

[35] The UNDL FOUNDATION is a private Swiss law Foundation with head office in Geneva, Switzerland, legally representing the United Nations Organization in protecting its property rights pertaining to the UNL language and system, and legally representing the EOLSS Publishers and the UNESCO Joint Committee in translating the EOLSS with the use of the UNL technology.

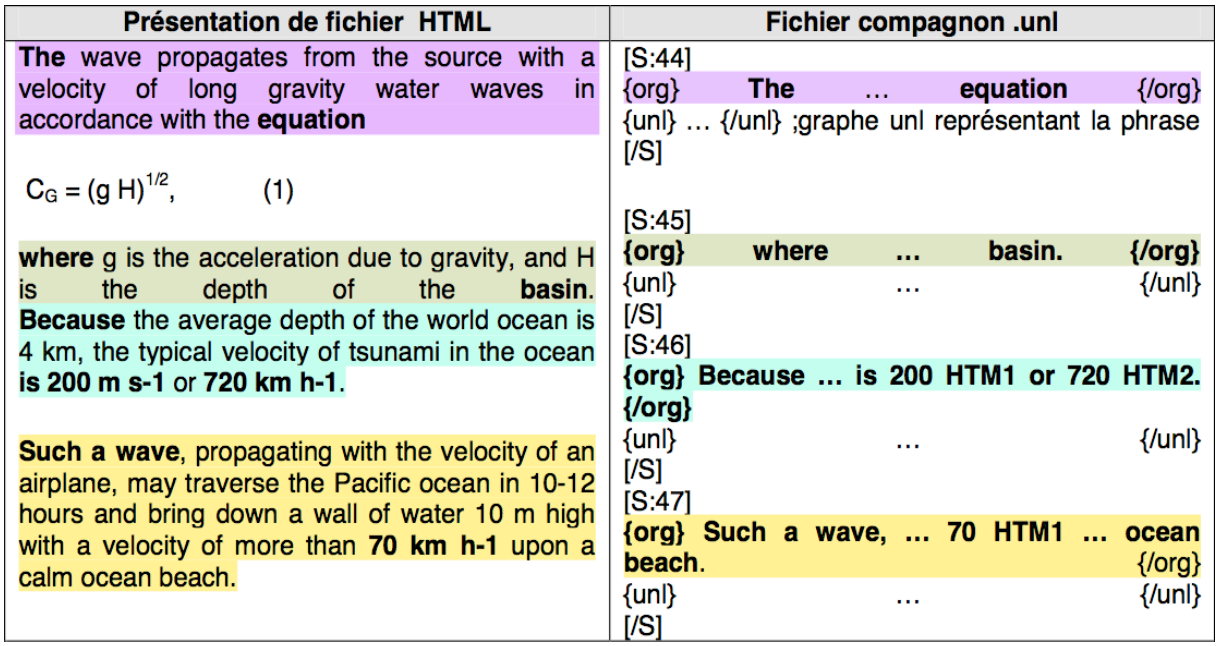


Figure 9 : Fichier HTML et fichier compagnon .unl

Pour adapter les documents à SECTRA_W, nous avons changé *.aspx* en *.html*. Nous avons utilisé le système de TA SYSTRAN pour produire les prétraductions, puis nous avons post-édité les sorties de TA. Dans la Figure 10, on présente un document (*GROUND AND SOIL WATER CHARACTERISTICS*) parallèle anglais-français. Après la post-édition du document, nous avons créé la page Web traduite en langue cible, et un fichier *.unl* en langue cible.

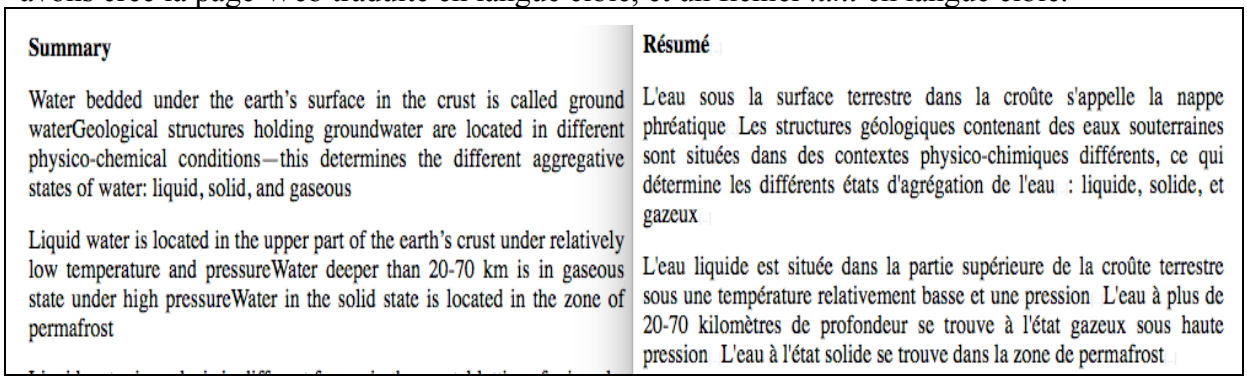
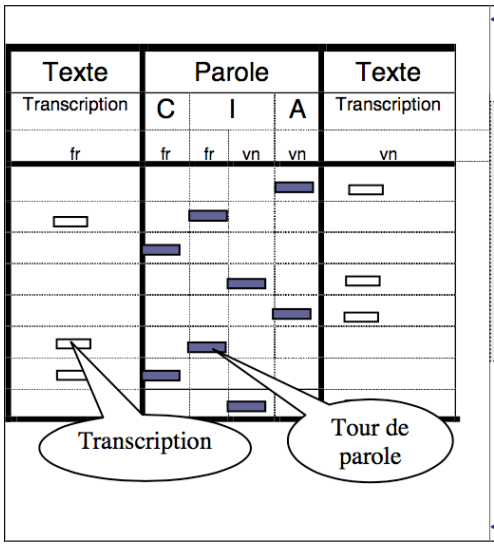


Figure 10 : Document 2 traduit de l'anglais vers le français (*GROUND AND SOIL WATER CHARACTERISTICS*)

II.1.1.2.4 Corpus ERIM

Le corpus ERIM est un corpus de dialogues bilingues (français↔chinois, vietnamien, hindi, tamil) oraux interprétés. Un segment est constitué d'un tour de parole, et éventuellement d'un texte descriptif et d'un texte transcrit.

Voici un type de descripteur d'un dialogue français↔vietnamien du corpus ERIM.




```
<Corpus nom="ERIM" structDc="erim.dtd">
  <dialog num="1" langs="Vi-Fr">
    <LOC nom="A">

    <TP num="1">
       <InfosTP lang="Vi" durée="13"
        Hdébut="10:20:00" Hfin="10:20:33">
       <InfosTP>
       <son>"A/TP1.wav"</son>
       <minitexte> "" </minitexte>
       <transc> Xin chào </transc>
       <trad lang= "fr"> Bonjour <trad>
     </TP>

    </LOC>
    <LOC nom="I"/>
     <TP num="2">
     .....
  </dialog>
</Corpus>
```

Figure 11 : Exemple de structure et de description d'un dialogue du corpus ERIM

La partie français↔vietnamien du corpus ERIM contient 4 séances de dialogue. Chaque séance contient 3 répertoires, 3 fichiers *.wpl* (par exemple, *french.wpl, vietamese.wpl,* et *dialogue.wpl*) et un fichier de transcription (comme le fichier *xml* dans la Figure 11). Chaque répertoire est nommé par le nom du locuteur (par exemple, *Cantona*, *Tung*, et *Interp*) ; il contient les tours de parole et les descripteurs. Les fichiers *.wpl* enregistrent les chemins des fichiers son. Par exemple, le fichier *french.wpl (*ou *vietamese.wpl)* enregistre les chemins des tours de parole en français (ou en vietnamien) émis/reçus par le locuteur français (ou vietnamien) et l'interprète (Figure 12). Le fichier *dialog.wpl* enregistre tous les tours de parole par ordre chronologique.

```
<?wpl version="1.0"?>
<smil>
<head><meta name="Generator"
content="Microsoft Windows Media Player"/></head>
<body><seq>
<media src="Cantona\TPE\TP1\SE_0.wav"/>
<media src="interp\TPE\TP2\SE_0.wav"/>
<media src="Cantona\TPE\TP2\SE_0.wav"/>
<media src="interp\TPE\TP4\SE_0.wav"/>
<media src="Cantona\TPE\TP3\SE_0.wav"/>
<media src="interp\TPE\TP6\SE_0.wav"/>
<media src="Cantona\TPE\TP4\SE_0.wav"/>
<media src="interp\TPE\TP8\SE_0.wav"/>
<media src="Cantona\TPE\TP5\SE_0.wav"/>
<media src="interp\TPE\TP11\SE_0.wav"/>
<media src="Cantona\TPE\TP6\SE_0.wav"/>
<media src="interp\TPE\TP13\SE_0.wav"/>
<media src="Cantona\TPE\TP7\SE_0.wav"/>
<media src="interp\TPE\TP15\SE_0.wav"/>
<media src="Cantona\TPE\TP8\SE_0.wav"/>
<media src="interp\TPE\TP17\SE_0.wav"/>
<media src="Cantona\TPE\TP9\SE_0.wav"/>
<media src="interp\TPE\TP19\SE_0.wav"/>
<media src="Cantona\TPE\TP10\SE_0.wav"/>
<media src="interp\TPE\TP21\SE_0.wav"/>
<media src="Cantona\TPE\TP11\SE_0.wav"/>
</seq></body>
</smil>
```

```
<?wpl version="1.0"?>
<smil>
<head><meta name="Generator"
content="Microsoft Windows Media Player"/></head>
<body><seq>
<media src="interp\TPE\TP1\SE_0.wav"/>
<media src="Tung\TPE\TP1\SE_0.wav"/>
<media src="interp\TPE\TP3\SE_0.wav"/>
<media src="Tung\TPE\TP2\SE_0.wav"/>
<media src="interp\TPE\TP5\SE_0.wav"/>
<media src="Tung\TPE\TP3\SE_0.wav"/>
<media src="interp\TPE\TP7\SE_0.wav"/>
<media src="Tung\TPE\TP4\SE_0.wav"/>
<media src="interp\TPE\TP9\SE_0.wav"/>
<media src="Tung\TPE\TP5\SE_0.wav"/>
<media src="interp\TPE\TP12\SE_0.wav"/>
<media src="Tung\TPE\TP6\SE_0.wav"/>
<media src="interp\TPE\TP14\SE_0.wav"/>
<media src="Tung\TPE\TP7\SE_0.wav"/>
<media src="interp\TPE\TP16\SE_0.wav"/>
<media src="Tung\TPE\TP8\SE_0.wav"/>
<media src="interp\TPE\TP18\SE_0.wav"/>
<media src="Tung\TPE\TP9\SE_0.wav"/>
<media src="interp\TPE\TP20\SE_0.wav"/>
<media src="Tung\TPE\TP10\SE_0.wav"/>
<media src="interp\TPE\TP22\SE_0.wav"/>
</seq></body>
</smil>
```

Figure 12 : Exemple du fichier *french.wpl* et *vietnamese.wpl*

Le répertoire du locuteur contient les fichiers *.wav* et 3 descripteurs (3 fichiers *.xml* : *DSD.xml, DSL.xml TIER.xml*).

- ***DSD.xml*** (*Description Session Dialogue*), enregistre l'information du session de dialogue.
- ***DSL.xml*** (*Description Session Locuteur*), enregistre l'information du locuteur.
- ***TIER.xml*** (*Traces des Informations Emises et Reçues*), enregistre (ou trace) la session de dialogue, ce qui en fournit un historique.

II.1.1.2.5 Corpus de brevets CLEF-IP 2011

Le corpus CLEF-IP contient des brevets, physiquement stockés comme une collection de fichiers XML encodant des documents de brevets. Voir le détail au VII.2.1.

II.1.1.3 Structure interne des corpus est également hétérogène

Tableau 5 : Comparaison de l'organisations logiques, physiques, et interne de quelque corpus

Corpus	Vue logique	Organisation physique	Organisation interne dans un corpus
TRANSAT	ensemble de multisegments	Fichiers de texte par langue.	L'entrée contient des segments avec des séparateurs de segments et des scores.
B@bel	Ensemble de multisegments	Fichiers de texte par langue	Listes de paires d'énoncés <Seg_Source_L1, Seg_Cible_L2>.
EOLSS	Ensemble de documents	Chaque document correspond à un dossier, contenant 2 fichiers, .html et .unl.	Le fichier .html est le « fichier principal », le fichier .unl est le « fichier compagnon », et est utilisé pour guider la segmentation du fichier .html.
ERIM	Ensemble de dialogues contenant des tours de parole avec leur descripteurs.	Chaque dialogue correspond à un répertoire, contenant des fichiers son (.wav), des fichiers transcrits (.xml), et des fichiers texte (.txt)	Un fichier son (.wav) attaché à un fichier de l'annotation selon une convention de nommage.
CLEF-IP 2011	Ensemble de documents (.xml)	Chaque document contient les segments multilingues.	Un fichier XML

II.1.2 Esquisse d'une méthode de modélisation des corpus de traductions

Pour modéliser la structure des données d'un corpus de traduction, nous nous sommes inspiré des bases lexicales et du modèle de données LEMON (*LExicon Model for ONtologies*). Nous souhaitons trouver une structure générique pouvant décrire tous les corpus de traductions, et une façon de décrire les relations entre de tels corpus. Nous distinguons les différents niveaux de structuration de ces corpus, et comparons les corpus existants pour identifier et lister les informations propres à chaque corpus de traductions (métadonnées).

II.1.2.1 Notions essentielles pour décrire un corpus

Pour décrire un corpus, nous utiliserons les notions de *segment document, corpus,* etc. définies dans la thèse de C. P. HUYNH et rappelées plus haut (pp. 20-23).

Les *métadonnées* sont par essence des informations sur les données. Elles servent le plus souvent à caractériser un objet par des informations homogènes, relativement à une collection d'objets. Pour formuler définir correctement les métadonnées, il faut donc bien connaître à la fois l'objet à décrire et les caractéristiques de la collection dans laquelle il doit être déposé.

Nous avons choisi de décrire les corpus de traductions avec les informations du DUBLIN CORE METADATA INITIATIVE[36]. Le DUBLIN CORE *(DC)* est un schéma de métadonnées générique, et il permet de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Il comprend officiellement 15 éléments relatifs à la forme de description (titre, créateur, éditeur), à la thématique (sujet, description, langue, etc.) et relatifs à la propriété intellectuelle.

II.1.2.2 Niveau des métadonnées

Pour chaque entité (*corpus*, *document*, et *segment*), il contient les métadonnées adaptées. Chaque entité peut contenir des métadonnées génériques comme le norme de métadonnées dans le DC.

1. *Title* : nom donné à la ressource (celui par lequel elle est connue officiellement).
2. *Subject* : sujet du contenu de la ressource, décrit par un ensemble de mots-clés, par des phrases, ou par un code de classification.
3. *Description* : une description du contenu de la ressource. Peut contenir un résumé, une table des matières, une référence à une représentation graphique du contenu, ou un texte libre présentant le contenu.
4. *Publisher* : une entité responsable de la diffusion de la ressource, dans sa forme actuelle.
5. *Contributor* : la ou les entités qui ont contribué à la création du contenu de la ressource.
6. *Date* : une date associée à un événement dans le cycle de vie de la ressource.
7. *Type* : la nature ou le genre du contenu de la ressource.
8. *Format* : la matérialisation physique ou numérique de la ressource.
9. *Identifier* : référence non ambiguë à la ressource dans un contexte donné.
10. *Source* : référence à une ou plusieurs ressources à partir de laquelle la ressource actuelle a été dérivée.
11. *Language* : la langue du contenu intellectuel de la ressource.
12. *Relation* : référence à une autre ressource qui a un rapport avec cette ressource.
13. *Coverage* : portée ou couverture spatio-temporelle de la ressource.
14. *Rights* : information sur les droits associés à la ressource.

La norme de métadonnées du DC peut être utilisée dans différentes entités. Par exemple, quand on fait la description d'un corpus dans *corpus*, il faut préciser des métadonnées comme le nom du corpus, le sujet du contenu de la ressource, une description du contenu de la ressource, etc. Pour *un document,* il y a aussi les métadonnées de présentation. En même temps, il existe les autres métadonnées dans différentes entités. Par exemple, pour un segment post-édité dans une MT, il a un attribut « *temps de post-édition* ».

II.1.2.3 Niveau structurel

Pour la proposition d'une structure générique, nous étudions différents corpus, et nous proposons 3 niveaux de structure abstract d'un corpus : marco-, micro-, mésostructure.

La *macrostructure* décrit l'organisation du corpus au niveau plus haut. Par exemple, un corpus contient les documents monolingues, ou un corpus contient deux documents parallèles. Il y a plusieurs types de *macrostructure.*

- *Macrostructure-monolingue.* le corpus contient un ou plusieurs documents monolingues.

[36] http://dublincore.org/

- *Macrostructure-multilingue.* le corpus contient un ou plusieurs documents multilingues.
    - *Macrostructure-parallèle.* Le corpus contient les documents parallèles.
    - *Macrostructure-comparable.* Le corpus contient les documents comparables.
    - *Macrostructure-mixte.* Le corpus contient des documents contenant des segments multilingues.

La *microstructure* décrit l'organisation des éléments dans un corpus. Par exemple, un document contient 100 segments français, et chaque segment est sauvegardé par ligne, ou un document du CLEF-IP contient les 3 champs <*title*>.

La *mésostructure* décrit les relations entre les éléments. Par exemple, "le segment B est la traduction du segment A", ou "le segment A a une description dans le document B".

II.1.3 Validation au niveau des métadonnées

II.1.3.1 Métadonnées pour le corpus d'évaluation à la TRANSAT

Les métadonnées du corpus d'évaluation à la TRANSAT contiennent 2 parties. La première concerne la description des segments monolingues extraits à partir du corpus BTEC, et l'autre concerne la description des résultats de TA avec les données d'évaluation.

Tableau 6 : Métadonnées du corpus BTEC (les segments extraits)

Métadonnées	**Valeur**	
Nom de corpus	BTEC	
Projet	TRANSAT	
Langue source	anglais	
Nombre de tours de parole	Partie I : 2223	4224 au total
	Partie II : 1998	
Nombre de phrases	Partie I : 2373	4486 au total
	Partie II : 2113	
Nombre de mots	Partie I : 14K (14001)	22419 au total
	Partie II : 10K (10418)	
Nombre de mots par tour de parole	moyenne = 6,2; min =1; max =30	
	moyenne = 5.2; min = 1; max = 22	
Nombre de mots par phrase	moyenne = 5,9; médiane = 5; min =1; max = 28	
	moyenne = 4,93; médiane = 5; min = 1; max = 19	
Domaine	Santé ; Accidents, perte, vol ; Itinéraire ; Transports, et Location de voiture ; Non spécifique	

Tableau 7 : Métadonnées des données d'évaluation à la TRANSAT

Métadonnées		**Valeur**
Langue cible		français
Nom du système de TA		Systran / Reverso
Version du système de TA		Version 4 (Systran) Version 10 Pack Monde (Reverso)
Évaluation subjective à la NIST	*Fluidité*	(F1) Formulation parfaitement compréhensible sans effort, que le style soit écrit ou oral. (F2) Formulation acceptable à l'oral, éventuellement compréhensible en faisant un effort. (F3) Formulation non acceptable.
	Adéquation	(A1) Toute l'information est transportée. (A2) Presque toute l'information est transportée. (A3) La moitié de l'information est transportée. (A4) Peu d'information est transportée. (A5) Aucune information n'est transportée, ou il y a un contresens.
Évaluation objective fondée sur la distance d'édition	*Distance d'édition en mots*	Par exemple : *Dc=20, Dw=7, D=9.6* Dc : Character distance Dw : word distance D : sentence distance
	Distance d'édition en caractères	
	Distance d'édition pondérée	
BLEU		Par exemple : 34.36%
NIST		Par exemple : 4,08
Évaluateur		Par exemple : Hervé Blanchon, Jean-Philippe Guillaud

II.1.3.2 Métadonnées pour le corpus Unesco/B@bel

Dans le cadre du projet UNESCO/B@BEL, on a traduit le texte du site Web B@BEL (équivalant 173 pages standard, 43200 mots) de l'anglais vers le chinois, le français, le russe, et l'espagnol. Ce corpus parallèle a été produit par la post-édition, donc il contient les métadonnées relatives à la description de la post-édition.

Tableau 8 : Métadonnées du corpus UNESCO-B@bel

Métadonnées		Valeur
Nom de corpus		B@bel
Nom du projet		B@bel UNESCO
Type de corpus		Corpus parallèle multilingue
Méthode de création		Post-édition
Langue source		Anglais
Langue(s) cible(s)		Chinois, français, russe, espagnol
Nombre de mots		43200
Système de TA		Systran v 5.0 Premium (en→fr/es/ch) ETAP3 (Labo IPPI) (en→ru)
Nombre de l'UNL graphe		906 graphes UNL
UNL graphe		Par exemple : voir l'Annexe 3
Échantillon de graphes UNL		50 graphes UNL (environ 5 pages standard)
Temps de post-édition sur les segments	*Temps de post-édition (par segment)*	Par exemple : voir l'Annexe 3
	Temps de post-édition (au total)	25 minutes / page standard (Systran) 20 minutes / page standard (ETAP3)

II.1.3.3 Métadonnées pour le corpus EOLSS

Le corpus EOLSS contient 25 documents. Chaque document est constitué de 2 fichiers *.aspx* (*.html*), et d'un fichier compagnons *.unl*, et de fichiers satellites (images, icônes, et autres hors-texte). Le fichier *.unl* contient les graphes UNL représentant des segments découpés à partir du fichier .html correspondant. Pour la description du corpus EOLSS, nous essayons de présenter les métadonnées en 3 niveaux de structure abstraite au niveau plus haut (macro-, micro-, mésostructure).

Les métadonnées de *macrostructure* ont pour but de décrire l'information globale du corpus (Voir Tableau 9).

Tableau 9 : Métadonnées du corpus EOLSS au niveau de la macrostructure

Métadonnées	**Valeur**
Nom du corpus	EOLSS (Encyclopedia of Life Support Systems)
Langue source	Anglais
Nombre des documents	25
Système de TA	Systran V6
Nombre des mots source	220K
Méthode de création	Post-édition
Lien original	http://www.eolss.net
Lien source	http://www.undl.org/unl-eolss/unldoc.html
Nom du document	Par exemple : D2_E2_03_05_TXT
Nom du fichier html	Par exemple : D2_E2_03_05_TXT_English.html D2_E2_03_05_TXT_French.html
Nom du fichier UNL	Par exemple : D2_E2_03_05_TXT.unl
Nom du fichier satellite	Par exemple : D2_E2_03_05_TXT_skeleton.txt

Les métadonnées de *microstructure* décrient la structure des données dans un document. Nous présentons un exemple du document 2 (*Ground and soil water characteristics*). Il y a les métadonnées pour les fichiers HTML, un fichier source avec son fichier cible a les mêmes métadonnées. Dans le Tableau 10, on prend le fichier source comme un exemple.

Tableau 10 : Métadonnées d'un fichier HTML au niveau de la microstructure

Métadonnées	**Valeur**
Nom du fichier source	D2_E2_03_05_TXT_English.html
Type du fichier	*.html*
Nombre de segments	502
Nombre de mots	6911
Titre	*"Ground and soil water characteristics"*
Langue source	anglais
Auteur	A.G. Kocharyan
Adresse	Laboratory of Water Quality, Water Problems Institute, Russian Academy of Sciences, Moscow, Russia
Mots-clés	*"physically bound water, free water, water structure, soil and ground water ....."*
Résumé	*"Water bedded under the earth's surface in the crust is called ground water……"*
Chapitre relatifs	*"Chapter 2"*
Glossaire	*"Artesian water : all ground water, excluding subsoil water, bedded between ……"*
Bibliographie	*"Bloch A.M. (1965). Water structure and geological processesLeningrad, "Nedra(This book reveals the links between the physical structure and chemical peculiarities of water)……"*
Esquisse biographique	*"Andrey G. Kocharyan graduated from the Geological Faculty, Institute of Chemistry and Oil, Baku, in 1960. He received his Ph. D. in geochemical techniques of deposits' search in 1968 at Baku State University, Geological faculty. ……."*
Date de livraison	*2007-12-03*
Date de la dernière mise à jour	*2008-06-26*

Dans le Tableau 11, on donne les métadonnées du fichier *.unl*.

Tableau 11 : Métadonnées d'un fichier UNL au niveau de la microstructure

Métadonnées	**Valeur**
Nom du fichier	D2_E2_03_05_TXT_en.unl
Type du fichier	*.unl*
ID	{S:1}
Phrase source	*{org} Agriculture water reuse and health {/org}*

UNL	`{unl}`
	`and(health(icl>state):0U.@entry, reuse(icl>act):0K)`
	`obj(reuse(icl>act):0K, water(icl>liquid):0E)`
	`mod(water(icl>liquid):0E, agriculture(icl>activity):02)`
	`{/unl}`

Les métadonnées de la *mésostructure* (Tableau 12) visent à montrer la relation entre les éléments du corpus, Par exemple, on montre la relation de traduction existant entre une phrase source et une phrase cible, avec éventuellement l'information complémentaire.

Tableau 12 : Métadonnées d'un corpus EOLSS au niveau de la mésostructure

Métadonnées	**Valeur**
ID	$$_sent_1
Phrase source	AGRICULTURE WATER REUSE AND HEALTH
Langue source	anglais
Phrase post-éditée	RÉUTILISATION DE L'EAU D'AGRICULTURE ET SANTÉ
Langue cible	français
Système de TA	Systran
Sortie de TA	RÉUTILISATION ET SANTÉ DE L'EAU D'AGRICULTURE
Post-éditeur	GETALP/XWikiGuest
Niveau de postéditeur	3
Date	2008-09-26 11:14:40.0
Temps de post-édition	17s
Score de qualité	17/20
Version	1

II.1.3.4 Métadonnées pour les corpus ERIM

Le Tableau 13 montre des métadonnées du corpus ERIM au niveau de *la macrostructure*.

Tableau 13 : Métadonnées d'un corpus ERIM au niveau de la macrostructure

Métadonnées	**Valeur**
Nom du corpus	ERIM français-vietnamien
Type du corpus	Parole + fichier accompagnons
Support	*.wav + .xml*
Langues de parole	français ↔ vietnamien
Nombre de tours de parole du corpus	Par exemple : 4 séances contenant 102 tours de parole dans "*Réservation d'hôtel*"
Date de création	*"2006-09-08"*
Date de la dernière modification	*"2006-09-10"*
Nombre de tour de parole par séance	Par exemple : 22 tours de parole dans la séance "*Réservation d'hôtel 060908-1344*"
Locuteur	Par exemple*, Tung / Eric Faffiote*
Langue du locuteur	français/vietnamien
Interpréteur	…

Le Tableau 14 montre des métadonnées du corpus ERIM au niveau de *la microstructure.*

Tableau 14 : Métadonnées de la séance dans le corpus ERIM a au niveau de *la microstructure*

Métadonnées	**Valeur**
Nom	"*Réservation d'hôtel 060908-1344*"
Sujet	"*Réservation d'hôtel"*
Langues de parole	A et B (Par exemple, français ↔ vietnamien)
Nombre de tour de parole	41
Date de création	*"2006-09-08"*
Date de la dernière modification	*"2006-09-10"*
Nombre de tours de parole en langue A	*11*
Nombre de tours de parole en langue B	*11*
Locuteur	Par exemple*, Tung et Cantona*
Langue du locuteur	Par exemple, français / vietnamien
Interprète	…
Taille de parole	06 : 0908 mn

Le Tableau 15 montre les métadonnées du corpus ERIM au niveau de la *mésostructure.* À ce niveau, les métadonnées ont pour but présenter l'information sur les fichiers *.wav* (parole).

Tableau 15 : Métadonnées d'un corpus ERIM au niveau de la mésostructure

Métadonnées	**Valeur**
Nom (de l'unité de parole)	"*TP1*" (Tour de parole 1)
Lien entre fichiers	TP1_fr.wav + TP1_fr.xml
Langues de parole	français
Rôle	Par exemple, *descripteur TP1_fr.xml*
Locuteur	Par exemple*, Cantona*
Sexe du locuteur	Par exemple*, male*
Nationalité	français
Langue du locuteur	français
Attributs	*Taux d'échantillonnage : 22000* *Bits par échantillonnage : 16* *Canaux audio : 1*
Heure de début	Par exemple*, 10 :20 :00*
Heure de fin	Par exemple*, 10 :20 :33*
Durée	13
Transcription associée	*TP1_fr.txt*
Date de création	*"2006-09-08"*
Date de la dernière modification	*"2006-09-10"*

II.1.3.5 Métadonnées pour les corpus de brevets CLEF-IP 2011

Du point de vue des métadonnées, le corpus CLEF-IP 2011 est un exemple spécifique. Il contient non seulement les métadonnées génériques, mais encore les métadonnées du document de brevet. On sépare des métadonnées en 2 parties. La première partie est constituée par les métadonnées du corpus CLEF-IP, et la seconde par celles concernant sur un document de brevet.

Tableau 16 : Métadonnées du corpus CLEF-IP 2011

Métadonnées	***Valeur***
Nom du corpus	*CLEF-IP 2011*
Auteur	*EPO*
Type de fichier	*Xml*
Langues	*anglais, français, allemand*
Langue source	*multilingues*
Langue cible	*multilingues*
Nombre de documents	*> 2,5M*
Type de document de brevets	*xml*
Nombre de brevets	*> 1,5M*
Nombre de fichiers	*3,5M fichiers xml*
Version	
Date de création	
License	*Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.*

Pour présenter les métadonnées d'un document de brevet, nous prenons un document (*EP-0000007-B2.xml*) comme exemple. Voici les métadonnées principales ; une version complète est donnée dans l'Annexe 6.

Tableau 17 : Métadonnées d'un document de brevet

Métadonnées	***Valeur***	
Nom du fichier	*EP-0000007-B2.xml*	
ucid (ID unique)	*EP-0000007-B2*	
Pays	*EP (EPO)*	
Numéro du fichier	*0000007*	
Genre	*B2*	
Langue	*fr*	
Date de publication	*19841121*	
Date de production	*20100220*	
ID de famille	*Identificateur de la famille DOCDB (utilisation pour le regroupement des familles simples)*	
ID de référence du fichier	*interne (utilisé par le fournisseur de données)*	
Nom du document de brevet	*EP-78200026-A*	
Retrait	*Retrait d'un document (oui est la seule valeur utilisée)*	
Page du document	*Utilisé dans les données EP/WO pour la section de recherche-rapport. Cet élément concerne les références à la description et les revendications (amendées) du document.*	
Relation	*"addition, division, continuation, continuation-in-part, continuing-reissue, reissue, us-divisional-reissue, reexamination, us-reexamination-reissue-merger, correction, utility-model-basis, us-provisional-application, related-publication"*	
document-id	*pays, numéro de document, genre,* *date, texte, langue*	
abstract	*langue et source*	
description	*invention-title* *technical-field* *background-art* *disclosure* *description-of-drawings*	*best-mode* *mode-for-invention* *industrial-applicability* *sequence-list-text* *heading|img|imgref|p*
claims	*langue, type* *déclaration, revendication* *texte de revendication* *référence de revendication*	
claim	*id, numéro* *type de revendication*	

II.2 Conception de nouvelles fonctionnalités et développements en cours

II.2.1 Nouvelles fonctionnalités

II.2.1.1 Intégration de « minidictionaires » associés aux segments

Pour faciliter l'utilisation des ressources lexicales, nous avons intégré des « minidictionaires » à l'interface d'IMAG et de SECTRA_W. Pour les ressources « finalisées », à savoir les traductions des mots et des termes qui sont présents dans des traductions (de segments) de qualité suffisante (pas les sorties brutes de TA, mais les post-éditions).

Parmi ces traductions, on peut encore distinguer celles qui sont « recommandées » pour le projet en cours. Pour chaque segment, on dispose aussi des informations lexicales potentiellement utiles, dans la liste des ressources indiquées, et on peut sauvegarder le résultat dans une structure de données (minidictionaires) associée à ce segment. La structure de minidictionaires est donnée dans l'Annexe 7.

II.2.1.2 Création d'une API pour appeler le système CREATDICO

CREATDICO est un intergiciel de consultation de dictionnaires. Avec cet intergiciel, on peut envoyer un texte ou un segment pour demander une consultation pour chaque lemme du texte ou du segment. Pour faire les lemmatisations, CREATDICO appelle LEXTOH (intergiciel de lemmatisation). L'API de CREATDICO a 10 paramètres :

Tableau 18 : 10 paramètres de l'API de CREATDICO

`typeEntry` (type d'entrée),	`input` (type d'entrée)
`output` (type de sortie)	`ls` (langue source en ISO 639-2)
`lc` (langue cible en ISO 639-2)	`serv` (serveur de service)
`dico` (nom de dictionnaire)	`lemmat` (outils d'analyse morphologique)
`trace` (trace pour debogage)	`formule` (affichage de l'interface de formulaire)

La sortie de CREATDICO est une sortie dédiée à la structure "MS (MINIDICO_SECTRA)" (Voir l'Annexe 7). Pour cette sortie, il y a un fichier de configuration qui permet de définir des valeurs par défaut des paramètres, par exemple :

Tableau 19 : Exemple de fichier de configuration de CREATDICO

```
typeEntry = txt
lemma = xip (condition : ls = {fra, eng})
lemma = jieba (condition : ls = {zho})
etc.
serv = pivax
dico = * // commentaire : tous les dictionnaires de pivax
```

Dans notre cas, on n'utilise jamais les paramètres : `trace` et `formule`, et ils ne sont pas obligatoires. Voici un exemple de l'utilisation de l'API de CREATDICO :

Tableau 20 : Exemple d'un lien pour l'utilisation de l'API de CREATDICO

http://atoum.imag.fr/getalp/Services/Web/CREATDICO/via-Ci-Hai/CREATDICO/index.php?output=MS&input=HelloWorld!&ls=eng&lc=zho

Pour faciliter l'utilisation des dictionnaires, nous avons ajouté un panneau de dictionnaires à l'interface de SECTRA_W. Quand l'utilisateur fait la post-édition, il peut double-cliquer sur un mot, puis choisir les dictionnaires existants dans ce panneau.

Figure 13 : Capture d'écran de panneau de dictionnaires ajouté à SECTra_w

Le résultat est affiché dans le champ en bas. Par exemple, si on double-clique sur le mot « *faisons* », il est ajouté automatiquement dans le champ de saisie. On choisit la langue source et le dictionnaire, et on clique sur le bouton « *Find* ». Le résultat est affiché en couleur rouge (Figure 13).

II.2.1.3 Manipulation de la MT

Sur un segment. Le résultat des systèmes de TA a évolué lors des mises à jour. Sur SECTRA_W, il faut permettre à l'utilisateur de supprimer les traductions de TA, et à de rappeler les systèmes de TA de la dernière version. Si la post-édition n'est pas bonne, s'il existe les erreurs sur le segment post-édité, il faut aussi permettre à l'utilisateur de le supprimer. Pour réaliser cette fonction, nous avons ajouté 2 boutons avant la traduction de TA. Le bouton « *Clean* » permet de supprimer la traduction de TA sur l'interface de SECTRA_W, et aussi sur la MT. Le deuxième bouton « *Get* » permet de recalculer (retraduire) le segment source par TA. Si la post-édition n'est pas bonne, on peut le supprimer par bouton « *Delete* ». (Voir Figure 14).

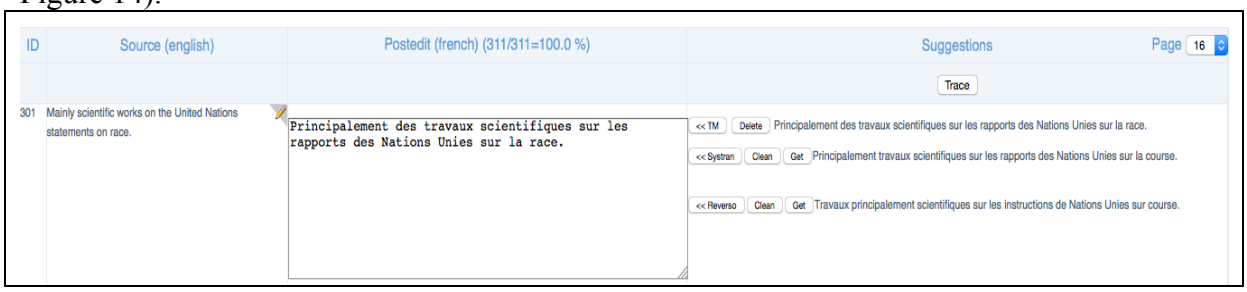


Figure 14 : Interface de SECTra_w intégrant les boutons « *Delete* », « *Clean* » et « *Get* »

Sur un corpus. On a besoin de (re)traduire une sélection de segments, éventuellement tous, depuis l'interface de SECTra_w. Pour réaliser cette fonction, nous avons créé une interface permettant de choisir les segments source et un ou plusieurs systèmes de TA. L'utilisateur peut cocher les segments sources, et les traduire par le système de TA choisi. Sur cette interface, il peut aussi importer des résultats de TA dans la MT associée. Il demande de préparer les segments source et les traductions, et les phrases sont alignées par ligne.

Voici une image d'écran (Figure 15).

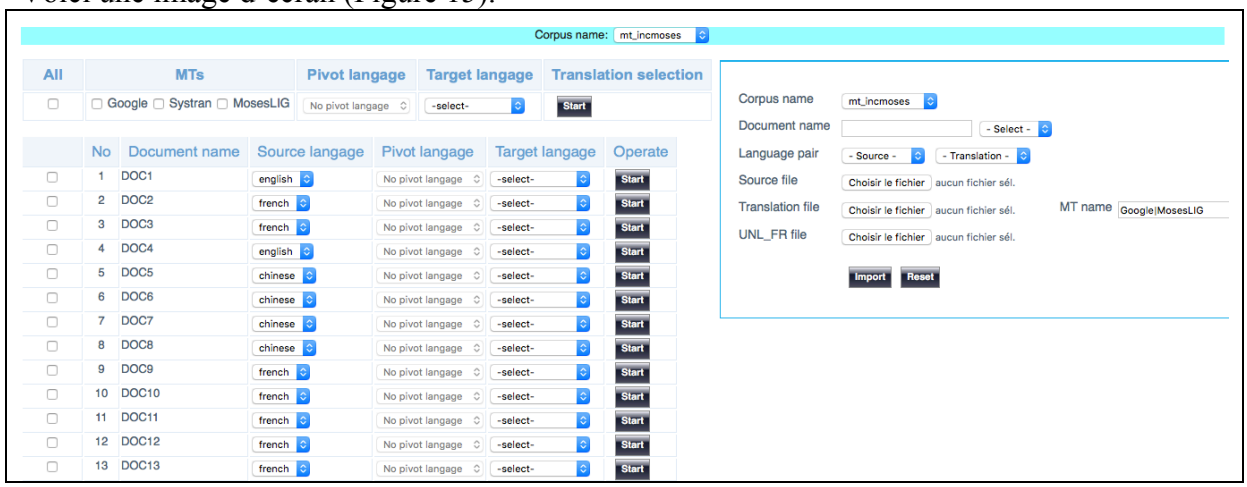


Figure 15 : Traduction des segments sélectionnés et ajout à la MT

II.2.2 Programmabilité

II.2.2.1 API avec CREATDICO, TRADOH, SEGDOC

SECTra_w a montré une possibilité de l'utilisation de l'API, et il permet l'accès à différents types de services. Pour appeler CREATDICO, Tradoh et SEGDOC, on a intégré l'API sur SECTRA_W, et les fonctions sont réalisées en Groovy. On peut suivre une instruction pour remplir les paramètres dans un lien. Quand on accède ce lien avec nos valeurs pour les paramètres, on peut recevoir le résultat immédiatement sur une nouvelle page Web. Par exemple, on peut appeler les systèmes de TA via TRADOH sur l'API de SECTRA_W (Figure 16). On remplit les paramètres (nom de TA, langue source/cible, et texte source), et le système génère un lien pour obtenir la traduction.

Figure 16 : Exemple de l'API « Call Tradoh »

II.2.2.2 Vers un langage de commandes pour SECTra_w

La définition d'un tel langage est particulièrement intéressante parce qu'il permettra de réaliser toutes les opérations sur les corpus, à la fois en ligne de commande et en une formulaire, et aussi de générer des codes pour les interfaces, ce qui accélérera leur développement tout en assurant leur homogénéité.

L'exemple le plus utilisé de langage déclaratif dédié à la sélection d'éléments et à la commande d'actions à exécuter sur toutes les données sélectionnées est SQL. Le langage à définir est bien de cette nature.

Par exemple, en MySQL, la commande qui sélectionne tous les enregistrements de la relation mémorisée dans la table `demo2`.`segments` et dont le champ `creation_date` est compris entre '2013-01-16' et '2013-01-18' est :

```
SELECT * FROM `demo2`.`segments` WHERE `creation_date` > '2013-01-16' AND `creation_date` < '2013-01-18'.
```

Le langage d'interrogation des données permettant les sélections, la définition de sélections dans les corpus doit s'en inspirer.

Par exemple :

```
Selection 3 = $seg \in demo2 \and $creation_data \in 2013-01-06 .. 2013-01-18 \and {{ $level = 3 \and $score > 13}}
```

L'analyseur du langage en question sera faite en ANTLR (Parr and Quong, 1995), générateur de compilateurs dont le GETALP a une grande expérience, à cause du projet Ariane-Y. Les règles de grammaire décrivant le langage et transmises à ANTLR peuvent définir des actions.

II.2.3 Modularité

II.2.3.1 Extraction de codes depuis XWiki vers Java dans des fichiers externes

Dans I.2.2.3, nous avons un peu parlé du problème de la programmation en XWIKI, et nous avons listé et classifié les fonctions de SECTRA_W/IMAG dans le cadre du projet Traouiero. Pour réaliser la modularité de SECTRA_W/IMAG, nous avons aussi extrait les codes de XWIKI (en Groovy). Comme Groovy est proche de Java, nous les avons reprogrammés en Java, et les avons publiés sous forme de packages de Java *(.jar*).

Par exemple, l'extraction de la fonction « Segmentation de la page Web » est détaillée à http://tools.aximag.fr/prepaTradSeg/_README.htm.

La version initiale de ce service avait seulement pour objectif de fournir aux IMAG le prétraitement d'une page Web. D'abord, on a l'URL de la page ou un contenu HTML (sérialisation du DOM fournie par un navigateur). Ensuite on appelle un outil de segmentation d'un contenu HTML en segments, avec le service http://tools.aximag.fr/enrichir/seghtml et un paramétrage de l'appel à ce service, tel que le résultat soit fourni dans un format attendu par l'uniformisateur d'appels à des systèmes de TA. Cela produit la même forme de sortie que celle qu'aurait fournie le service GT pour le contenu source, avec l'habillage de chaque segment source identifié par le traitement du segmenteur par l'intercalage de 2 nœuds imbriqués dans le DOM.

II.2.3.2 Vers un "serveur corporal pour la TA" avec lien en cours SECTra-Ariane

ARIANE-G5 est un générateur de systèmes de TA reposant sur 5 *langages spécialisés pour la programmation linguistique* (LSPL)[37]. Chacun de ces langages est compilé, et les tables internes produites sont données en paramètres au "moteur" du langage.

[37] Les LSPL sont conçus pour implémenter les outils linguistiques. Au CETA, l'idée de construire un système de TALN basé sur plusieurs LSPL est venue des travaux de B. Vauquois, G. Veillon, J. Veyrunes et leurs collègues des années 1962-1967. Cette idée s'est également imposée dans (Hutchins, 1986) ou encore (Arnold, Balkan et al., 1994) .

Bien que le système ARIANE-G5 ait déjà une fonction de post-édition et de gestion de corpus, on voudrait déléguer à un « serveur corporal » le soin de gérer les corpus. La motivation de cette délégation est identique à celle des connaissances lexicales où on pourrait profiter des avantages d'un système sur le Web comme SECTRA_W, et surtout avancer vers une unification des ressources corporales.

Dans le système ARIANE-G5, à l'entrée et à la sortie du processus de traduction, l'unité de traduction est une simple chaîne de caractères. Les chaînes doivent être codées en EBCDIC Dans l'implémentation ARIANE-H (HÉLOÏSE de V. Berment (BOITET AND BERMENT, 2012)), elle sont en ASCII.

En ARIANE-G5, l'organisation est la suivante :

- fichier texte source.
- fichier de description arborescente (utilisé pour calculer les unités de traduction).
- fichier de segmentation en unités de traduction, selon une fenêtre paramétrable par le linguiste (200-300 caractères).
- fichiers d'arbres de résultats intermédiaires possibles.
    - chaque fichier selon la chaîne d'exécution/production
    - fichiers selon les phrases. Chaque fichier contient la liste des arbres fournis par cette phase, un pour chaque unité de traduction.
- fichiers résultats de traduction automatique d'une ou plusieurs chaînes d'exécution. Le numéro de la chaîne d'exécution est inclus dans le nom de fichier.
- fichiers de révision.

Avec un corpus, on a 2 fichiers « spéciaux », *$DESCR* contenant la liste des textes de ce corpus avec le format <nom externe, nom interne>, et un autre contenant une liste hiérarchique de séparateurs (au maximum 8).

ARIANE-G5 a 3 types de fichier de résultat.

Résultat intermédiaire (liste d'arbres décorés) est un format d'échange entre les étapes du système. Ce format représente des arbres décorés, les formes et UL utilisées.

Résultat brut de la traduction est la concaténation des traductions des unités de traduction, et on n'a plus accès aux unités individuelles.

Révision. Pour un texte donné, il y a un seul fichier de révision par langue et par chaîne de traduction.

La version actuelle de SECTRA_W a été construite pour être utilisable aussi bien par des programmes que par des utilisateurs humains. Pour l'intégrer à ARIANE-G5, il faut :

- définir et implémenter certaines fonctions sur les corpus, existant en ARIANE-G5, mais pas encore intégrées à SECTRA_W ;
- définir une syntaxe appropriée pour échanger avec ARIANE-G5 des segments, des documents, des corpus et des mémoires de traductions (formats d'import-export)
- définir une structure interne dans Ariane-Y pour les données "corporales".

En particulier, la révision (en construction) Ariane-Y doit pouvoir envoyer des résultats de traitement,

- sur des segments (par exemple des arbres d'analyse ou des graphes UNL, à ranger comme des « annotations » au même titre que les TA et les PE),
- sur des documents (graphe de chaînes d'arbres, par exemple).

Chapitre III Variété des iMAG et de leurs usages : de l'accès multilingue à la création de bons corpus bilingues et à la traduction littéraire contributive de qualité

Introduction

Nous faisons ici le point sur l'ensemble des passerelles iMAG créées depuis 2009, et décrions leurs différents usages. En ce qui nous concerne, nous en avons créé une instance, essentiellement destinées à la création de bons corpus bilingues pour la TA français-chinois, à des expérimentations sur d'apprentissage incrémentale, et à un projet personnel de traduction littérature.

III.1 Liste (avec les MT associées)

Une passerelle iMAG peut être utilisée pour supporter non seulement la post-édition des sites Web et des documents, mais aussi l'expérimentation, la traduction, l'évaluation, et le support d'accès multilingue aux sites Web commerciaux.

Depuis 2009, nous avons créé 213 iMAG sur 20 MT, et notre plate-forme supporte plus de 8 langues source et 30 langues cible. Il y a plusieurs types de sites Web "élus" (accédés en multilingue), qui couvrent de nombreux domaines. Dans cette section, on montre des exemples (par liste) d'iMAG des différents types.

Laboratoire/Université. Nous avons créé des iMAG pour les sites Web de laboratoires et d'universités, semblable à celui de notre laboratoire (LIG), ISCC, LICIA[38], NICT[39], etc. et pour les sites Web d'universités sont comme l'Université Joseph Fourier (UJF), IITB, etc. (Tableau 21).

Tableau 21 : iMAG pour les sites Web de laboratoires et d'universités

ID	*Nom de site*	*Lien*	*Langues source*	*Langues Cible*	*TA*	*MT*
LIG-Lab	*Laboratoire d'Informatique de Grenoble*	*http://www.liglab.fr*	*French*	*Chinese ,|English ,|French ,|……*	*Moses+ Google*	*liglab*
IPPI-IITP	*Институт проблем передачи информации (Institute for Information Transmission Problems)*	*http://iitp.ru*	*Russian*	*Chinese ,|English ,|French ,|……*	*Google*	*demo*
ISCC	*Institut des sciences de la communication CNRS / Paris-Sorbonne / UPMC*	*http://www.iscc.cnrs.fr*	*French*	*Arabic ,|Bulgarian ,|Chinese ,|Croatian ,|Czech ,|Danish ,|Dutch ,|English ,|……*	*Systran*	*demo*
UJF	*Université Joseph Fourier*	*http://www.ujf-grenoble.fr*	*French*	*Arabic ,|Belarusian ,|Bulgarian ,|Catalan ,|Chinese ,|Czech,|…*	*Google*	*demo2*

[38] http://licia-lab.imag.fr/index.php/fr-FR

[39] http://www.nict.go.jp/

IITB	*Indian Institute of Technology BomBay*	*http://www.iitb.ac.in*	*English*	*...,|Hungarian ,|Icelandic,|Japanese ,|Korean ,| ...*	*Google*	*demo2*

Sociétés. Le site AXIMAG.FR[40] supporte des iMAG dédiées aux sites Web d'organismes et de sociétés (commerciales ou non), comme ACXYS, AMIES, IDB, etc. Ce type d'iMAG dédié aux sites contient souvent une MT spécifique pour un site Web.

Tableau 22 : iMAG pour les sites Web d'organismes et de sociétés

ID	***Nom de site***	***Lien***	***Langues source***	***Langues Cible***	***TA***	***MT***
AcXys	*AcXys*	*http://www.acxys.com/*	*English*	*Arabic ,|Chinese ,|Dutch ,|French ,|German ,|Greek ,|Hungarian ,|Italian ,|Japanese ,|Korean ,|Malay ,|Polish ,|Portuguese ,|Russian ,|Spanish ,|Thai ,|Turkish ,|Vietnamese*	*Google*	*mt_acxys*
iMAG-AMIES-privee	*AMIES*	*http://www.agence-maths-entreprises.fr/*	*French*	*Chinese ,|English ,|German ,|Japanese ,|Portuguese ,|Russian ,|Spanish*	*Google*	*mt_amies*
IDB	*Islamic Development Bank*	*http://www.isdb.org*	*English*	*Arabic ,|Chinese ,|French ,|German ,|Portuguese ,|Russian ,|Spanish ,|Swahili*	*Google+Systran*	*demo*
Bio-Clean	*Bio-Clean*	*http://projet-bioclean.3beesonline.com*	*French*	*Arabic ,|Chinese ,|German ,|Portuguese,|... ...*	*Google*	*demo2*

Projets et expérimentations. Le site aximag .fr peut aussi être utilisé comme une plate-forme pour la partie de post-édition ou d'évaluation de support multilingues d'un projet, ou pour des expériences de recherche. Les IMAG de projet sont EOLSS, AKENOU, HOMERICA, etc., et les IMAG d'expérimentation sont POWERS, BEMBOOK, etc.

Tableau 23 : iMAG pour des projets et des expérimentations

ID	***Nom de site***	***Lien***	***Langue source***	***Langues cible***	***TA***	***MT***
EOLSS	*EOLSS*	*http://www-clips.imag.fr/geta/User/christian.boitet/iMAGs-tests/EOLSS*	*English*	*Chinese ,|French ,|Japanese ,|Russian ,|Spanish*	*Google*	*demo2+ eolss*
Powers	*The book of me*	*http://www-clips.imag.fr/geod/User/laurent.besacier/TRANSLATION-EXP/The_Book_of_Me.html*	*English*	*French ,|Romanian*	*Moses*	*mt_powers*
BEMBOOK	*BEMBOOK*	*http://www.kenrico.com/media/bembook/21/21.htm*	*English*	*Arabic ,|Bengali ,|Bulgarian ,|Chinese ,|French ,|German ,|Hindi ,|Hungarian ,|Italian ,|Japanese ,|Malay ,|Marathi ,|Polish ,|Portuguese ,|Russian ,|Spanish*	*Google+ Systran*	*demo1*

[40] http://service.aximag.fr/xwiki/bin/view/home/imag

III.2 Commentaires sur les utilisations actuelles

III.2.1 Accès à des sites Web d'organismes ou de sociétés

Plusieurs sites Web de sociétés, de laboratoires, ou d'universités n'offrent pas d'accès multilingue, ou supportent un ou deux langues étrangères. Mais au laboratoire ou à l'université, il y a beaucoup de chercheurs ou étudiants étrangers, et le site Web devrait distribuer l'information à tous. Il y a une vraie demande pour des sites Web multilingues. Le serveur aximag.fr supporte plusieurs sites Web de sociétés, de laboratoires, ou d'universités. Par exemple, en 2011, nous avons créé une iMAG (*iMAG-LIG-LAB*) pour le site Web de notre laboratoire, en lui attachant une MT dédiée (*liglab*). Au départ de *iMAG-LIG-LAB*, nous avions environ 10K segments source en français. Nous avons fait la post-édition du français vers le chinois. Après première passe de post-édition, nous avons obtenu environ 2K segments post-édités en chinois, et les pages Web principales du laboratoire ont été bien présentées en chinois (en juin 2013)[41].

Pour aider les étudiants étrangers à accéder aux sites Web d'universités, nous avons expérimenté la technique iMAG pour traduire les sites Web de l'UJF. Les étudiants étrangers parlant (ou comprenant) le français ont aidé à post-éditer les pages Web du français vers leurs langues maternelles.

Des iMAG ont été créées non seulement pour des organismes d'enseignement (comme LIG, ISCC, ANRT, etc) et les universités, mais aussi pour les sites Web commerciaux. Pour les sites Web comme AcXys, nous traduisons les segments source de l'anglais vers 18 langues étrangères. Le site Web AcXys contient les segments traduits en français, allemand, russe, et polonais. Nous les ajoutons dans la MT dédiée (*mt_acxys*) pour la cohérence entre le site officiel et l'iMAG.

L'iMAG a aussi un cas particulier. Depuis mai 2010, la Métro propose l'intégralité de son site web en version chinoise et anglaise, les deux langues les plus parlées au monde. Nous avons créé une iMAG dédiée, et une mémoire dédiée (*lametro*) pour le site Web de La Métro, permettant d'obtenir des textes traduits en constante amélioration. Son atout majeur réside dans une mémoire spécifique au site : conservant les segments post-édités, celle-ci les réutilise ensuite en tenant compte du type de vocabulaire et des termes utilisés sur le site. La qualité de traduction est ainsi en constante amélioration.

III.2.2 Aide à la traduction de documents : rapports, de parties de thèse, manuels…

Le but essentiel d'une iMAG est d'offrir un bon support pour l'accès multilingue à un site Web. Mais cette technique peut aussi nous aider à la traduction de documents, si on fournit le document au format HTML. Nous avons travaillé sur la post-édition des rapports, de parties de thèse, manuels, etc.

Beaucoup d'universités publient leurs matériels éducatifs gratuitement, mais ils ne sont généralement pas disponibles dans plus d'une ou deux langues, ou disponibles moyennant un supplément. L'approche iMAG offre une alternative rapide, pratique et de très faible coût pour obtenir des versions multilingues de matériel éducatif, comme les manuels scolaires qui sont convertis dans un format compatible iMAG. Un bon exemple est l'iMAG de démonstration pour le livre « Bioelectromagnetism » de Jaakko Malmivuo et Robert Plonsey, visitable à http://service.aximag.fr//xwiki/bin/view/imag/BEMBOOK.

[41] Tout le travail s'effectue sur une MT dédiée (liglab). Après juillet 2014, on a créé une autre iMAG pour le site du LIG, utilisant la MT générique (partagée) *demo2*.

Un autre exemple, est l'iMAG *xan-fr*[42], utilisée pour traduire des documents du français vers l'anglais. Le dernier travail est la post-édition du résumé de ma thèse (voir Figure 17). On a utilisé GT pour prétraduire le texte, puis fait la post-édition sur le texte anglais.

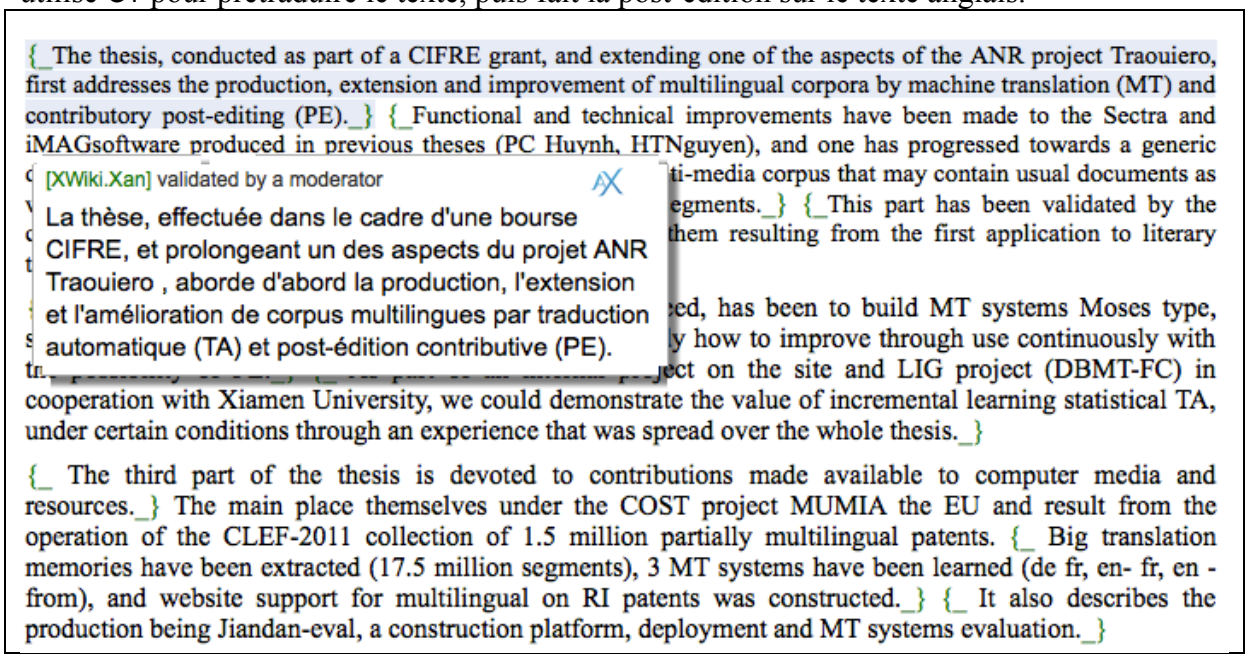


Figure 17 : Post-édition d'un document français accédé en anglais (résumé de la thèse de Lingxiao WANG)

III.2.3 Accès multilingue à des documents pédagogiques : MACAU

R. Kalitvianski et C. Boitet (Kalitvianski et al., 2012) ont lancé le projet MACAU (Multilingual Access & Contributive Appropriation for Universities) pour permettre aux étudiants étrangers de l'UJF d'accéder dans leur langue aux supports pédagogiques, par amélioration contributive des "prétraductions" par TA. Pour l'accès multilingue dans MACAU-OFI, nous avons créé une iMAG « dédiée » à ce projet.

> *L'idée de base est de ne pas chercher à construire différentes versions, une en chaque langue, mais de garder l'unicité de la version « originale », et de la « reconstruire » dans différentes langues en choisissant une « meilleure traduction » [ou suggestion de traduction] pour chaque segment dans une MT [mémoire de traductions] associée. Pour l'instant, nous supposons que la version originale de chaque document est dans une seule langue, mais, dans le futur, on utilisera un détecteur de langue, et il pourra y avoir des segments écrits en différentes langues dans le document original.*
>
> *Le projet MACAU-OFI consiste à essayer d'approfondir cette idée en allant dans deux directions : d'abord, permettre aux étudiants de participer en contribuant par des fragments de notes de cours ou d'exercices, et ensuite en enrichissant l'outil par l'intégration de ce que nous appellerons une « ontologie du domaine », permettant en particulier, dans le cas d'OFI, de viser à l'autoformation par utilisation d'outils de simulation d'automates, grammaires, et graphes.*
>
> *Au cours des deux dernières années universitaires, nous avons encouragé les étudiants de Master 1 d'informatique de l'UJF à produire des documents de cours sur la complexité calculatoire. Chacun a eu un compte sur aximag.fr. Certains ont beaucoup contribué, d'autre moins. Voici les nombres d'étudiants concernés, par langue maternelle.*

[42] http://service.aximag.fr/xwiki/bin/view/imag/xan-fr

Tableau 24 : Nombre de langue du projet MACAU (06/2013)

Langue	*Nombre*
Chinois	*7*
Arabe	*2*
Russe	*1*
Anglais	*2*

Les fichiers reçus étaient dans les formats doc, odt, pdf, tex, html. Certains étaient des cours complets, d'autres ne contenaient que quelques chapitres.

Tableau 25 : Statistiques de documents dans MACAU (06/2013)

	tex	*doc, docx*	*odt*	*pdf*	*html*
Nombre de fichiers	*16*	*5*	*5*	*19*	*7*
Cours complets	*2*	*1*	*0*	*0*	*0*
Chapitres disponibles (hors cours complets)	*1-8*	*1-8*	*4, 7*	*1-7*	*-*
Autres	*-*	*1 corrigé d'un quick, 1 fichier de notes explicatives*	*-*	*-*	*Site avec notes sur Pseudo-Pascal, RAM, 2SAT et exercices*

III.2.4 Évaluation

Depuis 2007, les fonctionnalités de SECTRA ont évolué pour l'adapter à des usages différents.

Le premier usage de SECTRA_W a été la « campagne d'évaluation » du projet TRANSAT. Il a été utilisé avec succès, fin 2007, dans le cadre du projet TRANSAT de FT R&D. Avec SECTRA_W, après avoir importé un corpus source, et éventuellement les traductions de référence, on peut appeler plusieurs systèmes de TA, stocker leurs résultats, et demander à des juges d'effectuer l'évaluation subjective (fluidité, adéquation). SECTra_w fournit plusieurs méthodes d'évaluation objective (NIST, BLEU, etc.), et permet aussi d'effectuer l'évaluation objective liée à la tâche, en permettant à des participants de post-éditer les résultats de systèmes de TA, et en mesurant un distance d'édition (et/ou le temps de post-édition). Les résultats post-édités peuvent être ajoutés à l'ensemble des traductions de référence, ou le constituer s'il n'y a pas de références.

Le deuxième usage est l'évaluation pour améliorer des systèmes (AI). Nous avons utilisé SECTRA_W/IMAG pour évaluer les résultats des systèmes de TA, puis nous faisons la post-édition pour créer une MT de bonne qualité. Nous avons utilisé cette méthode pour la construction des ressources français-chinois, comme la MT sur l'énergie et la MT du LIG-LAB. Notre système de TA a été amélioré par AI (Voir Chapitre IV et Chapitre V).

Enfin, SECTRA_W/IMAG a utilisé pour supporter l'expérimentation et la recherche d'une formule de prédiction de choix optimal par Haozhou WANG (Wang, 2015).

III.3 Utilisations plus novatrices : production de bons corpus parallèles, et post-édition de textes littéraires pour l'auto-apprentissage ou pour la traduction contributive

III.3.1 Production de « corpus parallèles » de qualité

*Thanks to SECTra_w in-built system of annotation of each translation or post-edition of a segment by a reliability level (from * to *****) and a quality score (0..20), one can extract from the TM associated to a website S a subset verifying any predicate based on levels and scores.*

To implement that, we have introduced and implemented into SECTra_w the notion of selection. *A selection is defined intentionally (by a predicate) or extensionally (by an explicit list), and can be named, for later recall.*

Take for example the TM of the website of Greater Grenoble (La Métro) that contains 2500 web pages, or about 30000 segments. More than half have been pre-translated and post-edited into Chinese for the Shanghai Expo in 2010. We may select a "quite good part" of this TM by creating the selection:

```
TM-lametro-extract-good = TM_select (lametro, [level=3 & score >=13 |
level=4 & score >=12 | level=5 & score >=11]).
```

The following example shows an even simpler extraction, from the French-Chinese part of the Demo2 TM associated with iMAG-Doc_Par_jour shown on. The predicate is simply [level=3 & score >=13], and its parameters can be directly chosen through the GUI.

	No	Pseudo Doc	Source	Cible	Stars	Notes
	1	DOC16	la salle du haut conseil située au 9ème étage de l\'institut du monde arabe, dans le 5ème.	位于巴黎5区的阿拉伯世界博物馆10楼高级理事会议厅。	3	20
	2	DOC16	le 24ème étage de la tour zamansky, l\'université pierre et marie curie, sur le campus de jussieu, dans le 5ème.	位于巴黎5区的皮埃尔和玛丽·居里大学加希耶校区的扎曼斯基大楼25楼。	3	20
	3	DOC16	le 18ème étage de la bibliothèque françois mitterrand, dans le 13ème.	位于巴黎13区的法国国家图书馆密特朗官19楼。	3	20
	4	DOC16	le 6ème étage de l\'hôtel industriel de dominique perrault, dans le 13ème.	位于巴黎13区的多米尼克·佩罗工业馆7楼。	3	20
	5	DOC16	la nuit blanche 2012 permettra aux visiteurs de découvrir la ville lumière d\'en haut grâce à 15 belvédères normalement fermés au public.	2012巴黎不眠夜将使参观者可以从15个平时不对外开放的平台发现欣赏巴黎这座光影之城。	3	20
	6	DOC16	jk rowling	j.k罗琳	3	20
	7	DOC16	en effet, jk rowling, qui a créé notre sorcier à lunettes, pourrait se replonger dans l\'univers d\'harry potter.	事实上，"创造了"我们那位眼镜魔法师的j.k罗琳，有可能会写哈利波特的魔法世界里的续集。	3	20
	8	DOC16	c\'est en 2011 que la saga de nos célèbres sorciers s\'est achevée, à l\'issue du septième livre intitulé « harry potter et les reliques de la mort » qui a été divisé en deux parties au cinéma.	2011年，第七本《哈利波特和死亡圣器》分为上下两部电影，上映结束之后，我们这著名的哈利波特系列魔法小说完结了。	3	20
	9	DOC16	cinq ans après le dernier opus de cette série à succès, jk rowling revient avec une nouvelle œuvre, pour adultes cette fois.	在这成功的系列小说最后一章完结的5年之后，jk罗琳带着她的新作品回归了。这次是给大人看的小说。	3	20
	10	DOC16	et la star des librairies a su entretenir le mystère.	这位图书史上的明星会将这光芒延续下去。	3	20
	11	DOC14	la joie	欢乐	3	20
	12	DOC16	« une bourgade apparemment idyllique mais qui va faire face "aux tourments les plus violents ».	"一个表面看起来非常美好的田园小镇，但是它将会面临最猛烈的动荡。"	3	20

Figure 18 : Extraction of a "good" TM from a TM produced by "natural" post-edition

*The selection obtained can then be exported, as 2 parallel files (source and post-edition) in a simple XML format (*Figure 19*). SECTra_w also provides additional information (TM, Last updated, Duration of post-editing, post-editor, etc.), and other available download formats (TMX, TXT, and CSV). These data can be used later to "feed" an empirical Moses-based MT system that will become specialized to that website*[43].

[43] We are running such an experiment but cannot describe it here for lack of space.

Figure 19 : Export of a « good » part of a TM

That possibility is very interesting in the current context. It has been proven that MT systems can be specialized to sublanguages and produce outputs of very high usage value (Chandioux, 1988) (Isabelle, 1987). That means that the outputs are quite readable, and very cheap to post-edit to produce professional quality output.

In recent experiments with a Paris-based multilingual content processing firm, a Moses instance built from a high proportion of a 300K bi-segment TM mixed with a standard parallel corpus extracted from EuroParl (Koehn, 2005) got a BLEU (Papineni et al., 2002) score of about 70%. At this high level, BLEU correlates with usage value: it takes typically 10-15 minutes only to post-edit the equivalent of 1 standard page (250 words, or 400 kanjis), instead of 1 hour to produce a draft translation. But that method works only if a parallel corpus specialized to the sublanguage at hand is available, and that is quite rare in practice.[44]

The situation is similar if the considered MT system is built by an "expert" method (as TAUM-METEO and then METEO).

For example, there is no available parallel Chinese↔French corpus for e-mails, chats, and short technical notes. Building a parallel corpus from scratch is not an option because of the cost of the operation and the scarcity of translators knowing both languages and the technical terms.

Using an iMAG offers a graceful way to solve that difficulty. Whatever MT systems are available, one can begin without any delay to start the bilingual service needed (a web-based chat, for example), routing messages and documents through web pages, and using iMAGs to make them accessible (and improvable) in the desired languages. After a while, the TM-S dedicated to the (empirically defined) sublanguage of S will contain enough "good" bi-segments to extract them and use them to build a specialized instance of an MT system (for example, a specialized Moses-S system[45] *).*

An important point here is that, in order to encourage end users to post-edit, post-editing should be made very simple and user-friendly. One should refrain from transforming it into a debugging environment for some MT systems. That would also go against the principle to be open to as many MT systems as possible.

[44] Remember: in 2001, Language Weaver (LW) claimed « to be able to produce an MT system overnight » from a large enough parallel corpus. While that was undoubtedly true, LW produced actually only 4 MT systems in 4 years… because parallel corpora corresponding to the translation needs of solvable clients were and are hard to find.

[45] We have built a French-Chinese Moses system for iMAG-LIG, based on 12000 already post-edited segments.

III.3.2 « Voyage au centre de la terre » de Jules Verne

Nous avons post-édité 21 chapitres du roman « VOYAGE AU CENTRE DE LA TERRE » de Jules Verne du français vers le chinois. Dans la Figure 20, nous présentons une capture d'écran de post-édition de ce roman.

Figure 20 : Capture d'écran de iMAG français→chinois pour « Voyage au centre de la terre »

Tableau 26 montre la statistique de la post-édition des 21 chapitres.

Tableau 26 : Statistique sur 21 chapitres de « Voyage au centre de la terre »

Chapitre	Segments	Mots	Mots/Seg (moyenne)	Page_std (250 mots)	Temps SECTra (secondes)	Temps/page_std en minutes.
CH1	76	1377	18,12	5,51	1009	3,10
CH2	104	1391	13,38	5,56	1565	4,69
CH3	44	592	13,45	2,37	742	5,22
CH4	106	1362	12,85	5,45	1477	4,52
CH 5-6	279	4160	14,91	16,64	3536	3,54
CH 7-9	399	6213	15,57	24,85	3534	2,37
CH 10-12	319	5154	16,16	20,62	4941	4,01
CH 13-15	254	5407	21,29	21,63	3190	2,46
CH16-18	333	4890	14,68	19,56	3252	2,77
CH19-21	202	2932	14,51	11,73	1942	2,76
Au total	2116	33478	15,49	133,91	27188	39,71

III.3.3 « The Book of Me » de Powers

L'expérimentation de post-édition de *« THE BOOKS OF ME »* de Richard Powers sur iMAG est pilotée par Laurent Besacier en 2014. Dans cette section, Nous montons la statistique, le résultat et l'évaluation. Citons ici (Besacier, 2014).

Corpus et statistiques de post-édition

L'œuvre, composée de 545 segments et 10731 mots est divisée en trois blocs identiques. Le Tableau 27 résume le nombre de mots des données source et cible (TA ou PE[46]). Sans surprises, un ratio supérieur à 1,2 est observé entre cible française (TA) et source anglaise. On constate cependant que ce ratio tend à diminuer après post-édition de la sortie française.

Tableau 27 : Corpus source, cible traduite et cible corrigée

Itération (nb. seg)	Anglais (nb. mots)	TA Français (nb. mots)	PE Français (nb. mots)
It.1 (184)	3593	4295	4013
It.2 (185)	3729	4593	4202
It.3 (176)	3409	4429	3912
Total (545)	10731	13317	12127

Le point de vue des lecteurs sur la traduction post-éditée

Neuf lecteurs ont accepté de lire l'œuvre traduite et ont répondu à un questionnaire, toujours ouvert sur fluidsurveys.com[47]. La version pdf de l'essai traduit ainsi que le fichier tableur rassemblant les résultats du sondage sont également rendus disponibles dans github. Après trois questions permettant de mieux cerner le profil du lecteur, une première partie (5 questions) interroge les lecteurs sur la lisibilité et la qualité du texte littéraire traduit. Une seconde partie (7 questions) vérifie que certaines subtilités du texte ont été bien comprises.

Le point de vue du traducteur de R. Powers

Pour finir cette étude pilote, un dixième lecteur a été sollicité : le traducteur français de l'auteur, J-Y Pellegrin, enseignant chercheur à Paris-Sorbonne. Son avis est résumé ici sous la forme de questions-réponses. Le manque de place ne nous permet pas de commenter ces remarques mais nous pensons qu'elles sont assez explicites pour être délivrées en l'état.

Lisibilité ?

"Le texte auquel vous êtes parvenu restitue une image fidèle du contenu de l'article de Powers. Le pari de la lisibilité est gagné et certains passages (notamment ceux qui portent sur les aspects scientifiques de l'expérience décrite) sont très convaincants."

Imperfections ?

"Il reste bien sûr des imperfections, des lourdeurs, voire des erreurs ponctuelles, qui appellent une correction"

Principales erreurs ?

"Le défaut le plus répétitif, celui dont souffre d'ailleurs le travail de tout traducteur débutant, est le calque syntaxique, là où le français structure différemment la phrase .../... On comprend, mais ça ne sonne pas vraiment français"

"Autre défaut assez fréquent, la perte des idiomatismes du français au profit d'anglicismes. Parfois ces anglicismes peuvent être plus dérangeants lorsqu'ils

[46] La post-édition utilisée ici est obtenue après chaque itération du processus; la dernière étape de révision n'est donc pas prise en compte à ce stade.

[47] https://fluidsurveys.com/surveys/manuela-cristina/un-livre-sur-moi-qualite-de-la-traduction/?TEST_DATA=

flirtent avec le franglais comme dans « connaissances actionnables » (p. 18) au lieu de « connaissances pratiques / utilisables ». "

"Un troisième défaut tient à la non prise en compte de certains repères culturels.../...Par exemple, Powers fait plusieurs références à la topographie de Boston qui donnent lieu à des inexactitudes dans la traduction : « la rivière Charles » par exemple (p. 12) qui n'est pas une rivière mais plutôt un fleuve ; c'est pourquoi on traduira par « la Charles River » ou simplement « la Charles »"

Ce texte pourrait-il servir de base de départ à un traducteur littéraire professionnel ?"

"Instinctivement, je serais tenté de répondre non pour l'instant, parce que, dès son premier jet, le traducteur possède des réflexes qui lui permettent de produire un texte plus « propre » que celui auquel vous êtes parvenu .../... Cependant, ce traducteur passera plus de 25 heures à produire les 42 feuillets de 1500 signes correspondant au texte de Powers. À raison de 7 feuillets par jour en moyenne, il faut 6 journées de 8h pour venir à bout du texte .../... Si, en revanche, je pouvais ne travailler que sur votre texte, en oubliant complètement celui de Powers parce que j'aurais la garantie que votre traduction ne comporte aucune erreur, ni oubli, ni aplatissement par rapport à l'original, mais qu'elle demande simplement à être améliorée, rendue plus fluide, dans un français plus authentique, les choses seraient différentes et le gain de temps sans doute considérable.".

III.3.4 « IITB : Monastery, Sanctuary, Laboratory » de Rohit Manchanda

Le livre retraçant cinq premières décennies de l'IIT-Bombay, « MONASTERY, SANCTUARY, LABORATORY », a été terminé et annoncé le 5 septembre 2008. Rédigé par Rohit Manchanda, et publié par Macmillan Inde, le livre retrace le parcours de l'IIT-Bombay depuis ses débuts jusqu'à présent sous la forme d'un récit historique.

EN 2012, à l'occasion d'un séjour à l'IIT Bombai professeur invité, C. Boitet a créé une IMAG pour traduire ce livre vers le hindi à l'aide de volontaires. Dans la Figure 21, on montre un chapitre étant traduit de l'anglais en hindi.

Figure 21 : Exemple de post-édition d'un chapitre de « Monastery, Sanctuary, Laboratory: 50 Years of IIT-Bombay » de Rohit Manchanda

Partie B Construction de systèmes de TA spécialisés à des sous-langages en français ↔ chinois

Introduction

Pour construire des systèmes de TA fr-zh adaptés aux besoins potentiels des clients de L&M (surtout EDF et RENAULT), nous avons d'abord cherché une mémoire de traductions pour au moins un des sous-langages envisagés, mais il n'y en avait aucune. Nous avons alors produit (par post-édition de TA produites par GT) une MT de bonne qualité de 9000 segments, à partir de laquelle nous avons construit un système initial de TA fr-zh, meilleur que GT sur ce sous-langage.

En contexte de recherche, nous avons d'autre part construit trois systèmes de TA dans le cadre du projet TABE-FC. Les données d'entraînement contiennent deux parties. Une partie a été extraite de segments parallèles trouvés à partir de sites boursiers, l'autre partie a été obtenue via la post-édition, par deux étudiantes chinoises, des prétraductions (produites par GT) de pages de sites boursiers publiés en français ou en anglais. Nous avons alors construit des systèmes de TA pour les couples où nous avons assez de données (français-chinois, fr-en, en-zh), et nous les utilisons pour accéder en chinois aux pages Web de ces sites via la plate-forme SECTRA_W/IMAG. Nous avons également créé des systèmes chinois-français à partir de MT utilisées « à l'envers », car nous n'avions personne capable de post-éditer du chinois→ français. Les résultats, comme on peut s'y attendre dans ce cas de figure, sont très mauvais pour la tâche de compréhension, et tout juste utilisables pour la post-édition par des francophones comprenant très bien le chinois.

Enfin, nous avons étudié les conditions dans lesquelles on peut utiliser avec profit la technique d'apprentissage incrémental proposée dans MOSES (Koehn et al., 2007). Pour des sites Web comme celui du LIG, il semble qu'il faille partir d'environ 30000 segments déjà post-édités (600K mots, ou 2400 pages), avec des incréments des 200 à 300 bisegments, et des « étapes » de 20 incréments.

Dans le Chapitre IV, nous passons revue système de TA français↔chinois en contexte industriel, et décisions nos efforts pour en construction un pour L&M. Dans le Chapitre V, nous présentons le suite de cette action, mimée cette fois-ici en contexte de recherche en coopération avec l'université de Xiamen. Enfin, dans le Chapitre VI, nous mettrons que la technique d'apprentissage incrémentale (AI) de MOSES utilisée avec profit, au moins dans certains conditions.

Chapitre IV Revue des systèmes TA français ↔ chinois en contexte industriel

Résumé

Ce chapitre présente une revue des systèmes de TA, directs ou passant par l'anglais, permettant de traduire entre chinois et français, ainsi que certains besoins dont nous avons eu connaissance à l'occasion de notre activité dans L&M.

Introduction

Beaucoup d'entreprises ont besoin de systèmes de TA français↔chinois pour aider à traduire des documents français ou des documents chinois. Elles demandent que le système de TA soit privé, et que la traduction soit vraiment fiable (pour leur sous-langages), même si la fluidité n'est pas parfaite. Pour pouvoir arriver à cela, nous avons commencé par passer en revue l'histoire du développement des systèmes de TA concernant le chinois, depuis les tout premiers travaux en Chine (dès 1957). Nous avons ensuite comparé 4 systèmes de TA opérationnels et disponibles sur le Web ou sur des serveurs privés sur la paire de langues français↔chinois. Aucun de ces systèmes n'est satisfaisant dans ces contextes, car (1) ou bien l'information « *sort* » de l'entreprise, (2) ou bien les licences en intranet sont trop chères, et (3) de toutes façons, la qualité des TA brutes est jugée insuffisante, et il n'y a pas moyen de l'améliorer en spécialisant le système au sous-langage concerné.

Après avoir présenté la demande de grosses sociétés et un état de l'art des systèmes actuels dont on peut penser qu'ils pourraient être utilisés par ce type de société (ou d'organisme) de façon opérationnelle, nous décrivons les travaux réalisés à L&M pour construire des systèmes de TA pour le chinois basés sur Moses, utilisables à terme par des clients potentiels.

IV.1 Demande de grosses sociétés

Lingua et Machina édite une application Web de gestion des contenus multilingues en entreprise appelée LIBELLEX[48]. Cette plate-forme intègre divers outils d'aide à la traduction (concordances bilingues, outils d'extraction et de gestion de terminologies, mémoires de traductions, systèmes de traduction automatique et outils de gestion de projets de traduction).

Les entreprises clientes de L&M, comme EDF et RENAULT, ont des filiales en Chine. Il y a beaucoup d'échanges internes aux entreprises entre la France et la Chine, comme les courriers électroniques, les comptes rendus, et les rapports. Ces textes doivent être traduits (en français ou en chinois), et pas en anglais, langue dans laquelle ni les uns ni les autres ne sont à l'aise.

Ces textes, et en particulière les rapports, contiennent généralement beaucoup de termes spécialisés. Dans de tels contextes, recourir à des traducteurs humains n'est pas une option, même si on avait de très bonnes mémoires de traductions. Il faudrait en effet qu'ils soient excellents et compétents dans la terminologie des domaines concernés, et donc très chers. Mais, de toutes façons, on n'arriverait pas à satisfaire la demande de « *temps réel* », et on n'a en réalité pas besoin de traductions de qualité professionnelle. Il suffit qu'elles soient assez compréhensibles et fiables, et surtout qu'on puisse améliorer « *en ligne* » celles qui contiennent des contresens et de faux termes.

[48] http://www.libellex.fr/

Pour certains textes quotidiens comme les courriers électroniques et les réunions à distance, on a simplement besoin d'une traduction « compréhensible » et en temps réel, mais pas d'une traduction parfaite.

Les entreprises clientes ont besoin d'une solution moins chère, plus rapide, et donc automatisée, mais les systèmes existants ne sont pas satisfaisants. Tout d'abord, l'information privilégiée de l'entreprise est considérée comme secrète, confidentielle. On ne peut donc pas utiliser les systèmes publics comme GT, parce que l'entreprise ne veut pas divulguer des informations propriétaires. Ensuite, un système de TA français-chinois commercial, comme Systran Enterprise Server, est très coûteux (15000€[49] pour déployer une instance de SYSTRAN ENTERPRISE SERVER 7 chez un client). Enfin, la qualité de traduction pour la paire de langues français↔chinois n'est pas suffisante. Le résultat de la TA est souvent incompréhensible.

Pour obtenir un système de TA dont le résultat brut (non post-édité) est jugé comme « *suffisant* » par une entreprise cliente potentielle de L&M, la seule solution est de « *personnaliser* » son système de TA français-chinois, et cela quel que soit le paradigme de TA utilisé. En ce qui concerne L&M, le choix s'est porté vers le paradigme de la TA statistique, et sur le développement à l'aide de l'outil MOSES.

IV.2 État de l'art de la TA du chinois

La recherche sur la TA du chinois a commencé depuis la deuxième moitié des années 1950. Les chercheurs se sont concentrés principalement sur les 4 paires de langues chinois↔anglais et chinois↔russe, mais il n'y pas eu alors de recherches sur les 2 paires français↔chinois. La première expérience concernant à la fois le français et le chinois a été faite par le professeur Feng Zhiwei en 1981-82 sur une maquette chinois→français/anglais/allemand/japonais en ARIANE-78 (Feng, 1981).

Les éditeurs de systèmes de TA qui proposent le couple français↔chinois utilisent l'approche dite du « pivote textuel », c'est-à-dire qu'ils appliquent successivement les « *paires* » français↔anglais et anglais↔chinois, en utilisant l'anglais comme "pivot textuel". Mais les résultats sont bien pires que ceux déjà jugés très peu satisfaisants obtenus avec l'anglais.

La section IV.2.1 propose un rapide historique de la recherche sur la TA du chinois, et dans la section IV.2.2, nous testons quatre systèmes de TA qui proposent la paire de langues français↔chinois.

IV.2.1 Historique

La Chine est le quatrième pays qui s'est lancé dans la traduction automatique (TA), à la suite des Etats-Unis (en 1951), du Royaume-Uni (1955), du Canada (Booth à Saskatoon, venant d'Angleterre) et de l'Union soviétique (1954, Lyapunov et Bagrinovskaya, Novosibirsk). En 1959, des chercheurs de l'Institut de technologie de l'informatique et de l'Institut de linguistique de l'Académie Chinoise des Sciences (ACS, 中国科学院, Chinese Academy of Sciences) menèrent la première expérience de traduction automatique en russe→chinois.

Le plus important des projets de TA chinois→anglais a commencé à l'Université d'État de l'Ohio en juillet 1961. C'était un projet de recherche mis en place avec le soutien de la « National Science Foundation » (NSF) sous la direction de William S. Y. Wang. Au milieu des années 1960, Wang rejoignit le groupe de Berkeley et continua sa recherche dans le cadre du projet POLA, toujours sur la TA chinois→anglais.

[49] http://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications

En même temps, le centre de recherche d'IBM à Yorktown travaillait sur un système chinois →anglais, fonctionnant sur les mêmes principes que le système russe→anglais, et utilisant aussi le tout nouveau et très fameux disque photoscopique (King and Chang, 1963).

En 1972, l'Université chinoise de Hong-Kong (香港中文大学, Chinese University of Hong Kong) proposa le système CULT (Loh and Kong, 1979), développé pour traduire des textes mathématiques du chinois vers l'anglais. Les résultats étaient très bons, grâce à une importante pré-édition manuelle.

Au début des années 1980, de nombreux instituts et universités ont été impliqués dans des recherches sur la TA. L'Institut de Technologie de Harbin (ITH, 哈尔滨工业大学, Harbin Institute of Technology) et l'Université du Nord-Est (UNE, 东北大学, Northeastern University) ont commencé leur recherches sur des systèmes de TA chinois↔anglais au milieu des années 1980. De plus, l'Université de Nankin (南京大学, Nanjing University) a commencé une recherche sur la TA japonais↔chinois durant cette période.

Pendant les années 1980, il y a eu une très grand activité commerciale au Japon, et au moins 30 sociétés informatiques (comme FUJITSU, HITACHI, NEC, SHARP, TOSHIBA, etc.) ont développé des logiciels de TA pour japonais↔anglais. Certains, notablement FUJITSU (avec ATLAS-II dû à H. Uchida), ont développé de gros prototypes pour d'autres langues, dont le chinois, mais ne les ont pas commercialisés à l'époque, car seul le couple anglais-japonais semblait pouvoir apporter un ROI[50] réel.

En 1987-93, beaucoup d'instituts et d'universités chinois ont aussi participé au projet « *Joint study for Multilingual Machine Translation* », qui a été financé par le gouvernement japonais[51], et visait à un objectif ambitieux : produire un système de TA à « pivot » sémantique de haute qualité sur le modèle d'ATLAS-II pour cinq langues asiatiques (japonais, chinois, thaï, malais et indonésien) et l'anglais.

Durant la période 1980-2005, tous les systèmes de TA adoptèrent approche "experte" à base de règles, tandis que certains d'entre eux utilisaient aussi une approche à base d'exemples comme complément.
Il y a trois systèmes très connus : (1) MT-IR-EC, un système de TA anglais→chinois pour traduire les titres et les catalogues des journaux, développé par le Research Institute of Post and Telecommunication Science, (2) KY-1, un système de TA anglais→chinois développé par l'académie des sciences militaires, qui est aussi le cœur du premier système commercial de TA, TRANSTAR (Dong, 1990), et (3) HUAJIAN, un système de TA chinois→anglais développé par HUAJIAN CO. LTD.

De 1995 à 2005, beaucoup d'équipes de recherche et d'entreprises ont publié des systèmes de TA, comme GAOLI (GAOLI CO. LTD), CCID (CCID GROUP), et KINGSOFT QUICK TRANSLATION (KINGSOFT CO. LTD).

La recherche sur la TA statistique a commencé en Chine à partir de 2004-2005. En 2006, cinq équipes de recherche (l'Institut de technologie de l'informatique de l'ACS, l'institut d'Automatique de l'ACS, l'Institut du logiciel de l'ACS, l'Université de Xiamen et ITH) ont publié un système de TA statistique en source ouvert, SILK ROAD (Silkroad, 2006). En 2011, UNE a publié un nouveau système de traduction automatique en source ouvert, NIUTRANS (Xiao et al., 2012).

[50] ROI = Return On Investment, ou "retour sur investissement"

[51] Essentiellement, par l'ODA (Overseas Development Agency) du METI.

Aujourd'hui, beaucoup de systèmes de TA proposent le couple français↔chinois, comme GT, SYSTRAN, REVERSO, BING, etc., mais il est difficile de trouver un système de TA français↔chinois de qualité d'usage correcte, sans doute car ils passent presque tous par l'anglais (« *pivot textuel* »).

IV.2.2 Expérimentations

IV.2.2.1 Systèmes étudiés

Nous avons fait des expériences sur quatre systèmes de TA français→chinois pour évaluer la qualité de TA. Nous avons d'abord testé GT, puis trois systèmes de TA statistique entraînés par les boîtes à outils disponibles en source ouvert (MOSES V2.0, JOSHUA V5.0, NIUTRANS V1.3).

Pour la construction des 3 systèmes de TA statistique, nous avons utilisé les mêmes données d'entraînement et les mêmes données de test. Ces données ont été extraites du corpus MULTIUN[52] (Eisele and Chen, 2010). Nous avons pris 1M de bisegments fr-zh comme données d'entraînement, 1K comme données de développement (*Tuning*), et aussi 1K comme données de test. Tous les textes chinois ont été segmentés par le segmenteur *Stanford*[53] (Chang et al., 2008), et nous avons utilisé le segmenteur de MOSES (un script perl *tokenizer.perl*) pour segmenter le texte français.

Tableau 28 : Statistique sur les données

	Nb de Segments	**Nb de mots français**	**Nb de caractères français**	**Nb de caractère chinois**
Données d'apprentissage	1M	28 951 255	199 486 574	35 109 257
Données de développement	1K	21 758	151 369	32 165
Données d'évaluation	1K	25 251	176 492	31 752

Notre expérimentation contient 2 étapes. Dans première, nous évaluons les systèmes de TA, par rapport avec le score BLEU. Dans la deuxième, nous utilisons SECTRA_W pour nous aider à évaluer les TA à la qualité d'usage du résultat et au degré d'automaticité.

La qualité d'usage et le degré d'automaticité sont obtenus à partir des temps de post-édition en utilisant les formules proposées par C. Boitet dans ses cours au NII (Boitet, 2009).

[52] http://opus.lingfil.uu.se/MultiUN.php

[53] http://nlp.stanford.edu/software/segmenter.shtml

Tableau 29 : Formule d'évaluation de l'automaticité et de la qualité d'un système de TA

1. Automaticity (MT module only): (taken from NII lecture notes by Boitet, 2009)

$$A = 1 - \frac{T(post_édition_hum)}{T(interaction_hum)}$$

Ex : A = 83,3% if first draft takes 1h per standard page (of 250 words or 1550 characters) and human interaction takes 10 mn/page (minutes per page).

2. Quality (wrt HT) in %

$$Q = 1 - (2/100 \times \frac{Tpe_{total}(for\ the\ task)}{Thum_{estim}(for\ the\ task)} \times Thum_{std_{page}}(\text{mn}))$$

Ex :

$Q = 40\%$ if $Tpe_{total} = 30mn/p$ $(8/20)$

$Q = 60\%$ if $Tpe_{total} = 20mn/p$ $(12/20)$

$Q = 90\%$ if $Tpe_{total} = 5mn/p$ $(18/20)$

$Thum_{estim}$est le temps moyen de traduction humain par page, estimé dans la tâche en cours. Sans plus d'information, on l'évalue à 60 mn.

IV.2.2.2 Expérimentation avec GT

Depuis 2007, GT fournit un service de traduction français↔chinois. Dans cette expérience, nous avons traduit un texte du français vers le chinois en utilisant GT. Le texte en entrée source contient 1K phrases françaises du corpus MULTIUN. Dans le Tableau 30, nous montrons 3 phrases extraites de ces 1K. Les références ont été produites par nous (par PE).

Tableau 30 : Exemple de traduction de GT

ID	Segment	Traduction de GT	Référence	Trace
1	Le Secrétaire général souhaite vivement que le plan-cadre d'équipement soit achevé d'ici à la mi-2014 et le nécessaire sera fait pour atteindre cet objectif, en contrôlant bien la portée du projet, en effectuant rapidement les réinstallations et en suivant de très près chaque activité de sorte qu'elle soit réalisée dans les délais prescrits.	秘书长 真诚 希望 ， 基本建设 总 计划 通过 2014 年 中期 完成 ， 必要 的 工作 将 实现 这 一 目标 ， 控制 项目 的 良好 范围 ， 迅速 进行 搬迁 及 以下 非常 接近 每 个 活动 ， 使 其 上 进行 的 时间 。	秘书长 的 坚定 目标 是 ， 到 2014 年年 中 完成 基本 建设 总 计划 ， 并 将 通过 控制 规模 、 加快 搬迁 和 加紧 监测 每 项 活动 的 时间表 ， 全力 实现 这 一 目标 。	秘书长 的 ~~秘书长~~ 坚定 ~~真诚~~ 目标 ~~希望~~ 是 ， ~~基本建设~~ 到 ~~总~~ 2014 ~~计划~~ 年年 通过 中 ~~2014~~ 完成 年 基本建设 ~~中期~~ 总 ~~完成~~ 计划 ， ~~必要~~ 并 的 ~~工作~~ 将 实现 通过 ~~这~~ ~~一~~ ~~目标~~ ~~，~~ 控制 ~~项~~ ~~目~~ 规模 的 ~~、~~ ~~良好~~ 加快 ~~范围~~ 搬迁 ~~，~~ 和 ~~迅速~~ 加紧 ~~进行~~ 监测 ~~搬迁~~ 每 ~~及~~ 项 ~~以下~~ 活动 ~~非常~~ 的 ~~接近~~ 时间表 ~~每个~~ ~~活动~~ ， ~~使~~ 全力 其 实现 ~~上~~ 这 ~~进行~~ 一 ~~的~~ 目标 ~~时间~~ 。
2	Au cours des six dernières années, la Commission s'est engagée dans des négociations complexes concernant ce projet, qui ne sera réalisé au mieux et dans les limites du budget imparti que si le financement est rapide.	在 过去 的 六 年 里 ， 该 委员会 已 从事 有关 项目 复杂 的 谈判 ， 这 将 是 最 好 的 实现 ， 并 在 规定 的 预算 范围 内 ， 如果 资金 是 快速 的 。	过去 六 年 来 ， 本 委员会 就 该 项目 进行 了 复杂 的 谈判 。 这个 项目 必须 及时 获得 资金 ， 才 能 有效 运作 不 超出 预算 。	过去 ~~的~~ 六 年 ~~里~~ 来 ， ~~该~~ 本 委员会 ~~已~~ 就 ~~从事~~ 该 ~~有关~~ 项目 进行 了 复杂 的 谈判 ~~，~~ ~~。~~ ~~这~~ 这个 ~~将~~ 项目 ~~是~~ 必须 ~~最~~ 及时 ~~好~~ 获得 ~~的~~ 资金 实现 ， ~~并~~ 才 ~~在~~ 能 ~~规定~~ 有效 ~~的~~ 运作 预算 不 ~~范围~~ 超出 ~~内~~ 预算 ~~，~~ ~~如果~~ ~~资金~~ ~~是~~ ~~快速~~ ~~的~~ 。
3	Afin de garantir que la	为 确保 满足 委员	为了 确保 委员会	为了 ~~为~~ 确保 ~~满足~~ 委员会 ~~的~~ 不

ID	Segment	Traduction de GT	Référence	Trace
	Commission réponde aux attentes et marque une réelle différence au Burundi, tous ceux qui sont impliqués doivent examiner la meilleure façon de soutenir la mise en œuvre des engagements identifiés en tenant compte des différences dans les capacités et l'expertise.	会 的 期望 ， 并 在 布隆迪 的 一个 真正 的 区别 ， 所有 参与 应 考虑 如何 最 好 地 支持 确定 的 承诺 的 执行 情况 ， 同时 考虑 到 在 能力 和 专业 知识 的 差异 。	不负众望 ， 给 布隆迪 带来 切实 的 变化 ， 所有 相关 方 都 必须 考虑 如何 才 能 最 好 地 支 助 承诺 的 执行 ， 并且 铭记 各方 在 能力 和 专门 知识 方面 的 差异 。	负众望 期望 ， ~~并~~ 给 ~~在~~ 布隆迪 ~~的~~ 带来 ~~一个~~ 切实 ~~真正~~ 的 ~~区别~~ 变化 ， 所有 ~~参与~~ 相关 ~~应~~ 方 都 必须 考虑 如何 才 能 最 好 ~~地~~ 地支 ~~支持~~ 助 ~~确定~~ ~~的~~ 承诺 的 执行 ~~情况~~ ， ~~同时~~ 并且 ~~考虑~~ 铭记 ~~到~~ 各方 在 能力 和 ~~专业~~ 专门 知识 方面 的 差异 。

Le score BLEU de GT est 38.25%.

IV.2.2.3 Expérimentation avec Moses

MOSES (Koehn, Hoang et al., 2007) propose l'ensemble des outils nécessaires à la construction d'un modèle de traduction. Un décodeur permet aussi d'utiliser ces outils afin de produire la traduction d'un texte source. C'est un outil sous licence libre.

Tout d'abord, nous calculons des alignements de mots en utilisant GIZA++ (Och, 2003), qui implémente les algorithmes des modèles IBM 1-5 (Brown et al., 1993) et HMM (Vogel et al., 1996). On utilise les alignements pour construire la table de traductions. Enfin, un modèle de réordonnancement est construit, contenant les informations sur les positions dans les phrases des mots traduits par rapport aux mots traduits précédemment. Le modèle de langue est construit à l'aide de l'outil IRSTLM (Federico et al., 2008).

La construction de le système de TA MOSES a pris 15 heurs (15h 32mn) pour finir la procédure d'entraînement (du prétraitement jusqu'à l'évaluation BLEU), Le score BLEU est 36,72%.

IV.2.2.4 Expérimentation avec Joshua

JOSHUA (Li et al., 2009) est un décodeur développé d'abord pour utiliser le modèle hiérarchique. Il est accompagné de l'ensemble des outils nécessaires à son fonctionnement : alignement (avec GIZA++), construction de la table de traductions, décodage, optimisation des poids, minimisation d'erreur, et calcul du modèle de langue cible. Depuis la version 6, il supporte le modèle à fragments (*chunks*). Il intègre un segmenteur du chinois, et on utilise donc son script *pipeline.pl*[54] pour entraîner le système de TA. JOSHUA demande d'écrire les paramètres dans un fichier de configuration. Les paramètres sont dans Tableau 31.

Tableau 31 : Paramètres de configuration de Joshua

```
$JOSHUA/scripts/training/pipeline.pl    // le script de pipeline
--rundir 1M    //le répertoire de travail
--source fr    // langue source
--target zh    // langue cible
--corpus 1M/train/train  // données d'entrainement
--tune 1M/tune/tune // données de développement
--test 1M/test/test // données de test
--lm-order 5
--aligner giza
```

[54] http://joshua-decoder.org/6.0/pipeline.html

La construction du système JOSHUA a pris 18 heures (18 h 12mn) pour la procédure d'entrainement (du prétraitement jusqu'à l'évaluation de BLEU). Le score BLEU est 32,18%.

IV.2.2.5 Expérimentation avec NiuTrans

NIUTRANS (Xiao, Zhu et al., 2012) est une boîte à outils en source ouvert permettant d'entrainer un système de TA statistique. Il est développé en C++ par l'UNE (东北大学). Actuellement, NiuTrans supporte déjà le modèle syntagmatique (PBMT) et le modèle hiérarchique.

Pour adapter des données à NIUTRANS, tout d'abord, on doit prétraiter les données. Nos données sont d'abord segmentées en mots (*tokenisation*) par le segmenteur STANFORD. Ensuite, on utilise les scripts perl, fournis par NIUTRANS, pour normaliser les segments[55]. Enfin on produit le fichier *aligment.txt*[56].

Après la préparation des données, nous avons pris 10 heures (10 h 37 mn, sans compte le temps d'alignement) pour entraîner le système. Nous avons obtenu un score BLEU de 33,19%.

IV.2.2.6 Description du résultat

Pour la comparaison de la qualité de traduction des systèmes de TA, nous avons entraîné les systèmes de TA dans ces mêmes conditions (les outils, les données et le matériel, sauf pour GT). Nous prenons en compte le temps d'entrainement, et le score BLEU. Le système de TA construit avec MOSES a la meilleure qualité de traduction parmi les systèmes de TA. C'est sans doute, grâce au corpus d'entrainement adapté au même domaine qu'il est un peu mieux que GT.

Le score BLEU n'est en pratique pas bien corrélé à la qualité d'usage de la TA. Pour l'évaluer, nous ajoutons les segments source et les résultats de TA dans SECTRA_W, puis nous post-éditons et calculons la distance de post-édition. Dans le Tableau 32, nous montrons un exemple de résultat de TA, et nous pouvons voir la « *Trace* » du calcul de distance entre la référence et le résultat de TA.

Tableau 32 : Comparaison d'exemples de traductions obtenues par TA et d'une référence

Source	le Secrétaire général souhaite vivement que le plan - cadre d ' équipement soit achevé d ' ici à la mi - 2014 et le nécessaire sera fait pour atteindre cet objectif , en contrôlant bien la portée du projet , en effectuant rapidement les réinstallations et en suivant de très près chaque activité de sorte qu ' elle soit réalisée dans les délais prescrits .
Reference	秘书长 的 坚定 目标 是 ， 到 2014 年 年 中 完成 基本建设 总 计划 ， 并 将 通过 控制 规模 、 加快 搬迁 和 加紧 监测 每 项 活动 的 时间表 ， 全力 实现 这 一 目标 。
Google	秘书长 真诚 希望 ， 基本建设 总 计划 通过 2014 年 中期 完成 ， 必要 的 工作 将 实现 这 一 目标 ， 控制 项目 的 良好 范围 ， 迅速 进行 搬迁 及 以下 非常 接近 每个 活动 ， 使 其 上 进行 的 时间 。
Trace (PE vers la référence)	秘书长 的 ~~秘书长~~ 坚定 ~~真诚~~ 目标 ~~希望~~ 是 ， ~~基本建设~~ 到 ~~总~~ ~~计划~~ ~~通过~~ 2014 年 ~~中期~~ 年 中 完成 ~~一~~ 基本建设 ~~必要~~ 总 ~~的~~ 计划 ~~工作~~ ， 并 将 ~~实现~~ 通过 ~~这~~ ~~一~~ ~~目标~~ ~~一~~ 控制 ~~项目~~ 规模 ~~的~~ ~~、~~ ~~良好~~ 加快 ~~范围~~ 搬迁 ~~一~~ 和 ~~迅速~~ 加紧 ~~进行~~ 监测 ~~搬迁~~ 每 ~~及~~ 项 ~~以下~~ 活动 ~~非常~~ 的 ~~接近~~ 时间表 ~~每个~~ ~~活动~~ ， ~~使~~ 全力 ~~其~~ 实现 ~~上~~ 这 ~~进行~~ 一 ~~的~~ 目标 ~~时间~~
Moses	秘书长 真诚 希望 这项 框架 计划 的将 于 Mi- 2014 年 完成，并 将 采取 步骤 实现 这 一 目标，通过 控制 的 项 目的 范围，迅速 重新 安置 或 采取 行动，不 影响 的 每 项 具体 活动 进行 规定 时限 内 提出 。
Trace (PE vers la référence)	秘书长 ~~真诚~~ 的 ~~希望~~ 坚定 ~~这项~~ 目标 ~~框架~~ 是 ~~计划~~ ， ~~的将~~ 到 ~~于~~ 2014 ~~Mi-~~ 年 ~~2014 年~~ 年 ~~完成，~~ 中 ~~并~~ 完成 ~~将~~ 基本建设 ~~采取~~ 总 ~~步骤~~ 计划 实现 ， ~~这~~ ~~并~~ ~~一~~ 将 ~~目标，~~ 通过 控制 ~~的~~ 规模 ~~项~~ ~~、~~ ~~目的~~ 加快 ~~范围~~ 搬迁 ~~，~~ 和 ~~迅速~~ 加紧 ~~重新~~ 监测 ~~安置~~ 每 ~~或~~ 项 ~~采取~~ 活动 ~~行动，不~~ ~~影响~~ 的 ~~每~~ 时间表 ~~项~~ ， ~~具体~~ 全力 ~~活动~~ 实现 ~~进行~~ 这 ~~规定~~ 一 ~~时限~~ 目标 ~~内~~ ~~提出~~ 。

[55] http://www.nlplab.com/NiuPlan/NiuTrans.YourData.html

[56] http://www.nlplab.cn/NiuTrans.Phrase.html

Joshua	秘书长 真诚 希望这项框架 计划 设备 完成 在 这里，mi 2014 年 之后 、 和 必要 的 工作 将 为 实现 这 一 目标，在 控制 两 个 项目 的 范围 、 执行 快速 的 重新 安置，通过 密切 关注 每个 活动，在 规定 时限 内 提出 。
Trace (PE vers la référence)	秘书长 ~~真诚~~ 的 ~~希望这项框架~~ 坚定 ~~计划~~ 目标 ~~设备~~ 是 ~~完成~~ ， ~~在~~ 到 ~~这里，mi~~ 2014 年 ~~之后~~ 年 ~~、~~ 中 ~~和~~ 完成 ~~必要~~ 基本建设 ~~的~~ 总 ~~工作~~ 计划 ~~将~~ ， ~~为~~ 并 实现 将 ~~这~~ 通过 ~~一~~ ~~目标，在~~ 控制 ~~两~~ 规模 ~~个~~ ~~项目~~ ~~的~~ ~~范围~~ 、 ~~执行~~ 加快 ~~快速~~ 搬迁 ~~的~~ 和 ~~重新~~ 加紧 ~~安置~~ 监测 ~~，~~ 每 项 ~~通过~~ 活动 ~~密切~~ 的 ~~关注~~ 时间表 ~~每个~~ ， ~~活动~~ 全力 ~~，~~ 实现 ~~在~~ 这 ~~规定~~ 一 ~~时限~~ 目标 ~~内~~ ~~提出~~ 。
NiuTrans	秘书长 强烈 希望 计划 框架 的 设备 或 已 完成 的 在 2014 年 Mi 必须 将 实现 这 一 目标，通过 检查 影响控制 两 个 项目 的 范围、执行 的 快速 的 重新 安装 又 密切 注视 活动，实现 的 规定 时限 。
Trace (PE vers la référence)	秘书长 ~~强烈~~ ~~希望~~ ~~计划~~ ~~框架~~ 的 ~~设备~~ 坚定 ~~或~~ 目标 ~~已~~ 是 ~~完成~~ ， ~~的~~ 到 ~~在~~ 2014 年 ~~Mi~~ 年 ~~必须~~ 中 ~~将~~ 完成 ~~实现~~ 基本建设 ~~这~~ 总 ~~一~~ 计划 ~~目标~~ ， ~~，~~ 并 将 通过 ~~检查~~ 控制 ~~影响控制~~ 规模 ~~两~~ ~~、~~ ~~个~~ 加快 ~~项目~~ 搬迁 ~~的~~ 和 ~~范围、执行~~ 加紧 ~~的~~ 监测 ~~快速~~ 每 ~~的~~ 项 ~~重新~~ 活动 ~~安装~~ ~~的~~ ~~又~~ 时间表 ~~密切~~ ， ~~注视~~ 全力 ~~活动~~ ~~，~~ 实现 ~~的~~ 这 ~~规定~~ 一 ~~时限~~ 目标 。

Dans l'Annexe 8, on donne 50 segments en « *vue SECTra/Post-édition* », montrant pour chaque segment le texte source, la PE, et les TA en mode « *Trace* ». Voici un exemple de résultat d'évaluation sur le segment présenté dans le Tableau 33.

Tableau 33 : Exemple de résultat d'évaluation

Mots	**TA**	**TPE**	**TPE/p.std**	**DistPE (α=0,2 ; ß=0,8)**	**Q**
53	Google	127s	16,0 mn	Dc: 112 ; Dw: 60 ; D= 70,4.	68%
	Moses	123s	15,5 mn	Dc: 126 ; Dw: 66 ; D= 78	69%
	Joshua	153s	19,2 mn	Dc: 137 Dw: 68 ; D= 81,8	61,6%
	NiuTrans	142s	17,8 mn	Dc: 117 Dw: 64 ; D= 70	64,4%

Conclusion

Nous avons comparé les 3 systèmes de TA et GT sur le BLEU et la qualité d'usage. Nous avons choisi le "meilleur" système, et il peut traduire les phrases français vers le chinois. Mais la qualité de traduction n'est pas satisfaisante. La qualité est limitée par la taille de corpus, le domaine de traduction, le lexique, etc. Comment construire un système de TA français-chinois en haute qualité ? Pour nous, c'est un vrai défi.

IV.3 Construction de systèmes de TA pour le chinois basés sur Moses en contexte industriel

L'étude précédente nous a montré qu'il n'y avait pas pour l'instant de système français↔chinois pouvant être utilisé tel quel, ou adapté rapidement, pour satisfaire les besoins des grandes sociétés en général, et des clients potentiels de L&M en particulier.

Nous avons donc essayé de construire nous-même un système français-chinois à partir d'une MT correspondant aux besoins d'au moins un client potentiel de L&M. Malheureusement, aucun n'avait de telle MT. Nous en avons donc construit une, mais, faute de ressources (en post-éditeurs), nous n'avons pas pu dépasser 9000 bisegments (dans ce cas, 112500 mots ou 450 pages standard). Les résultats ont été encore pires qu'avec les 4 systèmes étudiés plus haut. Notre hypothèse est qu'il aurait fallu disposer d'une MT d'au moins 30000 à 50000 segments.

IV.3.1 Choix du sous-langage et des couples à traiter

Comme le client potentiel le plus prometteur pour L&M était EDF, nous avons cherché à construire un système pour EDF. C'est une très grosse entreprise spécialisée dans le domaine

de l'énergie électrique, qu'elle soit produite dans des centrales nucléaires, hydrauliques, à charbon, à gaz, éoliennes, ou photovoltaïques.

Depuis une quarantaine d'années, EDF est implantée en Chine, où elle a construit des centrales nucléaires, et travaillé avec d'innombrables cadres, ouvriers et ingénieurs chinois. Beaucoup de documents ont été traduits, dans les deux sens, et nous espérions avoir accès à des documents parallèles, ou au moins à de grosses mémoires de traductions, en supposant que des outils comme SDL TRADOS ou DÉJÀ VU avaient été utilisés pour produire ces traductions.

Nous comptions bien sûr choisir comme sous-langage objet du système de TA à construire celui correspondant à la MT la meilleure en qualité et la plus grande en volume. Malheureusement, nous n'avons rien pu obtenir du tout. Peut-être ces textes parallèles ou ces MT existent-ils et sont-ils cachés, peut-être n'ont-ils jamais été créés, nous n'en savons rien. Sachant que la traduction n'est presque jamais consolidée dans les comptes des entreprises, et est le plus souvent sous-traitée de manière opportuniste, il est possible que la seconde hypothèse soit la bonne.

Quoi qu'il en soit, en 2013, L&M n'avait pas pu avoir accès à un corpus parallèle ou à une MT d'EDF français→chinois permettant de développer un système de TA, qu'il s'agisse d'un système MOSES (il aurait fallu entre 20K et 30K "bons" bisegments) ou d'un système ARIANE (à règles et dictionnaires), pour lequel il aurait fallu des corpus parallèles ou comparables de 2K à 3K segments (pour l'étude typologique) et un dictionnaire bilingue de 10K à 20K entrées.

Nous avons alors décidé d'essayer de construire un système MOSES à partir d'un corpus parallèle que nous construirions nous-même à partir d'un corpus bilingue le plus "adapté" ou "vraisemblable" possible, puis, s'il était trop petit, à partir d'un corpus monolingue complémentaire que nous traduirions.

Nous savions que les besoins d'EDF étaient dans les deux sens (français↔chinois). Nous nous sommes concentré sur le sens français→chinois, car nous savions que nous pourrions nous-même évaluer et post-éditer les résultats, alors que nous n'avions personne dans notre environnement qui comprenne bien le chinois technique et soit de langue maternelle française.

Cependant, nous avons aussi fait quelques essais en chinois→français, "pour voir", en nous disant que, si les résultats étaient encourageants, nous pourrions peut-être écrire des parties de cette thèse en chinois, les faire traduire par GT ou par notre système, et les faire ensuite réviser par des chercheurs du laboratoire, compensant leur ignorance du chinois par leur connaissance du domaine. Comme on pouvait s'y attendre, cet espoir a été totalement déçu, et nous ne nous étendrons pas sur cet essai.

IV.3.2 Recherche infructueuse de corpus parallèles adaptés

Notre première idée a été d'essayer d'extraire un corpus bilingue français→chinois concernant un des domaines d'EDF à partir des corpus parallèles librement disponibles sur le Web.

La performance d'un système de traduction automatique statistique (TAS) dépend fortement de la taille et de la qualité du corpus parallèle utilisé pour l'entraînement. Les ressources actuelles en corpus parallèles bilingues ou multilingues libres de droits proviennent généralement d'institutions internationales. C'est le cas du corpus « EUROPARL » (Koehn, 2005) extrait des délibérations du Parlement européen, du corpus « CANADIAN HANSARDS », contenant les transcriptions en français et en anglais des débats du Parlement canadien, et du corpus JRC-ACQUIS qui fournit une quantité comparable de textes législatifs européens en 22 langues (Steinberger et al., 2006). En ce qui concerne le chinois, il y a beaucoup de corpus parallèles

anglais→chinois et chinois→anglais, notamment celui du journal XINHUA NEWS (Graff et al., 2003), mais très peu de corpus français→chinois.

Or, pour construire un système de TAS français→chinois, il faut disposer d'un corpus parallèle français→chinois (dans le bon sens) pour entraîner les modèles. Nous nous sommes tourné vers le corpus parallèle MULTIUN français-chinois, qui a été construit par extraction du site Web des Nations-Unies, puis nettoyé et converti au format XML par Andreas Eisele et Yu Chen (Eisele and Chen, 2010) en 2010. En février 2013, la version alignée de ce corpus a été publiée sur le site Web OPUS[57] (Tiedemann, 2012). Voici les corpus que nous avions pu collecter à ce point (Tableau 34).

Tableau 34 : Corpus collectés en cherchant des corpus pour français→chinois

Nom	Direction	Nb segments	Nb mots source	Mots fr / segment	caractères zh / segment	Mots fr / caractères zh
MultiUN	en-zh	8,8M	220,4M	24,97	71,31	285,54%
MultiUN	fr-zh	8,7M	243,8M	27,94	71,94	257,33%

Que pouvait-on en espérer ? *A priori,* peu, car, quand on entraîne un système de TAS avec un tel corpus "généraliste" pour la traduction dans un domaine précis, par exemple l'énergie, on obtient d'habitude de mauvais résultats. Nous avons fait l'expérience, qui a confirmé cette crainte.

Nous en avons conclu qu'un système de TA français-chinois entraîné seulement avec le corpus MULTIUN ne pourrait pas répondre aux besoins des entreprises clientes de L&M. En nous inspirant de publications mentionnant la possibilité de mélanger un petit corpus spécialisé à un grand corpus généraliste, nous avons alors décidé d'essayer cela, et de construire un corpus parallèle français-chinois spécialisé au sous-langage des notes techniques et des courriels concernant le domaine de l'énergie.

Pour cela, nous sommes parti des sites Web d'EDF en France et en Chine, car nous avons remarqué qu'ils contiennent beaucoup de segments français et chinois presque parallèles.

Nous avons extrait des textes français et chinois à partir de ces sites Web à l'aide de l'outil BOILERPIPE (Kohlschütter et al., 2010). Mais ces textes ne peuvent pas être utilisés tels quels pour entraîner un système de TA, il faut d'abord les segmenter, les nettoyer, les aligner, et enfin extraire des bisegments réellement parallèles (en relation de traduction).

Tout d'abord, nous avons débruité les textes, en supprimant les segments inutiles comme les liens (par exemple, http://…), les chiffres, les dates, etc.

Ensuite, nous avons normalisé l'encodage des caractères en transformant tout en UTF-8 (certains textes français étaient en codage ASCII MAC ou WINDOWS, et les textes chinois étaient le plus souvent en GB-2312-80).

Après avoir nettoyé le "bruit", nous avons procédé à l'alignement au niveau des segments (phrases ou titres) en utilisant l'outil LF ALIGNER[58]. Nous avons finalement obtenu un corpus parallèle d'environ 3K bisegments. Un extrait en est donné à l'Annexe 9.

Cette petite quantité de données n'est pas suffisante pour entraîner un système Moses, mais, mélangée à ce que nous avions extrait de MULTIUN, elle a suffi pour améliorer un peu le

[57] http://opus.lingfil.uu.se/MultiUN.php

[58] http://sourceforge.net/projects/aligner/

système de TA, en lui faisant apprendre des termes comme « *le noyau des atomes* », « *l'hydraulique* », « *le charbon propre* », etc.

IV.3.3 Production de corpus par PE de résultats de Google

Pour augmenter la quantité des données d'entrainement, nous avons construit une MT par post-édition des sorties de la 1ère version de notre système. C'est un bon moyen de produire des MT de bonne qualité (Wang and Boitet, 2013).

Pour post-éditer plus vite et mieux, nous avons utilisé la plate-forme SECTRA_W/IMAG. Nous avons divisé notre MT en domaines, comme les nouvelles, les reportages et les pages Web de Wikipédia. Nous avons transformé tous les textes monolingues en des fichiers html, nous les avons placés dans une hiérarchie de fichiers mise sur le serveur du laboratoire, et nous avons créé une iMAG dédiée à ces fichiers.

Les prétraductions ont été fournies par GT. Après la post-édition, nous avons sélectionné, pour construire notre MT, les bisegments que nous estimions adaptés à notre besoin. Cette sélection était basée sur le niveau de fiabilité (d'une étoile "☆" à cinq étoiles "☆☆☆☆☆") et sur la note de qualité (de 0 à 20) associés à chaque segment et à chaque langue cible dans la MT. Le prédicat de sélection était:

```
(fiab = 3 && score ≥ 12) || (fiab = 4 && score ≥ 11) || (fiab = 5 && score ≥ 10)
```

Nous avons finalement obtenu 6000 segments parallèles. En les ajoutant aux 3000 segments parallèles extraits à partir des sites Web d'EDF, nous avons au total collecté 9000 bisegments (environ 450 pages standard) de qualité suffisante pour construire un système. Cependant, nous n'avons pas de garantie que ce "noyau" représente bien le sous-langage, qui pour nous reste inconnu et inconnaissable, des notes techniques d'EDF sur l'énergie (Tableau 35).

Tableau 35 : Exemples de bisegments français→chinois parmi les 9000 collectés ou produits

Français	**Chinois**
Charbon propre	清洁煤发电
EDF Asie	EDF 亚洲
Activités	业务概览
Charbon propre	洁净煤
En Chine, le charbon représente près de 80% de la production d'électricité et devrait continuer d'occuper une place majoritaire dans l'avenir (plus de 60 % à l'horizon 2020).	中国煤电约占全国总发电量的 80%，今后还会继续占有主导地位（预计 2020 年占 60%以上）。
Pour limiter les impacts sur l'environnement, la Chine développe des centrales à charbon à haut rendement moins polluantes. En s'appuyant sur ses compétences d'ingénierie, EDF prend part à ces projets. Ils permettent au Groupe de consolider et de développer son expérience pour faire face aux besoins qui pourraient émerger en Europe dans l'avenir.	为了减轻煤电对环境的影响，中国致力于发展高效、低污染的燃煤电厂。法国电力集团以专业技能为依托，参与中国洁净煤火电项目。通过参与项目建设，法国电力集团将巩固和发展其火电技术，应对欧洲未来可能出现的需求。
EDF a signé plusieurs accords de coopération avec des producteurs nationaux d'électricité, portant sur le développement conjoint de projets électriques, par exemple les Groupes de Trois Gorges, Guodian, Datang, etc.	法国电力集团已与国电、三峡集团、大唐等多家国有大型电力公司签署了多项电力合作协议。
French Investment Guangxi Laibin Electric power Co (FIGLEC) - Chine est une filiale à 100 % du groupe EDF. La société est propriétaire de la centrale de Laibin B (d'une puissance de 720 MW), exploitée par SYNERGIE, aussi filiale d'EDF.	广西来宾法资发电有限公司是法国电力集团的全资子公司，拥有两台单机容量为 360 兆瓦的机组，总装机容量 720 兆瓦。

IV.3.4 Construction de systèmes français→chinois

IV.3.4.1 Composants

Nos systèmes de TA sont construits avec MOSES, qui fournit des outils optimisés pour réaliser l'entraînement, mais qui ne contient pas d'outil pour traiter le chinois. À L&M, nous disposions aussi de la boîte à outils MYRIAM, qui intègre le segmenteur du chinois de XELDA[59] (Xerox Linguistic Development Architecture), et un programme java pour normaliser les phrases. Pour l'alignement des mots, nous avons utilisé l'outil MGIZA (Gao and Vogel, 2008), qui propose une implémentation efficace et parallèle de GIZA++.

IV.3.4.2 Paramétrisation et construction de 2 modèles de TAS

Les paramètres de ces systèmes ont été optimisés de manière usuelle avec l'outil MERT (Minimum Error Rate Training) (Och, 2003). Les traductions produites sont évaluées avec la mesure BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), ainsi qu'avec notre « distance mixte de PE » (MTER), qui est bien corrélée à la qualité d'usage.

Nous avons ainsi défini deux modèles. L'un est un modèle générique qui est entraîné avec le corpus parallèle MULTIUN, et l'autre est un modèle spécifique entraîné avec notre mémoire de bisegments post-édités. Quand on a de nouveaux segments post-édités, on n'a pas besoin de réentraîner le modèle avec toutes les données. On met seulement à jour le modèle spécifique. C'est très utile dans le cas où l'on a un très gros modèle générique qu'on souhaite utiliser dans un nouveau système, sans avoir à le réentraîner.

Tableau 36 : Comparaison des temps d'entraînement de Moses

Système de TA	Quantité de données	Temps d'entraînement	Temps du réentraînement avec 10K nouveaux bisegments
Un modèle	2M+9K bisegments	17 h 35mn	19 h 21mn
Modèle générique	2M+9K bisegments	13 h 10mn (10h30+2h40)	2 h 40mn
Modèle spécifique			

Nous avons ajouté 10K nouveaux bisegments, puis comparé les temps du réentraînement. Notre stratégie est de gagner 86,2% du temps sur le réentraînement avec 10K nouveaux segments. Le Tableau 37 présente les caractéristiques du serveur utilisé pour l'expérimentation.

Tableau 37 : Configuration de la machine

Nombre de processeurs	460
Nombre de cœurs par processeur	2
Thread	4
Mémoire	8G

IV.3.4.3 Choix des données pour le modèle général

Le corpus MultiUN français-chinois contient 9,7 M de phrases parallèles, composé d'environ 300 millions de mots français et environ 600 millions de caractères chinois, soit environ 315M mots du dictionnaire). Ce corpus contient des segments composés de chiffres, de

[59] http://www.xrce.xerox.com/About-XRCE/History/Historical-projects/XeLDA

[60] Intel(R) Core(TM) i7-3770 CPU @ 3,40GHz

numéros, ou encore de dates ; ces segments ne sont pas utiles pour notre système. Nous les avons donc éliminés. À la fin, il nous restait environs 8,3 M bisegments.

Nous n'avons pas utilisé tout le corpus pour entraîner le modèle générique, parce que nous avons d'abord fait une expérience pour trouver la quantité appropriée des segments. Nous avons testé l'entraînement du système Moses français-chinois avec des quantités différentes (1M, 2M, 3M, 4M, et 5M) de bisegments, et calculé le score BLEU pour chacune de ces quantités. Les résultats sont présentés dans le tableau ci-dessous (Tableau 38). Les scores de 2M vers 5M sont égale les mêmes. On prend en compte le temps d'entrainement, et on choisit 2M bisegments pour entraîner le modèle générique.

Tableau 38 : Scores BLEU pour différentes tailles du corpus d'entraînement

Nombre de bisegments	BLEU	Temps d'entrainement
1M	49,78%	16 h
2M	52,48%	18 h
3M	52,25%	21 h
4M	54,31%	32 h
5M	55,52%	42 h

Nous avons prétraité les bisegments (le corpus MULTIUN et la MT) en utilisant MYRIAM. Plus précisément,

- nous avons d'abord normalisé les segments (en convertissant les entités HTML, en séparant les ponctuations, et en ôtant la casse).
- nous avons ensuite utilisé XELDA pour marquer les segments français et chinois.

IV.3.4.4 Construction de systèmes de TA

Nous avons entraîné le modèle général construit uniquement à partir du corpus MultiUN (2M). Le tableau ci-dessous montre quelques chiffres concernant ce corpus.

Tableau 39 : Statistiques sur le corpus MultiUN

	Français	Chinois
Segment	2M	
Mots	58M	50M
Caractères	398M	90M
Pages_std	38K	226K

Le modèle de traduction est entraîné sur 2M segments tirés de MULTIUN. La partie cible du corpus a été utilisée pour produire un modèle de langue. Dans cette partie de l'entraînement, nous avons sauté l'étape usuelle d'optimisation des poids (tuning). Elle a été utilisée plus tard, pour construire un autre système, en association avec des modèles plus spécialisés.

Nous avons entraîné le modèle spécifique avec notre MT (9000 bisegments français-chinois), et collecté 1,7M segments monolingues chinois pour produire le modèle de langue cible.

Les paramètres utilisés ont été les mêmes que ceux utilisés pour l'entraînement du modèle général, sauf que nous avons utilisé la phase d'optimisation des poids (tuning).

Pour construire un système d'essai en chinois→français, nous avons utilisé les mêmes données à l'envers, en faisant comme si, étant donné un couple ($fr, $zh) de segments tels que $zh=trad($fr), on avait aussi $fr=trad($zh)). Nous savons pertinemment que c'est faux (la relation de traduction entre phrases n'est pas symétrique), mais… c'était "mieux que rien", puisque nous n'avions personne qui puisse post-éditer des résultats de TA zh→fr.

IV.3.5 Évaluations et perspectives

Pour évaluer la progression de notre système de TA "vers une cible" (pour chaque segment post-édité, la cible est sa post-édition), nous avons calculé le score BLEU pour chaque système de TA. Certes, BLEU ne mesure pas la "qualité" (ni au sens de qualité linguistique, ni au sens de qualité d'usage), et *ne peut pas* la mesurer, comme cela a été montré dans le fameux article d'Osborne, Callison-Burch et Koehn (Callison-Burch et al., 2006) "*Re-evaluating the Role of BLEU in Machine Translation Research*". Mais BLEU exprime bien une similarité textuelle, et peut être "raisonnablement" utilisé pour évaluer la progression d'un système vers une certaine "cible". Pour cela, il vaut d'ailleurs mieux n'avoir qu'une seule cible ("*traduction de référence*") par segment source, plutôt que 5, 10 ou 15 dans certaines campagnes d'évaluation. Notons que BLEU donne une mesure globale et ne donne vraiment pas d'indication fiable au niveau des segments individuels.

Nous évaluons aussi la qualité d'usage à partir du temps de post-édition (en minute par page, mn/p). Comme SECTRA associe un chronomètre à chaque segment, nous disposons du *temps primaire de post-édition,* `Tpe_1`, pour chaque segment édité à travers cette interface. Nous disposons aussi du *temps total de post-édition*, `Tpe_tot`, pour des sessions de post-édition[61]. Nous pouvons en déduire, pour chaque segment, le temps total `Tpe_tot` ainsi que le *temps secondaire de post-édition,* `Tpe_2`, qui correspond au temps passé à chercher des équivalents dans les lexiques et bases terminologiques, ou à communiquer avec d'autres personnes pour trouver une bonne traduction d'une expression "hors dictionnaire". Typiquement, ce temps représente les 2/3 ou les 3/4 du temps total en contexte de traduction professionnelle.

Au début, nous avons post-édité 300 segments (Voir un exemple des données de test dans le Tableau 40), du français vers le chinois (à partir des prétraductions fournies par GT), et les avons utilisées comme données de test pour notre évaluation.

Tableau 40 : Exemple de données de test

ID	Segment	Traduction de Google	PE par humaine	Trace
26	Nucléaire : la Chine adopte l'EPR	核电： 中国采用EPR	核电： 中国采用欧洲压水堆技术	核电 ：中国 采用 ~~EPR~~ 欧洲 压水 堆 技术
27	Au Laos, le projet de centrale hydraulique Nam Theun 2 (1070 MW) est porté par la société de projet Nam Theun 2 Power Company (NTPC), dont le groupe EDF est le premier actionnaire avec 40 % des parts.	在老挝， 液压动力项目南屯2 （1070兆瓦） 支持的项目公司南屯2电力公司 （NTPC），法国电力集团与40%的股权的第一大股东。	在老挝， 中央液压草案南屯2号项目 （1070兆瓦） 的支持， 该项目公司南屯2电力公司 （NTPC），法国电力集团是拥有40%股权的最大股东。	在老挝 ， 中央 液压 ~~动力~~ 草案 ~~项目~~ 南 屯 2 号 项目 （ 1070 兆瓦 ） 的 支持 ~~的~~ ， 该 项目 公司 南 屯 2 电力 公司 （ NTPC ） ， 法国 电力 集团 ~~与~~ 是 拥有 40% ~~的~~ 股权 的 ~~第一~~ 最 大 股东 。
28	D'une capacité de 715 MW, la centrale a été mise en service en février 2005. Elle bénéficie des technologies éprouvées des turbines les plus récentes, ainsi que des derniers retours d'expérience des centrales à gaz construites par EDF qui en a assuré la construction et la livraison « clé en main	容量为715兆瓦， 该厂已于2005年2月它已被证明的经验， 最后返回建造的EDF最新的涡轮机和燃气电厂保证了技术建设和交付“交钥匙”， 现在参与其运作。	该厂容量为715兆瓦， 2005年2月初投入运行。 这一项目的汽轮机采用了最新经过验证的技术， 吸取了EDF燃气机组最新的反馈经验，以"交钥匙"模式承担工程的建设， 和参与电厂的运行	该厂 容量 为 715 兆瓦 ， ~~该厂~~ 2005 ~~已~~ 年 ~~于~~ 2 ~~2005 年~~ 月初 ~~2~~ ~~月~~ 投入 ~~它~~ 运行 ~~已~~ ~~。~~ ~~被~~ 这 ~~证明~~ 一 项目 的 ~~经验~~ 汽轮机 ~~，~~ 采用 ~~最后~~ 了 ~~返回~~ 最新 ~~建造~~ 经过 验证 的 ~~EDF~~ 技术 ~~最~~ ， ~~新~~ 吸取 ~~的~~ 了 ~~涡轮机~~ EDF ~~和~~ 燃气 ~~电厂~~ 机组 ~~保证~~ 最新 ~~子~~ 的 ~~技术~~ 反馈 ~~建设~~ 经验 ~~和~~ ~~，~~ ~~交付~~ 以 ~~“~~ ~~"~~ 交 钥匙 ~~”~~ ~~"~~ 模式 承担 工程 的

[61] Quand nous et d'autres Chinois qui nous aident post-éditons, nous notons l'heure au début et à la fin d'une session, et aussi l'ensemble des segments post-édités durant la session. Nous en tirons un temps total moyen. Nous faisons l'hypothèse (qui semble vérifiée) que les temps sont proportionnels aux nombres de mots, et une simple règle de trois nous donne alors `Tpe_tot` pour chaque segment.

ID	Segment	Traduction de Google	PE par humaine	Trace
	» et qui participe maintenant à son exploitation.		。	建设 ， ~~现在~~ 和 参与 ~~其~~ 电厂 ~~运作~~ 的 运行 。
29	Le développement de l'énergie nucléaire est un enjeu majeur pour la Chine et le reste du monde dans le cadre de la préservation de l'environnement et de la réduction de l'effet de serre.	核电 的 发展 是 中国 和 重大 问题 世界 保护 环境 ， 减少 温室 效应 的 范围 内 。	发展 核能 对 中国 和 世界 都 具有 非常 重要 的 意义 ， 是 保护 环境 和 减缓 温室 效应 重要 的 途径 。	发展 核能 ~~核电~~ 对 ~~的~~ 中国 ~~发展~~ 和 ~~是~~ 世界 ~~中国~~ 都 ~~和~~ 具有 ~~重大~~ 非常 ~~问题~~ 重要 ~~世界~~ 的 意义 ，~~是~~ 保护 环境 ~~，~~ 和 ~~减少~~ 减缓 温室 效应 重要 的 ~~范围~~ 途径 ~~内~~ 。

Le Tableau 41présente les statistiques des données de test. Ensuite, nous avons continué à post-éditer des résultats de TA, mais plus ceux de GT : nous avons continué à mettre dans notre MT les résultats de GT, mais nous avons post-édité les résultats de nos systèmes, puis réinjecté ces nouvelles cibles comme des « références » dans le processus d'apprentissage, etc.

Tableau 41 : Statistiques des données de test

Nb de segments	Nb de mots Par segment (source)	Nb de p.std (source)	Nb de caractères par segment (cible)	Nb de p.std (cible)	Tpe_p.std
300	26,1	31,32	28,3	21,2	6,2 mn

Nous traduisons les segments source avec 3 systèmes, GT, système de TA entrainé avec le corpus MULTIUN (2M bisegments), et notre système combiné. Les segments post-édités sont utilisés comme les références. Les scores BLEU des systèmes de TA sont montrés dans le Tableau 42, et avec un exemple de traduction.

Tableau 42 : Score BLEU et exemples de sorties de systèmes de TA

Système	BLEU	Source : Être un leader du renouveau du nucléaire dans le monde Référence : 成为 全世界 核能 复兴 的 领导者	
GT	37.86%	Traduction	作为 世界 核 复兴 的 领导者
		Trace	成为 全世界 ~~作为~~ 核能 ~~世界~~ ~~核~~ 复兴 的 领导者
Système de TA (MultiUN)	21,52%	Traduction	受 教育 复兴 和 体面 工作 问题 的 一个 世界 核
		Trace	成为 全世界 ~~受~~ 核能 ~~教育~~ 复兴 ~~和~~ ~~体面~~ ~~工作~~ ~~问题~~ 的 领导者 ~~一个~~ ~~世界~~ ~~核~~
Système de TA (MultiUN+MT)	49,37%	Traduction	成为 一个 全球 核电 复兴 的 引领者
		Trace	成为 全世界 ~~一个~~ 核能 ~~全球~~ ~~核电~~ 复兴 的 领导者 ~~引领者~~

Conclusion

Nous avons essayé de créer un système de TA français→chinois en utilisant le corpus MULTIUN et la MT, traduisant le contenu dans le domaine de l'énergie.

Au début, nous nous sommes limité à ces ressources (très peu de corpus français-chinois, en particulier, le corpus français-chinois adapté au domaine) pour construire un système de TA statistique avec MOSES. La traduction de TA est "incompréhensible".

Ensuite, nous avons testé les systèmes existants comme GT, mais le résultat n'était pas satisfaisant. Pour construire un système ayant une qualité de traduction acceptable, nous avons commencé à construire un corpus parallèle spécialisé à notre domaine. Nous avons collecté et extrait des segments parallèles à partir de sites Web, mais la quantité de segments

parallèles n'était toujours pas suffisante. Nous avons alors créé des IMAG pour des sites Web correspondant à notre sujet, et nous les avons post-édités. Les résultats de TA ont été fournis d'abord par GT, puis par notre système de TA basé sur MOSES, en construction. Nous avons obtenu plusieurs milliers de segments parallèles.

Enfin, nous avons construit un système de TA français→chinois avec un corpus "*mixé*" (corpus extrait et MT), et la qualité de traduction s'est enfin acceptable. Notre hypothèse est que nous pourrions arriver à une qualité vraiment bonne si nous avions non pas 9000, mais entre 30000 et 50000 segments. Nous n'avons pas eu les ressources suffisantes pour la tester, et espérons trouver une situation le permettant dans le futur.

Chapitre V Construction de systèmes de TA pour le chinois avec Moses en contexte de recherche : le projet TABE-FC

Introduction

Au laboratoire, notre recherche sur la TA fr-zh a été surtout menée dans le cadre du projet TABE-FC, défini et commencé grâce à un séjour sabbatique d'un an du Pr Yidong Chen de XMU. Un des buts de ce projet est de pouvoir comparer différents « paradigmes » de TA, sans se limiter aux systèmes purement « MOSES ». Il s'agit aussi de systèmes « à règles » comme NEON (de XMU), ou de systèmes empiriques utilisant des statistiques ainsi que des informations sémantiques et pragmatiques (architecture du Pr. Yidong Chen (Chen et al., 2014)).

Pour construire des systèmes de TA français-chinois de bonne qualité en contexte de recherche, nous étions libre de sélectionner des sous-langages très restreints (comme des « *brèves* » de sites boursiers), ou pour lesquels nous aurions déjà un corpus parallèle de bonne qualité (comme pour le site du LIG). Nous avons travaillé sur les deux sous-langages mentionnés ci-dessus, et présentons ce qui a été fait pour celui des sites boursiers et économiques.

Après avoir extrait environ 1000 pages Web de ces sites, nous avons extrait environ 3000 bisegments, puis les avons filtrés et avons obtenu une MT de bonne qualité. La quantité de segments parallèles n'étant pas suffisante pour entraîner un système MOSES, nous avons créé 3 IMAG pour 3 sites Web boursiers, et nous avons post-édité les segments prétraduits par GT, grâce à la participation de 4 stagiaires chinois étudiants à l'UJF. Cela fait, nous avons pu construire des systèmes Moses, et les avons évalués (distance de post-édition, temps, ainsi que BLEU et NIST). Nous sommes arrivés à des résultats compréhensibles, mais pas encore très fiables.

V.1 Buts du projet TABE-FC

Le projet TABE-FC est un projet collaboratif mené par l'Université de Grenoble, LIG-GETALP et l'Université de Xiamen (XMU, 厦门大学, Xiamen University), et visant à créer des instances d'un nouveau type de système de traduction automatique statistique utilisant des ressources lexico-sémantiques et discursives.

V.1.1 Buts théoriques

Le but concret est de développer des systèmes de TAS français↔chinois pour des sites boursiers et économiques. Comme très peu de corpus ou de dictionnaires bilingues français↔chinois sont disponibles sur Internet, l'anglais est utilisé comme « pivot » pour construire les équivalents français↔chinois par transitivité. Outre la description générale de ce projet, nous décrivons les progrès sur deux axes de recherche lies à ce projet. Pour cela, nous utilisons une méthode, proposée par XMU, d'induction de probabilité fondée sur la similarité thématique, qui produit des tables de traduction français-chinois à partir de tables de traduction français-anglais et anglais-chinois. Pour disposer de bons corpus parallèles français-chinois, nous utilisons un système Web de post-édition collaborative qui peut déclencher l'amélioration incrémentale du système de TA en utilisant des mesures d'évaluation de TA et en extrayant la "meilleure partie" de la mémoire de traductions courante.

V.1.2 Buts pratiques

Ce projet se concentre sur deux objectifs pratiques. Nous prenons qu'une utilisation opérationnelle permettrait d'obtenir une excellente qualité. Le premier est de construire un service de systèmes de TA français↔chinois. Tout d'abord, nous nous restreignons à des sous-langages observés, puis nous récupérons et construisons la ressource français↔chinois, et créons finalement un système de TA.

À cause de la limitation des ressources français-chinois, nous essayons de construire le système de TA français-chinois via une langue pivot (anglais) pour faire l'extraction de la traduction équivalent.

Le deuxième objectif est intégré notre système de TA dans SECTRA_W/IMAG (via TRADOH). Pour cela, nous créons des IMAG dédiées pour les sites Web boursiers et économiques. Notre plate-forme permet d'accéder aux sites Web en multilingue, et fournit déjà une traduction français→chinois de bonne qualité.

V.1.3 Définition du projet

Nous avons organisé ce projet avec une architecture à trois niveaux, contenant sept « missions » (Figure 22). Tout d'abord, nous travaillons pour la construction de ressources français-chinois orientées vers l'économie (corpus parallèles, ressources sémantiques bilingues, banques d'annotations discursives, etc.) en vue du développement. Ensuite, nous construirons un nouveau modèle de TA basé sur la sémantique et l'information de discoure (niveau de la recherche fondamentale). Enfin, nous prévoyons de construire un système Web de TA, et de l'appeler par SECTRA_W/IMAG ou par JIANDAN-EVAL pour améliorer la qualité de traduction des sites Web économiques (niveau de l'application).

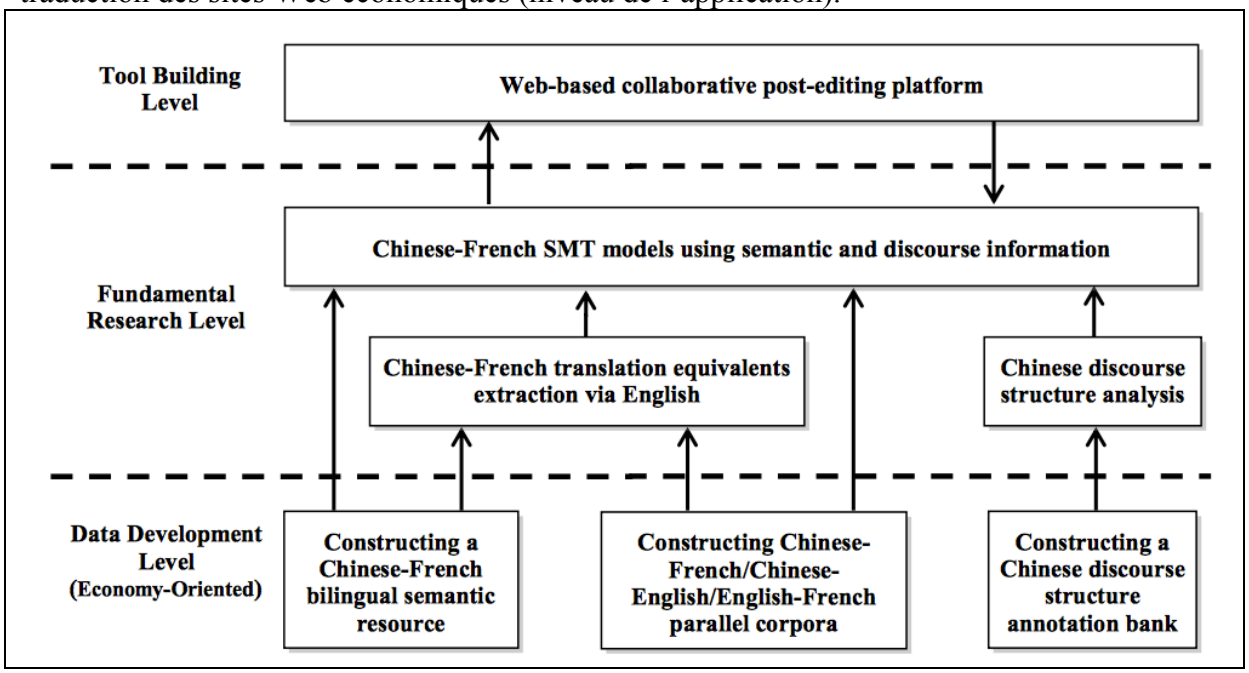


Figure 22 : Architecture à 3 niveaux et 7 « missions » du projet TABE-FC (Chen, Wang et al., 2014)

V.2 Constitution des corpus d'apprentissage

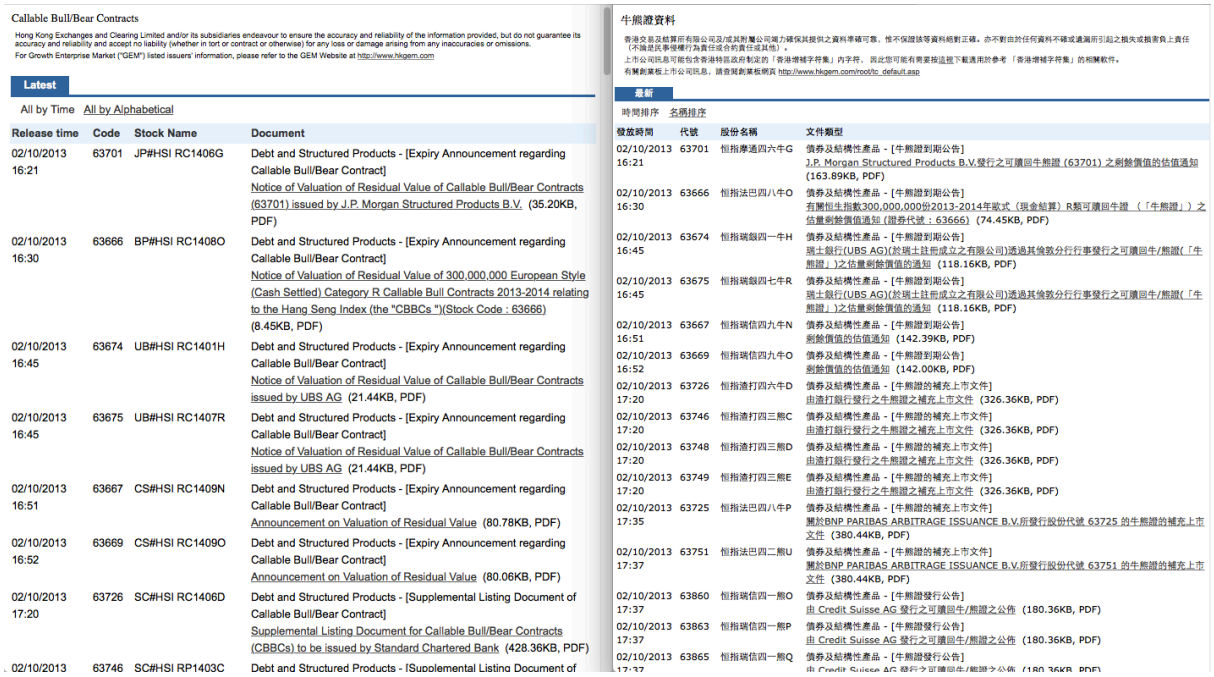
V.2.1 Recherche de sites et collecte de pages Web monolingues et bilingues

Comme déjà dit, il est difficile de construire un grand corpus parallèle chinois-français, en particulier pour le domaine économique.

Par conséquent, il est raisonnable de construire les ressources destinées à la construction de systèmes de TA chinois-français en utilisant l'anglais comme un pivot. Nous construisons un corpus anglais-chinois et le corpus anglais-français liés au domaine de l'économie. Ensuite, pour soutenir l'apprentissage des connaissances sur les transformations structurelles entre le chinois et le français, nous créons un corpus parallèle français-chinois par transitivité et révision (en PE).

Nous avons essayé de chercher une ressource existante pour construire des données d'apprentissage. Certains sites Web économiques contiennent des segments multilingues, par exemple, le site « BOURSE DE PARIS »[62] (français-anglais), le site « TMX »[63] de Toronto (anglais-français), « SZSE » de Shenzhen (chinois-anglais) （深圳证券交易所）, et le site « abnnewswire.net »[64] (multilingue).

Nous avons téléchargé les pages Web parallèles (*.html*) (Figure 23), et puis fait l'alignement et l'extraction des segments parallèles sur les pages Web en utilisant LF ALIGNER.



Callable Bull/Bear Contracts

Hong Kong Exchanges and Clearing Limited and/or its subsidiaries endeavour to ensure the accuracy and reliability of the information provided, but do not guarantee its accuracy and reliability and accept no liability (whether in tort or contract or otherwise) for any loss or damage arising from any inaccuracies or omissions.
For Growth Enterprise Market ("GEM") listed issuers' information, please refer to the GEM Website at http://www.hkgem.com

Latest

All by Time All by Alphabetical

Release time	Code	Stock Name	Document
02/10/2013 16:21	63701	JP#HSI RC1406G	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Notice of Valuation of Residual Value of Callable Bull/Bear Contracts (63701) issued by J.P. Morgan Structured Products B.V. (35.20KB, PDF)
02/10/2013 16:30	63666	BP#HSI RC1408O	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Notice of Valuation of Residual Value of 300,000,000 European Style (Cash Settled) Category R Callable Bull Contracts 2013-2014 relating to the Hang Seng Index (the "CBBCs ")(Stock Code : 63666) (8.45KB, PDF)
02/10/2013 16:45	63674	UB#HSI RC1401H	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Notice of Valuation of Residual Value of Callable Bull/Bear Contracts issued by UBS AG (21.44KB, PDF)
02/10/2013 16:45	63675	UB#HSI RC1407R	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Notice of Valuation of Residual Value of Callable Bull/Bear Contracts issued by UBS AG (21.44KB, PDF)
02/10/2013 16:51	63667	CS#HSI RC1409N	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Announcement on Valuation of Residual Value (80.78KB, PDF)
02/10/2013 16:52	63669	CS#HSI RC1409O	Debt and Structured Products - [Expiry Announcement regarding Callable Bull/Bear Contract] Announcement on Valuation of Residual Value (80.06KB, PDF)
02/10/2013 17:20	63726	SC#HSI RC1406D	Debt and Structured Products - [Supplemental Listing Document of Callable Bull/Bear Contract] Supplemental Listing Document for Callable Bull/Bear Contracts (CBBCs) to be issued by Standard Chartered Bank (428.36KB, PDF)
02/10/2013	63746	SC#HSI RP1403C	Debt and Structured Products - [Supplemental Listing Document of

牛熊證資料

香港交易及結算所有限公司及成其附屬公司竭力確保其提供之資料準確可靠，惟不保證該等資料絕對正確。亦不對由於任何資料不確或遺漏所引起之損失或損害負上責任（不論是民事侵權行為責任或合約責任或其他）。
上市公司訊息可能包含香港特區政府制定的「香港增補字符集」內字符，因此您可能有需要按這裡下載適用於參考「香港增補字符集」的相關軟件。
有關創業板上市公司訊息，請登閱創業板網頁 http://www.hkgem.com/root/tc_default.asp

最新

時間排序 名稱排序

發放時間	代號	股份名稱	文件類型
02/10/2013 16:21	63701	恒指摩通四六牛G	債券及結構性產品 - [牛熊證到期公告] J.P. Morgan Structured Products B.V.發行之可贖回牛熊證 (63701) 之剩餘價值的估值通知 (163.89KB, PDF)
02/10/2013 16:30	63666	恒指法巴四八牛O	債券及結構性產品 - [牛熊證到期公告] 有關恒生指數300,000,000份2013-2014年歐式（現金結算）R類可贖回牛證（「牛熊證」）之估量剩餘價值通知 (證券代號：63666) (74.45KB, PDF)
02/10/2013 16:45	63674	恒指瑞銀四一牛H	債券及結構性產品 - [牛熊證到期公告] 瑞士銀行(UBS AG)(於瑞士註冊成立之有限公司)透過其倫敦分行行事發行之可贖回牛/熊證(「牛熊證」)之估量剩餘價值的通知 (118.16KB, PDF)
02/10/2013 16:45	63675	恒指瑞銀四七牛R	債券及結構性產品 - [牛熊證到期公告] 瑞士銀行(UBS AG)(於瑞士註冊成立之有限公司)透過其倫敦分行行事發行之可贖回牛/熊證(「牛熊證」)之估量剩餘價值的通知 (118.16KB, PDF)
02/10/2013 16:51	63667	恒指瑞信四九牛N	債券及結構性產品 - [牛熊證到期公告] 剩餘價值的估值通知 (142.39KB, PDF)
02/10/2013 16:52	63669	恒指瑞信四九牛O	債券及結構性產品 - [牛熊證到期公告] 剩餘價值的估值通知 (142.00KB, PDF)
02/10/2013 17:20	63726	恒指滙打四六牛D	債券及結構性產品 - [牛熊證的補充上市文件] 由滙打銀行發行之牛熊證之補充上市文件 (326.36KB, PDF)
02/10/2013 17:20	63746	恒指滙打四三熊C	債券及結構性產品 - [牛熊證的補充上市文件] 由滙打銀行發行之牛熊證之補充上市文件 (326.36KB, PDF)
02/10/2013 17:20	63748	恒指滙打四三熊D	債券及結構性產品 - [牛熊證的補充上市文件] 由滙打銀行發行之牛熊證之補充上市文件 (326.36KB, PDF)
02/10/2013 17:20	63749	恒指滙打四三熊E	債券及結構性產品 - [牛熊證的補充上市文件] 由滙打銀行發行之牛熊證之補充上市文件 (326.36KB, PDF)
02/10/2013 17:35	63725	恒指法巴四八牛P	債券及結構性產品 - [牛熊證的補充上市文件] 關於BNP PARIBAS ARBITRAGE ISSUANCE B.V.所發行股份代號 63725 的牛熊證的補充上市文件 (380.44KB, PDF)
02/10/2013 17:37	63751	恒指法巴四二熊U	債券及結構性產品 - [牛熊證的補充上市文件] 關於BNP PARIBAS ARBITRAGE ISSUANCE B.V.所發行股份代號 63751 的牛熊證的補充上市文件 (380.44KB, PDF)
02/10/2013 17:37	63860	恒指瑞信四一熊O	債券及結構性產品 - [牛熊證發行公告] 由 Credit Suisse AG 發行之可贖回牛/熊證之公佈 (180.36KB, PDF)
02/10/2013 17:37	63863	恒指瑞信四一熊P	債券及結構性產品 - [牛熊證發行公告] 由 Credit Suisse AG 發行之可贖回牛/熊證之公佈 (180.36KB, PDF)
02/10/2013 17:37	63865	恒指瑞信四一熊Q	債券及結構性產品 - [牛熊證發行公告]

Figure 23 : Exemple de page Web économique parallèle

V.2.2 Nettoyage et filtrage

À partir des sites Web liés à l'échange d'actions, environ 1000 pages bilingues ont été explorées et traitées (le Tableau 43 présente les statistiques de cet ensemble de données).

[62] http://www.boursedeparis.fr

[63] http://www.tmx.com

[64] http://abnnewswire.net

Tableau 43 : Statistiques des pages Web collectées

Paire de langues	Nombre de pages	Taille
chinois-anglais	761	39,4Mo
anglais-français	250	12,5Mo

Pour obtenir un corpus parallèle propre, il est nécessaire de nettoyer les pages Web récupérées. Ce nettoyage est une opération informatique qui demande un paramétrage souvent poussé.

Nous avons présenté plus haut les problèmes à résoudre pour « filtrer » des corpus bilingues provenant de la faire du Web. Dans le cas des pages du site Web « Bourse de Hong Kong », une page Web en anglais contient le texte entouré de balises HTML, avec une valeur particulière de l'attribut *class*. Dans la Figure 24, le texte anglais et le texte chinois sont alignés, ils ont différentes valeurs de « *class* ». Nous choisissons d'abord les balises associées (<div class= '…'), puis nous comparons le nom de « *class* », et ajoutons un suffixe "_c" pour la partie chinoise. Nous extrayons enfin le texte anglais et le texte chinois. Nous trouvons qu'il y a un commentaire mélangé avec le texte, et nous ne pouvons pas être sûr que ce texte est bien traduit, donc nous l'éliminons.

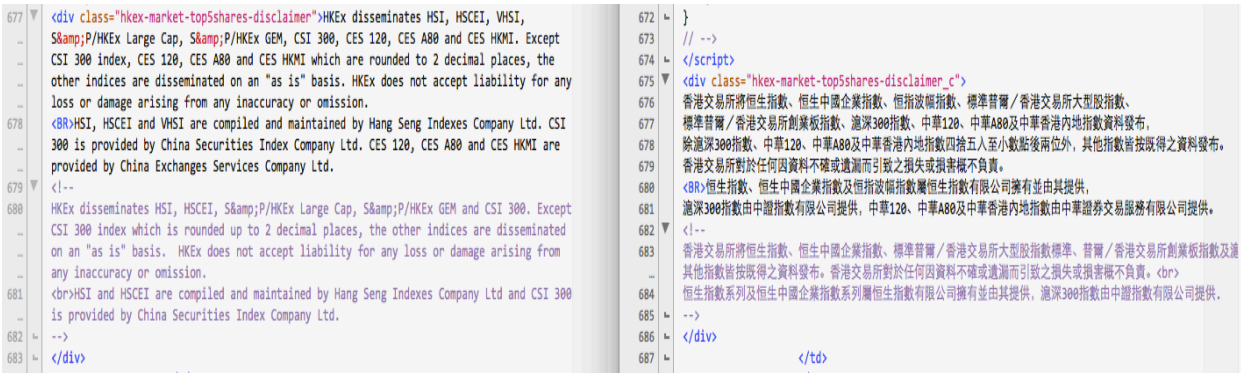


Figure 24 : Exemple d'une page Web du site de "Bourse de Hong Kong" en format html

À la fin de ce processus, nous avons obtenu 3000 bisegments « propres » extraits des sites Web boursiers, en anglais-chinois. Les segments chinois étaient en caractères traditionnels ; nous les avons convertis en caractères simplifiés, grâce au système de conversion intelligente des caractères chinois simplifiés et traditionnels (Shi et al., 2013)[65].

Tableau 44 : Exemple de conversion des caractères chinois du traditionnel vers le simplifié

Traditionnel	董事會整體負責確保集團的會計及財務匯報制度健全，並設有適當的內部監控及風險管理系統。
Simplifié	董事会整体负责确保集团的会计及财务汇报制度健全，并设有适当的内部监控及风险管理系统。
Traditionnel	財務匯報 董事會承諾以平衡及清晰可明的方式向股東及其他權益人評估集團的表現、財務狀況及前景。
Simplifié	财务汇报 董事会承诺以平衡及清晰可明的方式向股东及其他权益人评估集团的表现、财务状况及前景。
Traditionnel	這承諾包括所有發布資料，包括但不限於財務報表、按監管要求發出的公告和其他公司通訊。
Simplifié	这承诺包括所有发布资料，包括但不限于财务报表、按监管要求发出的公告和其他公司通讯。

[65] Précision : 99,991% (http://jf.cloudtranslation.cc/)

Nous les avons mis en format TXT et TMX. Voici un exemple des segments extraits au format TMX dans le Figure 25.

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
  <header
    datatype="text"
    encoding="UTF-8"
    segtype="sentence"
    adminlang="ZH"
    srclang="ZH"
    o-tmf="TW4Win 2.0 Format"
  >
  </header>
  <body>
<tu creationdate="20140116T095746Z" creationid="eco-hkex"><prop
type="Txt::Note">hkex-zh-en</prop>
<tuv xml:lang="ZH"><seg>董事会整体负责确保集团的会计及财务汇报制度健全，并设有适当的内部监控及风险管理系统。</seg></tuv>
<tuv xml:lang="EN"><seg>The Board has overall responsibility for ensuring the integrity of
the Group's accounting and financial reporting systems, and that appropriate systems of
internal control and risk management are in place.</seg></tuv> </tu>

<tu creationdate="20140116T095746Z" creationid="eco-hkex"><prop
type="Txt::Note">hkex-zh-en</prop>
<tuv xml:lang="ZH"><seg>财务汇报 董事会承诺以平衡及清晰可明的方式向股东及其他权益人评估集团的表现、财务状况及前景。</seg></tuv>
<tuv xml:lang="EN"><seg>Financial reporting The Board is committed to presenting a
balanced, clear and comprehensible assessment of the Group's performance, financial
position and prospects to our shareholders and other stakeholders.</seg></tuv> </tu>

<tu creationdate="20140116T095746Z" creationid="eco-hkex"><prop
type="Txt::Note">hkex-zh-en</prop>
<tuv xml:lang="ZH"><seg>这承诺包括所有发布资料，包括但不限于财务报表、按监管要求发出的公告和其他公司通讯。</seg></tuv>
<tuv xml:lang="EN"><seg>This commitment encompasses all published information, including
but not limited to the financial statements, regulatory announcements and other corporate
communications.</seg></tuv> </tu>
```

Figure 25 : Exemple de segments chinois-anglais extraits à partir de pages Web

Nous avons utilisé la même méthode pour extraire des segments parallèles anglais-français à partir du site Web « BOURSE DE PARIS ». Nous avons obtenu 2000 segments français-anglais (voir le Tableau 45).

V.2.3 TA par GT, puis PE (production d'un corpus parallèle)

Nous avons donc pu créer un corpus chinois-anglais par extraction de bisegments à partir des sites Web boursiers, et la traduction anglaise est d'assez une bonne qualité. Nous avons donc pu les utiliser pour entrainer le système de TA.

Pour la paire de langues français-chinois, nous n'avons pas de corpus parallèle, bien que nous ayons trouvé des sites Web boursiers et économiques disponibles en français et en anglais. C'est le cas du site « TMX » de Toronto, où le texte français est traduit par TA sans PE en anglais, mais pas en chinois.

Remarque : les utilisateurs semblent bien « couler » en anglais, mais il y a de nombreux contresens.

Pour construire le corpus français-chinois, nous créons une iMAG pour post-éditer les segments du français vers le chinois. Nous choisissons le site Web « BOURSE DE PARIS[66] » comme la ressource française. 2 étudiantes chinoises nous ont aidé à faire la post-édition dans le cadre de leurs stages TER de M1-info.

[66] http://www.boursedeparis.fr

Figure 26 : Capture d'écran de l'iMAG "Bourse de Paris" en chinois

Pour assurer la qualité des segments, il faut choisir quelques segments à post-éditer. Le site Web « Bourse de Paris » contient un centre d'apprentissage, et il présent les règles du marché, des définitions, une glossaire de bourse, etc. Nous avons post-édité les résultats de TA fournis par GT.

Dans la Figure 27, nous montrons un exemple de la traduction de GT et la post-édition faite dans SECTRA_W. GT traduit "Action" comme "Ce que l'on fait." (行动), mais ici, il faut le traduire par « stock » (股票).

Figure 27 : Comparaison de la traduction de GT et de la post-édition humaine

Les deux post-éditeurs étaient des étudiantes chinoises en informatique à l'Université Joseph Fourier, sans connaissances particulières en économie. Cependant, on voit dans la Figure 27 que les segments post-édités sont très compréhensibles et fidèles.

V.3 Construction de systèmes de TA

V.3.1 Construction de systèmes Moses "ligne de base"

Pour construire des systèmes de TA français-chinois adaptés au domaine économique et boursier, nous avons extrait des segments français-anglais, anglais-chinois à partir de sites Web boursiers et économique. Nous avons ainsi obtenu 2000 segments français-chinois via la post-édition (voir le Tableau 45).

Tableau 45 : Statistiques sur la ressource économique et boursière

Paire de langues	Nb de segments	Nb de mots (source)	Nb de mots (cible)	Nb de p_std (source)	Nb de p_std (cible)
anglais-chinois	3000	104K	128K (caractère chinois)	416	321
anglais-français	2000	60K	54K	241	215
français-chinois (PE)	2000	44K	46K (caractère chinois)	176	115

Dans le cadre du projet TABE-FC, nous avons d'abord construit deux systèmes de TA (anglais→français et anglais→chinois). Ensuite nous avons traduit les segments anglais extraits de « Bourse de Paris» vers le chinois, et les segments anglais extrait de « Shenzhen Stock Exchange (SZSE)[67] » vers le français. Enfin nous avons construit le système de TA français→chinois en utilisant les segments traduits par les systèmes de TA et les segments post-édités du français vers le chinois.

[67] http://www.szse.cn/main/

V.3.1.1 Système de TA anglais→ français

Nous sommes parti des 2000 segments parallèles anglais-français obtenus à partir du site Web « BOURSE DE PARIS », concernant le domaine boursier et économique. Comme ces données ne sont pas suffisantes pour entraîner un système de TA français→anglais avec MOSES, nous avons ajouté aux données d'entraînement 2M bisegments du corpus MultiUN anglais-français.

Le système de TA anglais→français a été utilisé pour traduire les segments anglais (3000 segments) extraits du site Web « SZSE » vers le français. Enfin, nous avons produit 3000 bisegments français-chinois, et la partie française a été traduite par le système de TA anglais→français.

V.3.1.2 Système de TA anglais→chinois

Pour construire le système de TA anglais→chinois, nous avons pris 3000 segments parallèles anglais→chinois extraits du site Web « SZSE », et nous avons ajouté aux données d'entraînement 2M bisegments du corpus MULTIUN anglais-chinois.

Le système de TA anglais→chinois a été utilisé pour traduire les segments anglais extraits du site Web « SZSE » vers le français. Enfin, nous avons produit 2000 bisegments français-chinois, et la partie chinoise a été traduite par notre système de TA anglais→chinois de bonne qualité.

V.3.1.3 Système de TA français→chinois

Nous avons traduit 3000 segments d'anglais en français, et 2000 segments d'anglais en chinois. Plus de 2000 segments ont été post-édités du français vers le chinois. Au total, nous avons obtenu 7000 bisegments français-chinois de bonne qualité.

V.3.1.4 Système de TA chinois→français

Pour l'instant, nous n'avons pas de corpus parallèle de bonne qualité pour entraîner un système de TA du chinois vers le français adapté au domaine économique et boursier, et nous n'avons pas non plus de MT post-éditée du chinois vers le français. Pour tester la traduction français→chinois, nous avons essayé d'utiliser la MT français→chinois « à l'envers ». Mais les résultats, comme on pourrait s'y attendre, sont assez désastreux.

V.3.2 Avancement de l'expérimentation

Pour obtenir plus de segments post-édités, nous avons construit deux nouvelles iMAG pour les sites Web « *SIX Swiss Exchange* (SIX)[68] » (français→chinois), et « *Hong Kong Stock Exchange* (HKEX)[69] » (anglais→chinois).

Cette expérience est toujours en cours au moment de la rédaction. Nous utilisons nos systèmes de TA et post-éditons leurs résultats sous SECTRA_W. Les systèmes de TA sont installés sur un serveur du laboratoire LIG, et on peut y accéder via TRADOH.

V.3.3 Résultats provisoires

Nous avons commencé à utiliser notre système pour post-éditer les pages Web boursières et économiques des sites « SIX », « HKEX », et « BOURSE DE PARIS ». Pour l'instant, nous n'avons pas encore assez de données pour présenter des statistiques et surtout une évaluation

[68] http://www.six-swiss-exchange.com/index.html
[69] http://www.hkex.com.hk/eng/index.htm

significatives. Nous espérons que nous pourrons fournir cela au moment de la soutenance, et que nous pourrons répondre à trois questions importantes dans ce contexte :

1. La qualité d'usage est-elle suffisante ou non sans intervention humaine ?
2. Le passage par l'anglais est-il une bonne chose (avec ou sans PE de l'anglais) ?
3. Peut-on réduire significativement le temps de recompilation d'un système MOSES avec de nouvelle données (bisegments) obtenue par PE.

Chapitre VI Démonstration de l'intérêt de l'apprentissage incrémental en TA statistique

Introduction

Dans l'objectif permanent d'améliorer la qualité de notre système de TA fr-zh, nous avons accumulé des bisegments de bonne qualité par post-édition sur la plate-forme SECTRA_W/IMAG. Étant donné la controverse qu'il y avait vers 2012 sur l'utilité de l'apprentissage incrémental pour des systèmes de type MOSES, nous avons cherché à monter une expérience pour déterminer les conditions favorables à l'utilisation de cette technique. Cette expérience a été menée en 3 phases. (1) J'ai d'abord construit un système appelé `Moses_fr-zh_v0` à partir de 10K bisegments post-édités, et j'ai travaillé seulement sur la MT du LIG-LAB, en post-éditant des segments choisis parmi les 6000 non post-édités, et en faisant des mesures sur des ensembles croissants de segments post-édités. Après 10 itérations d'apprentissage incrémental, les résultats étaient encourageants mais pas encore concluants. (2) Dans le cadre d'un stage d'été, deux étudiants chinois ont travaillé sur la post-édition en chinois d'articles en français et sur des supports de divers cours de Master 1 environ 13000 segments. (3) Enfin, disposant d'environ 30K bisegments de qualité (en fr-zh), nous avons recommencé l'expérience en améliorant son organisation et en l'automatisant.

Au total, nous avons prouvé que l'apprentissage incrémental peut améliorer la qualité de TA fr-zh, en tout cas si l'on vise un « sous-langage », et si de temps en temps, on réentraîne le système en utilisant toutes les données d'entraînement.

VI.1 Contexte

VI.1.1 Motivations

Pour obtenir une bonne qualité avec un modèle de type Moses sur un "sous-langage" précis, on n'a pas besoin de très grandes quantités de données d'entraînement (comme les 200M de bisegments du système généraliste de l'UE), mais il en faut quand même un certain nombre, en fonction de la taille et de la complexité du sous-langage en question. Une expérience précédente a montré que, à l'extrême, on pouvait se contenter de 1000 bisegments dans le cas d'un tout petit sous-langage (SMS de petites annonces en occasion automobile (Daoud, 2007), (Hajlaoui and Boitet, 2008). Dans le cas de notes techniques et de la paire français-anglais, la société L&M est arrivée à des taux de BLEU de l'ordre de 70% avec 300K exemples.

Si l'on s'attaque à un nouveau sous-langage, on n'a pas encore de corpus de bisegments de taille suffisante. On commence donc par utiliser un système généraliste existant, et à post-éditer ses résultats, jusqu'à obtenir une taille suffisante. On entraîne alors son propre système sur un corpus d'entraînement petit mais de bonne qualité, en le mélangeant, si les résultats sont meilleurs, avec une certain quantité de corpus général, et on met ce système en service. Au début, la qualité d'usage, mesurée par exemple par le temps de post-édition, n'est en général pas supérieure à celle d'un système généraliste (comme GT). Mais, si l'on produit de nouvelles versions de « son » système de TA en intégrant aux données d'entraînement les post-éditions qu'il produit sur de nouveaux segments, on y arrive.

Malheureusement, la création d'une nouvelle version consiste à "tout recompiler" et est extrêmement longue (plusieurs dizaines d'heures en temps partagé). Par exemple, on a fait une expérience avec 1,2M bisegments fr-zh sur un serveur à quatre cœurs (Intel i7-3770 CPU 3.40GHz), qui a pris 17~20h pour l'entraînement.

C'est un obstacle majeur à l'accroissement régulier de la qualité. Or, il est très important que les utilisateurs aient l'impression que « le système apprend », et, par exemple, que les contributions qu'ils ont faites dans leur activité de post-édition sont prises en compte

Fin 2011, une possibilité d'apprentissage incrémental venait d'être introduite dans MOSES, et L&M m'a demandé de l'étudier, ce que j'ai fait. C'était dans le cadre de la préparation d'une réponse à un appel d'offres, qui n'a pas été obtenu. L&M a alors provisoirement arrêté le travail sur ce point. Pour ma part, je continuais d'être intéressé par cette possibilité, et cela d'autant plus que mes expériences sur les corpus de L&M semblaient prouver que cette technique était prometteuse, même si les gains de qualité étaient nettement plus faibles que si l'on "recompilait tout", et aussi parce qu'il y avait dans le domaine diverses polémiques, l'une sur l'utilité potentielle de l'apprentissage incrémental dans MOSES, et l'autre sur l'intérêt pour les traducteurs de post-éditer des résultats de TA. J'avais en particulier lu un article mettant en doute l'utilité de cette technique (Mirkin, 2014).

Je me suis donc demandé comment je pourrais démontrer l'utilité de cette technique, et ce thème est devenu assez important dans ma recherche. J'ai expérimenté cette technique pour la première fois dans le cadre de la création d'un corpus bilingue français-chinois dans le domaine de l'énergie (voir IV.3.3).

J'avais aussi constaté que les améliorations obtenues en ajoutant des « mini-batch » et en exécutant l'AI tendaient à diminuer avec le temps, et étaient toujours bien inférieures aux améliorations obtenues par un réentraînement complet.

Comme L&M a abandonné ce projet, n'ayant pas obtenu le contrat espéré avec EDF, j'ai poursuivi cette recherche dans le cadre du laboratoire, en prenant comme base expérimentale le site du LIG, pour lequel nous avions déjà une MT 8000 segments dont 2000 post-édités.

VI.1.2 Expérience sur le site du LIG

VI.1.2.1 Motivations et conditions expérimentales

Pourquoi le site du LIG ? Le site Web LIG-Lab[70] est le site Web officiel du laboratoire LIG (Laboratoire d'Informatique de Grenoble). Le LIG rassemble près de 500 chercheurs, enseignants-chercheurs, doctorants et personnels en support à la recherche. C'est aussi le tout premier site Web pour lequel une IMAG a été créée (voir I.1.2.1.2). Un certain nombre de chercheurs, doctorants, ou stagiaires sont chinois, et ils souhaitent accéder notre site Web traduit en chinois.

Sous-langage envisagé. Le site Web du LIG contient la présentation du laboratoire, la présentation de ses 22 équipes de recherche, ses contributions du développement des aspects fondamentaux de la discipline (modèles, langages, méthodes, algorithmes) et les événements du laboratoire. Tous les contenus tournent autour de l'informatique. Une instance d'IMAG existe depuis fin 2008 et est dédiée à ce site, ou plutôt à son *sous-langage*.

Pré-condition de l'expérience. La Figure 28 montre la page d'accueil de cette IMAG dédiée (LIG-LAB) accédée en chinois. La langue source est le français, et il y a 34 langues d'accès possibles. Notre expérience porte sur la paire de langues français-chinois. L'IMAG du LIG contenait environ 8K segments français (en février 2013), et environ 2K segments avaient été post-édités.

[70] Site Web du laboratoire LIG : https://www.liglab.fr

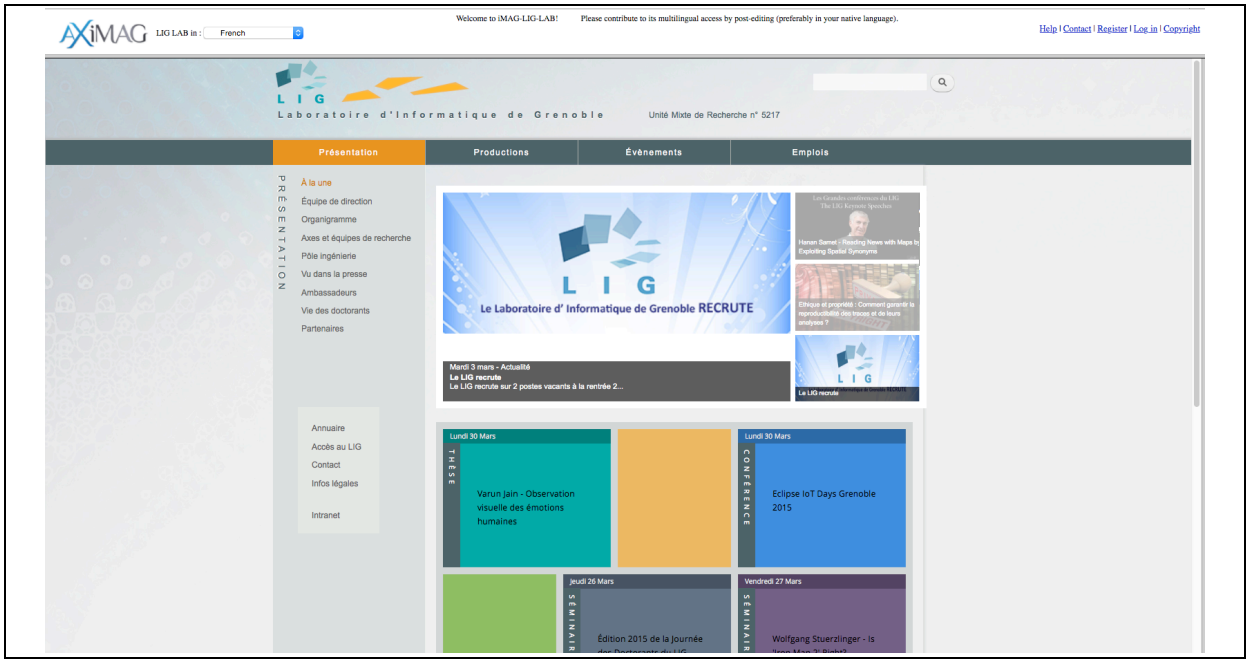


Figure 28 : Site du LIG vu en chinois à travers une iMAG

VI.2 Expérimentation

VI.2.1 Phase 1 (2-6/2013)

VI.2.1.1 Expérience sur le site du LIG

Notre corpus d'apprentissage initial a été construit en choisissant dans SECTRA deux parties tirées de deux MT différentes : une partie provenait de la MT de LIG-LAB[71], qui avait au total 8000 segments en français, dont 2000 étaient déjà post-édités de français en chinois. L'autre partie provenait de la MT DEMO2[72], qui avait été créée en 2012, et qui contenait 8000 segments post-édités de français en chinois. Ces 10K bisegments post-édités ont été utilisés pour créer la version initiale de notre système Moses français-chinois dédié au sous-langage du LIG. Nous l'avons appelée MOSES-LIG-FR-ZH-V_0 (Tableau 46).

Tableau 46 : Statistiques sur les données d'entraînement de la phase 1

	Nb de bisegments	Nb de mots (source)	Pages standard	Caractères chinois	Pages standard
Moses V_0 (initial)	10724	160K	644	182K	455

Dans l'expérience, nous avons travaillé seulement sur la MT du LIG-LAB, en post-éditant des segments choisis parmi les 6000 non post-édités, et en faisant les mesures sur les ensembles croissants de segments post-édités.

Nous utilisons les termes et les notations suivants.

- **Page standard :** texte contenant 250 mots (1400 caractères) en français, ou 400 caractères en chinois.
- **Segment :** unité de post-édition (phrase ou un titre).
- **Page logique :** page Web générée par SECTRA_W pour l'interface de PE et contenant N segments. La valeur par défaut est N=20.

[71] LIG-LAB : un projet d'iMAG. LIG-LAB est l'iMAG dédiée au site du LIG.

[72] Demo2 : c'est une MT dédiée au site du premier auteur.

- **MT_LIG :** MT de LIG-LAB.
- **MT_INC :** nouvelle partie de MT à extraire pour l'apprentissage incrémental.
- **NewPE :** ensemble des nouveaux segments post-édités.
- **Moses-LIG-fr-zh-V_N :** N-ième version de Moses-LIG-fr-zh (après N mises à jour de Moses-LIG).
- **Moses_INC :** opération d'apprentissage incrémental.
- **SEG[i] : s**egment i.
- **SEGPE[i] :** post-édition du segment i.
- **ApplyPE (Moses-LIG_N, SEG[i]) :** post-édition du segment i à partir de sa TA par Moses-LIG_N.

Dans cette expérience, j'ai post-édité les segments non post-édités, par pages logiques. SECTra note le temps de post-édition primaire (Tpe_1)[73] pour chaque segment. Lorsque 10 pages logiques (200 segments) ont été post-éditées, on les utilise pour mettre à jour Moses-LIG-fr-zh (de V_N à V_{N+1}). Nous répétons cela 10 fois de suite (2000 segments), et mesurons le temps moyen de post-édition. Enfin, nous utilisons les segments source du site LIG, leurs post-éditions à partir de la MT de LIG-LAB, et les traductions par Moses-LIG-fr-zh de V_1 à V_{10} pour évaluer la performance de chaque Moses-LIG-fr-zh-V_N par BLEU (Papineni, Roukos et al., 2002), NIST (Doddington, 2002) et par TER (Snover et al., 2009).

Il y a finalement 3 opérations essentielles

Opération 1 : post-édition des segments non post-édités, traduits par Moses-LIG-fr-zh-V_N.

```
SEGPE[i] := ApplyPE (Moses-LIG-fr-zh-V_N, SEG[i])  si SEGPE[i]∉MT_LIG_PE
```

Opération 2 : apprentissage incrémental de Moses-LIG_{N+1}.

```
MT_INCR:= Extraire (NewPE, Niveau, Score)  NewPE ⊆ MT_LIG
Moses-LIG-fr-zh-V_N+1 := Moses_INCR (Moses-LIG-fr-zh-V_N, MT_INCR);
```

Opération 3 : mesures (Voir VI.2.1.3).

VI.2.1.2 Processus d'AI

À l'itération N du processus, tous les segments de la MT sont traduits par la version N du TA (Moses-V_N), y compris, de façon continue, les nouveaux segments créés par le site. Certains segments non encore post-édités sont post-édités à cette itération (de façon opportuniste ou organisée).

Quand on a un certain nombre de nouveaux segments post-édités, jugé "bons" (200 dans notre expérience), on les traite (séparation des mots, nettoyage, traitement de la casse (*truecasing*), alignement, etc.), et on les ajoute à la table de traductions.

Moses-V_N est mis à jour vers Moses-V_{N+1}. À la fin de l'itération N, on met en service Moses-V_{N+1}, et on lui fait traduire tous les segments, post-édités ou non. Cela permet de mesurer la différence de qualité sur la partie déjà post-éditée. On passe alors à l'itération N+1. On obtient ainsi une suite de versions du STA (Moses-V_1, ..., Moses-V_N, Moses-V_{N+1}, ...), qui ne s'arrête (en usage) que quand tous les segments sont post-édités et qu'il n'en reste pas assez de "bons" pour procéder à l'itération suivante.

[73] Temps passé dans la zone de PE.

VI.2.1.3 Résultats

VI.2.1.3.1 Évolution du temps de post-édition

Ces temps sont ceux mesurés par SECTRA_W. Le temps moyen de post-édition (PE), pour chaque nouvel ensemble de 200 segments, diminue à chaque itération. Après la dixième, le temps moyen de PE par page standard (de 250 mots) a été réduit de 3,8 minutes sur 30,7 au départ, soit 12,4%. Si on le compare avec le temps de PE en partant de résultats de GT, on voit qu'il s'en rapproche. La forme de la courbe faisait penser (à la fin de cette phase 1 de l'expérience) qu'il passerait dessous après environ 30 itérations.

Tableau 47 : Évaluation du temps de post-édition (2-6/2013)

Fois	Segments	Mots/seg	page standard	temps de PE (min)	min/page standard
1	200	4,93	3,9	121,0	30,7
2	200	5,15	4,1	125,7	30,5
3	200	4,95	4,0	119,2	30,1
4	200	5,03	4,0	119,4	29,7
5	200	5,29	4,2	123,5	29,2
6	200	5,15	4,1	118,2	28,7
7	200	5,13	4,1	115,2	28,1
8	200	5,25	4,2	116,8	27,8
9	200	5,15	4,1	112,8	27,4
10	200	5,34	4,3	115,0	26,9
Total	2000		41,1	1186,8	

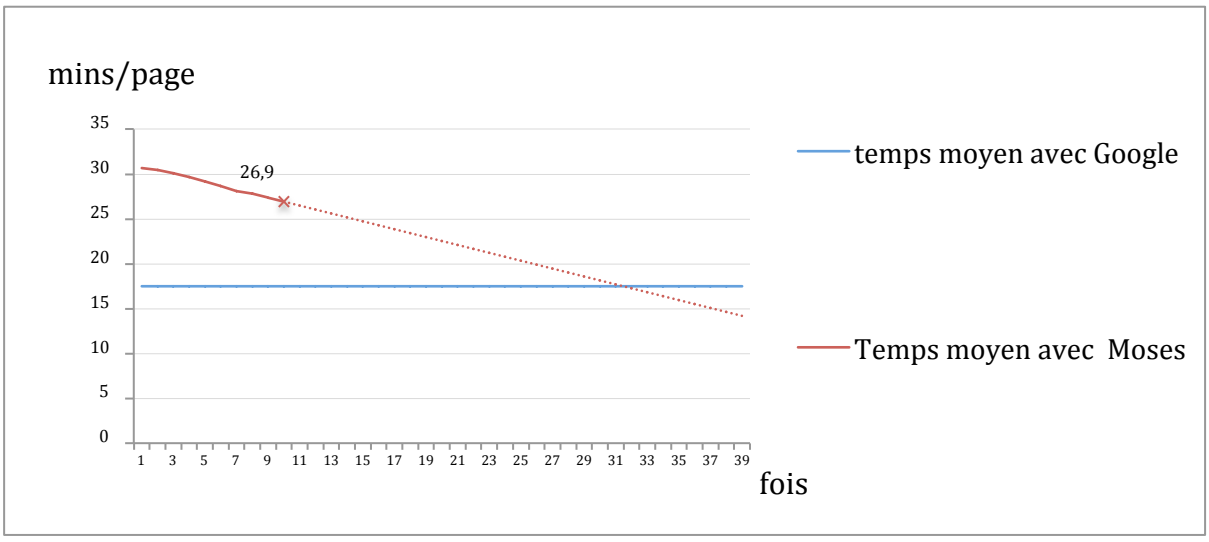


Figure 29 : Diminution de temps moyen de PE (par page standard) avec AI dans la phase 1 de l'expérience

VI.2.1.3.2 Évaluations basées sur des références

On a *M* segments post-édités, et on a la nouvelle version N de Moses-LIG. On traduit tous les segments avec elle, et on mesure BLEU, NIST et TER sur :

- **Origine (OR)** : les *M* segments post-édités.
- **Nouveau (NV)** : les 200 segments nouvellement post-édités.
- **Tout (TT):** tout ce qui a été post-édité.

Tableau 48 : Évaluations basées sur des références (BLEU, NIST, TER)

Version de Moses-LIG	BLEU			NIST			TER		
	OR	NV	TT	OR	NV	TT	OR	NV	TT
0	22,10	19,78	21,58	5,8782	5,5122	5,7655	64,32	65,51	65,32
1	22,12	23,31	21,79	5,8823	5,7356	5,8018	64,28	64,89	64,73
2	22,23	24,64	21,87	5,9125	5,8267	5,8230	64,14	64,62	64,21
3	22,37	20,82	21,53	5,9532	5,4112	5,6547	64,02	66,74	64,88
4	22,46	20,08	21,61	5,9708	5,6349	5,7743	63,98	65,63	64,43
5	22,59	21,33	21,65	5,9821	6,1183	5,8312	63,90	64,37	64,06
6	22,64	20,23	21,67	6,0246	5,4721	5,8414	63,85	65,32	64,52
7	22,71	18,92	21,43	6,0412	5,0537	5,5752	63,65	67,58	65,03
8	22,79	22,41	22,57	6,0856	5,8269	5,8763	63,59	64,73	64,31
9	22,84	20,23	21,41	6,0954	5,6412	5,8838	63,46	66,08	65,27
10	22,93			6,1139			63,34		

À la 10ième itération, il n'y avait pas de segments post-édités (assez bons). Au début de la Nième itération, M=2000+200*N, TT=2000+200*(N+1).

Conclusion

Nous avons fait une première expérience sur l'évaluation et l'amélioration incrémentale de la TA, réalisée avec le système MOSES, basé sur SECTRA_W et MOSES-LIG. Les résultats des mesures montrent que la méthode d'apprentissage incrémental permet de réduire le temps de post-édition de 12,4% en 10 itérations, et d'améliorer les mesures basées sur les références (BLEU : ⇑ 0,83, NIST : ⇑ 0,2357, TER : ⇓ 0,98).

VI.2.2 Phase 2 (7-9/2013)

VI.2.2.1 PE par deux étudiants chinois dans le cadre d'un stage d'été (Wu Jiang et Chen Shi)

Wu Jiang et Chen Shi étaient deux étudiants chinois de l'UJF qui, en juin 2013, étaient en fin de leur 1ère année de master en informatique. Ils ont fait un stage d'été en binôme au laboratoire, durant lequel ils ont fait de la post-édition sur des textes et sur des supports de divers cours de Master 1, du français vers le chinois.

Ils ont travaillé sur 2 IMAG dédiées. (1) « *Corpus par jour* »[74] est un site Web que j'avais créé pour récupérer des ressources français-chinois. J'ai sélectionné des textes en français à partir de sites de journaux LE MONDE et 20 MINUTES. Je les ai mis en format HTML, et les ai ajoutés dans mon site Web. La Figure 30 montre une capture d'écran de post-édition sur l'IMAG associée.

[74] http://service.aximag.fr/xwiki/bin/view/imag/Corpus-lx

Figure 30 : Capture d'écran de l'iMAG « Corpus par jour »

À la fin de cette période (04/07/2013-13/09/2013), nous avions post-édité 21 articles français. La statistique est présentée dans le Tableau 49.

Tableau 49 : Statistiques de post-édition sur 21 articles français 4/7-13/9/2013

Nb de doc	Segments	Mots	Mots/seg	Page_std	Temps en sec.	Temps/page_std en sec.
21	4775	115564	24,20	462,26	159408	344,85

(2) « *MACAU* »[75], Cette iMAG a été créée dans le cadre du projet MACAU-OFI. Elle permet aux étudiants étrangers de l'UJF d'accéder dans leur langue à des supports pédagogiques, par l'amélioration contributive des "pré-traductions" obtenue par TA (Figure 31). (Voir la présentation de MACAU au III.2.3).

[75] http://46.105.41.94/macau/chamilo-macau/index.php

Figure 31 : Capture d'écran de Chamilo affichant le lien AXiMAG

En septembre 2013, nous avons post-édité les supports de cours de l'UJF comme « Complexité en M1 », « IHM en M1», « *Logique en L2* », etc. Le Tableau 50 donne la statistique de post-édition sur les supports de cours.

Tableau 50 : Statistiques de post-édition sur les supports de cours

Nb de doc	Segments	Mots	Mots/seg	page_std	Temps en sec.	Temps/page_std (calculé)
236	16069	136573	8,50	546,29	412784	755,61s =12,6 mn

VI.2.2.2 Processus d'AI

Dans cette phase, nous utilisons la méthode présentée au IV.3.4 pour construire notre système. Tout d'abord, nous utilisons le système de TA français→chinois, qui est entraîné dans la phase 1, comme système de base (« *baseline* »). Les nouveaux segments post-édités sont utilisés pour entraîner la nouvelle partie de TA. Les paramètres sont les même que ceux de la phase I.

Après la construction du système de TA français→chinois avec tous les segments post-édités (Phase I + Phase II)[76], nous avons fait une expérimentation pour comparer le temps de post-édition parmi les différentes des système de TA dans le cadre d'un stage M1 TER.

VI.2.2.3 Résultats

Pour évaluer notre système de TA, nous avons choisi 50 segments à partir du site Web du LIG en langue française, et un étudiant chinois a ensuite post-édité les prétraductions produites par GT et MOSESLIG français-chinois (100 observations) dans SECTRA_W. En fait, dans cette situation, nous ne pouvons pas analyser directement le Tpe_{total}, parce que, pour chacune des prétraductions, le segment source est le même, et le temps de recherche lexicale (l'essentiel de

[76] Dans la phase II, on n'a pas de façon pour accéder le réseau intranet du labo, donc nous faisons un fois d'apprentissage incrémental.

Tpe_2) est souvent passé sur la première prétraduction. Néanmoins, nous pouvons analyser le Tpe_1 qui est fortement corrélé avec le Tpe_{total}.

VI.2.2.3.1 Évolution du temps de post-édition

Dans le Tableau 51, le nombre moyen de mots par segment des prétraductions de MOSESLIG est presque le même qu'avec GT. Nous pouvons constater que, pour post-éditer un segment prétraduit, le Tpe_1 moyen de MOSESLIG est inférieur à celui de GT.

Tableau 51 : Résultat de l'expérimentation (en français-chinois)

TA	Nb moyen de mots par segment (source)	Nb moyen de mots par segment (cible)	Tpe_1 moyen par segment	Tpe_1 moyen par page_std[77]
MosesLIG	26	25,4	23,4 s	6,1 mn
Google	26	27,1	25,3 s	6,3 mn

VI.2.2.3.2 Conclusions provisoires

Dans la phase 2, nous n'avons pas fait l'AI comme dans la phase 1, mais nous avons réentraîné notre système de TA français-chinois avec une plus grosse MT. Notre expérimentation a produit un bon résultat. Du point de vue du temps de post-édition, le résultat est claire : notre système est maintenant un peu meilleur que GT.

VI.2.3 Phase 3 (9-12/14 et 7-11/15)

VI.2.3.1 Motivation

Dans la phase 1, nous avons fait le premier essaie de l'AI, et nous avons obtenu un résultat encourageant. Mais notre système de TA était limité par la quantité de segments post-édités, et la longueur moyenne du segments (7 ou 9 mots).

Dans la phase 2, nous n'avons pas pu faire d'AI. Nous avons réentraîné notre système avec une grande MT français-chinois, mais cela ne prouve pas que notre méthode (AI) peut améliorer le système de TA, mais ne permet pas non plus de trouver le point de « réentrainement du système ». Nous voulions donc refaire notre expérimentation pour répondre nos « doutes ».

VI.2.3.2 Poursuite de l'expérience en français-chinois sur le site du LIG

Dans cette phase, nous avons exporté tous les segments post-édités de SECTRA_W. Dans le Tableau 52, nous montrons la quantité de segments post-édités.

Tableau 52 : Nombre de segments dans chaque MT

Nom du corpus / de la MT	**Quantité de segments PE**
Demo2	28K
MT_INC	7K
MT_Macau	16K
MT_TED	14K
MT_énergie	9K

Préparation des données. Nous avons mélangé toutes les données, environ 80K bisegments, puis nous les avons nettoyées et les avons séparées en 16 parties. Chaque partie contient 5000 segments. Les statistiques sur les données nettoyées sont données dans le Tableau 53.

[77] Une page standard contient 250 mots français ou 400 caractères chinois

Conception de l'expérience. Nous avons fait 16 itérations. Le processus de l'AI a été présenté dans le VI.2.2.2. Le système MOSES-LIG V_0 a été construit avec 100K bisegments extraits à partir du corpus MULTIUN.

Tableau 53 : Statistiques sur les données pour l'AI (phase 3 de l'expérience)

Incrémental	**Longueur moyenne de segment (source)**	**Longueur moyenne de segment (cible)**	**Nombre de mots (source)**	**Nombre de caractères chinois (cible)**	**Pages standard (source)**	**Pages standard (cible)**
Base (MultiUN)	52,7	41,8	263 K	209 K	1 054	522,50
1	42,9	27,3	214 K	136 K	858	341,25
2	43,3	26,7	216 K	133 K	866	333,75
3	39,6	25,3	198 K	126 K	792	316,25
4	34,5	23,3	172 K	116 K	690	291,25
5	31,7	22,3	158 K	111 K	634	278,75
6	30,1	21,9	150 K	109 K	602	273,75
7	28,8	21,5	144 K	107 K	576	268,75
8	27,7	21,1	138 K	105 K	554	263,75
9	27,1	20,9	135 K	104 K	542	261,25
10	26,5	20,7	132 K	103 K	530	258,75
11	25,9	20,5	129 K	102 K	518	256,25
12	25,5	20,4	127 K	102 K	510	255,00
13	25,1	20,2	125 K	101 K	502	252,50
14	24,8	20,1	124 K	100 K	496	251,25
15	23,6	19,2	118 K	96 K	472	240,00
16	22,3	19,1	111 K	95 K	446	238,75
Au total	**Longueur moyenne de segment**	**Longueur moyenne de segment**	**Nombre de mots (source)**	**Nombre de caractère chinois (cible)**	**Pages standard (source)**	**Pages standard (cible)**
	31,3	23,1	2 660 K	1 961 K	10 642,00	4 903,75

VI.2.3.3 Résultats

VI.2.3.3.1 Évolution du temps de post-édition

Pour évaluer le temps de post-édition, nous continuons à post-éditer les nouveaux segments sur l'iMAG LIG-LAB comme nous l'avons présenté au VI.2.1.3.1. Les segments sont sélectionnés dans la MT LIG-LAB. Nous avons filtré sur la longueur, et supprimé les segments très courts.

Tableau 54 : Évaluation du temps de post-édition (9-12/2014)

Fois	Segments	Mots/seg	Page standard	Tpe (min)	Min/page standard
0	200	16,7	13,36	60,4	45,21
1	200	17,5	14	39,9	27,52
2	200	18,6	14,88	40,4	26,83
3	200	21,2	16,96	45,1	26,62
4	200	24,6	19,68	49,0	24,89
5	200	16,8	13,44	32,3	24,03
6	200	18,2	14,56	34,4	23,62
7	200	16,5	13,2	29,2	22,12
8	200	16,1	12,88	27,8	21,56
9	200	14,3	11,44	23,7	20,72
10	200	20,5	16,4	32,9	20,06
11	200	27,1	21,68	41,1	18,98
12	200	13,8	11,04	18,8	17,04
13	200	15,7	12,56	21,0	16,73
14	200	19,1	15,28	24,5	16,06
15	200	13,7	10,96	16,7	15,24
16	200	18,6	14,88	22,5	15,11

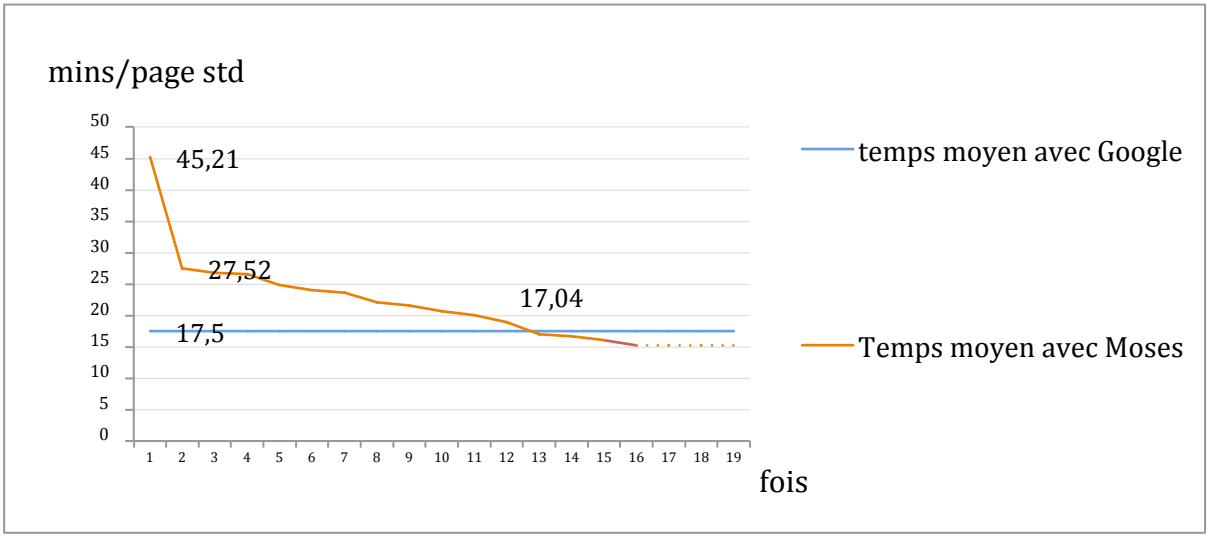


Figure 32 : Diminution du temps moyen de PE (par page) avec AI dans la phase 3 de l'expérience

VI.2.3.3.2 Évaluations basées sur des références

Pour évaluer notre système de TA V_N, nous avons choisi 500 bisegments à partir des données d'entraînement V_{N+1} comme données de test. Par exemple, nous avons fait l'AI avec les données « *Incrémental 10* », et choisi 500 bisegments des données d'entraînement « *Incrémental 11* » pour évaluer notre système. Les statistiques sur les données de test et le score BLEU sont données dans le Tableau 55.

Tableau 55 : Données de test et scores BLEU

	Longueur moyenne de segment (source)	Longueur moyenne de segment (cible)	Longueur moyenne de segment (résultat)	Nombre de mots (source)	Nombre de caractère (cible)	Nombre de caractère (résultat)	Pages standard (source)	Pages standard (ref)	Pages standard (résultat)	BLEU
Eval1	37,2	24,8	23,1	18612	12406	11544	74,4	31,0	28,9	0,1382
Eval2	40,6	25,8	27,6	20309	12912	13817	81,2	32,3	34,5	0,1428
Eval3	20,2	18,5	15,9	10098	9244	7974	40,4	23,1	19,9	0,1603
Eval4	20,7	18,3	16,3	10336	9141	8141	41,3	22,9	20,4	0,2357
Eval5	25,4	22,9	20,3	12687	11462	10165	50,7	28,7	25,4	0,2823
Eval6	19,1	17,3	15,4	9541	8646	7692	38,2	21,6	19,2	0,3346
Eval7	23,1	21,0	18,7	11543	10518	9356	46,2	26,3	23,4	0,3770
Eval8	20,7	19,3	17,2	10364	9651	8582	41,5	24,1	21,5	0,4285
Eval9	19,5	17,5	15,8	9725	8766	7901	38,9	21,9	19,8	0,4331
Eval10	21,8	19,8	17,8	10899	9921	8897	43,6	24,8	22,2	0,4597
Eval11	20,5	18,4	16,9	10248	9194	8459	41,0	23,0	21,1	0,4682
Eval12	23,0	20,5	18,5	11501	10251	9260	46,0	25,6	23,2	0,4694
Eval13	21,9	20,0	17,9	10973	10004	8964	43,9	25,0	22,4	0,4734
Eval14	14,5	13,0	11,3	7242	6481	5655	29,0	16,2	14,1	0,4792
Eval15	16,2	17,4	17,2	8107	8734	8622	32,4	21,8	21,6	0,4831

VI.3 Analyse des résultats

Après 16 itérations, les résultats des mesures montrent que la méthode d'apprentissage incrémental permet de réduire le temps de post-édition et d'améliorer les mesures basées sur les références (BLEU).

La V_0 de notre système avait été entraînée sur le corpus MULTIUN, et ce système trop généraliste ne pouvait pas nous donner une bonne qualité de traduction. Avec 45mn/ page_std, on avait en effet Q=10%=2/20.

Cependant en l'utilisant comme point de d épart, grâce à l'AI, nous sommes arrivé assez rapidement (16 itérations et environ 90h de calcul pour l'AI, sans jamais tout recompiler) à un système de bonne qualité (avec 15mn/page_std de PE en moyenne, notre formule donne Q=70%=14/20).

Les temps de PE primaire (Tpe_1) et les scores BLEU sont donnés dans le Tableau 54 et dans le Tableau 55.

Partie C Contribution d'outils et de ressources

Introduction

Pendant ma thèse, j'ai participé au projet *MUMIA*[78] (*Multilingual and Multifaceted Interactive Information Access*). Tout d'abord, j'ai travaillé sur le corpus de brevets *CLEF-IP 2011*[79]. J'ai construit les mémoires de traductions à partir de ce corpus pour les paires des langues allemand↔anglais, anglais↔français, et français↔allemand. Ensuite, j'ai construit plusieurs systèmes de TA avec ces mémoires de traductions. Enfin, j'ai extrait les segments monolingues dans ce corpus. Avec l'aide de Huanan SUN, étudiant de M1, j'ai utilisé les IMAG dédiées à trois sites Web pour post-éditer ces segments monolingues à l'aide de nos systèmes de TA.

Dans le cadre de ce projet, on a demandé de mettre à disposition des participants à MUMIA, dont une bonne partie travaille sur la RI translingue sur les brevets.

Ce travail m'a montré l'intérêt de telles mises à disposition, non seulement pour leurs utilisateurs potentiels, mais aussi du point de vue de l'ingénierie linguistique. J'ai donc essayé de mettre à disposition le plus possible d'outils et de ressources dérivables de mon travail de thèse. Ce faisant, j'ai pu dégager un certain nombre de problèmes liés à ce type de tâche, et leur trouver des solutions assez génériques. Ce dernière partie est consacrée à cet aspect.

[78] http://www.mumia-network.eu/index.php/the-action/objectives et http://www.cost.eu/COST_Actions/ict/Actions/IC1002

[79] http://ifs.tuwien.ac.at/~clef-ip/

Chapitre VII Construction de systèmes de TA et support à la RI multilingue pour MUMIA

Résumé

Le corpus CLEF-IP 2011 contient 3,5 millions de documents de brevet ; il a été distribué dans le cadre du projet MUMIA. Nous en avons extrait les segments parallèles, et puis nous avons construit les mémoires de traductions correspondantes. Pour augmenter la quantité de segments multilingues, tout d'abord, nous avons filtré les fichiers XML pour supprimer les segments traduits, ensuite nous avons transformé les fichiers XML en format HTML, et les avons importé dans notre plate-forme SECTRA_W/IMAG. Enfin, nous avons fait la post-édition pour l'élargissement à d'autres langues.

VII.1 Contexte et motivations

VII.1.1 Description du projet MUMIA et du WG2

Le projet MUMIA (EU COST ACTION) a pour but de construire et mettre en réseau une communauté de chercheurs travaillant dans le domaine de l'accès à l'information multilingue, multimodale, et multifacette, et plus particulièrement aux documents de brevet. Il encourage la recherche et la communication de la technologie dans les domaines généraux de la traduction automatique, de la recherche d'information, et de l'accès multimodèle, multilingues à l'information interactive, visant à faciliter le développement de systèmes de recherche de prochaine génération.

Pour augmenter le synergie entre les communautés de RI et de TA, une série d'activités ont été organisées (par exemple réunion interdisciplinaire MUMIA GT à Tallinn et à Amsterdam, réunion WG/atelier ouvert à Hissar) pour discuter des ressources, et des outils linguistiques pour la recherche de développement des systèmes, et aussi pour soutenir la production des prochains systèmes de recherche multilingue.

Du point de vue de l'ACTION COST MUMIA, la principale motivation de ces activités était d'accroître la communication entre les scientifiques RI/TAL/TA sur les problèmes et les défis principaux suivant : (1) le développement/l'aide/l'intégration des ressources et des outils linguistiques pour le développement des systèmes de RI, et (2) la clarification des ressources et des solutions technologiques pour la mise en œuvre de RI multilingue dans les systèmes de recherche professionnels modernes.

Par exemple, un progrès technologique est le travail sur la production de corpus parallèles de haute qualité et sur l'évaluation permanente basée sur les tâches de plusieurs systèmes de TA pour les brevets. Une vitrine des corpus parallèles développés en utilisant la MT de CLEF-IP est disponible sur la page web de MUMIA. L'autre exemple est l'intégration de divers services de TA pour soutenir la recherche multilingue dans un système fédératif de RI sur les brevets (PERFEDPAT).

L'Action MUMIA COST contient cinq groupes de travail (GT) (*Integrating and Managing Language Resources, Processing Infrastructures for IR and MT, User Centered Aspects of MUMIA, Semantic Search and Faceted Search, Visualization, Distributed and Social Search*). J'ai participé au WG2. L'objectif de ce groupe de travail est l'étude de nouvelles infrastructures pour la recherche plus efficace dans les grands environnements numériques.

VII.1.2 PerFedPat et Khresmoi

Le système PERFEDPAT (Salampasis and Hanbury, 2014), basé sur le cadre EZDL, fournit des services basiques de RI en utilisant une méthode qui fédère l'accès à de multiples systèmes (en ligne) de RI concernant les brevets (actuellement ESPACENET [80], GOOGLE PATENTS, PATENTSCOPE[81] et MAREC[82]), et qui repose sur un accès parallèle à de multiples sources de brevet.

PERFEDPAT cache la complexité à l'utilisateur qui utilise un outil unique pour interroger tous les ensembles de données de brevets. PERFEDPAT fournit des services comme la recherche par champ, la fusion, le regroupement et le filtrage des résultats. Il offre un soutien pour l'historique de la requête en cours et des sessions de recherche. La deuxième caractéristique innovante de PERFEDPAT est qu'il a une architecture enfichable (par plugins) et extensible.

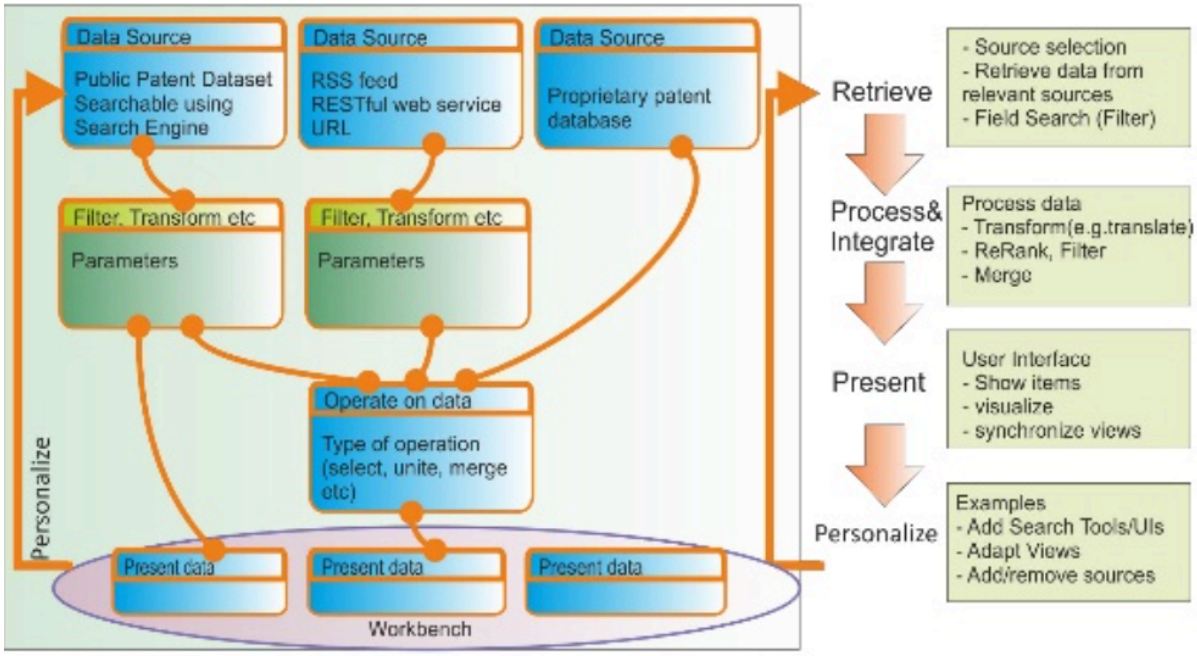


Figure 33 : Architecture de PerFedPat[83]

Actuellement les outils intégrés concernent la recherche CIB (Classification internationale des brevets), la navigation à facettes, le regroupement des résultats de la recherche et la TA des requêtes. Le système est disponible en téléchargement à partir d'Internet[84].

KHRESMOI[85] (Hanbury et al., 2011) est un système de recherche multilingue/multimodèle, et d'accès d'information pour l'information biomédicale, construit dans le cadre du projet éponyme KHRESMOI. KHRESMOI utilise des données disponible en ligne, comme les sites de santé certifiés par *Health on the Net*[86] et des revues en accès ouvert.

[80] http://worldwide.espacenet.com
[81] http://www.wipo.int/patentscope/fr/
[82] http://www.ifs.tuwien.ac.at/imp/marec.shtml
[83] L'image est à partir du transparent de « PerFedPat patent search system »
[84] http://www.perfedpat.eu/index.php/download-perfedpat
[85] http://www.khresmoi.eu
[86] https://www.healthonnet.org

Figure 34 : KHRESMOI

VII.1.3 Objectif poursuivi

Nous travaillons pour supporter la recherche en RI multilingue sur les brevet. Les systèmes de recherche existant, comme PERFEDPAT et KHRESMOI, utilisent des systèmes de TA pour réaliser la fonction de recherche en contexte multilingue. La qualité du système de TA détermine la qualité du résultat de la recherche. Par exemple, PERFEDPAT utilise GT pour traduire les requêtes, mais le résultat n'est pas satisfaisant, et il est impossible d'améliorer la qualité de traduction.

Un document de brevet contient des « segments multilingues » au sens défini plus haut (p.21) : un élément XML correspondant à un segment contient ses versions en plusieurs langues). Si nous pouvions construire des systèmes de TA (STA) avec ces segments parallèles, nous pourrions améliorer la qualité de la RI translingue. C'est ce que nous avons fait.

VII.2 Construction de MT et de STA à partir des corpus CLEF-IP 2011

VII.2.1 Description du corpus CLEF-IP

La collection CLEF-IP contient des brevets, physiquement stockés comme une collection de fichiers XML encodant des documents de brevets. Un document de brevet peut être un document de demande de brevet, un rapport de recherche, ou un document de brevet accordé. À chaque document de brevet est attribué un code de type, qui apparaît comme un suffixe attaché à l'identifiant du brevet (par exemple, EP-nnnnnnn-A1, WO-nnnnnnnnnn-A2). Dans le cas de l'OEB (EPO), les documents de demande de brevet qui incluent un rapport de recherche portent le code A1, les rapports de recherche de demandes de brevet portent le code A3, les documents de brevets accordés portent le code B1, etc.[87].

La collection de données CLEF-IP 2011 est basée sur les données de 2010, et est extraite à partir du corpus de données MAREC. La collection CLEF-IP contient principalement des documents brevets publiés par l'OEB (EPO).

Deux ajouts importants ont été apportés au corpus de la collection 2010. Le premier a été d'inclure dans le corpus distribué certains brevets publiés par l'Organisation mondiale de brevets intellectuels OMPI (WIPO). Un pourcentage élevé des brevets de l'OEB (EPO) figurant dans le corpus CLEF-IP sont des demandes de brevet déposées à l'échelle internationale dans le cadre du Traité de coopération en matière de brevets (PCT), auquel cas l'OEB ne republie pas la demande de brevet en totalité, mais seulement une entrée bibliographique faisant référence à la demande originale. Pour ces brevets, CLEF-IP 2011 a ajouté leur équivalent de l'OMPI à la

[87] https://register.epo.org/help?topic=kindcodes

collection, afin de fournir une collection qui soit à la fois plus grande et plus réaliste. Le deuxième ajout à la collection CLEF-IP concerne l'une des nouvelles tâches basées sur les images contenues dans les brevets, à savoir la tâche de recherche basée sur les images des brevets. Pour cette tâche, nous avons ajouté à la collection cible CLEF-IP les images des brevets pour trois classes d'ICP: A43B, A61B, et H01L.

Dans nombre de documents, ajouter les documents de brevets de l'OMPI (WIPO) au corpus de la collection l'a augmentée de 1,2 million de documents de brevets, arrivant à un nombre final d'environ 3,5 millions de documents XML, correspondant à environ 1,5 million de brevets. Les images correspondant aux 47 000 documents XML dans les trois classes de la CIB (ICP) dans la tâche IMG PAC occupent 5,4 Go, pour 290 880 fichiers TIFF.

Comme dans les années précédentes, le corpus de la collection de test a été livré aux participants sans fusionner les documents relatifs à un même brevet en un seul document. Chaque brevet est identifié par un numéro de brevet unique, une chaîne identifiant le bureau éditeur («EP» pour l'OEB et «WO» pour l'OMPI), suivie par une série de chiffres. Pour chaque brevet, il y a un répertoire contenant les documents XML représentant les documents de brevet liés à ce brevet. Pour les brevets de l'EP, le format du nom est 00000n/nn/nn/nn/*.xml, où la séquence de chiffres dans le nom du répertoire correspond à celle du nom du brevet. Pour les brevets WO, le format du nom est 00000n/nn/nn/nn/*.xml, où les 4 premiers chiffres (après « 00 ») représentent l'année de publication du document.

Par exemple, le brevet EP 09 81 2 01 correspond aux répertoire contenant les fichiers EP-0981201-A2.xml, EP-0981201-A3.xml, et EP-0981201-B1.xml:

EP/000000/98/12/01

Au brevet WO 1994030029 correspond le répertoire contenant le fichier WO-1994030029-A1.xml:

WO/001994/03/00/29 → WO-1994030029-A1.xml

Les fichiers d'images des brevets sont stockés comme des fichiers TIF dans un dossier séparé, la correspondance entre le fichier d'images et son fichier XML est établie par un petit ensemble de règles.

Tous les documents textuels de la collection CLEF-IP contiennent les principaux champs XML suivants: données bibliographiques, des résumés, résumé et revendications. Tous les documents n'ont pas nécessairement de contenu dans ces champs. Le contenu des différents champs XML peut être de l'anglais, de l'allemand ou du français. Certains champs peuvent apparaître plus d'une fois, chaque fois avec une langue différente. Un fichier de brevet XML a également une langue du document (anglais, allemand ou français), ce qui n'exclut pas que ses sous-champs apparaissent avec un attribut de langue différent de la langue du document. Par exemple, les documents de brevet EP (EP-nnnnnnn-Bn.xml) accordé doivent contenir les revendications en trois langues (anglais, allemand et français).

La dernière version de la collection est la même que celle utilisée dans le CLEF-IP 2011 LAB (le corpus de données utilisé en 2012 et 2013 est le même que celui utilisé en 2011), de sorte que notre travail est basé sur la collection CLEF-IP 2011.

Figure 35 : Exemple de fichier XML Dans CLEF-IP

VII.2.2 Extraction de MT à partir de CLEF-IP 2011

VII.2.2.1 Analyse du corpus

Nous avons analysé les brevets en ce qui concerne la structure de leurs champs XML. Nous avons constaté que quatre grands champs peuvent avoir des segments parallèles: <*invention-title*>, <*abstract*>, <*description*>, et <*claims*>. Chaque champ peut avoir des sous-champs, par exemple, un champ <*claims*> peut contenir 6 sous-champs <*claim*> dans *EP-0260000-B1.xml* (Figure 36).

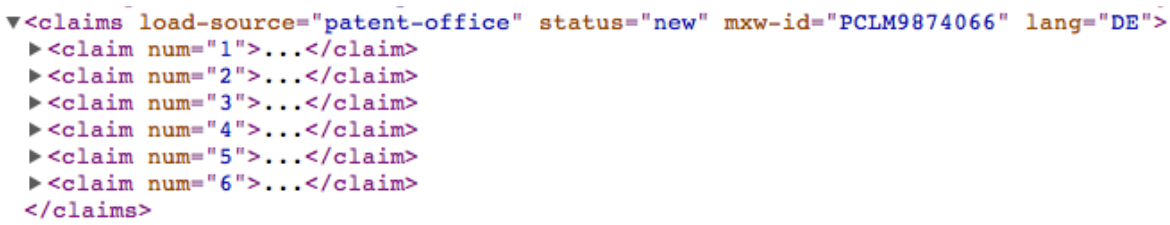


Figure 36 : Exemple de champ <*claims*> contenant 6 sous-champs <*claim*> dans *EP-0260000-B1.xml*

Nous commençons par ces champs, en cherchant des champs qui apparaissent plus d'une fois dans le document de brevets et tels que chaque champ ait un attribut de langue différente. Par exemple, Figure 37 montre un champ <*invention-title*> avec 3 attributs de langue différents (lang = "DE", lang = "EN", et lang = "FR"). Chaque champ contient également du contenu, dans la langue qui correspond à son attribut.

```
▼<invention-title lang="DE" load-source="ep" status="new">
   Verfahren zur Herstellung von Polyimidestern der Trimellitsäure
 </invention-title>
▼<invention-title lang="EN" load-source="ep" status="new">
   PROCESS FOR THE FABRICATION OF POLYIMIDE-ESTERS OF TRIMELLITIC ACID
 </invention-title>
▼<invention-title lang="FR" load-source="ep" status="new">
   Procédé pour la fabrication de polyimide-esters de l'acide trimellitique
 </invention-title>
```

Figure 37 : Exemple d'un champ *<invention-title>* avec 3 attributs de langue différents et les contenus correspondants en 3 langues différentes

Il est bien connu que les corpus parallèles ont aussi un problème avec la direction de la traduction, car la relation de traduction est symétrique au niveau des termes, mais pas des phrases. Si une MT contient des directions de traduction différentes, cela affecte directement notre travail. Lorsque nous traitons la collection CLEF-IP 2011, nous devons donc également tenir compte de la langue source de chaque document de brevet. Chaque document XML de CLEF-IP 2011 comporte une indication de sa langue de départ. Par exemple, la Figure 38 montre un document de brevet en langue anglaise. Au cours de notre processus d'extraction, nous considérons la langue du document comme état la langue source de tous ses segments.

```
<patent-document ucid="EP-0071719-B1" country="EP" doc-
number="0071719" kind="B1" lang="EN" date="19881102" family-
id="23112365" date-produced="20100220" status="new">
```

Figure 38 : Un champ *<patent-document>* avec attribut lang = "EN"

VII.2.2.2 Traitement du corpus

(Utiyama and Isahara, 2007) ont utilisé les parties du champ de description "description détaillée des modes de réalisation préférés" (Detailed Description of the Preferred Embodiments) et "Contexte de l'invention" de chaque brevet pour trouver des segments parallèles (japonais-anglais), car ils ont constaté que ces deux parties ont plus de traductions que d'autres. Parce qu'ils avaient moins de paires de brevets, (Lu et al., 2009) ont utilisé toutes les parties des documents de brevet pour trouver des segments parallèles (chinois-anglais). Dans notre travail, nous avons extrait toutes les parties des documents de brevet, mais dans le but d'assurer la qualité du corpus parallèle, nous avons rendu le champ <*invention-title*> et les parties <*claims*> disponibles dans la première version du corpus parallèle CLEF; les autres parties du corpus parallèle seront disponibles dans la prochaine version.

Notre travail est basé sur 3,5 millions de documents de brevet (fichiers XML), et nous voulons en extraire autant de segments parallèles utiles que possible. Tout d'abord, nous parcourons chaque document de brevet. Pour chaque document de brevet, nous sélectionnons la langue source à partir du champ <*patent-document*>, selon l'attribut de langue de ce champ. Deuxièmement, nous recherchons les segments parallèles contenus dans les quatre champs principaux (<*invention-title*>, <*abstract*>, <*description*>, et <*claims*>). Parfois, certains champs ont un attribut de langue différent de la langue du document. Par exemple, dans *EP-0260000-B1.xml,* l'anglais est la langue du document, mais <*claims*> segments ne existent pas en anglais, seules les versions allemandes et françaises sont disponibles. Même s'il est toujours souhaitable de collecter autant de texte que possible, il est encore plus important de veiller à la qualité des textes, de sorte que, dans ce cas, nous ne stockons pas les parties en allemand et en français comme un segment parallèle, parce que nous ne savons pas laquelle est la source.

Tous les champs qui apparaissent plus d'une fois dans un document de brevet et qui ont différents attributs de langue sont traités comme une collection. En général, un document de brevet OEB (IPO) a un maximum de 3 langues (anglais, français et allemand). Nous avons choisi comme segment source un segment dont l'attribut langue est compatible avec la langue source, puis avons ensuite extrait le segment parallèle cible à partir des autres champs. Par exemple, dans EP-0301015-B1.xml, la langue source est l'anglais, et <*revendications*> champ apparaît 3 fois. Donc, nous utilisons la partie anglaise des champs de revendications comme segments source, et nous considérons les parties en français et en allemand comme les segments cibles. Le segment source et les segments cibles sont ensuite stockés séparément dans des fichiers différents. Dans l'exemple ci-dessus, le segment source a été stocké dans *CLEF_claims_en-fr.en* et *CLEF_ claims_en-de.en,* et les segments cibles dans *CLEF_ claims_en-fr.fr* et *CLEF_claims_en-de.de,* respectivement. Afin de réduire le bruit dans les données, nous ne gardons que le texte extrait, et enlevons toutes les balises.

Toutes les données extraites ne sont pas entièrement adaptées à une utilisation directe pour les applications de TAL (NLP). Nous devons nettoyer les données extraites et éliminer un peu de bruit. Pour l'alignement, nous avons utilisé LF ALIGNER[88], un outil open-source basé sur HUNALIGN (Varga et al., 2007) développé par András Farkas, qui, surtout, a la couverture linguistique la plus large (un total de 32 langues), et permet la génération automatique des dictionnaires dans une combinaison quelconque de ces langues. Les segments alignés sont préparés de façon bilingue pour 4 types (titre, résumé, description et revendications), et toutes les 6 paires de langues (de_en, de_fr, en_de, en_fr, fr_de, fr_en).

Le Tableau 56 montre le nombre de segments et de mots qui sont extraits à partir des champs de titre et de revendications en source et en cible après l'alignement des segments. Toutes les phrases parallèles extraites sont enregistrées dans les formats TMX et TXT, et peuvent être trouvées à http://membres-liglab.imag.fr/wang/downloads.

Tableau 56 : Nombre de segments extraits comme source et cible après l'alignement de segments dans les champs <*title*> et <*claims*>

Pairs de langues		**Titre**		**Revendications**	
		Segments	**Mots**	**Segments**	**Mots**
de-en	De	311,298	2,038,785	1,696,498	62 M
	En		2,582,703		71 M
de-fr	De	311,184	2,036,112	1,661,419	79 M
	Fr		2,482,257		86 M
en-de	En	884,759	6,661,481	5,218,024	332 M
	De		5,508,289		296 M
en-fr	En	884,727	6,661,322	5,373,452	330 M
	Fr		8,538,012		380 M
fr-de	fr	106,211	963,508	572,356	36 M
	de		1,204,439		37 M
fr-en	fr	106,246	1,285,467	586,498	38 M
	en		1,048,374		37 M

[88] http://sourceforge.net/projects/aligner/

VII.2.3 Construction des systèmes de TA

Nous avons utilisé notre corpus parallèle extrait (le titre et les champs de revendication) pour construire des systèmes de TA avec le Toolkit MOSES. Tout d'abord, pour la préparation de l'ensemble du développement et de l'ensemble de test, nous avons extrait 2000 phrases pour la le réglage des poids de fonctionnalités de Moses, et extrait 1000 phrases pour les tests. Ensuite, nous avons utilisé le reste pour former les modèles de traduction de Moses. Nous avons en fait construit des systèmes de TA pour seulement 3 directions: de-fr, de-fr, et en-fr.

Les systèmes comprennent également des modèles de langage 5-grammes formés sur le côté cible correspondant des textes parallèles à l'aide de IRSTLM. Les poids de fonctionnalités requises par le décodeur de MOSES ont encore été déterminés avec MERT en optimisant les scores BLEU sur l'ensemble de développement (1000 phrases). Les ensembles de test ont été traduits par les systèmes qui en résultent et ensuite utilisés pour évaluer les systèmes en termes de scores BLEU, comme indiqué dans Tableau 57.

Tableau 57 : Scores BLEU des systèmes de TA tirés de CLEF-IP

Paires de langues	Jeu de développement	Jeu de test
de-en	37.46	31.67
de-fr	35.41	28.72
en-de	43.16	36.01
en-fr	42.59	38.82
fr-en	44.12	42.61
fr-de	34.85	30.14

VII.3 Expérimentation et élargissement à d'autres langues

VII.3.1 Reconstruction de trois sites Web de brevets monolingues

Le corpus de brevets CLEF-IP contient certain nombre segments monolingues, par exemple, dans la Figure 39, ce fichier XML est en anglais (dans la balise `<document>`, l'attribut de langue est `lang="EN"`), et son titre anglais est traduit en français et en allemand, mais le champ `<claims>` n'est pas traduit. CLEF-IP contient environs 500K fichiers monolingues.

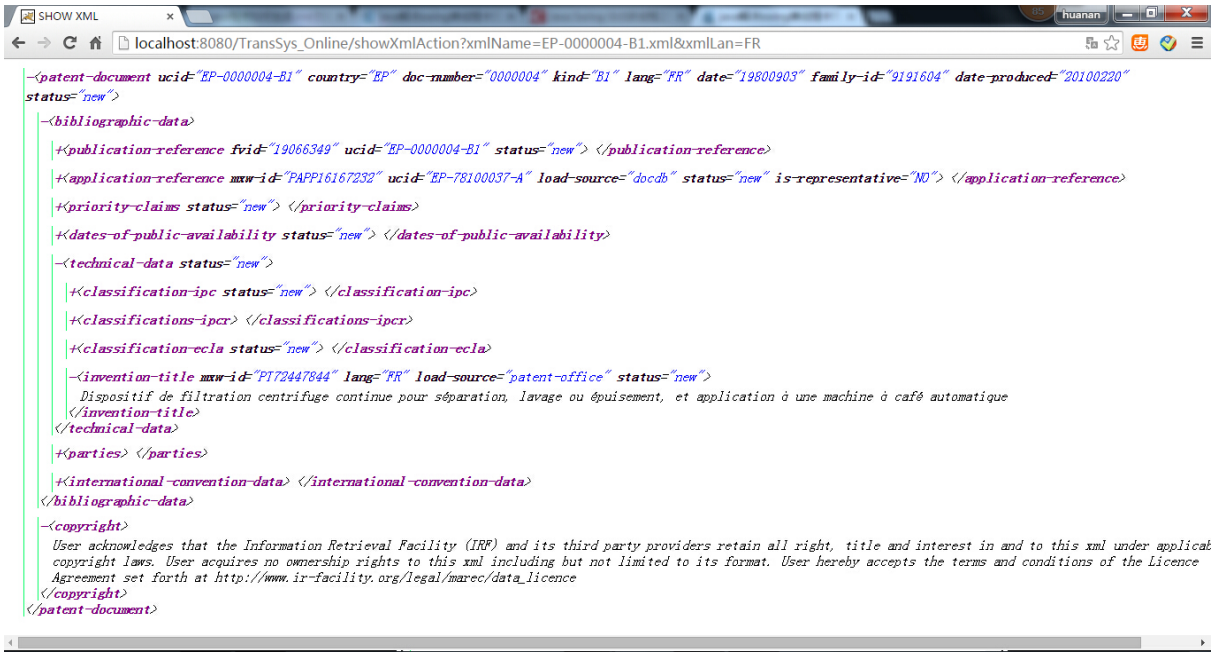


Figure 39 : Exemple de fichier XML monolingue

L'expression des phrases du document de brevet est très similaire dans un document, et beaucoup de mots-clés sont présentés plusieurs fois. Par exemple, nous voyons les revendications du fichier XML *EP0203923B1.xml*. Chaque revendication dans le champ `<claims>` change quelques mots entre eux, mais la structure de phrase est presque la même.

CLAVIER À TOUCHES CAPACITIVES SELON LA R1 CARACTÉRISÉ EN CE QUE LE CHOIX DE FRÉQUENCES ÉLEVÉES (SUPÉRIEURES À 500 KHZ) POUR LES OSCILLATEURS DE TOUCHE PERMET UN FONCTIONNEMENT FIABLE JUSQU'À UNE ÉPAISSEUR DE LA PAROI DIÉLECTRIQUE (2) ATTEIGNANT 30 MM DE VERRE COURANT.
CLAVIER À TOUCHES CAPACITIVES SELON LES R1, R2, R3 CARACTÉRISÉ EN CE QU'UN RÉGLAGE IMMÉDIAT ET SIMPLE PERMET D'ADAPTER LE FONCTIONNEMENT DU CLAVIER À L'ÉPAISSEUR DE LA PAROI DIÉLECTRIQUE (2).
CLAVIER À TOUCHES CAPACITIVES SELON LES R1 ET R2 CARACTÉRISÉ EN CE QUE LA CIRCUITERIE ASSOCIÉE AUX TOUCHES COMPREND UN COMPTEUR D'ADRESSE, UN DÉCODEUR ET LES OSCILLATEURS, ET PEUT ÊTRE RELIEE AU MICROPROCESSEUR PAR UN CÂBLE LIMITÉ À 5 FILS ET POUVANT ATTEINDRE 10 MÊTRES DE LONGUEUR.

Figure 40 : Exemple de revendication dans le fichier EP0203923B1.xml

Un système de TA entraîné avec notre MT pourra fournir des traductions de haute qualité, ce qui facilitera beaucoup la post-édition des segments ne contenant pas leur version dans la langue cible. Nous avons post-édité les segments monolingues et bilingues (nous avions 3 langues au total) pour agrandir la MT.

D'abord, nous filtrons les fichiers XML du corpus CLEF-IP, et utilisons *JSoup*[89] pour analyser la langue source. Nous trouvons la langue dans le champ `<document>`. Ensuite, nous cherchons les balises qui contiennent l'attribut `lang`, et où la valeur de l'attribut `lang` est différente de la langue source du document. Enfin, nous supprimons les balises et leurs contenus. Par exemple, dans le fichier EP-0000004-B1.xml, l'attribut `lang` dans le champ `<document>` est l'anglais. Les trois champs `<title>` sont en anglais, en français, et en allemand, et le champ `<claims>` (revendications) est en anglais. Donc nous supprimons le champ `<title>` en français et en allemand, et conservons le champ `<claims>`.

[89] http://jsoup.org

Après le prétraitement, il faut adapter nos fichiers XML monolingues à SECTRA_W/IMAG. Le fichier XML doit avoir l'aide de l'outil pour la visualisation de contexte. Nous le transformons au format HTML. Pour conserver la structure XML du document de brevet, nous reconstruirons les balises du fichier XML, et nous utilisons les balises HTML/CSS pour décorer les balises XML (avec les couleurs et le style). Dans le Figure 41, on donne un exemple de fichier XML monolingue présenté en format HTML.

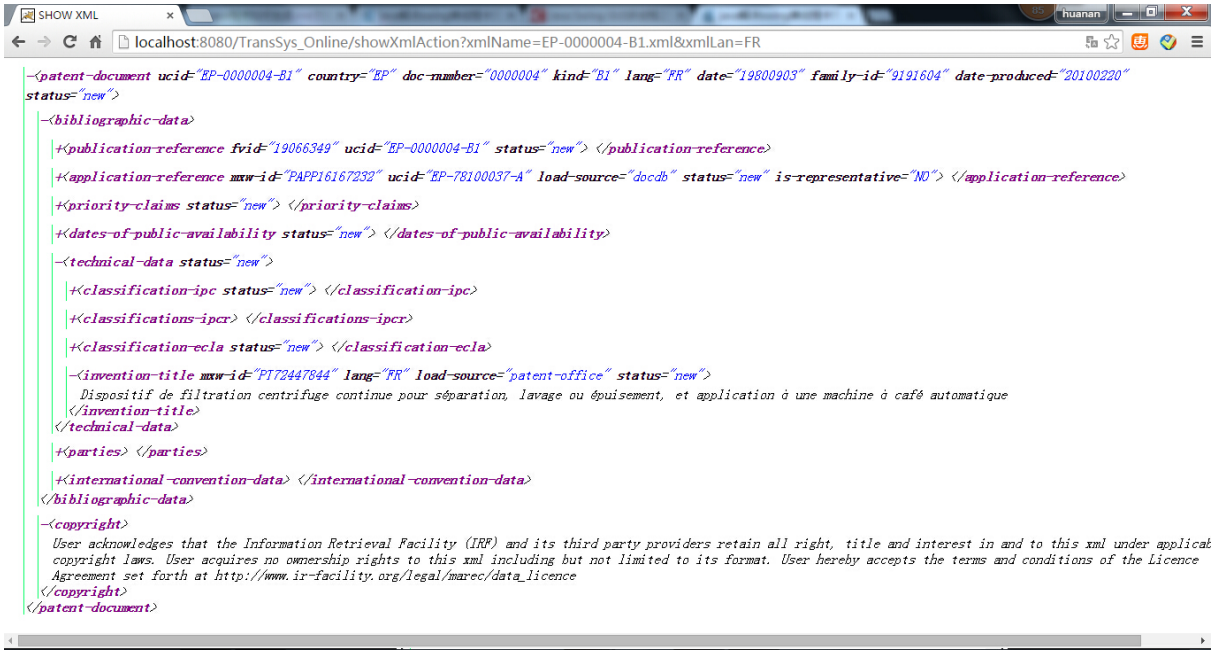


Figure 41 : Exemple de fichier HTML décoré

VII.3.2 Accès multilingue en utilisant les systèmes de TA créés

Nous avons construit 3 sites Web pour les fichiers monolingues, et chaque site Web contient tous les fichiers monolingues traités. Nous avons aussi créé 3 IMAG dédiées à ces 3 sites Web monolingues. Le site Web anglais est présenté dans AXIMAG. Ces 3 IMAG dédiées partagent la MT CLEF-IP. Nous importons les segments parallèles dans la MT CLEF-IP. Quand une iMAG dédiée à un site Web monolingue de CLEF-IP (une page Web monolingue) est accédée par l'utilisateur, l'IMAG d'abord cherche d'abord sa traduction dans la MT CLEF-IP. Si elle existe, il l'affiche dans la page Web traduite, sinon elle appelle nos systèmes de TA via Tradoh.

Nous avons entraîné 3 systèmes de TA (anglais→français, français→anglais, et allemand→français) en utilisant les MT extraites du corpus CLEF-IP. Pour fournir le service de traduction, nous avons intégré le plugin MOSES dans TRADOH, puis traduit les segments monolingues.

VII.3.3 Accès multilingue utilisant d'autres systèmes et pour d'autres langues

Sur SECTRA, nous pouvons ajoutons des différents systèmes de TA pour traduire les segments source. Il nous aide à évaluer la qualité des différents systèmes de TA.

VII.3.3.1 Comparaison sur les mêmes langues

Pour comparer la qualité du système de TA sur les même langues, nous utilisons GT à retraduire les 3 sites Web monolingues Les segments traduits sont sauvegardés dans la MT CLEF-IP, et ils sont présentés dans SECTRA.

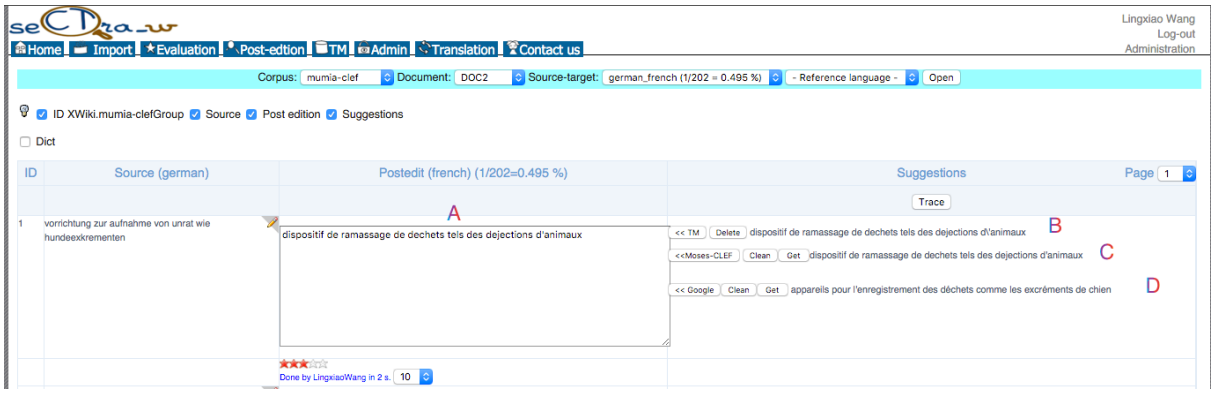


Figure 44 : PE en mode avancé, avec pseudo-trace montrant les différences entre les sorties de TA, la post-édition (utilisée comme référence), et la MT.

Dans le carré rouge de la Figure 44, il y a 4 segments. Le premier segment (A) est la post-édition de l'utilisateur, le deuxième segment (B) est la MT de CLEF-IP, le troisième (C) est la traduction du système de TA entraîné avec la MT CLEF-IP, et le dernier (D) est la traduction de GT. Quand on clique sur le bouton « *Trace* », la distance entre les segments et la référence peut être affiché dans la colonne « *Suggestion* ».

On remarque que le résultat de traduction de GT ne correspondant pas au segment source (). Il est très différent de la référence. Si l'utilisateur fait la post-édition sur ce résultat, il aura besoin de plus de temps.

VII.3.3.2 Utilisation pour d'autres langues

Nous ajoutons les nouvelles langues dans la MT de CLEF-IP. Nous utilisons GT et les systèmes de TA pour traduire les segments monolingues vers d'autres langues.

Les pages Web sont segmentées, puis leurs segments sont sauvegardés dans la MT. Pour ajouter un nouveau système de TA ou ajouter une nouvelle langue, nous retraduisons tous les segments sous SECTRA. Par exemple, pour ajouter la traduction chinoise, on coche notre système de TA français→chinois MOSESLIG dans la colonne « *MTs* ». La Figure 45 montre comment nous retraduisons les segments du français vers le chinois dans le pseudo-document DOC6.

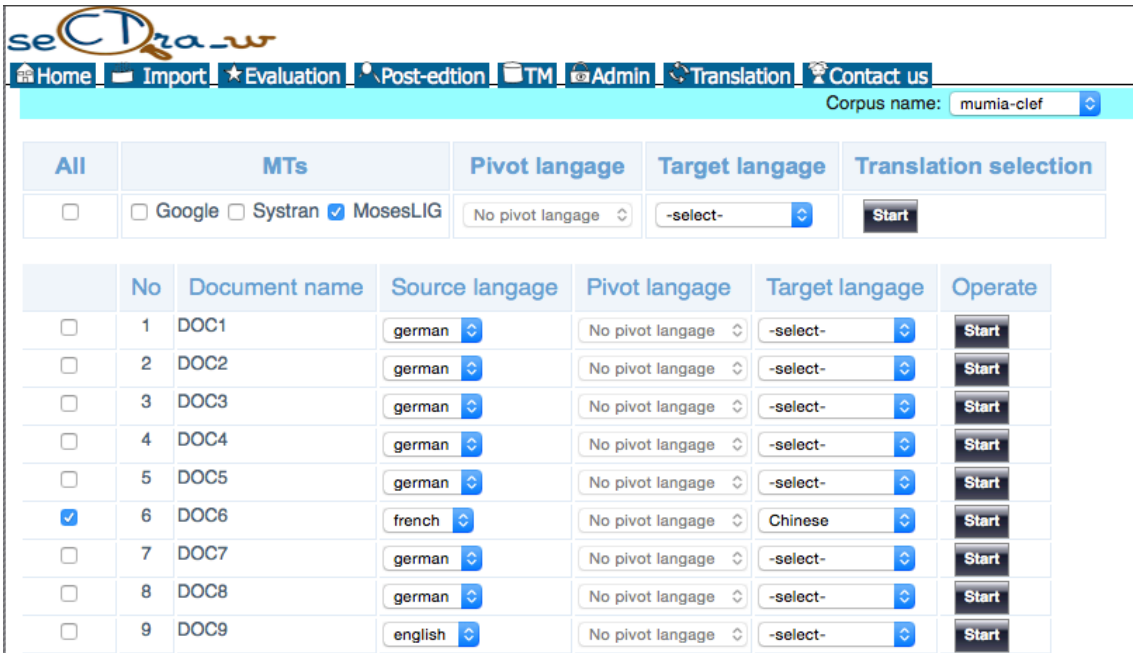


Figure 45 : Retraduction des segments du français vers le chinois pour DOC6 avec le système de TA français→chinois MosesLIG

Chapitre VIII Mise à disposition de ressources

Résumé

Dans le cadre de ma thèse, je me suis aussi attaché à mettre à disposition de la communauté des chercheurs les ressources et outils que j'ai développés, sous des formes « statiques » ou « dynamiques ».

Les ressources et outils « statiques » sont mis à disposition sous forme de fichiers téléchargeables.

La mise à disposition dynamique consiste à proposer des serveurs Web d'utilisation de systèmes de TA et de post-édition collaborative de MT associées à des sites Web, via les iMAG associées.

Introduction

Nos ressources « statiques » sont particulièrement intéressantes à cause de leur taille, et parce qu'elles sont utilisables pour construire on expérimenter des système de TA.

Nous avons mis à disposition des corpus multilingues post-édités dans des formats usuels comme TXT et XML. Ces formats sont bien adaptés aux systèmes existants : les segments source et les segments post-édités sont alignés, et sauvegardés dans deux fichiers TXT. Ils sont utilisables directement par les boîtes à outils comme MOSES et JOSHUA pour entraîner des systèmes de TA. Le format XML peut être importé dans plusieurs plates-formes comme DÉJÀ VU, MEMOQ, SDL TRADOS, etc.

Une grande partie de ces ressources statiques a été construite en utilisant la plate-forme SECTra_w/iMAG. Les segments post-édités sont munis de l'information associée, comme le temps de PE, le système de TA initial, la référence, etc. L'utilisateur en a besoin pour certains travaux, l'évaluation de systèmes de TA, par exemple.

Les MT tirées de la collection de brevets CLEF-IP sont les plus grosses (17,5M segments en fr-de-en). Suivent les MT fr-zh : 9000 pour le sous-langage d'EDF, 15000 pour celui du LIG, 23000 pour les documents pédagogiques de MACAU-UJF, et 10000 pour le roman de Jules Verne « Voyage au centre de la Terre ».

Les systèmes de TA téléchargeables sont ceux déduits de CLEF-IP (fr-de, de-fr, fr-en) et quelques variantes de notre système fr-zh MOSESLIG-FR-ZH.

Les services Web en ligne sont ceux des quelques iMAG parmi lesquelles deux servent actuellement à construire de bonnes MT du français vers le somali et vers le comorien.

VIII.1 Contribution de ressources statiques sous forme de MT

VIII.1.1 Formats choisis

Nous avons choisi 2 types de format pour sauvegarder et partager nos mémoires de traduction, nous avons choisi 2 formats documentaires, TXT et XML.

TXT est le format plus utilisable pour l'apprentissage du système de TA. La boite à l'outil, comme MOSES et JOSHUA, ils supportent des données en format txt (*raw text*) à entrainer les modèle. Figure 46 présente un exemple de la MT aligné CLEF-IP anglais-français. Voir l'exemple avec 100 bisegments anglais-français dans l'Annexe 11.

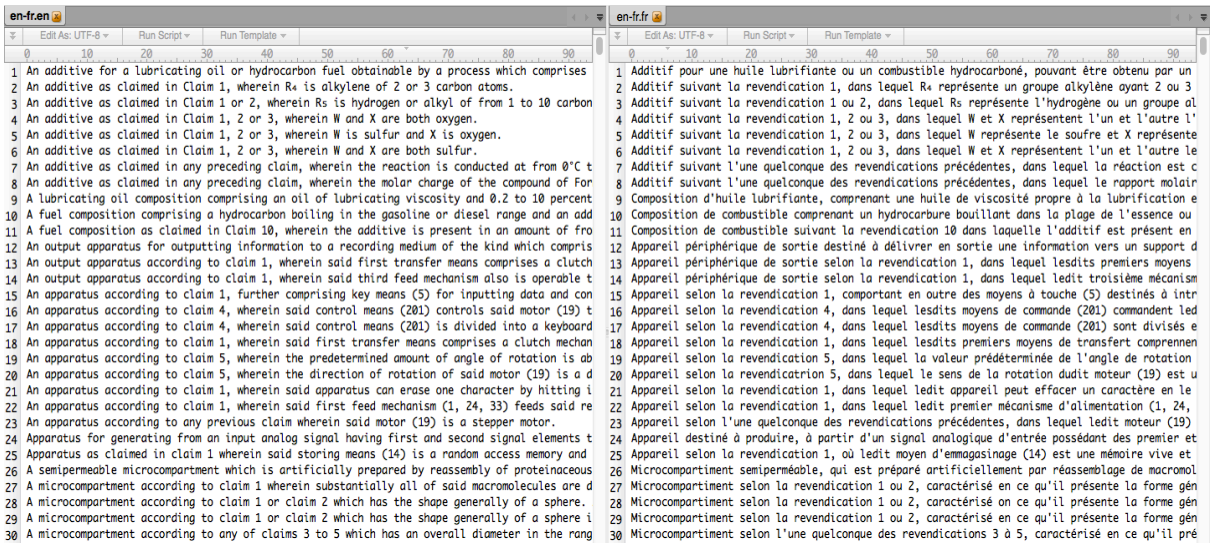


Figure 46 : Exemple des données en format TXT (MT CLEF-IP anglais-français)

Pour adapter aux système de l'aide de traduction comme SDL TRADOS et MEMOQ, nous avons choisi le format XML (fichier TMX ou fichier XML avec fichier accompagne) pour partager ces données (Figure 47). Voir l'exemple avec 20 bisegments anglais-français en format TMX dans l'Annexe 13.

```
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
  <header creationtool="LF Aligner" creationtoolversion="3.11"
    datatype="patent document" segtype="sentence" adminlang="EN"
    srclang="EN"
    o-tmf="TW4Win 2.0 Format"
  >
  </header>
  <body>
<tu creationdate="20131025T135933Z" creationid="clef-ip-2011"><prop
type="Txt::Note">CLEF-claim_en-fr.p1.en-CLEF-claim_en-fr.p1.fr</prop>
<tuv xml:lang="EN"><seg>An additive for a lubricating oil or hydrocarbon fuel obtainable
by a process which comprises reacting a polyamino alkenyl or alkyl succinimide with a
compound of the general formula: wherein W is oxygen or sulfur; X is oxygen or sulfur; R4
is an alkylene group of 2 or 3 carbon atoms optionally substituted by from 1 to 3 alkyl
groups of 1 or 2 carbon atoms each; and R5 is hydrogen or alkyl of from 1 to 20 carbon
atoms.</seg></tuv>
<tuv xml:lang="FR"><seg>Additif pour une huile lubrifiante ou un combustible hydrocarboné,
pouvant être obtenu par un procédé qui consiste à faire réagir un polyamino-alcényl- ou
alkyl-succinimide avec un composé de formule générale dans laquelle W représente
l&apos;oxygène ou le soufre ; X représente l&apos;oxygène ou le soufre ; R4 représente un
groupe alkylène ayant 2 ou 3 atomes de carbone, facultativement substitué avec 1 à 3
groupes alkyle ayant chacun 1 ou 2 atomes de carbone ; et R5 représente l&apos;hydrogène
ou un groupe alkyle ayant 1 à 20 atomes de carbone.</seg></tuv> </tu>

<tu creationdate="20131025T135933Z" creationid="clef-ip-2011"><prop
type="Txt::Note">CLEF-claim_en-fr.p1.en-CLEF-claim_en-fr.p1.fr</prop>
<tuv xml:lang="EN"><seg>An additive as claimed in Claim 1, wherein R4 is alkylene of 2 or
3 carbon atoms.</seg></tuv>
<tuv xml:lang="FR"><seg>Additif suivant la revendication 1, dans lequel R4 représente un
groupe alkylène ayant 2 ou 3 atomes de carbone.</seg></tuv> </tu>

<tu creationdate="20131025T135933Z" creationid="clef-ip-2011"><prop
type="Txt::Note">CLEF-claim_en-fr.p1.en-CLEF-claim_en-fr.p1.fr</prop>
<tuv xml:lang="EN"><seg>An additive as claimed in Claim 1 or 2, wherein R5 is hydrogen or
alkyl of from 1 to 10 carbon atoms.</seg></tuv>
<tuv xml:lang="FR"><seg>Additif suivant la revendication 1 ou 2, dans lequel R5 représente
l&apos;hydrogène ou un groupe alkyle ayant 1 à 10 atomes de carbone.</seg></tuv> </tu>
```

Figure 47 : Exemple des données en format TMX (MT CLEF-IP anglais-français)

VIII.1.2 Méthode de création

Nous avons construit les ressources statiques en utilisant 2 méthodes. Grâce à SECTra_w, qui offre un système d'annotation de chaque traduction ou post-édition d'un segment par un niveau de fiabilité (de * à *****) et un score (de 0 à 20), il est possible d'extraire de la mémoire de traductions, associée à un site Web, une sous-MT vérifiant n'importe quel prédicat basé sur les niveaux et les scores. Le résultat de post-édition peut être exporté en format XML et TXT. Nous avons choisi les segments post-édités de haute qualité (par les critères de SECTra).

L'exemple suivant montre une extraction à partir de la partie français-chinois de la MT-demo2 associée à la IMAG-DOC_PAR_JOUR. Le prédicat est tout simplement [Level = 3 & score> = 13], et ses paramètres peuvent être modifiés directement via l'interface graphique.

	No	Pseudo Doc	Source	Cible	Stars	Notes
	1	DOC16	la salle du haut conseil située au 9ème étage de l\'institut du monde arabe, dans le 5ème.	位于巴黎5区的阿拉伯世界博物馆10楼高级理事会议厅。	3	20
	2	DOC16	le 24ème étage de la tour zamansky, l\'université pierre et marie curie, sur le campus de jussieu, dans le 5ème.	位于巴黎5区的皮埃尔和玛丽·居里大学加希耶校区的扎曼斯基大楼25楼。	3	20
	3	DOC16	le 18ème étage de la bibliothèque françois mitterrand, dans le 13ème.	位于巴黎13区的法国国家图书馆密特朗官19楼。	3	20
	4	DOC16	le 6ème étage de l\'hôtel industriel de dominique perrault, dans le 13ème.	位于巴黎13区的多米尼克·佩罗工业馆7楼。	3	20
	5	DOC16	la nuit blanche 2012 permettra aux visiteurs de découvrir la ville lumière d\'en haut grâce à 15 belvédères normalement fermés au public.	2012巴黎不眠夜将使参观者可以从15个平时不对外开放的平台发现欣赏巴黎这座光影之城。	3	20
	6	DOC16	jk rowling	j.k罗琳	3	20
	7	DOC16	en effet, jk rowling, qui a créé notre sorcier à lunettes, pourrait se replonger dans l\'univers d\'harry potter.	事实上，"创造了"我们那位眼镜魔法师的j.k罗琳，有可能会写哈利波特的魔法世界里的续集。	3	20
	8	DOC16	c\'est en 2011 que la saga de nos célèbres sorciers s\'est achevée, à l\'issue du septième livre intitulé « harry potter et les reliques de la mort » qui a été divisé en deux parties au cinéma.	2011年，第七本《哈利波特和死亡圣器》分为上下两部电影，上映结束之后，我们这著名的哈利波特系列魔法小说完结了。	3	20
	9	DOC16	cinq ans après le dernier opus de cette série à succès, jk rowling revient avec une nouvelle œuvre, pour adultes cette fois.	在这成功的系列小说最后一章完结的5年之后, jk罗琳带着她的新作品回归了。这次是给大人看的小说。	3	20
	10	DOC16	et la star des librairies a su entretenir le mystère.	这位图书史上的明星会将这光芒延续下去。	3	20
	11	DOC14	la joie	欢乐	3	20
	12	DOC16	« une bourgade apparemment idyllique mais qui va faire face "aux tourments les plus violents ».	"一个表面看起来非常美好的田园小镇，但是它将会面临最猛烈的动荡。"	3	20

Figure 48 : Extraction d'une "bonne" MT de la MT produite par post-édition "naturelle"

Avec l'autre méthode, nous récupérons d'abord les documents multilingues comme le corpus CLEF-IP 2011, et analysons les documents.

Ensuite, nous extrayons les segments parallèles, et les nettoyons.

Dans un troisième temps, nous les alignons et les mettons en format TXT et TMX.

VIII.1.3 Résultats

Après les 4 premières années d'utilisation (09/2011-09/2015), il y avait environ 80 sites Web accédés par des IMAG. Il y en a maintenant plus de 200. Les sites visités ont plus de 8 langues comme langue source, et plus de 10 langues comme langue cible avec de la post-édition. Il y a plus de 820.000 segments, et environ 45% (370.000+) des segments ont été post-édités par des contributeurs (payés ou non, organisés ou occasionnels). La plupart des segments parallèles sont en anglais-français, en anglais-chinois, et en français-chinois.

Voici quelques données chiffrées.

Dans le Tableau 58, on donne la longueur du texte source des segments post-édités, le balisage XML personnalisé est supprimé, et on compte le nombre de caractères chinois pour le chinois.

Tableau 58 : Segments post-édités dans SECTra_w à partir de 3 langues source

Langue	Phrase	Mots
anglais	350 174	5 252 610
français	55 642	834 621
chinois	270 108	3 241 296

Le Tableau 59 présente les segments post-édités avec la direction de post-édition et la taille de la MT.

Tableau 59 : Segments parallèles obtenus à partir des MT (mêmes remarques)

Paire de langues (L1→L2)	Phrase	Mots L1 (ou caractères)	Mots L2 (ou caractères)	Taille L1	Taille L2
anglais→français	121 074	2 542 731 9 685,9 pages	2 613 351 9862,1 pages	10,1Mo	10,4Mo
anglais→chinois	208 106	4 370 530 16 648,5 pages	6 063 942 151 159,8 pages	19,1Mo	17,6Mo
français→anglais	29 079	627 661 2 326,3 pages	610 098 2 210,1 pages	4Mo	3,9Mo
français→chinois	10 890	228 703 871,2 pages	317 322 793,3 pages	1,5Mo	1,25Mo
chinois→anglais	2 013	58 656 146,6 pages	42 275 161,1 pages	240Ko	263Ko
chinois →français	10 062	291 192 727,9 page	211 185 804,7 page	874Ko	1Mo

VIII.1.3.1 Français-chinois pour l'énergie

Dans le cadre de ma thèse, j'ai travaillé sur la construction d'un système de TA français→ chinois pour un client potentiel de L&M. J'ai collecté la ressource française concernant le domaine de l'énergie, et puis nous l'avons post-éditée en chinois à l'aide de GT. La Figure 49 affiche 2 fichiers TXT exportés de SECTra. Le fichier en haut est celui des segments source, le fichier du bas est celui des segments post-édités.

Répondre à la demande croissante d'énergie tout en parant aux risques climatiques et à la raréfaction des ressources : c'est le défi lancé aux énergéticiens. C'est pourquoi le développement durable est au cœur de notre stratégie. Notre politique traduit la volonté du Groupe de « changer l'énergie ensemble » en apportant des solutions réalistes.
La lutte contre le changement climatique commence par la maîtrise de nos impacts environnementaux.
Nous nous sommes fixé 9 engagements pour répondre à 3 enjeux prioritaires
La lutte contre le changement climatique et la préservation de la biodiversité :
Un enjeu environnemental
rester, en tant que Groupe, le moins émetteur de CO des grands énergéticiens européens,
1. adapter notre parc de production et nos offres au changement climatique,
2. réduire notre impact environnemental, notamment sur la biodiversité.
3. Faciliter l'accès à l'énergie et développer des liens de proximité avec les territoires :
Un enjeu sociétal
favoriser l'accès à l'énergie et l'éco-efficacité énergétique,
4. développer dans la durée la proximité avec les territoires où nous opérons,
5. contribuer à l'effort éducatif sur les questions liées à l'énergie.
6. Contribuer au débat sur le développement durable par le dialogue, l'information et la communication :
Un enjeu de gouvernance
poursuivre le développement des politiques et le partage des valeurs au sein du Groupe, en relation avec les parties prenantes,
7. communiquer et rendre compte des activités et résultats du Groupe en matière de développement durable,
8. participer au débat sur le développement durable au niveau national et international.

既能满足日益增长的能源需求，又要应对气候变化风险和资源枯竭问题，这就是能源企业面临的挑战。为此，可持续发展成为法国电力集团的战略核心。"共同改变能源结构"体现了法国电力集团的意愿，为此，
我们将提供切实可行的技术方案。
应对气候变化首先应从控制环境影响做起。
我们制定了九项承诺，优先应对以下三方面的挑战：
应对气候变化，保护生物的多样性：
环境的挑战
法国电力集团将继续保持欧洲大型能源公司二氧化碳排放最少企业的地位；
1. 电力生产设施和服务项目适应气候变化的要求；
2. 减少对环境，特别是对生物多样性的影响。
3. 为普及能源使用提供方便，与地方合作，推行就近供电政策：
社会的挑战
为普及能源使用提供便利，发挥能源项目的生态效益；
4. 与地方合作，在集团经营区域内长期实施就近供电政策；
5. 支持能源教育的各种活动。
6. 支持关于可持续发展的辩论，可采用对话、信息与交流等各种形式：
管理方面的挑战
与有关方面合作，在集团内部继续制定相关政策，共享企业价值观；
7. 交流和通报集团在可持续发展方面开展的工作和取得的成果；
8. 参与法国和国际有关可持续发展的讨论。

Figure 49 : Segments post-édités pour la ressource énergie

Il y a environ 10K segments français, que nous divisons en 500 pages Web. Après la post-édition, nous avons récupéré 9000 segments français→chinois (Tableau 61).

Tableau 61 : Statistique des données pour la ressource énergie

	Nb de segments	**Longueur moyenne de segment**	**Nb de mot (ou caractère chinois)**	**Page_std**
français	9000	41	368K	263
chinois	9000	36	324K	810

VIII.2 Contribution sous forme de système de TA

Il s'agit soit de systèmes téléchargables, soit de système utilisables comme des services Web.

VIII.2.1 Systèmes de TA téléchargeables

Nous les avons mis à l'url http://…. Il s'agit des systèmes créés en français-chinois et des systèmes créés à partir de la collection de brevets CLEF-IP 2011.

Tableau 62 : Systèmes de TA téléchargeables

Nom du système de TA	Langues	Ressources utilisées	Commentaire
Moses-L&M	fr-zh	MT de 9000 segments créée par L&M	Qualité assez basse BLEU TER Tpe estimé
Moses-LIG	fr-zh	MT de 80K segments PE + 100K MultiUN	
Moses-CLEF-IP	fr-de fr-en de-fr		Très bien sur les parties de brevets BLEU :

VIII.2.2 Systèmes de TA utilisables comme des service Web

Certains des systèmes précédents sont des services Web utilisables directement via l'API de TRADOH, ou la plate-forme JIANDAN-EVAL. Certains sites Web boursiers et économiques sont aussi utilisables via une IMAG, ce qui permet l'amélioration par AI.

Tableau 63 : Systèmes de TA utilisables comme des services Web

Nom du système de TA	Langues	Ressources utilisées	Commentaire
Moses-L&M	fr-zh	MT de 9000 segments créée par L&M	Qualité assez basse
Moses-LIG	fr-zh	MT de 80K segments PE + 100K MultiUN	
Moses-CLEF-IP	fr-de fr-en de-fr		Très bien sur les parties de brevets

VIII.3 Passerelles iMAG vers des sites Web statiques ou dynamiques

Une ressource de ce type est simplement une IMAG-S dédiée à un site Web S. La MT associée peut être propre ou partagée, cela n'a pas d'importance pour les visiteurs. Nous distinguons les sites accédés S selon qu'ils sont statiques ou dynamiques.

VIII.3.1 Passerelles iMAG pour des sites statiques

Il s'agit essentiellement de sites contenant des documents mis sous forme HTML, et pour lesquels on vise à obtenir une assez bonne qualité de traduction.

Tableau 64 : Passerelles iMAG pour des sites statiques

iMAG	Documents accédés	Commentaires
BemBook	Le BemBook (de livre 700 p.) Site : http://www.bembook.ibpsa.us/	Post-édité vers le français de façon occasionnelle Expérimentation organisée pour PE du chapitre 21 en 7~8 langues.
Corpus par jour	« Voyage au centre de la terre » de Jules Verne	Post-édition de ce roman du français vers le chinois
Manchanda	« IITB : Monastery, Sanctuary, Laboratory »	Préparé pour la PE vers français et hindi
Macau	Support de cours	Pour les étudiants étrangers
Powers	« The book of me » de Powers	Expérimentation de L. Besacier
CLEF-IP	Corpus de brevets « CLEF-IP 2011 »	PE inutile pour les langues de la collection, mais possible vers le chinois

VIII.3.2 Passerelles iMAG pour des sites dynamiques

Il s'agit ici de sites dont certaines pages changent régulièrement, par exemple celles des nouvelles dans des sites Web d'organismes ou de firmes.

Il y a aussi des sites de journaux, utilisés pour la construction de ressources (corpus parallèles, dictionnaires bilingues) utilisables pour construire ensuite des systèmes de TA (empiriques ou experts ou les deux) vers des langues peu dotées. C'est le cas des iMAG créées pour la Nation de Djibouti (thèse sur le somali) et pour La Gazette des Comores et Al-Watwan (thèse sur le comorien et plus particulièrement le shingazidja) visant à créer un système de lecture active de textes administratifs et juridiques écrits en français (appris à l'école) à l'intention des Comoriens qui parlent presque toujours le comorien en fait.

Tableau 65 : iMAG pour des sites dynamiques

iMAG	sites accédés	Commentaires
LIG-LAB	Site Web du LIG	Post-édité vers le chinois (avec expérimentation) et l'anglais. Un peu vers l'allemand.
lanationdedjibouti	Site Web du journal La Nation de Djibouti. Environ 500 nouveaux segments par jour, 3 dernières années en ligne.	Construction par PE (à partir des résultats de GT) d'une bonne MT français-somali. But : créer un système de TA Moses meilleure que GT pour d'accéder à ce journal.
lametro	Site Web de La Métro à Grenoble	Site de La Métro. Gros travail (fr→zh/en) en 2010.
alwatwan	Sites Web des 2 journeaux Al-Watwan et La Gazette des Comores	Construction par PE à partir des traductions en wahili de GT d'une bonne MT français-comorien

VIII.3.3 Structure d'une contribution « dynamique » par iMAG

Il s'agit ici d'iMAG dédiées à des sites Web dynamiques. Les utilisateurs post-éditent le contexte sur SECTRA_W/iMAG. Les segments post-édités sont sauvegardés dans les MT.

Il y a 2 types de sites Web dynamiques. (1) Un site Web (par exemple le site Web *lamétro*[90]) met à jour son contexte en temps réel. Pendant que nous accédons son iMAG dédiée dans une langue cible, nous choisissons les phrases intéressantes, et les post-éditons. Un segment peut être post-édité (vers une ou plusieurs langues) par un ou plusieurs post-éditeurs, donc chaque segment source peut avoir plusieurs versions de segments post-édités. Quand nous avons un certain nombre de segments post-édités, nous les exportons pour construire la MT. (2) Nous collectons des ressources spécifiques pour construire une MT adaptée au sous-langage. Tout d'abord, nous récupérons les textes, et les nettoyons. Ensuite nous les mettons en format HTML (Figure 50). Enfin, nous les ajoutons dans SECTRA_W/iMAG.

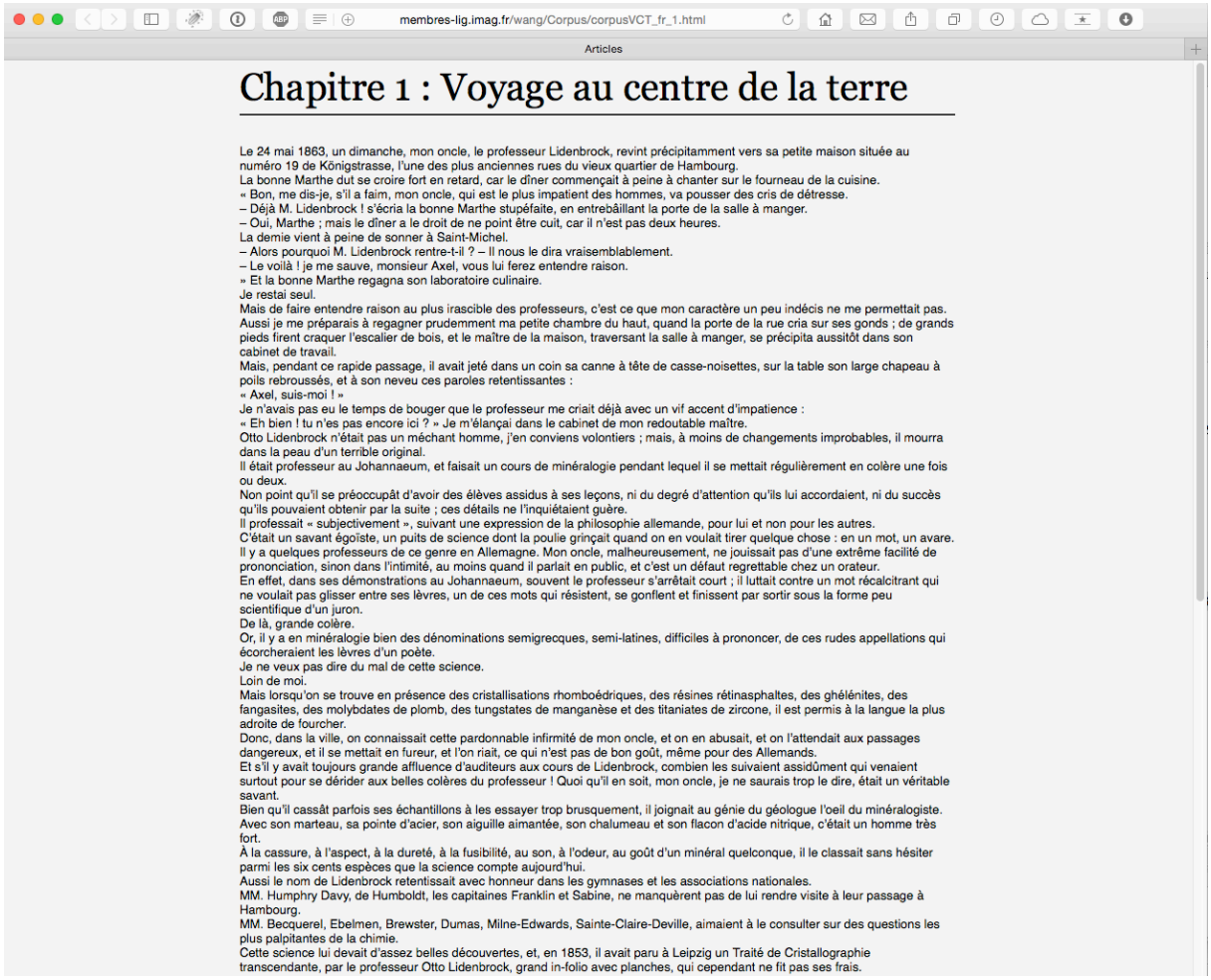


Figure 50 : Exemple de contribution au format HTML (Chapitre 1 : Voyage au centre de la Terre)

[90] http://www.lametro.fr

VIII.3.4 Remarques sur la création de certains des sites « contribués »

Certaines de ces contributions ont demandé des travaux de préparation spécifiques.

VIII.3.4.1 La Métro

Le site Web de La Métro de Grenoble contient 2500 pages Web, soit environ 30000 segments. Plus de la moitié ont été post-édités en chinois pour l'Exposition universelle de Shanghai en 2010. Nous avons créé un système français-chinois spécifique en choisissant une partie "assez bonne" de cette MT du 26/06/2015 au 30/06/2015 par la sélection:

```
TM-lametro-extract-good = TM_select (lametro, [level=3 & score >=13 |
level=4 & score >=13 | level=5 & score >=11.5], [langage_pair=fr-zh |
langage_pair=fr-en | langage_pair=zh-fr], [Begin_date : 20150626,
End_date : 20150630]).
```

VIII.3.4.2 MUMIA-CLEF

Avce Huanan Sun (M1-TER), nous avons construit 3 sites Web monolingues pour le corpus CLEF-IP. Chaque site Web contient environ 500K pages Web. Chaque page Web représente un fichier XML de CLEF-IP (Figure 51). J'ai créé 3 iMAG dédiées pour accéder à ces sites en utilisant les systèmes que nous avons créés pour les couples fr→en, de→fr, fr→de, et les services de TA disponibles pour les autres couples.

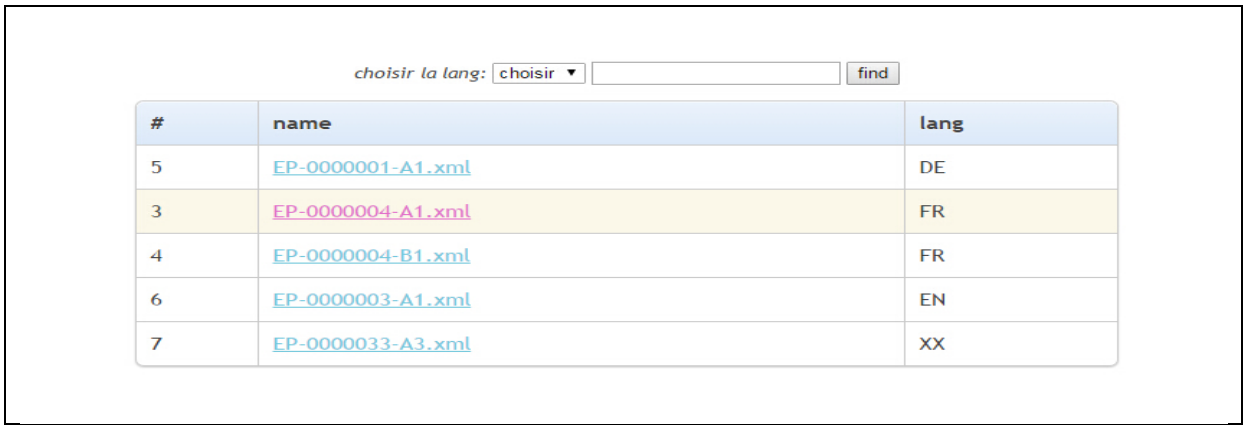


Figure 51 : Capture d'écran du site Web monolingue de CLEF-IP

La Figure 52 montre une capture d'écran de l'iMAG CLEF-EN utilisée pour accéder dans d'autres langues aux brevets écrits initialement en anglais. Les segments entre parenthèses vertes ont été trouvés dans la MT CLEF-IP, et les segments entre parenthèses rouges ont été traduits par le système de TA anglais→français.

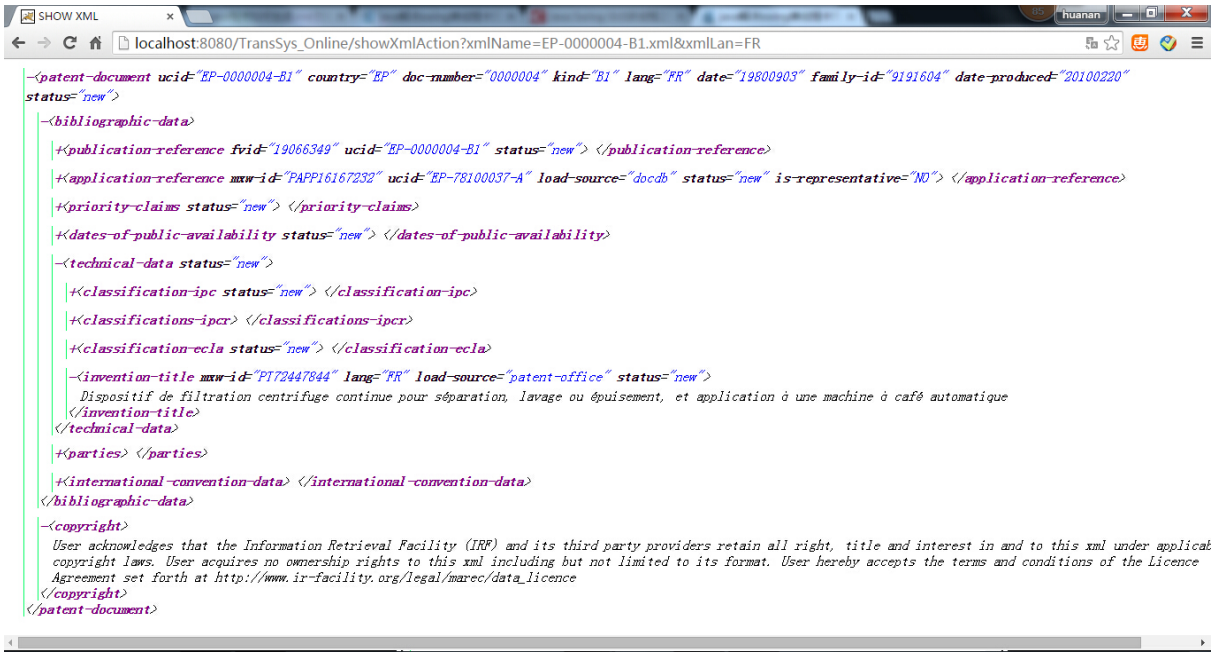


Figure 52 : Capture d'écran de l'iMAG dédiée CLEF-IP

Chapitre IX Vers une plate-forme de construction, déploiement et évaluation de systèmes de TA: JianDan-eval

Cette partie présente la construction en cours de la plate-forme JianDan-eval. Elle contient pour l'instant le cahier des charges et l'architecture logicielle. Ce système est partiellement implémenté, et a déjà servi à un étudiant de M2R pour construire et évaluer des systèmes MOSES, mais les spécifications externes de certaines fonctionnalités restent à compléter, tandis que d'autres aspects sont déjà opérationnels (par exemple, la saisie des paramètres d'entraînement et d'évaluation). Nous les présenterons en annexe ou les incluerons dans la version finale de ce mémoire.

IX.1 Buts du système JIANDAN-EVAL

JIANDAN-EVAL est une plate-forme (en construction) d'évaluation, de création et d'utilisation de systèmes de traduction automatique (TA). JIANDAN-EVAL permettra la construction de TA en ligne, l'évaluation de la qualité de systèmes de TA (support de campagnes ponctuelles ou permanentes), et la création d'une infrastructure pour l 'évaluation humaine (objective par PE, et subjective).

La conception de JIANDAN-EVAL a été inspirée par KANTANMT[91], et, comme dans KANTANMT, la partie « construction de TA » est limitée aux systèmes MOSES. Par contre, la partie « évaluation » est ouverte à tous les types de systèmes de TA, du moment qu'on peut les appeler par TRADOH (il s'agit donc de versions en ligne et gratuites).

JIANDAN-EVAL a deux groupes de fonctionnalités principales. D'abord, la création ou l'intégration de TA dans le nuage ; ce point contient les fonctions suivantes:

- création de systèmes de TA sur le nuage
- intégration de l'appel à des systèmes de TA créés par des utilisateurs
- création de scripts pour l'entraînement de systèmes de TA
- service de traduction (serveurs de TA)
- surveillance de l'entraînement

Le second groupe de fonctionnalités concerne l'évaluation de différents systèmes de TA (comme MOSES, JOSHUA, NIUTRANS, APERTIUM, GT, SYSTRAN, BING, REVERSO, etc.) et la génération de sites Web pour l'évaluation. JIANDIAN-EVAL doit supporter des campagnes d'évaluation ponctuelles (comme IWSLT) ou permanentes:

- paramétrabilité des TA et des données
- paramétrabilité des méthodes « objectives » : BLEU, NIST, TER, MTER, temps de PE…
- paramétrabilité des méthodes « subjectives » : fluidité, parallélisme/adéquation, valeur d'usage, fiabilité, exactitude terminologique, et phraséologique …
- création d'une infrastructure pour l'évaluation humaine (objective par PE, et subjective).

JDE intégrera partout l'aspect contributif et historique.

[91] KantanMT : http://www.kantanmt.com

IX.2 Architecture logicielle

L'architecture de JianDan-eval, présenté ci-dessus, est composée de 3 parties.

Tout d'abord, il y a un serveur d'applications,

- qui assure le rôle de passerelle vers la base des données de tous les utilisateurs,
- qui contient une base de données,
- qui fournit l'interface utilisateur (client léger) pour accéder aux pages Web,
- qui gère les appels à la TA.

Ensuite, il y a un serveur de TA (ou plusieurs), qui contient les décodeurs et les modèles de traduction, et qui fournit le service de traduction. Ce serveur n'est pas visible par l'utilisateur, mais l'interface de service de traduction est accessible par l'API de TRADOH.

Enfin, il y a un serveur d'entraînement, qui est utilisé pour la création de systèmes de TA de type MOSES.

Serveur d'applications

Le serveur d'applications assure un double rôle dans l'architecture de JIANDAN-EVAL.

D'abord, il hébergera l'application Web d'édition des disponibilités (client léger). Afin d'éviter un nombre important de problèmes de sécurité, les communications entre le serveur et l'utilisateur pourront être chiffrées.

Le serveur d'applications permettra également de fournir les services permettant aux utilisateurs d'accéder aux informations de la base de données. L'avantage est qu'ainsi les interactions (consultation, modification ou suppression) ne se feront que depuis un seul et unique point d'accès, ce qui facilitera grandement la maintenance. De cette manière, nous ne dépendrons pas du choix de la base de données. De plus, cela permettra d'assurer un contrôle d'accès plus strict aux informations applicatives, et simplifiera notamment les mises à jour du schéma de la base.

Serveur de TA

Le serveur de TA permettra de fournir les services de traduction par TA au serveur d'applications. Il contiendra donc des décodeurs de TA, comme le décodeur de Moses ou le décodeur de Joshua. Les modèles de traduction seront déployés dans ce serveur.

Serveur d'entraînement

Le serveur d'entraînement permettra de créer les tables d'un système de TA. Tout d'abord, il recevra les données d'entraînement à partir du serveur d'applications. Ensuite, il préparera les données, et démarrera une procédure d'entraînement. Enfin, le serveur d'entraînement enverra le nouveau modèle au serveur de TA.

Interconnexion avec les systèmes extérieurs

- **Authentification avec Facebook/Google+/Twitter**

Pour l'authentification des utilisateurs, il sera intéressant de proposer des alternatives de connexion rapide via les réseaux sociaux : FACEBOOK, GOOGLE PLUS ou TWITTER.

L'authentification sociale permet de gérer à la fois l'authentification sur les réseaux, et aussi de récupérer facilement des données depuis ces derniers.

- **Envoi de courriels avec SMTP**

Les envois de courriels se feront par l'intermédiaire du serveur SMTP de chaque machine.

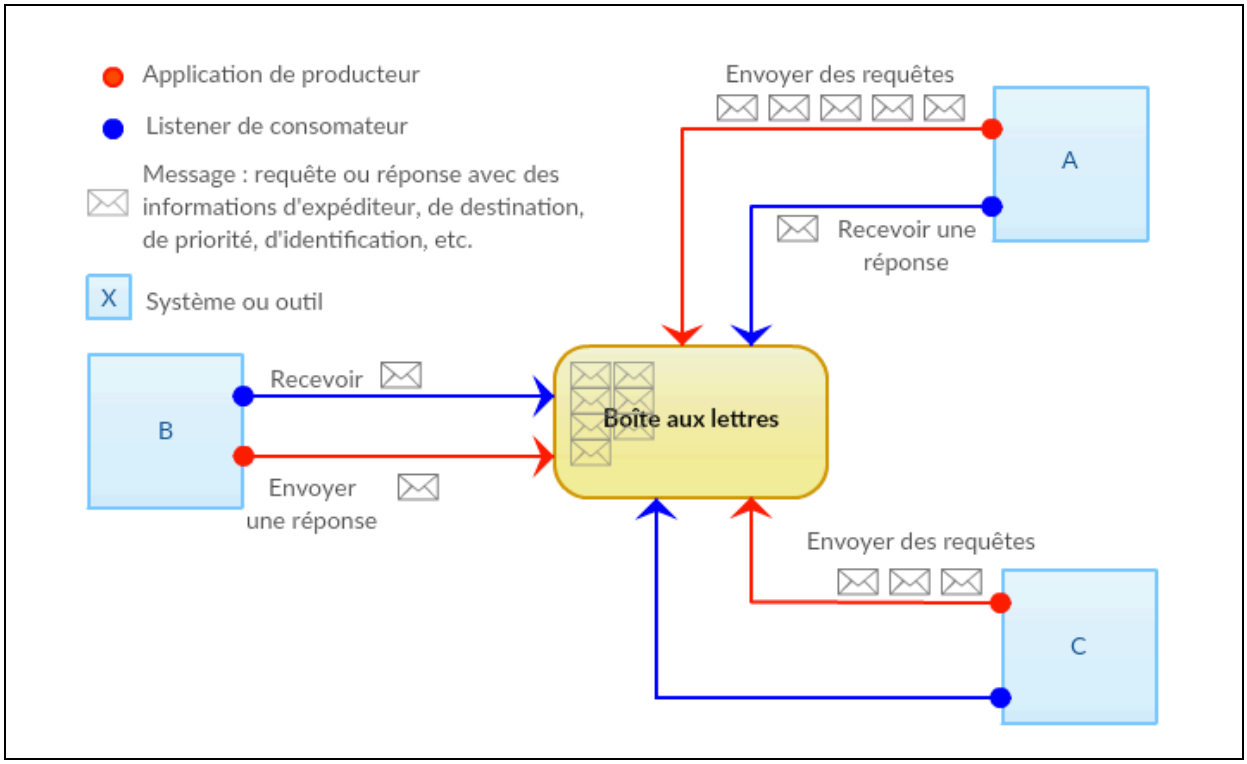


Figure 53 : Architecture initiale de gestion de travaux

- **Files d'attente & tâches**

Après une brève études de l'existant, nous nous proposons d'utiliser l'intergiciel ACTIVEMQ pour gérer les requêtes. Cette méthode permet non seulement de contrôler l'ordre des requêtes en ordre, mais aussi de mettre en place une gestion de priorité en donnant une importance particulière à certaines tâches, ou à certains clients. De la même manière, en qualifiant correctement les traitements, il sera facile de mettre en place des serveurs dédiés à certains processus.

Acteurs du système

Les acteurs sont les différentes entités interagissant avec le système. Nous avons identifié différentes catégories d'acteurs : utilisateur, administrateur, et évaluateur.

- **Utilisateur**. L'acteur « utilisateur » représente tout utilisateur physique souhaitant se servir de JIANDAN-EVAL. Il devra s'identifier dans le système, afin d'avoir par la suite les droits et privilèges correspondant à son profil.
- **Administrateur.** L'acteur « administrateur » gère les utilisateurs, et surveille les privilèges des utilisateurs.
- **Évaluateur**. L'acteur « évaluateur » évalue la qualité de TA, et les post-éditions par les utilisateurs. Il y a 2 types d'évaluateur, « évaluateur permanent » et « évaluateur temporaire ». Le rôle d'évaluateur permanent, avec ses privilèges, peut être attribué par l'administrateur de JIANDAN-EVAL. Le rôle d'évaluateur temporaire peut être attribué par l'administrateur, et aussi par l'organisateur d'un projet, dans un ou plusieurs projets.

Conclusions et perspectives

Dans le cadre de cette thèse, effectuée dans le cadre d'une bourse CIFRE, et prolongeant un des aspects du projet ANR TRAOUIERO, nous avons d'abord abordé la production, l'extension et l'amélioration de corpus multilingues par traduction automatique (TA) et post-édition contributive (PE). Nous avons apporté des améliorations fonctionnelles et techniques aux logiciels SECTRA_W et IMAG produits lors de thèses antérieures (P.C. Huynh, H.T. Nguyen), et nous pensons avoir progressé vers une définition générique de la structure d'un corpus multilingue, multi-annoté et multimédia, pouvant contenir des documents classiques aussi bien que des pseudo-documents (comme des pages Web) et des méta-segments. Cette partie a été validée par la création de bons corpus bilingues français-chinois, l'un d'eux résultant de la toute première application à la traduction littéraire (un roman de Jules Verne que nous avons traduit en chinois pour améliorer notre connaissance du français).

La seconde partie de cette thèse a initialement été motivée par un besoin industriel. Elle a consisté à construire des systèmes de TA de type Moses, spécialisés à des sous-langages, en français↔chinois, et à étudier la façon de les améliorer dans le cadre d'un usage en continu avec possibilité de PE. Dans le cadre d'un projet interne sur le site du LIG et d'un projet (TABE-FC) en coopération avec l'université de Xiamen, nous avons pu démontrer l'intérêt de l'apprentissage incrémental en TA statistique, sous certaines conditions, grâce à une expérience qui s'est étalée sur toute la thèse. Lors de la deuxième phase de cette expérience, nous n'avons pas pu faire d'apprentissage incrémental, à cause de l'indisponibilité d'un serveur, et n'avons « recompilé » le système de TA qu'à la fin. Cependant, cette phase a permis de démontrer le gain de qualité obtenable par spécialisation à un sous-langage. La troisième phase, toujours en cours, démontre vraiment l'apport de l'AI, toutes choses égales par ailleurs.

La troisième partie de la thèse a été consacrée à des contributions et mises à disposition de supports informatiques et de ressources. Cet aspect n'est pas seulement intéressant par ses résultats, mais aussi parce que la spécification et surtout l'implémentation de ressources et outils aussi variés et de grande taille pose des problèmes spécifiques, auxquels nous avons trouvé des solutions assez génériques et efficaces.

Nos principales contributions se placent dans le cadre du projet COST MUMIA de l'EU et résultent de l'exploitation de la collection CLEF-IP 2011 de 1,5M de brevets partiellement multilingues. De grosses mémoires de traductions en ont été extraites (17,5 M segments), 3 systèmes de TA en ont été tirés (de-fr, en-fr, fr-de), et un site Web de support à la RI multilingue sur les brevets a été construit.

Enfin, nous avons construit en 2015 deux IMAG utilisées par des doctorants pour construire des MT de bonne qualité en français-somali et français-comorien par post-édition de journaux de Djibouti et des Comores, et ensuite en dériver des systèmes de TA spécialisés aux sous-langages de ces journaux, en extraire du vocabulaire bilingue, etc.

Dans le tout dernier chapitre, nous avons décrit la réalisation en cours de JIANDAN-EVAL, une plate-forme de construction, déploiement et évaluation de systèmes de TA. Ce travail s'est révélé plus lourd que nous ne le pensions initialement, et n'est pas achevé au moment où nous rédigeons. Nous espérons cependant qu'il le sera avant 2016. Il faut dire aussi qu'une version α a déjà été utilisée depuis début 2015 par H. Wang pour son M2R IdL sur l'estimation a priori (QE) de la qualité (d'usage) des résultats de systèmes de TA variés, toujours en fr-zh.

Les perspectives de cette recherche sont multiples. D'abord, nous voulons terminer l'implémentation de JIANDAN-EVAL, le mettre à disposition de la communauté, et l'évaluer.

Ensuite, nous voudrions continuer notre travail sur la modélisation des corpus de traductions, et arriver à une nouvelle implémentation de SECTRA dans laquelle on pourrait définir un corpus en s'appuyant sur la description de sa structure (*macrostructure*, *microstructure* et *mésostructure*), écrite dans un langage qui reste encore à définir (sans doute par une DTD XML, puisque nous avons déjà décrit toutes les métadonnées de nos différents corpus en XML).

Enfin, nous souhaitons trouver un cadre approprié pour valoriser notre expertise en TA français→chinois, et pour développer aussi des systèmes chinois→français. Un cadre approprié pourrait être celui de l'accès (de HQ) à des sites boursiers et économiques en français par des sinophones, puis à des sites de même type en chinois par des francophones.

Bibliographie

[1] (Alabau, Bonk et al., 2013) Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L. and Mesa-Lao, B. (2013). CASMACAT: An open source workbench for advanced computer-aided translation. The Prague Bulletin of Mathematical Linguistics 100: 101-112.

[2] (Allen, 1999) Allen, J. (1999). The Postediting Special Interest Group (SIG). MT News International 21.

[3] (Allen, 2003) Allen, J. (2003). Post-editing. Benjamins Translation Library 35: 297–318.

[4] (Arnold, Balkan et al., 1994) Arnold, D., Balkan, L., Meijer, S., Humphreys, R. and Sadler, L. (1994). Machine translation: An introductory guide. NCC Blackwell. 200 p.

[5] (Bertoldi, Simianer et al., 2014) Bertoldi, N., Simianer, P., Cettolo, M., Wäschle, K., Federico, M. and Riezler, S. (2014). Online adaptation to post-edits for phrase-based statistical machine translation. Machine Translation 28(3-4): 309-339.

[6] (Besacier, 2014) Besacier, L. (2014). Traduction automatisée d'une œuvre littéraire: une étude pilote. Proc. *Traitement Automatique du Langage Naturel (TALN)*. Marseille, France.

[7] (Blanchon, Boitet et al., 1999) Blanchon, H., Boitet, C. and Caelen, J. (1999). Participation francophone au consortium C-STAR II.

[8] (Boitet, 1996) Boitet, C. (1996). La synergie entre THAM, réseau et TA comme facteur de progrès théoriques et pratiques en TAO. Proc. *TAL+ AI-96 (le traitement du langage et ses applications industrielles)*. Moncton. 4-6.

[9] (Boitet, 2009) Boitet, C. (2009). Machine Translation (MT) and Computer-Aided Translation (CAT) — Lecture 1: Linguistic, computational and operational architectures of MT systems. Proc. *NII lecture series*. Tokyo, Japan. Slides available on NII web site. 44 p.

[10] (Boitet, Bellynck et al., 2008) Boitet, C., Bellynck, V., Mangeot, M. and Ramisch, C. (2008). Towards Higher Quality Internal and External Multilingualization of Web Sites. Proc. *ONII-08 (Summer Workshop on Ontology, NLP, Personalization and IE/IR)* IITB, Mumbai, Inde. 8 p. .

[11] (Boitet and Berment, 2012) Boitet, C. and Berment, V. (2012). Heloise—An Ariane-G5 compatible environment for developing expert MT systems online. Proc. *24th International Conference on Computational Linguistics (COLING)*. Mumbai. 9 p.

[12] (Boitet, Bey et al., 2005) Boitet, C., Bey, Y. and Kageura, K. (2005). Main research issues in building web services for mutualized, non-commercial translation. Proc. *the 6th Symposium on Natural Language Processing (SNLP)*. pp. 451–454.

[13] (Boitet, Blanchon et al., 1994) Boitet, C., Blanchon, H., Seligman, M. and Bellynck, V. (1994). Evolution of MT with the Web. Proc. *Conference "Machine Translation 25 Years On"*. pp. 1-13.

[14] (Boitet, Boguslavskij et al., 2007) Boitet, C., Boguslavskij, I. M. and Cardeñosa, J. (2007). An evaluation of UNL usability for high quality multilingualization and projections for a future UNL++ language. Proc. *Computational Linguistics and Intelligent Text Processing*. Springer. pp. 361–373.

[15] (Boitet, Huynh et al., 2010) Boitet, C., Huynh, C.-P., Nguyen, H.-T. and Bellynck, V. (2010). The iMAG concept: multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations. Proc. *Traitement Automatique du Langage Naturel (TALN)*. Montréal, Canada.

[16] (Brown, Pietra et al., 1993) Brown, P. F., Pietra, V. J. D., Pietra, S. a. D. and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2): pp. 263–311.

[17] (Callison-Burch, Osborne et al., 2006) Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. Proc. *EACL*. 249-256.

[18] (Cettolo, Federico et al., 2010) Cettolo, M., Federico, M. and Bertoldi, N. (2010). Mining parallel fragments from comparable texts. Proc. *IWSLT*. 227–234.

[19] (Chang, Galley et al., 2008) Chang, P.-C., Galley, M. and Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation performance. Proc. *third workshop on statistical machine translation* Association for Computational Linguistics (ACL-SMT). pp.224-232.

[20] (Chen, Wang et al., 2014) Chen, Y., Wang, L., Boitet, C. and Shi, X. (2014). On-going Cooperative Research towards Developing Economy-Oriented Chinese-French SMT Systems with a New SMT Framework. Proc. *Traitement Automatique du Langage Naturel (TALN)*. Maresille, France. pp. 401- 408.

[21] (Chiang, 2012) Chiang, D. (2012). Hope and fear for discriminative training of statistical translation models. The Journal of Machine Learning Research 13(1): 1159-1187.

[22] (Damerau, 1964) Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM 7(3): 171-176.

[23] (Daoud, 2007) Daoud, M. (2007). Towards interactive Multilingual Access Gateways (iMAG). Université Joseph Fourier, Grenoble, France. Rapport de M2R. 60 p.

[24] (Déjean and Gaussier, 2002) Déjean, H. and Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables.

[25] (Delpech, 2013) Delpech, E. (2013). Traduction assistée par ordinateur et corpus comparables: contributions à la traduction compositionnelle. Thèse de doctorat, Université de Nantes**:** 266 p.

[26] (Do, 2011) Do, T. N. D. (2011). Extraction de corpus parallèles pour la traduction automatique depuis et vers une langue peu dotée. Thèse de doctorat, Université de Grenoble.

[27] (Doddington, 2002) Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Morgan Kaufmann Publishers. 138–145.

[28] (Dong, 1990) Dong, D. Z. (1990). TRANSTAR: a commercial English-Chinese MT system. Proc. *Conf of the Association for Computational Linguistics (ACL)*. Pittsburgh, Pennsylvania, USA. 339–341.

[29] (Eisele and Chen, 2010) Eisele, A. and Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. Proc. *International conference Language Resources and Evaluation 2011(LREC 2011)*. Istanbul, Turkey. 2868-2872.

[30] (Esplá-Gomis, 2009) Esplá-Gomis, M. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. Proc. *Beyond Translation Memories Workshop (MT Summit XII)*. Ottawa, Ontario, Canada.

[31] (Fafiotte, 2004) Fafiotte, G. (2004). Interprétariat à distance et collecte de dialogues spontanés bilingues, sur une plate-forme générique multifonctionnelle. Proc. *11ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2004),* . Fès (Maroc). 10 p.

[32] (Federico, Bertoldi et al., 2008) Federico, M., Bertoldi, N. and Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. Proc. *9th Annual Conference of the International Speech Communication Association 2008 (INTERSPEECH 2008)*. Brisbane, Australia. 1618–1621.

[33] (Federmann, Giannopoulou et al., 2012) Federmann, C., Giannopoulou, I., Girardi, C., Hamon, O., Mavroeidis, D., Minutoli, S. and Schröder, M. (2012). META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. 3300–3303.

[34] (Feng, 1981) Feng, Z. (1981). Mémoire pour une tentative de traduction multilingue du chinois en français, anglais, japonais, russe et allemand. Doc. GETA, Grenoble, 40p. + annexes.

[35] (Forcada, Ginestí-Rosell et al., 2011) Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine Translation 25: 127–144.

[36] (Ganitkevitch, Cao et al., 2012) Ganitkevitch, J., Cao, Y., Weese, J., Post, M. and Callison-Burch, C. (2012). Joshua 4.0: Packing, PRO, and paraphrases. Proc. *50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea. 283–291.

[37] (Gao and Vogel, 2008) Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. Proc. *Workshop: Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. 49–57.

[38] (Graff, Kong et al., 2003) Graff, D., Kong, J., Chen, K. and Maeda, K. (2003). English gigaword. Linguistic Data Consortium, Philadelphia.

[39] (Guilbaud, 1999) Guilbaud, J.-P. (1999). Ariane-G5: Environnement de développement et d'exécution de systèmes (linguiciels) de traduction automatique. GDR I3 ATALA, Paris.

[40] (Hajlaoui and Boitet, 2008) Hajlaoui, N. and Boitet, C. (2008). TA statistique à petits corpus pour des petits sous-langages. Proc. *TOTh 2008 Conférence sur la Terminologie & Ontologie: Théories et Applications*. 20 p.

[41] (Hanbury, Boyer et al., 2011) Hanbury, A., Boyer, C., Gschwandtner, M. and Müller, H. (2011). KHRESMOI: towards a multi-lingual search and access system for biomedical information. Med-e-Tel, Luxembourg: 412–416.

[42] (Hasler, Haddow et al., 2011) Hasler, E., Haddow, B. and Koehn, P. (2011). Margin infused relaxed algorithm for Moses. The Prague Bulletin of Mathematical Linguistics 96: 69-78.

[43] (Hsu and Su, 1995) Hsu, Y.-L. U. and Su, K.-Y. (1995). The New Generation BehaviorTran: Design Philosophy and system architecture. Proc. *ROCLING*. Taiwan. 65-79.

[44] (Hutchins, 1986) Hutchins, W. J. (1986). Machine translation: past, present, future, Ellis Horwood: Chichester.

[45] (Huynh, 2010) Huynh, C.-P. (2010). Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimédia. Thèse de doctorat, Université Joseph Fourier.

[46] (Kalitvianski, Boitet et al., 2012) Kalitvianski, R., Boitet, C. and Bellynck, V. (2012). Collaborative Computer-Assisted Translation Applied to Pedagogical Documents and Literary Works. COLING (Demos): 255–260.

[47] (King and Chang, 1963) King, G. W. and Chang, H.-W. (1963). Machine translation of Chinese. Scientific American 208: 124-135.

[48] (Koehn, 2005) Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. 79–86.

[49] (Koehn, Hoang et al., 2007) Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. and Others (2007). Moses: Open source toolkit for statistical machine translation. Proc. *Conf of the Association for Computational Linguistics*. 177–180.

[50] (Kohlschütter, Fankhauser et al., 2010) Kohlschütter, C., Fankhauser, P. and Nejdl, W. (2010). Boilerplate detection using shallow text features. Proc. *third ACM international conference on Web search and data mining (WSDM)*. New York City, NY, USA. 441–450.

[51] (Levenshtein, 1966) Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Proc. *Soviet physics doklady*. 707-710.

[52] (Li, Callison-Burch et al., 2009) Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W. N., Weese, J. and Zaidan, O. F. (2009). Joshua: An open source toolkit for parsing-based machine translation. Proc. *Association for Computational Linguistics*. Singapore. 135–139.

[53] (Loh and Kong, 1979) Loh, S.-C. and Kong, L. (1979). An interactive on-line machine translation system (Chinese into English). Translating and the Computer. North-Holland, Amsterdam: 135-148.

[54] (Lu, Tsou et al., 2009) Lu, B., Tsou, B. K., Zhu, J., Jiang, T. and Kwong, O. Y. (2009). The construction of a chinese-english patent parallel corpus. Proceedings of the MT Summit XII: 17–24.

[55] (Mangeot, 2010) Mangeot, M. (2010). Environnements pour lexicographes et lexicologues: Environnements génériques, centralisés et distribués en ligne de construction collaborative de bases lexicales en contexte multilingue. Thèse, Université Joseph Fourier 280 p.

[56] (Mirkin, 2014) Mirkin, S. (2014). Incrementally Updating the SMT Reordering Model. Proc. *Proceedings of The 28th Pacific Asia Conference on Language, Information and Computing (PACLIC), Phuket, Thailand*.

[57] (Munteanu and Marcu, 2005) Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics 31(4): 477–504.

[58] (Munteanu and Marcu, 2006) Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. Proc. *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Sydney, Australia. Association for Computational Linguistics. 81–88.

[59] (Nguyen, 2010) Nguyen, H.-T. (2010). Des systèmes de TA homogènes aux systèmes de TAO hétérogènes. Thèse de doctorat, Université de Grenoble

[60] (Och, 2003) Och, F. J. (2003). Minimum error rate training in statistical machine translation. Proc. *41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan. Association for Computational Linguistics. 160-167.

[61] (Papineni, Roukos et al., 2002) Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. Proc. *Association for Computational Linguistics*. Philadephia, Pennsylvania, USA. 311–318.

[62] (Parr and Quong, 1995) Parr, T. J. and Quong, R. W. (1995). ANTLR: A predicated‐LL (k) parser generator. Software: Practice and Experience 25(7): 789-810.

[63] (Pineau, 2004) Pineau, M. (2004). Comparaison de résultats de traduction (automatique ou non). Université Joseph Fourier, Rapport de TER.

[64] (Pouliquen, Elizalde et al., 2013) Pouliquen, B., Elizalde, C., Junczys-Dowmunt, M., Mazenc, C. and García-Verdugo, J. (2013). Large-scale multiple language translation accelerator at the United Nations. Proc. *MT Summit*. Nice, France.

[65] (Pouliquen and Mazenc, 2011) Pouliquen, B. and Mazenc, C. (2011). COPPA, CLIR and TAPTA: three tools to assist in overcoming the Pat-ent language barrier at WIPO. Proceedings of the 13th Machine Translation Summit: 24–30.

[66] (Ramisch, 2008) Ramisch, C. (2008). Développement d'un site Web iMAG générique et instanciation sur des sites iMAG concrets. Rapport de PFE Ensimag, 2008. 29 p.

[67] (Resnik, Olsen et al., 1999) Resnik, P., Olsen, M. B. and Diab, M. (1999). The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. Computers and the Humanities 33: 129–153.

[68] (Resnik and Smith, 2003) Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. Computational Linguistics 29: 349–380.

[69] (Salampasis and Hanbury, 2014) Salampasis, M. and Hanbury, A. (2014). PerFedPat: An integrated federated system for patent search. World Patent Information 38: 4–11.

[70] (San Vicente and Manterola, 2012) San Vicente, I. and Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. Proc. *Eighth edition of the Language Resources and Evaluation Conference (LREC)* Istanbul, Turkey. 1–6.

[71] (Schwenk, Barrault et al., 2015) Schwenk, H., Barrault, L., Blain, F., Bougares, F., Hazem, A. and Servan, C. (2015). MateCat - An Open Source CAT Tool with closely integrated User Specific Statistical Machine Translation. Proc. *TAO-CAT 2015*. Angers, France.

[72] (Senez, 1998) Senez, D. (1998). Post-editing service for machine translation users at the European Commission. Translating and the Computer 20.

[73] (Shi, Chen et al., 2013) Shi, X., Chen, Y. and Huang, X. (2013). Key Problems in Conversion from Simplified to Traditional Chinese Characters. Proc. *XIV Machine Translation Summit* Nice, France. p. 287-293.

[74] (Silkroad, 2006) Silkroad (2006) 基于短语的统计机器翻译系统"丝路" 1.0 版 (SilkRoad V1.0) 设计与使用说明.

[75] (Smith, Quirk et al., 2010) Smith, J. R., Quirk, C. and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. Proc. *Conf of the Association for Computational Linguistics*. Uppsala, Sweden. 403–411.

[76] (Snover, Madnani et al., 2009) Snover, M. G., Madnani, N., Dorr, B. and Schwartz, R. (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. Machine Translation 23: 117–127.

[77] (Steinberger, Pouliquen et al., 2006) Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. arXiv preprint cs/0609058.

[78] (Tian, Wong et al., 2014) Tian, L., Wong, D., Chao, L., Quaresma, P., Oliveira, F., Li, S., Wang, Y. and Lu, Y. (2014). UM-Corpus: a large English-Chinese parallel corpus for statistical machine translation. Proc. *LREC 2014*. ELRA Reykjavik, Iceland.

[79] (Tiedemann, 2012) Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. Proc. *eighth international conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey. 2214–2218.

[80] (Uchino, Shirai et al., 2001) Uchino, H., Shirai, S., Yokoo, A., Oyama, Y. and Furuse, O. (2001). ALTFLASH: A Japanese-English Machine Translation System for Market Flash Reports. EICE Transactions on Information and Systems, Pt. 2 (Japanese Edition): 1167-1174.

[81] (Utiyama and Isahara, 2007) Utiyama, M. and Isahara, H. (2007). A Japanese-English patent parallel corpus. Proceedings of MT summit XI: 475–482.

[82] (Varga, Halácsy et al., 2007) Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. and Trón, V. (2007). Parallel corpora for medium density languages. Amsterdam studies in the theory and history of linguistic science series 4. 247.

[83] (Vasconcellos and León, 1985) Vasconcellos, M. and León, M. (1985). SPANAM and ENGSPAN: machine translation at the Pan American Health Organization. Computational Linguistics 11(2-3): 122-136.

[84] (Vo Trung, 2004) Vo Trung, H. (2004). Méthodes et outils pour utilisateurs, développeurs et traducteurs de logiciels en contexte multilingue. Thèse de doctorat, Université Stendhal.

[85] (Vogel, Ney et al., 1996) Vogel, S., Ney, H. and Tillmann, C. (1996). HMM-based word alignment in statistical translation. Association for Computational Linguistics. 836–841.

[86] (Wagner and Fischer, 1974) Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. Journal of the ACM (JACM) 21(1): 168-173.

[87] (Wang, 2015) Wang, H. (2015). Évaluation comparative de la qualité d'usage de plusieurs systèmes de TA français-chinois en fonction de la tâche de post-édition. Mémoire de M2R.

[88] (Wang and Boitet, 2013) Wang, L. and Boitet, C. (2013). Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. Proc. *MT Summit XIV Workshop on Post-editing Technology and Practice*. Nice, France.

[89] (Xiao, Zhu et al., 2012) Xiao, T., Zhu, J., Zhang, H. and Li, Q. (2012). NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation. Proc. *ACL 2012 System Demonstrations*. Jeju Island, South Korea. 19–24.

[90] (Zhao and Vogel, 2002) Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. Proc. *IEEE International Conference on*. IEEE. pp. 745–748.

Table des définitions

Définition 1. Une *phrase* (*sentence* en anglais) est l'unité élémentaire d'un énoncé, formée de plusieurs mots ou groupes de mots, et qui présente un sens complet. (*TheFreeDictionary*).

Définition 2.* Un *syntagme* ou *groupe* (*phrase* en anglais) est un constituant possible d'une phrase. En général, un *titre* est un groupe nominal.

Définition 3. Un *segment* est l'unité de traduction de base des traducteurs humains. Il s'agit d'une phrase, d'un titre, ou d'un terme dans une nomenclature.

Définition 4.* Un *fragment* (*chunk* en anglais) est une partie d'un segment, qui peut être un *groupe syntaxique* (*phrase* en anglais) ou un simple n-gramme.

Définition 5.* Un segment peut contenir des *éléments non textuels*, ou *hors-texte,* comme des images, des formules, ou des balises, qui ont un rôle linguistique et une valeur non linguistique.

Définition 6. Un *segment monolingue* est un segment dont le contenu textuel est en une seule langue.

Définition 7.* Un *segment multilingue* est un segment dont le contenu textuel est dans plusieurs langues, chaque version étant considérée comme contenant exactement la même information, exprimée de façon correcte.

Définition 8. Un *segment monolingue multilingualisé (annoté)* est un objet contenant un segment « source » primaire, une ou plusieurs traductions (automatiques ou humaines ou automatiques post-éditées) pour une ou plusieurs langues, et des annotations, en général des objets, comme des arbres linguistiques, des graphes UNL, des résultats d'évaluation(s), et des références aux contributions ayant produit chaque objet non primaire.

Définition 9.* Un *segment multilingue multilingualisé (annoté)* est un objet contenant un segment multilingue, dans N langues « sources », et, dans M autres langues, une ou plusieurs traductions (automatiques ou humaines ou automatiques post-éditées), ainsi que des annotations, comme celles d'un segment monolingue multilingualisé et annoté.

Définition 10.* Le *chemin traductionnel* d'une annotation, en particulier d'une traduction ou d'une post-édition, est la suite des opérations l'ayant produite, ainsi que les intervenants humains impliqués, et les éventuels objets auxiliaires utilisés.

Définition 11*. Un *métasegment* est un segment comportant une ou plusieurs *variables*, éventuellement typées (nombre, date, balise faible…).

Définition 12*. Un *document* est un ensemble formé par un support et une information, celle-ci enregistrée de manière persistante. Nous nous intéressons aux documents textuels, qui contiennent des "segments" textuels.

Définition 13*. Un *métadocument* est un document pouvant contenir des métasegments.

Définition 14*. Un *pseudo-document* est défini par une référence (nom de fichier, url, uri) à un document qui peut varier au cours du temps.

Définition 15. *Contexte* : Le contexte *m-n* d'segment source par rapport à un document, ou plus généralement à une instance d'un pseudo-document, est défini par :.....
- la liste des *m* segments (de même langue) qui le précèdent.
- la liste des *n* segments (de même langue) qui le suivent.
- l'instanciation des variables, s'il y en a, dans ces *m+n+1* segments. Le contexte *m-n* d'un segment (cible) dans une version résultat de TA ou de PE, dans un segment monolingue, ou multilingue multilingualisé, est défini par
- la liste des *m* segments (de même langue et de même version) qui le précèdent.
- la liste des *n* segments (de même langue et de même version) qui le suivant.
- l'instanciation des variables, s'il y en a, dans ces *m+n+1* segments pour la même version.

Définition 16.* Un *corpus* est un ensemble usuellement fermé de documents homogènes du point de vue de leur structure, de leur(s) langue(s), de leur genre et de leur domaine.

Définition 17.* Un *corpus de traductions* est un corpus au sens précédent, contenant les traductions de tout ou partie de ses segments, dans une ou plusieurs langues.