Aides informatisées à la traduction collaborative bénévole sur le Web

Youcef Bey

UNIVERSITÉ JOSEPH FOURIER DE GRENOBLE

N° attribué par la bibliothèque

/__/__/__/__/__/__/__/__/__/

THESE

pour obtenir le grade de

DOCTEUR DE l'Université Joseph Fourier

Spécialité : "INFORMATIQUE : SYSTEMES ET COMMUNICATION"

préparée au GETA (CLIPS, IMAG), puis au GETALP (LIG),

ainsi qu'au NII (Japon) et à l'Université de Tokyo

dans le cadre de **l'École Doctorale**

"Mathématiques, Sciences et Technologie de l'Information"

présentée et soutenue publiquement
par

Youcef BEY

le 2 juin 2008

Aides informatisées à la traduction collaborative bénévole sur le Web

JURY

Marie-Christine Rousset	, Président
Eric Wehrli	, Rapporteur
Jacques Chauché	, Rapporteur
Emmanuel Planas	, Rapporteur
Françoise Létoublon	, Examinateur
Mathieu Mangeot-Nagata	, Examinateur
Kyo Kageura	, Co-directeur de thèse
Christian Boitet	, Directeur de thèse

Remerciements

Je tiens à remercier en tout premier lieu le Professeur Christian Boitet, mon directeur de thèse, qui m'a conseillé scientifiquement et aidé pratiquement tout au long de cette thèse. Il m'a fait apprendre les fondements de la recherche et m'a donné la volonté d'avancer dans les moments les plus difficiles. Je lui suis très reconnaissant du savoir qu'il m'a transmis que je considère comme un trésor exploitable tout au long de ma future carrière de la recherche, en particulier, dans le domaine de la traduction automatique.

Je tiens à remercier le Prof. Kyo Kageura pour m'avoir accueilli au NII (National Institute of Informatics) en stage avec une bourse attribuée par l'association japonaise JASSO (Japan Student Services Organization), ainsi que pour son encadrement et soutien durant mon séjour à l'Université de Tokyo à la « Graduate School of Education » dans le cadre du programme CDFJ (Collège Doctoral Franco-Japonais). Son expérience dans le domaine de la traduction des droits de l'homme a été pour moi le « pilier » à partir duquel j'ai identifié les problèmes réels que rencontrent les traducteurs bénévoles.

Je voudrais adresser un remerciement particulier au Professeur Akiko Aizawa (NII) pour son soutien scientifique et financier durant mon séjour au NII, et pour m'avoir accepté en deuxième stage de 6 mois dans le cadre du programme de stages internationaux offerts par le NII aux étudiants étrangers.

J'ai eu également le plaisir de collaborer au projet DEMGOL avec deux équipes, une en Italie et l'autre en France. En France, je tiens à remercier Mme le Professeur Françoise Letoublon et le docteur Valérie Bellynck pour leur aide précieuse durant mes expériences. En Italie, je tiens à remercier aussi dans le même groupe le docteur Francesca Marzari et Giovanni Zorzetti, grâce à qui nous avons mené ensemble des expériences avec succès sur notre environnement.

Je remercie mes rapporteurs, les Professeurs Jacques Chauché et Eric Wehrli, ainsi que le Dr. Emmanuel Planas, qui ont accepté d'être rapporteurs de ma thèse à une période très

chargée. Je voudrais adresser un remerciement particulier au professeur Marie-Christine Rousset pour avoir accepté de présider mon jury.

Table des matières

Table des figures

Table des tables

Introduction générale

Après un DEA consacré à la génération pseudo-aléatoire de très grands corpus plus au ou moins richement étiquetés, dans le but de développer des méthodes « empiriques » (par apprentissage) d'identification de relations sémantiques, j'ai recherché un sujet aussi intéressant mais offrant aussi une perspective d'application rapide et utile.

Depuis près de 10 ans, le GETA[1] (GETALP depuis 2007) avait proposé l'idée de mettre à disposition des traducteurs « occasionnels » (on pensait surtout à la communauté scientifique francophone), via la Toile, des outils et ressources jusque là réservés aux traducteurs professionnels. Une communauté aurait partagé une mémoire de traductions et des dictionnaires construits par elle-même et pour elle-même, en utilisant (chez soi, sur PC ou Mac) un outil gratuit dérivé d'un outil professionnel tel que TM2 (IBM), Trados, Déjà Vu (Atril), Eurolang Optimizer (Site/Eurolang), XMS (Xerox), ou Similis (Lingua & Machina).

Parlant de cette idée avec le Professeur Kyo Kageura durant l'été 2004, il est apparu qu'il venait de lancer un projet de ce type, à titre de chercheur en informatique et de traducteur bénévole de textes sur les droits de l'homme. De notre côté, nous avons proposé deux idées nouvelles, rendues possibles par le développement du Web et des services de traduction automatique (TA) gratuits :

1. Pourquoi ne pas proposer aux traducteurs des « prétraductions » par TA, à côté des classiques « suggestions » des mémoires de traduction (MT) ?

2. Pourquoi ne pas proposer aux traducteurs de traduire « en ligne », comme avec translationwiki.net (Translationwiki, 2006), mais avec en plus des « aides traductionnelles », et pas seulement « hors ligne », avec un simple partage de ressources via un serveur ?

La première idée (utilisation de la TA) ne pose que des problèmes techniques simples, tant qu'on ne cherche pas à développer un système pour une communauté donnée.

[1] En 1995, Ch. Boitet, B. Oudet et Y. Chiaramella proposèrent sur ce concept le projet « Montaigne » post-Eurolang, qui ne put malheureusement pas être financé par le MRI de l'époque.

Par contre, la deuxième idée suppose un changement du mode de travail des traducteurs bénévoles, et cela est apparu comme un problème majeur. C'est pourquoi le Professeur Kageura nous a conseillé de commencer par participer à son projet QRLex, de façon à bien comprendre les besoins (actuels et potentiels) des traducteurs bénévoles, et à résoudre les problèmes indépendants du mode de travail (hors ligne ou en ligne), à savoir la réutilisation et le partage de ressources traductionnelles (dictionnaires, lexiques, mémoires de traduction, documents ou fragments déjà traduits), et la communication entre traducteurs. Cette première étape de notre recherche a duré un peu plus d'un an et fait l'objet de la première partie de cette thèse.

Dans un second temps, commençant en parallèle avec la fin du travail précédent, nous nous sommes concentré sur les problèmes posés par la traduction bénévole « en ligne », avec 3 axes principaux :

1. Comment concevoir et implémenter un « éditeur traductionnel » utilisable à travers tout navigateur Web, en mode Wiki, i.e. en permettant à plusieurs contributeurs de participer en même temps à la traduction d'un même document, et ce de façon conviviale ?

2. Comment y intégrer des aides traductionnelles « proactives », i.e. comment proposer des équivalents lexicaux et des suggestions de traduction pour des « segments » complets au moment où on veut les traduire ?

3. Comment offrir un « support » à la communauté et au travail coopératif entre traducteurs d'une même communauté en utilisant au mieux les possibilités liées à la situation « en ligne », sans toutefois « forcer » des contacts immédiats de façon intrusive ?

Ce travail fait l'objet de notre deuxième partie, dans laquelle nous décrivons la solution des problèmes en question, telle qu'implémentée dans notre plate-forme Web BEYTrans (Better Environment for Your Translation). Nous présentons en détail l'expérimentation faite sur le projet DEMGOL en italien-français (traduction du Dictionnaire Étymologique de la Mythologie Grecque On Line), et un second protocole d'évaluation mesurant mieux l'apport spécifique de l'aspet communautaire et Wiki, et dont la mise en œuvre va commencer sur la localisation de Tikiwiki.

Enfin, nous nous sommes demandé si BEYTrans ne pourrait pas être utilisé de façon « réflexive », pour améliorer sa ou ses mémoires de traductions, en les assimilant à de très

gros documents. Cela nous a conduit à l'idée de l'étendre (du point de vue du « passage à l'échelle ») aux très gros corpus parallèles multilingues utilisés pour la construction et l'évaluation de systèmes de TAO statistiques (ou plutôt probabilistes). Nous décrivons dans la troisième partie la solution des problèmes associés et plusieurs expérimentations sur des corpus utilisés lors de diverses campagnes d'évaluation de systèmes de TA.

Partie I

Ressources linguistiques multilingues destinées à la traduction collaborative bénévole en ligne

Introduction

Nous présentons dans cette partie l'état des pratiques en traduction bénévole et professionnelle. Nous comparons les deux et montrons l'intérêt actuel de la traduction faite par des bénévoles en ligne. À travers cette étude, les besoins des traducteurs bénévoles seront recensés, en se basant sur les expériences de communautés actives sur le Web, qui se concentrent sur la dissémination des connaissances multilingues.

Ensuite, nous décrivons notre expérience dans le cadre du projet QRLex (un système japonais destiné aux communautés de traducteurs bénévoles sur les droits de l'homme). Lors de notre premier séjour au Japon, nous avons participé au développement de l'outil QRselect visant à recycler des traductions sur le Web. Nous présenterons ce système et montrerons comment nous l'avons adapté au recyclage de documents variés en plusieurs langues.

Ensuite, nous avons défini et implémenté une méthodologie simple et efficace pour importer les ressources lexicales libres disponibles, pour aider les communautés de traducteurs. Nous avons proposé et utilisé une structure XML *ad hoc* qui nous a aidé à structurer des données hétérogènes en provenance de différentes sources (banques terminologiques, dictionnaires, glossaires, etc.).

Enfin, nous avons participé à l'intégration de nos outils dans l'environnement QRlex que nous présenterons de façon détaillée. La conception et l'architecture ainsi que les modules et les flux de données seront expliqués, et illustrés par application de cet environnement aux communautés de traducteurs Japonais faisant la traduction de documents concernant les droits de l'homme.

Le composant le plus novateur est l'éditeur, auquel nous donnons une importance particulière durant le processus de développement. Il sera présenté avec les fonctionnalités linguistiques qui y sont intégrées pour l'aide linguistique durant le processus de traduction.

Chapitre 1

Situation et problématique de la traduction bénévole

Introduction

La traduction a été pour longtemps réservée aux professionnels. Il y a de plus en plus d'outils divers d'aide à la traduction développés à leur intention. On trouve des banques terminologiques riches, des dictionnaires sous différentes formes, et des mémoires de traductions. La plupart de ces outils sont payants et l'investissement pour leur développement est très coûteux. Les outils libres d'aide à la traduction sont rares, et, quand ils existent, n'offrent que des fonctionnalités classiques, éventuellement un éditeur basique, et aucune aide dictionnairique ou traductionnelle.

Cependant, un mouvement très actif de traducteurs bénévoles dissémine des connaissances en une centaine de langues gratuitement sur la Toile, et ne cesse de se développer. Les sites *Wiki*[2] en sont des exemples. Le succès de Wikipedia[3] et de son compagnon linguistique Wiktionary est un travail collaboratif bénévole remarquable. On compte des millions de pages très riches et diverses, de bonne qualité de traduction. S'ajoutent à cela des communautés plus ou moins petites comportant quelques centaines de traducteurs bénévoles, telles que la communauté des standards Web W3C[4], avec plus de 300 traducteurs traduisant en 41 langues.

Malheureusement, ces communautés ne bénéficient pas d'outils d'aide à la traduction, ni gratuits ni payants. Elles ne disposent en fait d'aucune ressource linguistique accompagnée d'un minimum de fonctionnalités (ne serait-ce qu'un dictionnaire avec une interface adéquate) ! Ces communautés ont donc un besoin urgent et immédiat d'outils en ligne leur permettant de traduire en groupe et efficacement avec une productivité accrue.

Nous présentons dans ce chapitre les pratiques des traducteurs professionnels et les comparons à celles des bénévoles. Nous illustrons notre étude des communautés de bénévoles par des cas de figure extraits de situations réelles en montrant les problèmes à résoudre.

[2] http://fr.wikipedia.org/wiki/Wiki
[3] http://www.wikipedia.org/
[4] http://www.w3.org/

1.1 Traduction professionnelle commerciale et traduction « libre »

Dans cette section, nous présentons les pratiques de la traduction dans les services professionnels ainsi que celles des bénévoles. La taille et la complexité des tâches de traduction et de localisation sont présentées, pour montrer la position de nos travaux de recherche, et aussi l'intérêt général de la construction d'un environnement collaboratif générique orienté vers l'aide aux communautés de traducteurs bénévoles.

1.1.1 La traduction professionnelle

1.1.1 Contextes de travail

1.1.1.a Bureaux de traduction

Les compagnies les plus concernées par la traduction sont celles qui ont un grand marché international. La croissance des échanges culturels, la mondialisation et l'augmentation de la communication à travers Internet les ont obligées à monter des groupes de traduction, où chacun dispose d'outils divers et de nombreuses ressources linguistiques. Le plus souvent, un petit groupe interne organise le travail de traduction, et le sous-traite à des traducteurs indépendants, auxquels il envoie des « kits » (document à traduire, avec mémoire de traduction et lexique terminologique bilingue adaptés), à utiliser avec un des outils du marché (Trados[5], Déjà vu[6], Similis[7], TM2[8], etc.)

Les services de traduction d'IBM en sont un exemple. Selon Christophe Chenon (responsable des outils pour la francisation chez IBM-France), la traduction à IBM se caractérise en particulier par son volume très important : environ 20 millions de mots anglais sont traduits chaque année en 25 langues. Les textes traduits chez IBM sont essentiellement techniques ; ils concernent principalement les produits de l'entreprise, et très secondairement d'entreprises partenaires ou clientes. Environ 50% du volume traduit consiste en des documents techniques (entre autres, les manuels en PDF et HTML), 34% constituent l'aide en ligne, et les 16% restants regroupent les messages des programmes et les éléments d'interface. On trouve aussi d'autres types de textes, par exemple juridiques et promotionnels, en très petite quantité. La « localisation » de logiciels représente donc environ la moitié de l'effort de traduction d'IBM.

[5] http://www.trados.com/en/
[6] http://www.atril.com/
[7] http://www.similis.org/linguaetmachina.www/index.php
[8] http://www.lim.nl/monitor/ibm-tm2-1.html

La gestion multilingue de la terminologie est assurée par de nombreuses équipes spécialisées, coopérant à travers le monde. En France, la base terminologique bilingue contient 35 000 entrées. Depuis l'été 2004, la base de données terminologique multilingue compte environs 20 000 termes anglais. La proportion traduite est variable en fonction des langues et des acteurs. Actuellement, la langue la mieux représentée dans cette base après l'anglais est le français, avec la moitié des entrées traduites, suivie du hongrois, de l'allemand, de l'espagnol, etc. (Chenon, 2005).

En ce qui concerne les services de traduction étatiques, le bon exemple est celui de la DGT[9] (direction générale de la traduction) de la Commission Européenne. Les bureaux de traduction occupent à plein temps 1 750 traducteurs ; ils sont aidés dans leur travail par 600 personnes affectées à des tâches de gestion et de secrétariat et à la communication. Par exemple, en 2006, la DGT a traduit 1 541 518 pages ; 72 % des textes originaux (dont ceux qui provenaient de l'extérieur) étaient rédigés en anglais, 14% en français, 2,7 % en allemand et 10,8 % dans l'une des vingt autres langues. L'anglais et le français sont prédominants parce que ce sont les principales langues de rédaction de la commission.

De son côté, la DGT cherche depuis des années, et avec tous les moyens, à produire des traductions de qualité, tout en augmentant la productivité, la cohérence et l'efficacité. L'outil le plus important de cette organisation est la base terminologique Eurodicautom (devenue IATE[10]), une des plus volumineuses au monde. Cette base a été créée en 1975 et mise en ligne en 1980 pour les institutions de la Commission Européenne. Et comme la Communauté croit, elle a été étendue à 6, puis à 9 et à 11 langues. Aujourd'hui, la base terminologique s'est élargie à 24 langues (cf. *Glossaires et dictionnaires*, p. 26).

L'organisation vise l'intégration de cette base dans les environnements de travail des traducteurs, pour augmenter la productivité, accroître la qualité, et faciliter l'harmonisation de la terminologie utilisée dans les textes.

Le second outil le plus utilisé à la DGT est la traduction automatique. Les premières licences de Systran (anglais-français et français-anglais) ont été achetées en 1976. De nombreux couples de langues ont été ajoutés par Systran, et la DGT a construit des dictionnaires spécialisés importants. La plupart du temps, les traductions automatiques ne sont pas révisées, mais utilisées telles quelles, en parallèle avec l'original, pour la lecture.

[9] http://ec.europa.eu/dgs/translation/index_fr.htm
[10] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load

Certains des outils et des ressources employés par les organismes publics sont consultables et exploitables en ligne, mais ne sont pas libres et sont régis par la législation sur les droits d'auteur. S'y ajoute l'impossibilité des les télécharger, même à des fins de recherche. Il est donc impossible de les exploiter en TAL. D'autre part, il est pour l'instant impossible de contribuer à l'évolution des ressources par des mises à jour ou des améliorations d'aucune sorte.

1.1.1.b Traducteurs indépendants

La catégorie des traducteurs *indépendants* regroupe les services de traduction proposés par les professionnels individuels ou organisés en petits groupes. Ils n'appartiennent à aucune organisation et proposent des services traductionnels payants. Les organisations de traduction leur affectent des tâches qui dépendent de chaque projet de traduction (par exemple, le service de traduction des manuels HP à Grenoble). Le nombre de traducteurs de cette catégorie varie selon la grandeur de la tâche de traduction et les projets de traduction conduits.

Certains groupes proposent même la gestion de ressources linguistiques en ligne. Par exemple, le groupe *ProZ*, avec plus 1 700 traducteurs dispersés sur 100 pays, propose des services de traduction avec une gestion centralisée des ressources linguistiques propres à chaque organisation (ProZ, 2008).

La majorité de ces groupes sont formés de professionnels, et leurs ressources linguistiques sont souvent réservées à un usage interne et propre à chaque groupe.

1.1.1.c Traducteurs occasionnels

Les personnes ayant des compétences en traduction et participant occasionnellement à des tâches de traduction rentrent dans cette catégorie. Elle couvre, entre autres, les traducteurs bénévoles qui produisent des traductions dans le cadre d'organisations à but non lucratif telles que le W3C (consortium des standards Web)[11], Traduc (traduction de la documentation technique « How to »)[12], etc. D'autres traducteurs font de la traduction professionnelle à la maison, comme l'encourage la Commission des Communautés Européennes par le recrutement à domicile de traducteurs occasionnels.

[11] http://www.w3.org/Consortium/Translation/
[12] http://www.traduc.org/

1.1.2 Traduction professionnelle en réseau

1.1.2.a Sous-traitance à distance

La sous-traitance à distance est pratiquée par tous les « donneurs d'ordres ». Par exemple, les langues dans lesquelles sont traduits les documents à IBM sont : allemand, arabe, bulgare, catalan, chinois, coréen, danois, espagnol, français, grec, hébreu, hongrois, italien, japonais, néerlandais, norvégien, finnois, polonais, portugais/brésilien, roumain, russe, serbo-croate, suédois, thaï, tchèque, turc. Chaque centre IBM dans un pays donné s'occupe de la traduction dans sa propre langue. Selon Christophe Chenon (Chenon, 2005), les traductions s'effectuent en même temps dans le monde entier (dans tous les centres de traduction) à partir du moment où est disponible le texte en anglais, le plus souvent produit aux États-Unis. Une conséquence est que les échanges d'information, les questions et les réponses, ne sont pas seulement axés vers l'anglais : les équipes travaillant sur les mêmes projets au même moment communiquent entre elles depuis les différents centres de traduction.

Un autre exemple est le cas de la branche de la compagnie HP (Hewlett-Packard), qui est installée sur le pôle industriel d'Eybens (près de Grenoble) et qui traduit en 40 langues. Pourtant, la traduction n'occupe aucun traducteur de la société (pas plus qu'à IBM) ; toute la documentation est envoyée à des groupes de traduction professionnels et spécialisés dans la traduction, chacun dans une langue.

Les besoins en sous-traitance reflètent les besoins constants et croissants en traduction et localisation. Il faut enfin remarquer que les grandes compagnies ne sont pas capables de traduire dans toutes les langues dans lesquelles il faudrait le faire, et font sur ce point nettement moins bien que le libre. La compagnie Microsoft traduit dans 45 langues, HP (Hewlett-Packard) dans 40, Adobe dans 32 et IBM dans 25 ou 26, alors que le projet Mozilla est traduit par des bénévoles en 70 langues.

Malheureusement, les ressources linguistiques ainsi que les environnements exploités par les traducteurs en sous-traitance sont aussi fermées (« propriétaires »). Il est impossible de savoir quel volume de données linguistiques est exploité durant la traduction, et *a fortiori* d'y avoir accès.

1.1.2.b Traduction en local et ressources partagées

La raison d'utiliser des ressources partagées est de permettre leur utilisation par plusieurs traducteurs. La centralisation de ces ressources assure un haut niveau de qualité et une cohérence accrue lors d'une tâche de traduction. Ces ressources centralisées deviennent alors

une référence. Les raisons pour lesquelles le partage des ressources est nécessaire sont les suivantes.

- Diminuer le coût. Les ressources partagées permettent d'orienter les traducteurs et de leur éviter des traductions multiples qui sont souvent la source d'un travail énorme et en fait inutile pour les réviseurs.
- Permettre aux traducteurs spécialistes de domaines différents (logiciels, sites Web, commerce, etc.) de partager un même contenu traductionnel.
- Répartir les tâches de traduction sur plusieurs centres géographiquement éloignés.
- Avoir le sentiment d'être en phase avec une base centrale de ressources.

Ces solutions sont déjà proposées dans certains outils commerciaux d'aide à la traduction tels que Trados TM Server qui permet le partage des mémoires de traductions.

1.1.3 Usage d'outils d'aide à la traduction

1.1.3.a Glossaires et dictionnaires

Les traducteurs ont souvent besoin de ressources linguistiques spécialisées leur permettant de trouver des traductions de bonne qualité, en particulier pour les termes qui apparaissent dans les documents à traduire. C'est ainsi que, généralement, la plupart des grands services de traduction mettent à la disposition de leurs traducteurs des ressources consultables hors ligne ou en ligne. En prenant à nouveau l'exemple de la traduction à la Commission Européenne, on constate que c'est l'une des raisons pour lesquelles l'UE a proposé de construire la grande base terminologique « IATE » (InterActive Terminology for Europe) qui couvre 23 langues (en 2007, la base terminologique « Eurodicautom » a été remplacée par cette grande base) (EU, 2008).

L'IATE combine les bases de données terminologiques de différentes institutions européennes dans une seule base de données. Elle contient 8,7 millions de termes et couvre les 23 langues officielles de l'UE[13]. Elle est utilisée par les services de traduction l'UE et joue déjà

[13] L'IATE ne contient pas seulement 8,7 millions de termes, mais aussi 500 000 abréviations et 100 000 phrases couvrant l'ensemble des 23 langues officielles de l'UE. L'ensemble des coûts de développement de la base terminologique de 1999 à 2003 est estimé à 1,41 millions d'euros. Les coûts annuels de maintenance en 2007 sont estimés à 627 000 euros. Ces coûts sont couverts par les budgets de toutes les institutions de l'UE.

un rôle important en assurant la qualité des communications écrites entre ses institutions. De plus, elle garantit la cohérence et la fiabilité de la terminologie, ce qui est indispensable pour produire des textes clairs et les moins ambigus possibles. Cela permet aux institutions, en particulier la DGT, d'assurer à la fois la validité et la transparence du processus législatif, et la communication efficace avec les citoyens de l'UE.

L'utilisation de la base est ouverte au grand public. L'utilité potentielle de l'ouverture de l'IATE pour le grand public de l'UE a été identifiée dès le départ. En permettant l'accès libre à l'IATE pour tous les citoyens de l'UE, les institutions de l'UE offrent ainsi une ressource unique et incomparable à la disposition de quiconque souhaitant utiliser son contenu (non seulement pour les professionnels de langues en dehors des institutions de l'UE, mais aussi pour les traducteurs indépendants ou bénévoles, les chercheurs et les étudiants en langues).

L'IATE est exploitable via une interface interactive au moyen de laquelle tous les traducteurs des institutions de l'UE peuvent ajouter et mettre à jour des informations (Figure 1). Cette interface est conviviale et permet aussi d'effectuer des recherches de termes spécifiques dans une langue source et de trouver leurs équivalents dans les langues cibles sélectionnées.

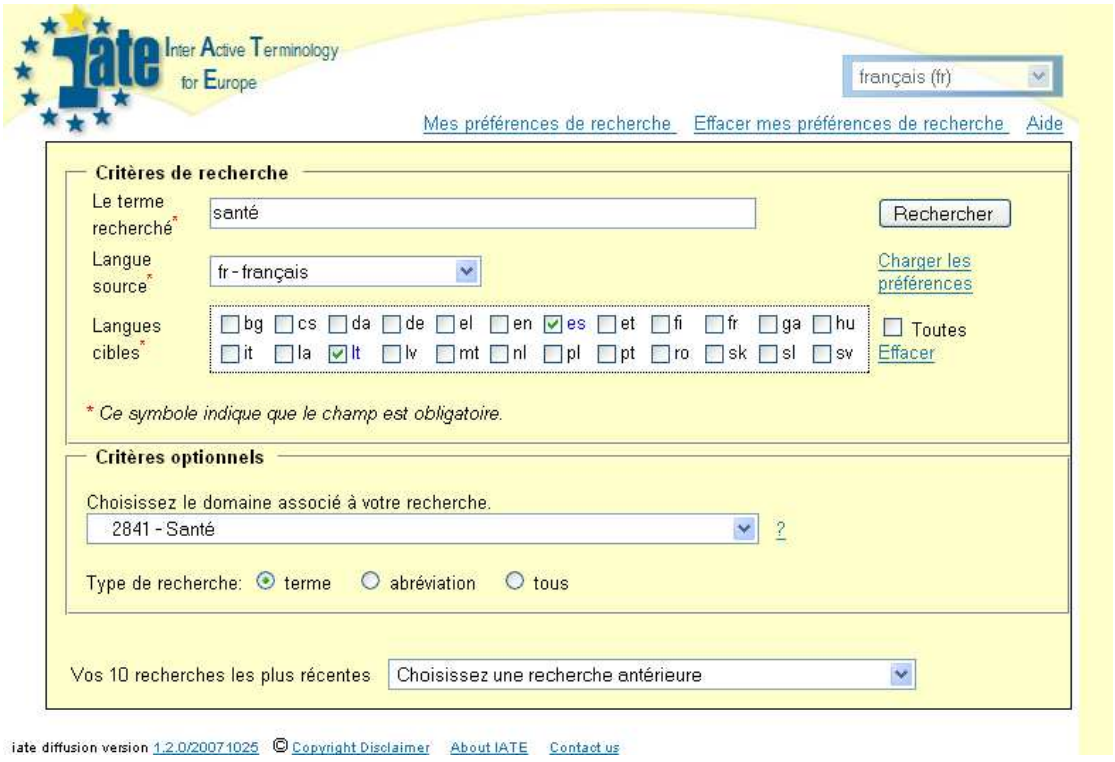


Figure 1 : Interface en ligne de la base terminologique « IATE »

Afin d'assurer une bonne qualité des contributions individuelles, un changement dans la base de données est suivi par le lancement d'un **cycle de validation**, dans lequel les terminologues des institutions de l'UE interviennent pour valider les modifications.

Cependant, les bases terminologiques, les glossaires et les dictionnaires ne sont pas les seules ressources demandées par les traducteurs. Les mémoires de traductions sont aussi des ressources très importantes et sont utilisées par la plupart des services de traduction.

1.1.3.b Mémoires de traduction

Les mémoires de traductions ont été introduites après qu'on a constaté l'impossibilité de produire des traductions automatiques à 100% correctes sans limitation stricte à un « sous-langage » restreint. Parmi les mémoires de traductions commerciales les plus répandues, on trouve : Translation Manager™ d'IBM, Déjà vu™ d'Atril, Trados™ de Trados, Similis™ de Lingua et Machina, etc.

Le fonctionnement d'une MT est basé sur l'utilisation des traductions déjà existantes. Les segments traduits sont stockés au moment de la traduction et sont retrouvés et proposés aux traducteurs sous forme de suggestions. Les suggestions se font par des calculs de similarité entre les segments déjà stockés et le segment source en cours de traduction.

Des entreprises comme IBM ont une mémoire de traduction très riche et volumineuse résultant des traductions de 20 millions de mots par an en 25 langues. Malheureusement, il est impossible d'y avoir librement accès.

Enfin, les ressources qui existent, telles que EuroParl et EURAMIS, sont importantes en quantité mais restent pauvres du point de vue richesse et spécialisation.

1.1.3.c Traduction automatique

La traduction automatique (TA) est le premier problème non numérique auquel l'informatique naissante s'est attaquée, dès 1950. Dans toute sa généralité, le problème consiste à faire traduire par une machine des documents écrits (puis des dialogues et des discours oraux) d'une langue naturelle « source » dans une langue naturelle « cible ». En faisant varier le degré d'automatisation de l'opération traductionnelle proprement dite, on arrive à plusieurs intermédiaires.

À un extrême, on trouve la traduction humaine assistée par la machine (THAM, dans laquelle c'est une personne bilingue qui traduit, éventuellement à partir de « prétraductions » produites automatiquement), à et l'autre extrême, la traduction entièrement automatique de

haute qualité (TAHQ) ou *Fully Automatic Human Quality Machine Translation* (FAHQMT), telle que produite par le système canadien MÉTÉO sur le sous-langage des bulletins météorologiques.

Entre les deux, on a différents types de traduction automatique aidée par l'homme (TAAH), dans laquelle l'opération de traduction proprement dite est effectuée par la machine : TAFD (TA augmentée d'une désambiguïsation interactive en langue source, et éventuellement d'une désambiguïsation interactive en langue cible, par des personnes monolingues) et TA-R (traduction automatique avec révision, dite aussi post-édition, effectuée selon la qualité des traductions automatiques et du résultat souhaité par des personnes connaissant les deux langues, ou seulement la langue cible). L'ensemble de ces approches a été regroupé sous le nom de « TAO » en 1981, à l'occasion du lancement par l'ADI (agence de l'informatique) du « projet national de TAO » (ou PN-TAO). On utilise aujourd'hui presque indifféremment « TA » ou « TAO ».

L'intérêt de la TA sera abordé dans les paragraphes suivants. Nous montrerons aussi comment exploiter les services de TA gratuits sur le Web.

1.1.3.c.i Bureaux : contextes restreints et TA spécialisée

Depuis longtemps (1976 avec le système TAUM-Météo), les chercheurs en TA ont constaté qu'il est possible de produire des traductions de qualité en construisant des systèmes de TA spécialisés à des sous-langages restreints, avec des vocabulaires limités et un nombre réduit de règles de grammaires. Par exemple, le système MÉTÉO (successeur opérationnel de TAUM-Météo) produit des traductions de très bonne qualité des bulletins météorologiques[14].

Il sert à produire des traductions de qualité professionnelle avec une intervention humaine minimale (3 opérations de l'éditeur pour 100 mots traduits).

Selon la société « John Chandioux experts-conseils », l'U.S. « National Weather Service » a adopté le système MÉTÉO pour assurer la disponibilité en français de toutes les prévisions, les avertissements, les veilles et les avis météorologiques pendant les Jeux Olympiques d'Atlanta en 1996. Pour cet événement, le système a été complètement redéveloppé pour Windows/OS2 et pour le sous-language américain des bulletins météo. Il a assuré la traduction de 305 000 mots en seize jours avec un taux d'exactitude de plus de 93 %, soit l'équivalent de *sept mois et demi* de travail par un traducteur humain. Les sorties ont été

[14] MÉTÉO fut mis en servir opérationnel en mai 1977.

révisées par trois météorologues Canadiens[15]. De plus, selon la même source, un traducteur met en moyenne actuellement *3 minutes et 8 secondes* pour réviser la traduction d'un bulletin produite par MÉTÉO. Dans le cas d'une traduction purement humaine, cette durée serait comprise entre 30 et 40 minutes (Tremblay, 2000), 10 fois plus.

Bien que les traductions produites soient de très bonne qualité, le système MÉTÉO ne peut être appliqué à la traduction de documents dans des domaines différents. Dans ce cas, on fait appel à d'autres systèmes tels que Systran et Reverso pour produire des prétraductions qui seront ensuite post-éditées (on peut y gagner en moyenne 2/3 du temps) (MTPostEditing, 2008).

1.1.3.c.ii Les STA en ligne : services gratuits de traduction automatique

Quelques fournisseurs de traducteurs automatiques offrent des services gratuits pour la traduction sur le Web. La TA Web peut être donc appelée pour traduire des textes limités ou des documents Web complets (1 document à la fois). Bien qu'il existe plusieurs STA Web, nous ne présentons dans les paragraphes suivants que trois systèmes : Systran Web, Reverso et WebSphere Translation Server d'IBM.

I.1.1.1.3.c.ii.1 Systran Web

Systran est un fournisseur de traducteurs automatiques très connu et sa technologie alimente des solutions de traduction pour Internet, PC et infrastructures de réseau avec 52 couples de langues dans 20 domaines spécialisés[16]. Ces systèmes de traduction automatique (STA) sont utilisés depuis longtemps par de grands comptes, comme la Communauté Européenne ou l'US NAIC (National Air Intelligence Center) et, plus récemment, par des portails Internet avec les trafics les plus importants, comme Altavista, AOL, Google (jusqu'au 1/11/2007), etc.

Le site Web de Systran propose un service de traduction gratuite pour traduire des textes de petite taille et des pages Web par URL[17].

I.1.1.1.3.c.ii.2 Reverso Web

Le produit initial s'appelait Stylus (produit par la société ProMT de Saint-Pétersbourg). Il a été diffusé par la société Softissimo pour la traduction automatique en ligne de textes ou

[15] http://www.chandioux.com/press_meteo20.html
[16] Les 52 couples n'ont été mis en service sur les PC que depuis 09/2007, mais à la même date, on ne trouve sur Internet que 32 couples.
[17] http://www.systran.fr/.

pages Web[18] sous le nom de Reverso de 1994 à 2006. Depuis 2007, Softissimo a un autre fournisseur de STA (également une société russe). Reverso est utilisable sur réseau (Internet, intranet), ou sur PC comme une application autonome.

I.1.1.1.3.c.ii.3 WebSphere Translation Server

Le système WebSphere Translation Server est un traducteur automatique (dérivé de LMT qui a été commercialisé sur PC sous le nom de "PT", ou Personal Translator) développé par IBM. Il peut fonctionner sur plusieurs plates-formes comme NT, AIX, Solaris et Linux. La version en ligne ne permet de traduire que des pages Web. Les internautes peuvent traduire des pages Web en 12 paires de langues.

La TA gratuite en ligne illustrée par ces trois exemples ne peut être utilisée telle quelle pour aider les traducteurs, car elle est limitée à un nombre restreint de mots et de documents par session. Cette situation a suscité la combinaison de traductions de morceaux traduits par la TA Web pour produire une traduction complète d'un document. Vo-Trung Hung (2004) a montré à l'aide de son système TRADOH (développé au GETA à Grenoble) comment il est possible de traduire des documents volumineux par la TA en regroupant plusieurs traductions produites par les STA Web (Vo-Trung, 2004).

Nous utiliserons cette technique pour traduire (ou plutôt « prétraduire ») de la même façon les documents multilingues et les corpus parallèles.

1.1.2 La traduction bénévole avec partage de ressources sur le Web

On observe récemment la création en continu de communautés de traducteurs bénévoles. Ces traducteurs se regroupent, partagent des documents, des connaissances, et communiquent via Internet pour traduire des projets. Les traductions produites sur des sites Web spécifiques de bénévoles sont de plus ou moins de bonne qualité (Figure 20).

Notre étude des communautés existantes nous a montré que la pratique de la traduction par ces communautés se manifeste dans trois situations différentes :

1. localisation de logiciels libres (par exemple, Mozilla, W3C, Arabeyes, etc.).
2. traduction de documents culturels et humanistes (par exemple, relatifs aux, comme droits de l'homme, tel que TeaNotWar, PaxHumana, etc.).
3. traduction occasionnelle en ligne (par exemple, translationwiki.net).

[18] http://www.reverso.net/text_translation.asp.

Nous présentons dans les sections suivantes ces situations en détail en nous focalisant sur les pratiques des traducteurs bénévoles, relativement aux ressources linguistiques d'aide à la traduction. Nous identifions les problèmes que rencontrent les traducteurs et leurs besoins en outils d'aide plus avancés.

1.2.1 Traducteurs de logiciels libres : cas de documents techniques

Les communautés de traducteurs bénévoles s'intéressant à la traduction de documents techniques, ou à la localisation de logiciels libres, sont nombreuses.

La communauté de traduction Arabeyes en est un exemple. Elle s'intéresse à la traduction en arabe des documents techniques et à la localisation des logiciels libres de la suite Mozilla (Thunderbird, navigateur Firefox, navigateur Mozilla, et leurs add-on). Elle est constituée d'entre 10 et 20 traducteurs, traduisant bénévolement (ArabicMozilla, 2007). Chacun des traducteurs vérifie constamment si de nouvelles versions anglaises des logiciels sont disponibles (la version source étant en anglais) ; ensuite ils se distribuent la traduction et la localisation en arabe.

Le déroulement de ce processus nécessite beaucoup d'efforts, en particulier du côté technique, car les traducteurs ont besoin de scripts et d'outils pour automatiser certaines tâches. Par exemple, le système d'écriture de la langue arabe est RTL (Right To Left), ce qui oblige les traducteurs à régler les logiciels localisés avec une nouvelle compilation du code source. Ce changement du sens d'écriture nécessite des connaissances techniques. En fait, la plupart des traducteurs-localiseurs de cette communauté sont des développeurs et ont des connaissances assez avancées en informatique et en gestion de projets, comme le projet de traduction dans lequel ils sont impliqués.

Le directeur de ce même projet estime que la traduction de 100 chaînes peut être faite en une heure de travail. Par exemple, la version 2.18 – la plus récente – de *Gnome* contient environ 6 000 nouvelles chaînes par rapport à sa version précédente. Ces chaînes ont été traduites en 60 heures. La traduction est suivie – comme n'importe quelle traduction professionnelle – par un processus de révision qui est 100 % humain. Donc la révision de qualité est un processus obligatoire. Dans la même version du système « Gnome », ce processus aurait été réalisé par 6 traducteurs sur une durée d'un mois. Cette durée inclut aussi la révision avec les nouvelles chaînes des anciennes chaînes – le nombre total des chaînes est estimé à 35 000.

Les traducteurs se réfèrent à des ressources internes souvent spécialisées (par exemple, les versions sur papier et électroniques des dictionnaires) et aux ressources externes souvent

disponibles en libre service sur la Toile (construites par des traducteurs). Certaines communautés ayant des compétences en informatique proposent des ressources en ligne, mais cela n'est pas le cas pour toutes les autres communautés.

Par exemple, le site d'Arabeyes propose aux traducteurs bénévoles un dictionnaire technique collaboratif fonctionnant en mode Wiki. Le point remarquable dans cette ressource est qu'elle est riche et englobe presque tous les termes techniques des logiciels à localiser. La dernière version (en 2007) contenait environ 18 000 entrées construites à 100% par des traducteurs bénévoles. Cette ressource est considérée comme la plus grande en taille par rapport à d'autres communautés de langue du même groupe. S'ajoute à cela la consultation par les traducteurs des versions commerciales papier ou en ligne de dictionnaires, tels que le dictionnaire anglais-arabe « Al-warid », qui est largement considéré comme faisant autorité en traduction anglais-arabe.

Un autre cas est celui du consortium W3C, qui a ouvert un service de traduction en ligne pour les traducteurs bénévoles s'intéressant aux documents techniques. Pendant plusieurs années, ces standards ainsi que la documentation ont été écrits et disséminés uniquement en langue anglaise. Les internautes demandent de plus en plus qu'on diffuse les documents libres dans leurs langues maternelles.

Cependant, lancer une traduction professionnelle sur la documentation de ce site présenterait deux inconvénients majeurs :

(1) le coût en serait très élevé, car elle doit être faite suivant toutes les phases professionnelles, extracttion de données, traduction, révision, … De plus, le suivi de l'évolution de traduction d'une version à une autre nécessite un contrôle permanent, ce qui rend la traduction effectivement très coûteuse. Notons aussi que cette communauté n'a pas les moyens de lancer un projet avec des professionnels.

(2) il est quasi-impossible que la traduction professionnelle puisse traduire la documentation dans toutes les langues existant sur la planète. C'est pour cette raison que le consortium a ouvert la traduction de documents aux bénévoles. Les communautés de bénévoles peuvent de leur côté s'organiser virtuellement sur la Toile, chacune dans sa langue, et peuvent produire des traductions de bonne qualité.

Il est important de noter que les traductions ne se font pas en ligne sur le site du W3C, mais dans des environnements personnels qui dépendent de chacun des traducteurs.

1.2.1.a Modes de travail et de coopération

La localisation est orientée, a priori, vers une mission bien définie. Elle est fortement coordonnée et disciplinée, car les traducteurs adoptent des outils informatiques (non linguistiques) facilitant la traduction et l'organisation de la mission. Ils sont donc tantôt traducteurs tantôt développeurs. La documentation est uniquement technique (pas commerciale). La traduction se fait sur des chaînes ou segments de message et d'aide, intégrés dans chaque version des logiciels (FrenchMozilla, 2005) (Traduct, 2006) (W3C, 2007). Dans le cas de FrenchMozilla, la production d'une nouvelle version traduite ne nécessite pas que des compétences en traduction, mais aussi des compétences techniques (Table 1).

Les fichiers du module de traduction pour Mozilla		
Type de fichiers	**Extension**	**Description**
fichier de description d'un module, ou d'un dossier dans un module	**.rdf**	Il s'agit d'un fichier XML qui gère diverses propriétés d'un module ou d'un dossier d'un module : entre autres, le numéro de version du module, d'où il vient et, pour une traduction, de quelle langue il s'agit. Dans l'aide en ligne, l'index, la liste des entrées du glossaire et une partie de la fonction de recherche prennent leurs données dans ces fichiers. Ils sont donc à modifier après que les fichiers HTML ont été mis à jour. Certains de ces fichiers doivent être mis à jour à chaque nouvelle version.
fichier texte lié à l'interface utilisateur	***.dtd**	Il s'agit d'un fichier XML. C'est dans ce type de fichiers qu'une grande partie de la traduction de l'interface est concentrée. Ces fichiers contiennent le texte affiché dans les boîtes de dialogue et dans les menus de Mozilla. Les raccourcis clavier (souvent Ctrl+lettre et caractères soulignés dans les boîtes de dialogue et les menus) sont également définis dans ce type de fichier.
fichier de propriétés JavaScript	***.properties**	Le contenu de ces fichiers est principalement utilisé dans les boîtes de messages et autres endroits où des messages personnalisés peuvent apparaître de façon ponctuelle, par opposition à une boîte de dialogue dans laquelle le texte est plus statique. Par exemple, les messages d'information à propos des cookies ou des certificats, les messages d'erreur, et les messages de la barre des tâches sont définis dans ce type de fichier.
fichier HTML	***.html**	Ces fichiers sont le cœur de l'aide en ligne. Les instructions d'utilisation sont contenues dans ces fichiers. Mais il en existe également dans d'autres parties de Mozilla.
...	...	...

Table 1 : Nature des documents à traduire dans le projet FrenchMozilla

Les documents à traduire sont organisés dans un arbre CVS[20] (Concurrent Versions System) où les documents sont répertoriés avec les informations qu'ils contiennent. Les traducteurs doivent savoir comment exploiter ce type de serveur de version (récupération, mise à jour, etc.) pour envoyer et récupérer les documents à traduire.

1.2.1.b Ressources linguistiques

La plupart des communautés citées ci-dessus mettent à disposition de leurs traducteurs bénévoles des ressources lexicales dans un état en pratique non utilisable. Malheureusement, aucune communauté n'est dotée d'outils d'aide à la traduction. À part quelques tentatives comme celle d'Arabeyes, il est rare de trouver des interfaces dédiées à une communauté et permettant la mise à jour directe des articles, à la Wiktionary (Wiktionary, 2007).

Les traducteurs sont souvent amenés à consulter un contenu textuel après un téléchargement ou à parcourir séquentiellement la liste des articles en ligne selon l'ordre alphabétique (FrenchMozilla, 2005).

Une première consultation de ces ressources nous aide à clarifier éventuellement les problèmes de la traduction bénévole :

(1) manque de convivialité,

(2) impossibilité de mettre les dictionnaires à jour,

(3) absence de fonction de recherche/remplacement, etc.

Ces manques de fonctionnalités de base sont dus aux problèmes de manque de temps et de personnes compétentes en technologie de traduction.

Le projet FrenchMozilla (FrenchMozilla, 2005) en est un bon exemple : le site FrenchMozilla offre aux traducteurs un glossaire consultable en ligne qui contient quelques centaines d'entrées classées par ordre alphabétique. La recherche se fait tout simplement par introduction du mot vedette.

[20] http://fr.wikipedia.org/wiki/Concurrent_versions_system

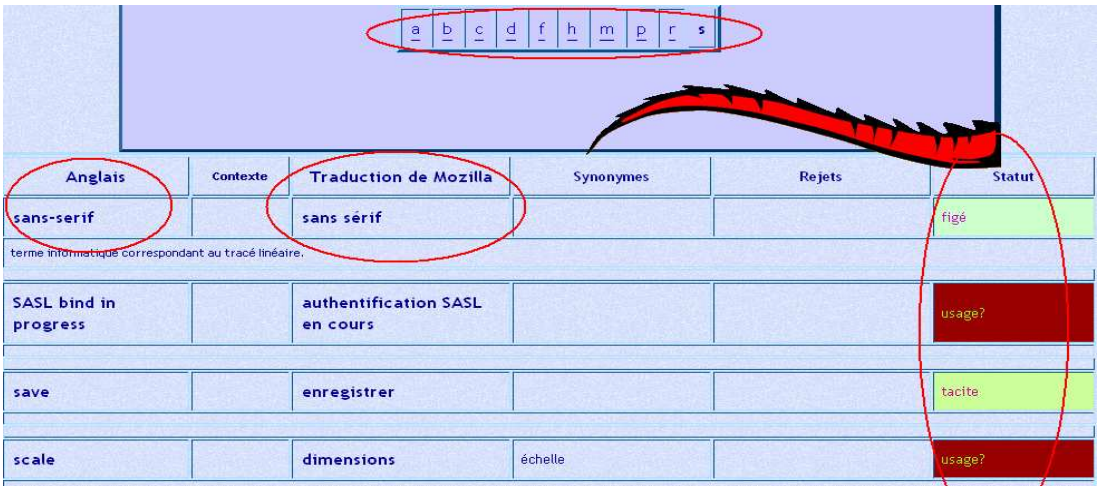


Figure 2 : Interface de consultation du glossaire français de Mozilla

La mise à jour des données de ce site est faite par un coordinateur humain qui, après avoir mené un échange entre traducteurs, met à jour un article de ce glossaire. Ce glossaire est orienté vers la traduction technique et ne contient que les termes ou expressions techniques relatifs à Mozilla. Il est assez petit et n'intègre aucune fonctionnalité avancée permettant d'aider les traducteurs. Par exemple, on ne peut pas mettre à jour les entrées, ni faire de « chercher/remplacer » sur un ensemble d'articles. Ainsi, la mise à jour est lente à cause de l'intervention permanente de l'administrateur qui doit suivre toutes les discussions de ses homologues pour confirmer la traduction de chaque terme.

D'après nos discussions directes et par courriel avec les responsables de cette communauté, il s'avère que cette ressource est rarement utilisée durant la traduction par les traducteurs, et que son évolution est également lente.

1.2.2 Communautés pour la traduction culturelle et « humaniste »

Les traducteurs de cette catégorie ont des missions non définies d'avance (Teanotwar, 2005) (Paxhumana, 2006), mais ils partagent certaines opinions concernant une variété de sujets : pacifisme, paix, droits de l'homme, aide humanitaire, etc.

Nous avons étudié ces communautés et observé leur méthode de travail.

1.2.2.a Modes de travail et de coopération

Un traducteur choisit un document en ligne qui l'intéresse (par exemple, un document d'un journal). Ensuite, il procède à sa traduction localement sur son PC par exploitation de différents outils linguistiques, d'un éditeur de texte (Emacs, Notepad, MS Word, etc.), et de

quelques ressources lexicales (dictionnaires, glossaires, etc.). À la fin de la traduction, le document cible est mis sur le site de la communauté.

Dans une telle pratique, Internet joue un rôle important pour l'aide à la traduction. Les traducteurs l'utilisent pour la consultation de ressources linguistiques et la recherche de traductions existantes pour éviter de faire deux fois la traduction d'un même document.

1.2.2.b Besoins apparents

Les environnements et les ressources libres de droits n'aident en fait pas beaucoup les communautés de traducteurs et ne sont pas satisfaisants au regard de ce qui peut être réalisé aujourd'hui par les technologies avancées.

Les besoins et les manques exprimés par les traducteurs sont les suivants :

- Ils auraient besoin d'un environnement unifié de gestion des documents et des ressources linguistiques (dictionnaires, glossaires, mémoires de traductions, etc.).
- Ils souhaiteraient des fonctionnalités pour le partage de la traduction de documents en mode collaboratif, et des ressources linguistiques. De plus, ils souffrent du manque d'outils collaboratifs qui leur permettraientt de partager leurs compétences traductionnelles et de résoudre différents problèmes durant le processus de la traduction.
- Les paires de documents déjà traduits ne peuvent pas être visualisées efficacement et systématiquement durant la traduction d'un nouveau document.
- Les traductions disséminées en ligne sont dispersées sur des environnements personnels, ce qui entraîne que, en pratique, la traduction d'un même document ne peut pas être partagée entre plusieurs traducteurs de la communauté.
- Les traducteurs expriment le besoin d'un *éditeur multilingue en ligne* offrant des aides linguistiques diverses. Ils s'attendent à un éditeur qui leur permette l'amélioration des traductions de façon conviviale par des suggestions traductionnelles.
- Selon les traducteurs bénévoles, les données traductionnelles disséminées sur le Web nécessitent un recyclage pour être réutilisée efficacement. Cela à un impact autant sur les suggestions que sur

les duplications. Cette technique est généralisable, que se soit sur des documents ou des segments semblables.

Si tous ces besoins sont exprimés par les traducteurs, certains sont plus importants que d'autres. L'éditeur en ligne d'aide à la traduction restera l'une des composantes les plus importantes, surtout lorsqu'il s'agit de résoudre les problèmes de collaboration et d'offrir des fonctionnalités avancées (par exemple, des suggestions proactives). Ensuite, nous remarquons que les traductions déjà disséminées ne peuvent pas actuellement être exploitées efficacement par d'autres traducteurs. Or, quand ils disposent de ces données, les traducteurs peuvent produire des traductions plus cohérentes et économisent énormément de temps.

1.2.3 Traduction occasionnelle en ligne

Dans cette catégorie, les traductions dépendent des technologies employées. Les bénévoles pratiquent la traduction de façon occasionnelle, complète ou partielle. Ils exploitent généralement des technologies collaboratives telles que les Wiki.

1.2.3.a Un site avec TA et aides linguistiques : Yakushite.net

Yakushite.net (de la société OKi Electric) est un environnement collaboratif en ligne offrant des services traductionnels divers. Au départ, il a été mis en ligne pour les traducteurs professionnels et étendu par la suite, pour un usage par les bénévoles.

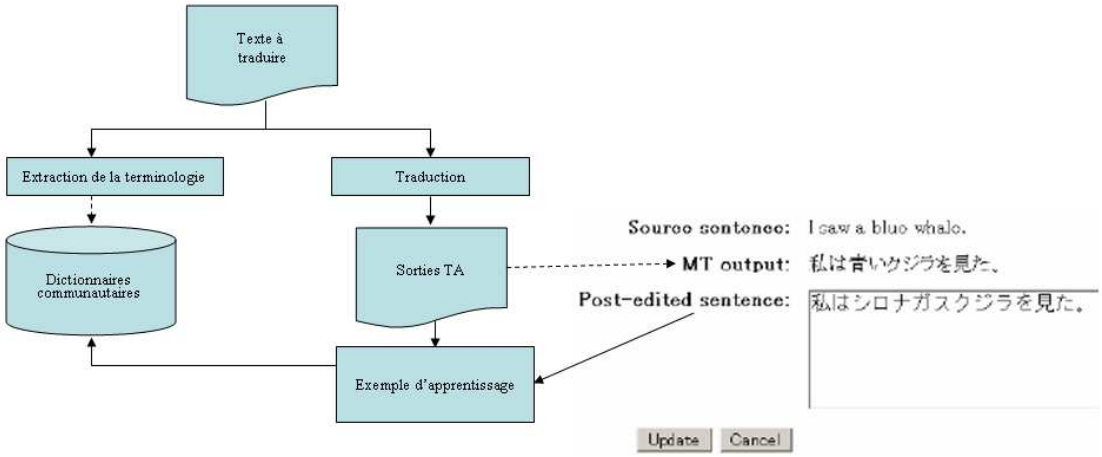


Figure 3 : Traduction avec post-édition dans Yakushite.net

L'un des buts majeurs de cet environnement est d'aider les traducteurs en ligne et d'améliorer la qualité des traductions par l'augmentation et la spécialisation des connaissances internes du traducteur automatique PENSEE (PENSEE MT, 2007). L'environnement permet

la construction d'une hiérarchie de dictionnaires par les traducteurs eux-mêmes et l'amélioration de la traduction automatique (Yakushite.net, 2007).

Pour enrichir les connaissances du système PENSEE, les auteurs ont intégré un module d'extraction terminologique spécialisé et basé sur des méthodes statistiques. Des expérimentations internes pour l'évaluation de l'efficacité de Yakushite.Net montrent que 30% des termes détectés couvrent 50% des termes qui doivent être saisis. Les prétraductions produites par le système PENSEE sont soumises à une phase de post-édition faite par les traducteurs bénévoles. Ensuite, Yakushite.Net ajuste les connaissances du système PENSEE en fonction des modifications effectuées durant la révision (Kitamura, *et al.*, 2003) (Shimohata, *et al.*, 2003).

Le but essentiel de cet environnement est de pouvoir enrichir le traducteur automatique PENSEE par des contributions bénévoles sur le Web. La plupart des contributions vont vers l'ajustement de la traduction automatique basée sur des schémas (« patterns » ou « «patrons »). Cependant, on remarque que l'environnement n'est pas équipé de fonctions permettant la collaboration et la communication entre les traducteurs durant le processus de la traduction. Parmi les manques relevés, les principaux sont les suivants :

- Pas de manipulation « libre » des dictionnaires.
- Absence d'organisation des traductions en communautés.
- Manque de gestion des documents multilingues en différents formats.
- Pas d'intégration de fonctionnalités d'aide linguistique pendant la traduction humaine (dictionnaires, MT, etc.).

Ces remarques coïncident avec les besoins identifiés plus haut. De plus, les traducteurs professionnels sont souvent réticents à l'ouverture des ressources linguistiques, car ils travaillent en groupe et souhaitent avoir plus de contrôle sur la qualité des données linguistiques introduites par les bénévoles.

L'environnement de Yakushite.net n'a pas été conçu spécialement pour aider les traducteurs bénévoles, car on note l'absence presque totale des fonctionnalités (comme la MT) généralement présentes dans tous les outils d'aide à la traduction libres ou commerciaux (exemple : MT dans Omega-T, outils libre, et Trados, outil commercial).

La principale composante manquant à Yakushite.net est un éditeur multilingue. Cela génère un manque de fonctionnalités d'aide à la traduction multilingues, car Yakushite.net est conçu spécifiquement pour ne gérer que le japonais et l'anglais.

La question qui se pose est alors : que faudrait-il améliorer pour le transformer en un environnement réellement utile aux traducteurs bénévoles ?

L'environnement Translationwiki.net nous apporte quelques réponses.

1.2.3.b Un environnement collaboratif en ligne : Translationwiki.net

L'environnement collaboratif libre Translationwiki.Net est un environnement de traduction disponible sur le Web et utilisable par tout traducteur bénévole[21]. Les documents traduits dans cet environnement sont de nature textuelle et sont limités aux documents d'information, extraits directement des sites d'informations.

L'environnement offre des services de traduction simples en quelques étapes :

(1) Choix du document source, normalisation du format et codage des caractères.

(2) Segmentation automatique et identification des unités de traduction (UT), qui peuvent être des paragraphes ou des phrases.

(3) Support à la traduction humaine (éditeur bilingue de segments, à la Wiki).

(4) Publication immédiate de la traduction dans la langue cible, dès qu'elle est commencée.

Le chargement des documents est fait par les traducteurs eux-mêmes. Après le chargement, le contenu textuel est segmenté automatiquement en une liste d'UT (unité de traduction). Mais si les traducteurs ou lecteurs notent du vandalisme, ce qui est possible, car l'environnement est totalement ouvert et à accès libre, une modification par une source suspecte peut être facilement rectifiée par les « vrais » traducteurs utilisant cet environnement. Le contrôle de qualité n'est donc pas fait par le gestionnaire du site, qui se limite à gérer l'environnement et n'a vraisemblablement pas de temps ni de connaissances suffisantes des langues.

[21] Lancé vers 2003, il semble avoir été abandonné courant 2007. Nous n'avons pas encore trouvé s'il a un successeur. Sinon, ce pourrait être une version ultérieure de notre système (cf., Partie II, p. 109).

Durant le processus de traduction, une seule UT est manipulable à un instant donné, ce qui se traduit par l'invisibilité du document complet. En effet, chaque UT est actuellement traitée comme un petit document Wiki. Les traducteurs ne peuvent pas avoir une vue générale sur le document, ce qui les empêche d'en vérifier la cohérence, d'où un grand risque d'obtenir des traductions différentes d'une même expression.

Cet environnement est limité à cinq langues (allemand, arabe, chinois, français, et italien). Les traducteurs peuvent trier et chercher dans une seule langue à la fois. En ce qui concerne la direction de la traduction, un document chargé ne peut être traduit que dans l'une des langues proposées par l'environnement. Les traducteurs n'ont donc pas la possibilité de traduire un document donné en plusieurs langues dans la même interface. Il faut à chaque fois recharger le document pour pouvoir le traduire à nouveau dans une langue.

Les documents sont accessibles directement à partir de la liste principale des documents. Les traducteurs sélectionnent un document et passent ensuite en édition dans un éditeur simple basé sur des composants HTML où il n'est possible de traduire qu'un segment à la fois (Figure 4). Dans la même interface d'édition, une seule paire « source et cible » est présentée et traduite à la fois.

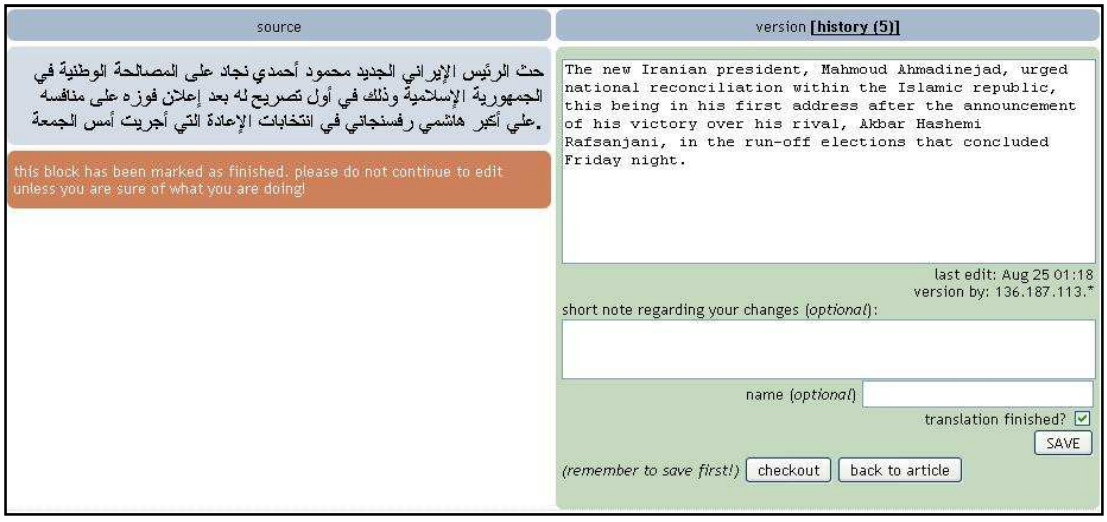


Figure 4 : Éditeur de traduction de Translationwiki.net

Les versions de traduction sont gardées après les modifications successives. Cette propriété existe dans presque tous les environnements Wiki. Les traducteurs sont capables de comparer les différentes modifications de traduction pour, d'une part, avoir une idée sur les contributions, et d'autre part, pouvoir améliorer ou rectifier les erreurs commises par d'autres traducteurs (Figure 5).

Cet environnement présente un intérêt dans sa conception et dans la technologie adoptée. La traduction segment par segment s'avère intéressante quand les MT sont gérées en phrases (UT, unité de traduction). La traduction en mode collaboratif à la Wiki est une originalité de cet environnement. Il permet de passer de la lecture à l'édition sans changer d'interface. La gestion de versions permet aussi de suivre les contributions faites par plusieurs bénévoles en cas de collaboration.

Les inconvénients ou manques de cet environnement sont les suivants :

- l'éditeur multilingue n'intègre aucune fonctionnalité linguistique.
- il n'y a pas de gestion documentaire. Les documents sont traduits et visualisés comme une MT, i.e. comme une liste de phrases et de paragraphes.
- il n'y a aucun moyen de rechercher les UT déjà traduites dans d'autres documents ou même dans le document en cours de traduction.
- l'environnement ne propose que la gestion d'un seul type de document.
- un document ne peut être visualisé intégralement durant la traduction, ce qui est un inconvénient majeur pour le contrôle de la cohérence, parce que la traduction ne se fait pas qu'en une seule passe.

Figure 5 : Gestion des versions de traduction à la Wiki

Les traductions sont réalisées à la volée de façon « incrémentale » et distribuées : la traduction d'un document cible est généralement diffusée sur Internet dans un état

partiellement complet. Les lecteurs peuvent lire dans leurs langues les traductions et peuvent contribuer s'ils le souhaitent en même temps que d'autres traducteurs.

1.1.3 Perspectives et enjeux de la traduction bénévole

1.3.1 Construction collaborative de ressources linguistiques

Les méthodes de traduction exposées ci-dessus montrent que les ressources linguistiques sont peu utilisées par les traducteurs bénévoles. Cela est dû principalement au manque d'interface ou à leur pauvreté. Cela s'explique aisément : le coût de leur construction est très élevé, même pour de relativement petites tailles, et les traducteurs ne peuvent pas les améliorer. Pour étudier de près les méthodes de construction de ressources linguistiques, nous avons étudié trois environnements différents : (1) Wiktionary, (2) Papillon et (3) XNLRDF.

(1) Le dictionnaire Wiktionary est un environnement collaboratif libre ouvert à toutes les contributions des bénévoles. Il a été construit entièrement par des bénévoles[22].

(2) La base lexicale multilingue « Papillon » est une ressource linguistique libre. Elle permet d'enregistrer les contributions des bénévoles pour l'amélioration du contenu. (Figure 6). Son principe de fonctionnement est basé sur des pivots reliant les différentes entrées des dictionnaires monolingues[23].

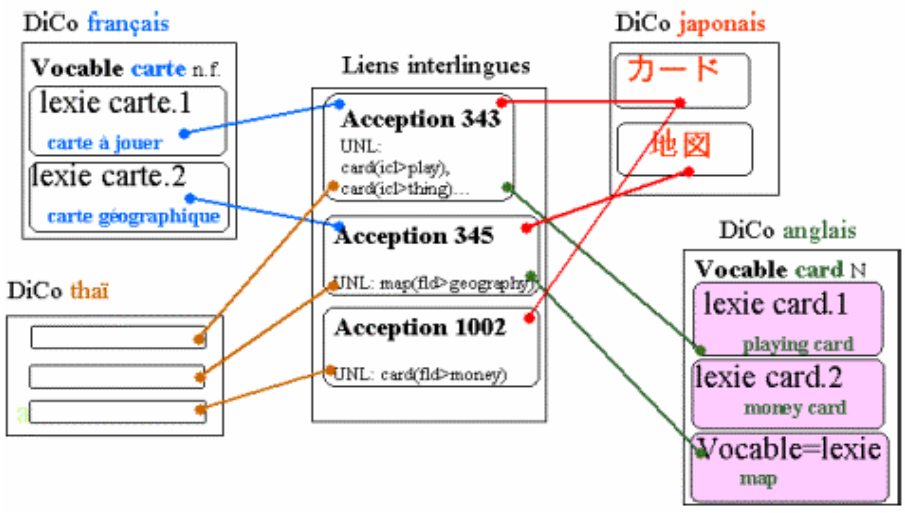


Figure 6 : Structure du dictionnaire Papillon (partie Papillon-NADIA)

[22] http://www.wiktionary.org/
[23] http://www.papillon-dictionary.org/Home.po

(3) L'environnement collaboratif XNLRDF a été développé pour permettre la collecte de systèmes d'écriture et de scripts de langues. Il est doté de fonctionnalités permettant la description collaborative de tous les systèmes d'écriture (Streiter, *et al.*, 2005)[24].

Ces trois environnements montrent l'intérêt de la construction collaborative de ressources linguistiques diverses par les bénévoles. Pour avoir une idée sur le volume de chacune de ces ressources, nous présentons dans la sous-section suivante quelques statistiques.

1.3.2 Quelques statistiques

Des bénévoles ont participé à améliorer la partie Papillon-NADIA de la base lexicale Web Papillon par l'ajout de : 253 entrées en langue française, 138 en langue japonaise, 186 en anglais et 68 en malais.

Cette partie de la base Papillon est destinée à la recherche en lexicographie multilingue avancée. Chaque volume monolingue est divisé en « lexies » (sens de mots) à structure très complexe (celle du DiCo [25] de Mel'čuk et Polguère). Un volume d'axies (acceptions interlingues) relie les lexies synonymes.

L'autre partie, Papillon-CDM, présente de nombreux dictionnaires libres de façon uniforme. Elle contient actuellement plus de 2M entrées dans 8 langues différentes. Elle contient en particulier 90 000 entrées en japonais-anglais du dictionnaire bilingue JMDict de Jim Breen et 250 000 entrées allemand-japonais et japonais-allemand collectées par Ulrich Apel dans son dictionnaire WaDokuJiTen[26].

Le Wiktionary inclut plusieurs dictionnaires unilingues, bilingues et multilingues. Le dictionnaire anglais contenait fin 2007 345 214 entrées traduites en 387 langues (ces chiffres viennent de l'analyse du dictionnaire en décembre 2006). Son contenu ne cesse d'évoluer par des contributions collaboratives (Figure 9). Depuis 2003 (date de sa création), il y a eu environ 1 891 363 éditions faites par la contribution de 29 042 bénévoles à travers la Toile (Wiktionary, 2007).

Enfin, le projet XNLRDF tente de collecter et de décrire des systèmes d'écriture multiples et des scripts pour des langues peu dotées, ou « langues-π » (Berment, 2004) (Streiter, *et al.*,

[24] http://140.127.211.214/research/nlrdf.html
[25] http://olst.ling.umontreal.ca/recherche/linguistique/dico/lang-pref/fr/
[26] http://www.wadoku.de/

2006). Les contributions ont donné lieu à 23 000 systèmes d'écriture, 8 2000 langues, 7 600 exemples de système d'écriture, et 150 scripts (Streiter, *et al.*, 2006).

Ces environnements montrent l'intérêt de la construction des ressources linguistiques en ligne par la contribution de bénévoles. Bien que chacun des environnements cités ait un intérêt particulier, Wiktionary reste le plus attractif pour les bénévoles. L'aspect ergonomique et la possibilité de modification instantanée offerte par les Wiki ont permis non seulement de créer des ressources, mais aussi d'attirer un très grand nombre de bénévoles. Nous aborderons les différents aspects des Wiki dans la deuxième partie pour la construction de notre environnement d'aide à la traduction. À l'inverse des ressources citées ci-dessus, nous montrerons aussi comment intégrer les ressources linguistiques dans le processus de traduction.

1.2 Problèmes de la traduction bénévole en ligne

1.2.1 Communautés de traducteurs « bénévole » : pratiques et besoins

2.1.1 Deux études des usages

Plusieurs facteurs interviennent dans la conception d'un environnement d'aide à la traduction pour bénévoles. Il est important de tenir compte du fait que les traducteurs bénévoles ont des besoins spécifiques par rapport aux professionnels et qu'il faut considérer leurs besoins pour construire un environnement de traduction utilisable par eux.

Étapes	Communiés	Type de traducteurs
Observations directes	Mozilla (arabe/français)	Traduction-localisation
	W3C	Traduction
	Traduct (anglais/français)	Traduction
Interviews	Mozilla (arabe/français)	Traduction-localisation
	Traduct (anglais/français)	Traduction
	TeaNotWar (anglais/ japonais)	Humaniste
Questionnaires	Mozilla (arabe/français)	Traduction-localisation
	TeaNotWar (anglais/ japonais)	Humaniste

Table 2 : Identification des besoins des traducteurs bénévoles

Nous avons employé des méthodes empiriques telles que les observations directes, les interviews et les questionnaires (Teoh, *et al.*, 2004). Pour certaines communautés, nous nous sommes limité à l'observation directe (Table 2).

Enfin, nous avons procédé à l'observation et à l'expérimentation de certains environnements tels que (Translationwiki, 2006) et (Yakushite.net, 2007).

2.1.1.a Traducteurs-localiseurs des logiciels libres

La méthode appliquée pour connaître les besoins réels de ces communautés est informelle. Elle a consisté, entre autres, en des échanges de courriels avec les responsables de quelques projets (Arabeyes, FrenchMozilla, Traduct, Paxhumana, etc.). Ces échanges ont été entrepris après avoir étudié la méthode de traduction sur le serveur de chaque communauté.

Par exemple, lors de notre conversation avec le responsable de FrenchMozilla, nous avons constaté l'absence quasi-totale de tous les outils d'aide à la traduction (MT, dictionnaires, bases terminologiques spécialisées, etc.) et l'ignorance du fait que les traductions déjà faites peuvent fortement aider la traduction d'autres textes et donc d'autres logiciels.

Par exemple, on peut plus rapidement traduire Firefox et Thunderbird à partir de la traduction de Mozilla (ce passage est jusqu'à présent fait par un script Shell).

D'ailleurs, d'après le responsable de FrenchMozilla, les logiciels Thunderbird et Firefox partageront dans le futur la même base de ressources, car ils contiennent beaucoup de chaînes semblables à traduire.

Les révisions et les corrections terminologiques se font à la main. Par exemple, le remplacement d'un terme par un autre dans un document traduit se fait par l'ouverture des fichiers et le lancement de la fonction « chercher/remplacer ». Nos discussions et contacts avec les responsables ont montré leur grand intérêt pour avoir des environnements collaboratifs en ligne d'aide à la traduction.

2.1.1.b Traducteurs humanistes

Durant notre stage au NII (National Institute of Informatics) où nous avons commencé la conception de l'environnement QRLex, nous avons réalisé une deuxième étude concernant les communautés de traducteurs partageant plutôt des opinions que des missions. Nous avons été amené à communiquer avec les traducteurs humanistes japonais qui s'intéressent à la traduction de documents sur les droits de l'homme (Teanotwar, 2005).

Des questionnaires ont été envoyés à 15 bénévoles japonais, ce qui nous a permis de faire quelques observations.

La plupart des traducteurs bénévoles n'ont pas un niveau très élevé dans la langue source qui est généralement l'anglais). La traduction d'un document nécessite souvent la consultation de plusieurs ressources linguistiques (par exemple Wikipedia et Wiktionary).

Pour pouvoir les aider, il faut que les outils d'aide demandés soient adaptés à leurs besoins : ils doivent faciliter le processus de la traduction en réduisant l'effort impliqué dans la consultation de ressources, qui représente jusqu'à 75% du temps total de traduction.

Les traducteurs donnent une importance particulière aux unités linguistiques suivantes : les idiomes, les citations, les expressions, les noms propres et les termes techniques. En ce qui concerne les idiomes, on peut faire les remarques suivantes :

- beaucoup de traducteurs ont moins de connaissances des idiomes que des mots.
- certains idiomes ne doivent pas être identifiés comme tels, car ils ont une interprétation directe à côté de l'interprétation idiomatique, et cela peut facilement mener à de fausses traductions.

De plus, les besoins diffèrent d'une communauté à une autre. Il est difficile de cerner tous leurs problèmes sans les catégoriser et regrouper leurs besoins. Dans les paragraphes suivants, nous présentons les différentes catégories de traducteurs, ce qui nous permettra de mieux associer les problèmes rencontrés à chaque catégorie.

2.1.2 Similarités et différences dans les pratiques observées

Notre étude sur les communautés de traducteurs bénévoles existantes montre qu'il existe deux grands groupes, chacun ayant ses pratiques propres dans la traduction en ligne. Ces deux groupes sont définis comme suit :

> *Communautés orientées par la mission (catégorie A) :* ce sont des groupes fortement coordonnés et impliqués dans la traduction d'un ensemble de documents définis d'avance.
>
> Elles couvrent principalement la traduction de documents techniques, Linux (Traduct, 2006), spécifications (W3C, 2007), et la localisation de logiciels libres comme (ArabicMozilla, 2007). Généralement, la traduction s'étend aux messages et textes d'interface des composants logiciels.

La majorité des traducteurs coordonnent ensemble les traductions[27].

Peu de ressources linguistiques sont proposées dans les sites de ces communautés. Celles qui sont disponibles manque d'interfaces de consultation et de mise à jour. Elles sont généralement proposées en format textuel et exploitées en local via des éditeurs de texte. Les modifications faites en local par les traducteurs ne sont pas prises en compte par la communauté, car aucun outil n'est proposé pour permettre la mise à jour centralisée.

Communautés orientées par le thème « catégorie B » : ce sont des traducteurs qui traduisent des documents en ligne tels que des informations, des analyses, et des rapports. Ils font en sorte que les traductions soient disponibles dans des sites Web personnels ou de groupe.

Ils forment des groupes de traducteurs sans aucune orientation connue d'avance, mais qui partagent des opinions similaires concernant des événements (militants contre la guerre, pour les droits de l'homme, pour la diffusion d'information sur la santé, pour l'aide humanitaire, etc.).

Ces communautés ne sont pas organisées et ne disposent d'aucune ressource linguistique utilisable. Certains traducteurs ne savent même pas qu'il existe des outils d'aide à la traduction. On remarque aussi que la traduction est rarement partagée et que la méthode de traduction est spécifique à chaque traducteur.

Les trois situations de pratique décrites précédemment (p. 31) sont regroupées dans ces deux catégories. Les traducteurs localiseurs de logiciels libres appartiennent à la *catégorie A.* Cependant, la *catégorie B* regroupe les deux autres situations (traduction de documents culturels, humaniste, et les traductions occasionnelles).

Les traducteurs impliqués dans la *catégorie A* ont des compétences diverses, traductionnelles ou techniques. Ce sont des communautés organisées qui profitent de beaucoup d'outils informatiques leur permettant de faciliter plusieurs tâches (gestion des documents, édition, gestion de ressources linguistiques, etc.). Par contre, les traducteurs de la *catégorie B* n'ont pas assez de connaissances techniques. Ils choisissent leurs documents selon leurs intérêts et selon les points communs qu'ils partagent avec les autres traducteurs.

[27] Pour le contrôle de la progression, les traducteurs disposent de certains fichiers qui balisent les points d'arrêt des traductions pour que de nouveaux traducteurs intéressés puissent suivre ces traductions. Quelques scripts sont mis à la disposition des traducteurs pour la manipulation des fichiers en provenance de différents logiciels (chercher/remplacer, mixage de traduction des logiciel à localiser, etc.).

2.1.2.a Récupération des documents à traduire

Les traducteurs-localiseurs de logiciels libres sont amenés à télécharger les documents sources depuis un site Web dédié aux projets de chaque logiciel, tel qu'il est, dans son format original (textuel ou structuré). Pour illustrer clairement la méthode de récupération des documents sources à traduire, nous reprenons le cas d'Arabeyes[28].

Un traducteur dans Arabeyes doit tout d'abord savoir maîtriser correctement l'accès aux serveurs CVS (Concurrent Versions System)[29]. Il doit en outre disposer d'un accès lui permettant la modification des documents, car il y a deux types d'accès à CVS : un accès anonyme (accès en lecture seule ; ouvert à tous) et un accès réservé aux développeurs (lecture/écriture ; restreint).

Le dernier type d'accès permet de gérer les documents et les historiques des modifications via des techniques offerts par l'outil CVS. Le document source est récupéré à l'aide de l'opération « check out » et envoyé à nouveau (après la traduction) au serveur grâce à l'opération « commit ».

La question qui se pose maintenant est : quelle est la nature des documents à traduire, et les structures de données adéquates ?

```
#: finddialog.cpp:55
msgid "Galaxies"
msgstr "مجرّات"

msgid ""
"_: do not use a target symbol\n"
"No Symbol"
msgstr ""

#: src/acme.h:80
#, fuzzy
msgid "Eject key"
msgstr "مفتاح الإخراج"
```

Figure 7 : Structure des entrées dans un fichier « po »

Les données à traduire sont stockées dans des fichiers *po*[30] où le texte est présenté sous forme d'un ensemble de couples (des segments sources et cibles) balisés par des mots-clés (*msgid* pour le segment source et *msgstr* pour la cible). Les balises facilitent la récupération des segments par les programmes lors de la visualisation des messages (Figure 7).

[28] http://www.arabeyes.org/
[29] http://en.wikipedia.org/wiki/Concurrent_Versions_System
[30] http://en.wikipedia.org/wiki/Gettext

C’est grâce à cette structure que les programmes (comme le navigateur Mozilla) récupèrent les chaînes traduites et visualisent la version traduite et localisée. Malheureusement, cette méthode n’est pas totalement générale, car dans le cas de l’arabe, il faut résoudre les problèmes liés au *système d’écriture* qui est un système fonctionnant de la droite vers la gauche. Les chaînes à traduire sont extraites à partir des fichiers source de type « po », souvent par l’outil Gettext[31] et sont traduites une par une dans la langue cible par des outils spécifiques tels que poedit[32].

2.1.2.b Méthode de traduction

Comme déjà dit, les traducteurs bénévoles ont leurs propres environnements de traduction qui diffèrent souvent de l’un à l’autre.

Ces environnements consistent en un ensemble d’outils : simple éditeur de texte, dictionnaires (version électronique, version papier, version en ligne, etc.), glossaires, banque terminologique, et quelquefois des mémoires de traduction. De plus, il faut noter l’importance d’Internet qui devient une ressource linguistique incontournable pour les traducteurs : durant les traductions, les bénévoles se réfèrent beaucoup aux ressources libres et aux traductions déjà existantes.

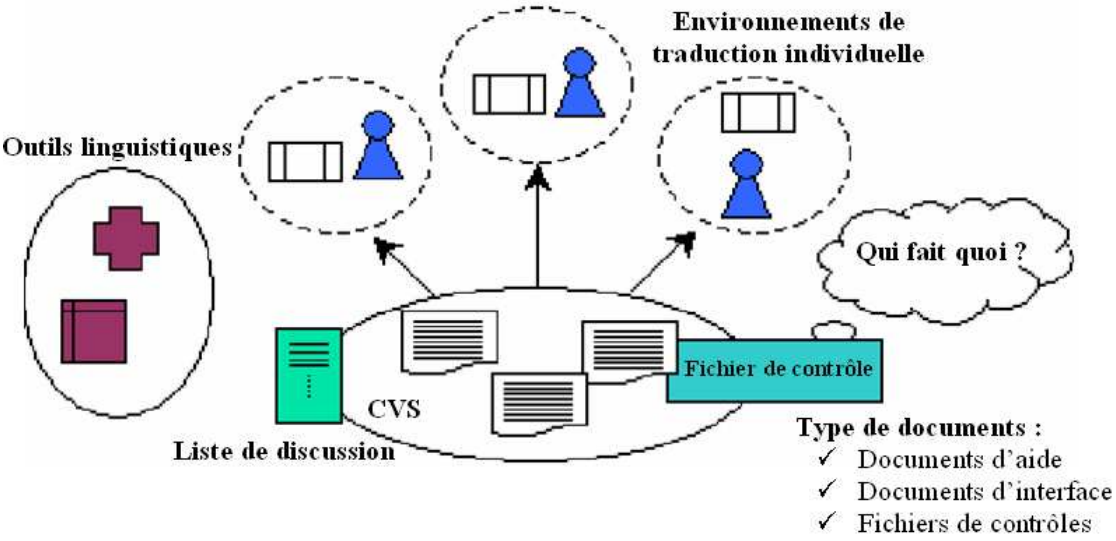


Figure 8 : Pratique de la traduction bénévole

[31] Pour avoir plus de détails sur la traduction et l’utilisation des fichiers « po » dans les serveurs CVS, on peut se référer deux sites : http://www.arabeyes.org et http://www.gnu.org/software/gettext/.
[32]http://gettingeek.com/translation-with-poedit-internationalize-localize-wp-themes-guide-part-3-74.html

Les documents sont traduits en respectant le format original, car, à la fin de la traduction, le document produit doit avoir le même format que l'original. Ces formats sont variés et dépendent des projets de traduction. Par exemple, dans le projet Traduct[33], les documents sont structurés en XML DocBook (DocBook, 2007) et les traductions se font entre les balises XML pour ne pas altérer le format original (Figure 8).

Retournons maintenant aux propriétés des environnements utilisés. Le fait d'identifier la méthode de traduction n'est en effet pas suffisant. Étudier les problèmes des outils exploités dans les environnements est aussi utile pour comprendre les habitudes traductionnelles des traducteurs bénévoles.

Cette étude nous a montré que les outils utilisés ont les caractéristiques suivantes :

- Les éditeurs sont « basiques » et ne proposent aucune aide linguistique durant la traduction. Très peu de traducteurs utilisent des outils tels qu'Excel mettant le texte source et le texte cible en visualisation parallèle et de façon synchronisée.

- Les communautés de traducteurs bénévoles sont apparemment allergiques aux outils commerciaux. La plupart des traducteurs bénévoles n'exploitent pas les outils professionnels tels que Trados™, Déjà Vu™, Transit™ ou Similis™. La première raison est qu'ils sont payants, alors que ces communautés n'utilisent que des outils gratuits (de GNU, de SmartOffice, NeoOffice, Mozilla,…). La seconde est que leurs volumes de traduction sont trop petits pour constituer des mémoires de traductions réellement utiles.

- Les traducteurs-localiseurs de logiciels libres envoient le document traduit sur leur site Web, à l'endroit indiqué, exactement dans le format du document source. Ce document est ensuite intégré automatiquement au prochain assemblage (« build ») de la version localisée du logiciel. Cette pratique est courante dans la *catégorie A*, où les traductions sont compilées avec le code source de l'application pour produire une nouvelle version localisée.

- Les traducteurs culturels et humanistes (traduisant pour des causes) ne traduisent souvent que du HTML ou du texte (sur des sites Web personnels). Ils ne sont pas concernés par la diversité des formats. En effet, la grande majorité ne sait pas qu'il y a des outils d'aide à la traduction. Bien que les documents soient importés de

[33] http://www.traduc.org/

journaux ayant des formats différents, ils sont transformés en html, d'où la normalisation de fait du format des textes à traduire.

Ces situations traductionnelles ainsi que la pratique des bénévoles, seront prises en compte durant le processus de développement de nos outils.

2.1.3 Facteurs humains et besoins ressentis

2.1.3.a Besoins de communication et d'organisation des communautés

Le besoin de collaboration est ressenti dans les projets de traduction de la *catégorie A*. De façon naturelle, les traducteurs communiquent souvent entre eux pour achever leurs projets dans les délais. Les moyens de communication peuvent prendre plusieurs formes. Les forums marchent très bien, mais ils ont aussi recours au courrier électronique, tchat, ou aux commentaires directs. Une autre forme de communication, existant dans les Wiki, consiste à comparer les versions des données par rapport aux dernières modifications.

Dans Mozilla, les traducteurs cherchent à savoir qui a fait quoi. La gestion de versions *à la Wiki* peut aider à implémenter cette fonctionnalité.

2.1.3.b Besoins en ressources linguistiques

Les besoins des traducteurs bénévoles en ressources linguistiques sont exprimés explicitement, mais certains traducteurs ne savent pas qu'il est possible d'automatiser un grand nombre de fonctionnalités du processus de traduction ainsi que la gestion des ressources linguistiques. Ces besoins en ressources linguistiques sont de deux sortes, les besoins exprimés et les besoins potentiels.

- *Les besoins exprimés* sont ceux que les traducteurs évoquent souvent. Il s'agit de l'accès à des dictionnaires et glossaires libres, et de la présence de fonctionnalités intégrées.

- *Les besoins potentiels* ne sont pas évoqués par les traducteurs. Il nous revient à les détecter et de faire des propositions adéquates. Il s'agit de MT liées aux types de textes traduits, et aussi de la possibilité d'appeler des systèmes de TA.

Ces besoins peuvent être satisfais par le développement de composants intégrables dans un éditeur multilingue en ligne.

1.2.2 Mise en ligne de ressources linguistiques « libres »

La construction de ressources linguistiques libres (gratuits) est de première importance dans notre contexte. Nous montrons en quoi ce qui existe est insuffisant, et finalement comment il est possible de surmonter ces insuffisances en se limitant au cadre de la traduction bénévole en ligne.

2.2.1 Glossaires, dictionnaires et bases lexicales

2.2.1.a Glossaires et bases de données terminologiques

Les traducteurs ont besoin de leurs propres ressources. Bien qu'il existe beaucoup de ressources librement consultables sur le Web, il s'agit presque toujours de terminologie technique. C'est utile, mais souvent insuffisant pour les domaines visés par les traducteurs « humanistes » et « culturels ». Les termes sont hors des terminologies recensées et, en ce qui concerne les logiciels libres, les néologismes et les termes spécifiques n'y sont pas. Malheureusement, les listes terminologiques établies par les communautés de traduction bénévole deviennent très rapidement obsolètes, car il n'y a pas de processus pour les mettre à jour et les améliorer en continu. Au bout d'un moment, elles ne sont plus beaucoup utilisées ou plus du tout (FrenchMozilla, 2005).

2.2.1.b Dictionnaires

2.2.1.b.i Dictionnaires disponibles

D'autre part, nous nous intéressons aux dictionnaires libres qui peuvent être téléchargés et intégrés dans un environnement d'aide à la traduction, et pas aux dictionnaires commerciaux.

Dans les paragraphes suivants, nous les analysons en détail et montrons les avantages et les inconvénients de chacun.

Des bénévoles ont contribué activement à améliorer l'encyclopédie *Wikipedia* en y ajoutant beaucoup d'entrées dictionnairiques accompagnées d'informations telles que les antonymes, les synonymes, l'étymologie, etc. Ces ajouts ont donné l'idée aux administrateurs du *Wikipedia* de créer *Wiktionary* pour regrouper les entrées dictionnairiques en les séparant des entrées du corps du Wikipedia (Jimmy, 2007).

Le dictionnaire a commencé à présenter un intérêt majeur, car le nombre d'entrées en plusieurs langues est devenu important et ne cesse d'augmenter chaque jour.

À part ces avantages, le dictionnaire ne peut pas être exporté vers d'autres applications ou usages, surtout si l'on veut intégrer l'accès à son contenu durant le processus de traduction comme nous voulons le faire dans notre environnement. De plus, on ne peut pas le transformer en des structures telles que *CDM* du projet *Papillon* et TBX, de façon à le rendre exploitable par les applications, et en particulier par celles destinées à aider les traducteurs bénévoles dans leur processus quotidien de traduction.

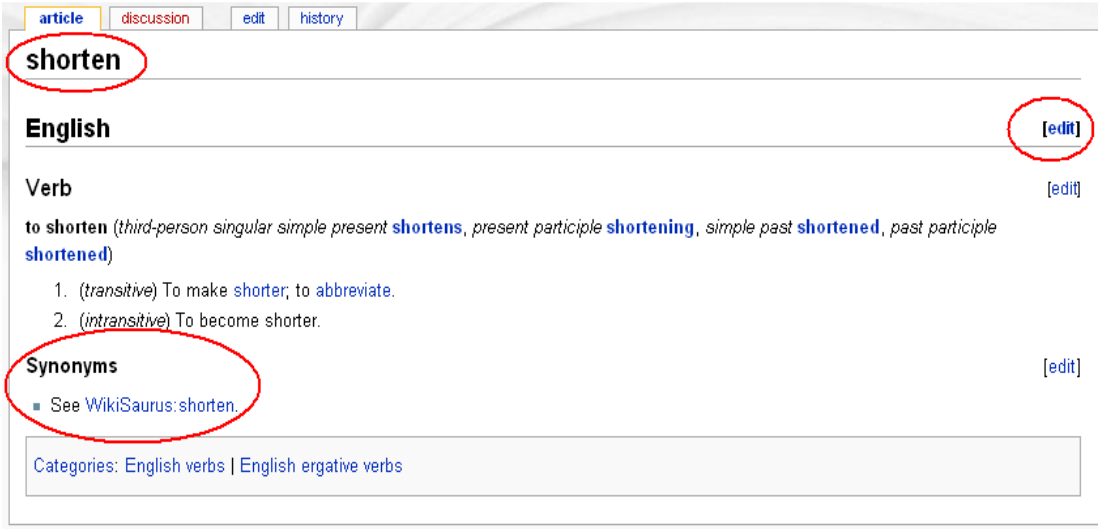


Figure 9 : L'article « Shorten » du dictionnaire « Wiktionary »

L'autre environnement déjà évoqué (cf. *Construction collaborative de ressources linguistiques*, p. 43) est le projet Papillon. Il a deux parties, Papillon-NADIA, orienté vers la recherche lexicographique (construction d'une base lexicale reliant des dictionnaires monolingues DiCo, format simplifié du DEC de Mel'čuk), et Papillon-CDM. La partie Papillon-CDM se présente comme une base lexicale multilingue sur le Web qui permet de collecter des contributions de bénévoles en ligne. Il met en œuvre un serveur pour le travail collaboratif, où chaque contributeur a son espace propre, de façon que ses contributions puissent être validées et intégrées dans la base par un groupe d'experts. Un des buts du projet est de fournir un outil couplé à des méthodes génériques de fabrication de ressources lexicales riches. (Mangeot-Lerebours, 2001).

La base lexicale Papillon-CDM contient certes environ 2M entrées (dans 9 langues au total), mais évolue très peu, en comparaison de Wiktionary, et son interface n'est pas non plus utilisable directement pendant la traduction, car trop riche. Il faudrait en fait la « compiler » vers un format adapté, comme nous le ferons plus loin pour le dictionnaire Eijiro, ou l'appeler par programme, de façon semblable à l'appel des services de TA en ligne.

2.2.1.b.ii Problèmes recensés

Ce tour d'horizon sur les environnements de construction bénévole de ressources linguistiques, et sur les méthodes de traduction, nous a permis de recenser quelques problèmes intéressants :

- Un premier problème rencontré est que la plupart des dictionnaires libres n'ont pas une structure unifiée, et qu'il est difficile de les intégrer sans les prétraiter à la main.

- L'intégration de la consultation de ressources dans l'éditeur de traduction doit tenir compte de la « proactivité » désirée : les traductions des mots et des expressions doivent être proposées sans l'intervention des traducteurs. En effet, dans le processus de la traduction classique, les traducteurs consultent eux-mêmes plusieurs ressources linguistiques pour trouver les traductions et perdent ainsi beaucoup de temps.

- La plupart des méthodes existantes de recyclage du Web ont été développées pour la construction automatique de ressources linguistiques (corpus, terminologie, etc.). Un problème intéressant est d'améliorer ces méthodes pour trouver les paires de documents qui constituent des traductions mutuelles. Ensuite, à partir de ces paires, il faudrait diminuer la granularité, pour offrir aux traducteurs des données plus fines (segments, idiomes, collocations, termes, etc.) correspondant au contexte de traduction en cours.

Toutes les ressources linguistiques doivent être intégrées au processus de traduction durant lequel il doit être possible de les augmenter et les mettre à jour.

2.2.1.b.iii Nos ambitions

Nous voulons donc concevoir et développer des outils permettant un accès unifié à plusieurs dictionnaires libres, intégrés dans un éditeur Web complet de traduction. Par exemple, l'accès au dictionnaire dans l'éditeur devrait se faire de façon proactive : les suggestions dictionnairiques de la TA et des MT doivent être proposées sans aucune intervention de la part des utilisateurs.

Un environnement d'aide à la traduction ne peut maintenant être dit « avancé » que s'il peut présenter « tout seul » les informations utiles pour la traduction d'un texte en cours. Ces idées ont déjà été proposées par d'autres chercheurs. Par exemple, la notion de « dictionnaire

actif » a été introduite par (Martin, 1990) où il fait des réflexions sur l'utilisation des dictionnaires :

« *...the use of dictionary can be seen as a typical problem-solving activity, and user-orientation should involve both static and dynamic features of the intended user...* »

(Agirre, *et al.*, 2000) ajoutent :

« *Furthermore, along with the usual information about the meaning of the entries, dictionaries should show how to use words in context. In other words, we advocate that dictionaries should actively cooperate in finding the correct translation.* »

Cela fait donc plus de 15 ans qu'on en parle, mais réaliser une telle proactivité est assez difficile. Eurolang Optimizer le faisait dès 1992 mais a disparu du marché. C'est sur ce type de problèmes que nous focalisons la conception de l'éditeur et les fonctionnalités de suggestion.

2.2.1.c Mémoires de traduction

La majorité des systèmes à mémoire de traductions (MT) sont coûteux.Il existe de moins en moins de MT libres.

Prenons l'exemple d'Omega-T. Il y a une MT permettant la gestion des documents sous forme structurée TMX et offrant la traduction en plusieurs langues (Lisa, 2008) (OmegaT, 2007), mais elle fonctionne seulement en local, et il n'est pas possible de travailler sur les traductions de documents en mode collaboratif. Cette MT ne peut répondre aux attentes des traducteurs bénévoles impliqués dans des projets de localisation de logiciels libres, car elle n'est pas équipée de mécanismes de partage et de communication, et n'offre aucune possibilité d'utilisation et d'exploitation en ligne.

En ce qui concerne les MT commerciales, des entreprises comme IBM ont une mémoire de traductions très riche et volumineuse résultant des traductions de 20 millions de mots par an en 25 langues. Malheureusement, il est impossible de l'obtenir en version libre avec une taille et une qualité similaire.

Enfin, les ressources qui existent, telles que EuroParl et EURAMIS, sont importantes en quantité mais restent pauvres des points de vue richesse et spécialisation.

Que faire pour combler ce manque ?

Ce manque ne peut être résolu que par le recyclage semi-automatique des traductions existantes pour la construction de MT libres. Cette idée peut être analysée sous deux angles : (1) développement de mécanismes d'automatisation du recyclage, et (2) révision humaine pour améliorer la qualité.

Le processus de recyclage est un robot qui fonctionne selon le même principe que Google, mais il est adapté à la recherche des équivalences entre des documents déjà traduits. La recherche pourra se faire, cependant, sur un domaine Web précis ou sur la totalité du Web.

Comme nous avons besoin de gérer des MT, nous proposons dans le chapitre suivant une méthode pour le recyclage de documents déjà traduits pour la construction des MT. Nous aurons aussi recours aux algorithmes de calcul de similarité pour implémenter les suggestions proactives.

2.2.2 Outils

Les besoins des traducteurs bénévoles ne se limitent pas qu'aux ressources linguistiques, d'autres fonctionnalités sont aussi utiles pour mener à bien les traductions.

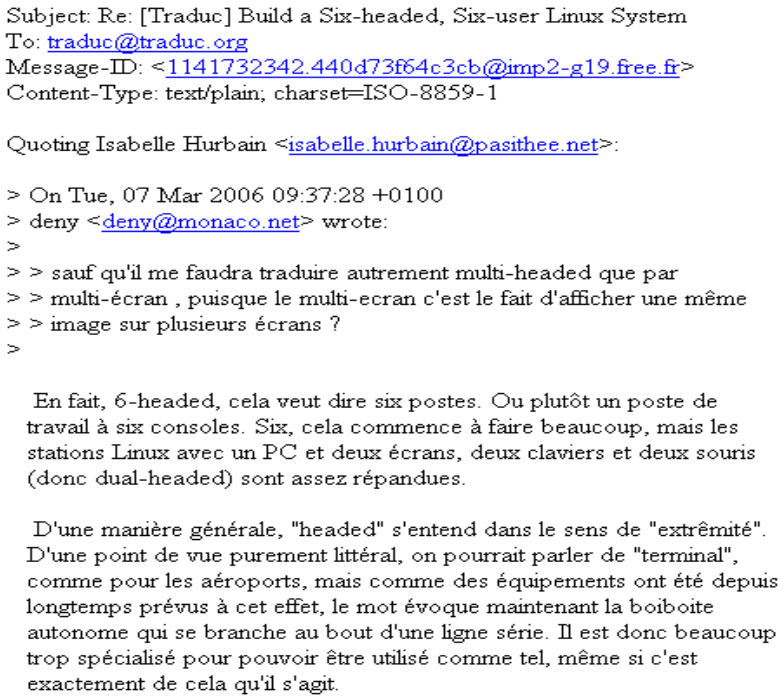
Par exemple, les traducteurs bénévoles de la *catégorie A* (cf. *Similarités et différences dans les pratiques observées*, p. 47) ont besoin de communiquer pour s'organiser autour d'un projet de traduction (par exemple localisation de Mozilla). Un environnement collaboratif doit donc avoir des moyens efficaces pour favoriser la communication entre les traducteurs et propose des mécanismes souples de gestion des versions des traductions.

Les communautés de bénévoles pensent cependant que les outils dont elles disposent actuellement (scripts, ressources textuelles, etc.) sont suffisants. Cela s'explique par le manque de connaissances dans le domaine de la TA. Or, il est maintenant possible de les aider par l'exploitation des avancées en TALN, en TA et en technologie Web. Il faudra donc, non pas les adapté, mais les faire connaître, et trouver comment convaincre les traducteurs bénévoles de les utiliser.

2.2.2.a Communication entre traducteurs

Les moyens de communication préférés par les traducteurs bénévoles sont le courriel, les blogs, les forums et les CVS. Mais les traducteurs « humanistes » et culturels, ainsi que, de façon surprenante, les traducteurs à la Wiki, ne profitent d'aucun, sans doute pas parce qu'ils ne se sentent pas faire partie d'une communauté, mais plutôt parce qu'ils n'imaginent pas que cela existe et puisse leur être utile.

À titre d'illustration, la Figure 10 montre un courriel dans lequel un traducteur appartenant à la communauté « Traduc » explique à un autre traducteur une interprétation possible du terme technique « six headed » en langue française.



Subject: Re: [Traduc] Build a Six-headed, Six-user Linux System
To: traduc@traduc.org
Message-ID: <1141732342.440d73f64c3cb@imp2-g19.free.fr>
Content-Type: text/plain; charset=ISO-8859-1

Quoting Isabelle Hurbain <isabelle.hurbain@pasithee.net>:

> On Tue, 07 Mar 2006 09:37:28 +0100
> deny <deny@monaco.net> wrote:
>
> > sauf qu'il me faudra traduire autrement multi-headed que par
> > multi-écran , puisque le multi-ecran c'est le fait d'afficher une même
> > image sur plusieurs écrans ?
>

En fait, 6-headed, cela veut dire six postes. Ou plutôt un poste de travail à six consoles. Six, cela commence à faire beaucoup, mais les stations Linux avec un PC et deux écrans, deux claviers et deux souris (donc dual-headed) sont assez répandues.

D'une manière générale, "headed" s'entend dans le sens de "extrêmité". D'une point de vue purement littéral, on pourrait parler de "terminal", comme pour les aéroports, mais comme des équipements ont été depuis longtemps prévus à cet effet, le mot évoque maintenant la boiboite autonome qui se branche au bout d'une ligne série. Il est donc beaucoup trop spécialisé pour pouvoir être utilisé comme tel, même si c'est exactement de cela qu'il s'agit.

Figure 10 : Communication par courriel dans la communauté « Traduc »

Dans la plupart des communautés de traduction, le courriel et les blogs sont les seuls moyens de communication entre les bénévoles.

2.2.2.b Aide à la traduction

Les serveurs gérant les traductions bénévoles proposent des services linguistiques minimaux qui consistent en :

- des dictionnaires et glossaires sous forme textuelle, et des indications de liens Web, généralement non mis à jour.
- une liste de discussion utilisée pour l'échange de connaissances et pour résoudre les problèmes de traduction des documents.
- des fichiers de contrôle pour vérifier qui fait quoi. Dans la traduction collaborative, avant de commencer la traduction, les traducteurs marquent des points d'arrêt des traductions pour que d'autres traducteurs sachent d'où repartir.

Les exemples suivants illustrent la situation d'aide par les ressources linguistiques tirées de quelques sites de traducteurs bénévoles :

- Le projet *Traduc* propose un glossaire français-anglais contenant environ 25 000 entrées. La mise à jour se fait par un coordinateur humain à partir des échanges sur la liste de discussion.

Dans Mozilla, deux cas concrets doivent être cités :

- *FrenchMozilla* propose un glossaire de quelques centaines d'entrées,
- *Arabeyes* propose un glossaire contenant 18 000 entrées gérées en mode Wiki.
- Par contre, les communautés de traducteurs humanistes ne proposent pas d'aides linguistiques, donc il n'y a pas de ressources mises à la disposition des traducteurs (*Paxhumana*, *TeaNotWar*, etc.).

1.2.3 Discussion

Les aides linguistiques ne peuvent être proposées séparément du processus de traduction. Les dictionnaires doivent être importés facilement, puis intégrés, de façon à pouvoir les modifier en collaboration, et cela durant le processus de traduction. Cependant, l'utilisation de nouveaux outils n'est pas évidente, parce que les traducteurs bénévoles sont mariés avec leurs outils informatiques favoris.

Pour motiver les traducteurs à utiliser un nouvel environnement, il faut qu'il ne soit pas vide de contenu. Il ne doit pas être dans l'état dans lequel sont vendus les outils d'aide professionnels (i.e. sans mémoire de traduction et sans dictionnaires).

Pour arriver à une solution rapide et attractive, le mieux serait de réaliser deux choses importantes :

- importer le plus possible de ressources linguistiques libres pour permettre aux traducteurs de démarrer rapidement.
- récupérer les documents déjà traduits et essayer d'affiner l'alignement pour construire des mémoires de traductions exploitables lors de la traduction.

L'import de ressources libres et la compilation de traductions existantes seront des avantages pour notre environnement par rapport aux outils commerciaux d'aide à la

traduction. Notre environnement devra donc être proposé avec des ressources linguistiques, et sera prêt à être utilisé par les traducteurs bénévoles.

1.2.4 Conclusion

Il ressort ce tour d'horizon sur les communautés de traducteurs bénévoles qu'elles sont privées de tout outil d'aide à la traduction, les outils existants n'étant en fait largement exploités que par les professionnels travaillant dans de grands groupes de traducteurs et non de façon isolée.

Les communautés de traducteurs bénévoles veulent produire et produisent effectivement des traductions de grande qualité, ce qui, sans aides traductionnelles, rend leur travail nettement plus lent et pénible qu'il ne pourrait l'être. Un certain nombre d'entre elles en sont conscientes et appellent à l'aide pour qu'on leur fournisse le plus rapidement possible un outil gratuit et collaboratif leur permettant de travailler plus et en groupe, et de disséminer des documents de qualité et de cohérence encore meilleures.

1.3 Traduction mutualisée : problèmes intéressants

Dans les sous-sections précédentes, nous avons présenté la situation et les besoins en matière d'aide à la traduction par des bénévoles, ainsi que les problèmes à résoudre pour satisfaire ces besoins. Certains de ces problèmes ont été déjà étudiés et résolus (par exemple, l'import et la construction d'une base lexicale multilingue).

Les problèmes sur lesquels nous nous sommes concentré sont ceux qui nous ont semblé présenter le plus d'intérêt théorique et pratique :

- l'exploration du Web traductionnel,
- la construction d'un éditeur multilingue couplé à des fonctionnalités avancées,
- la traduction de masses de données, en particulier les mémoires de traductions elles-mêmes, et les gros corpus de phrases pour la TA.

1.3.1 Exploration du Web traductionnel

La détection des documents déjà traduits est un problème difficile. Nous désirons de plus l'étendre à la détection des expressions plus ou moins complexes telles que les idiomes et les citations.

Cette fonctionnalité doit être intégrée à l'environnement de traduction, dans lequel on trouvera, calculées d'avance (proactivité), des suggestions systématiques répondant au contexte de traduction précis (segment en cours, domaine, etc.).

1.3.2 Éditeur multilingue offrant des fonctionnalités avancées

Il y a trois points importants dans la conception de l'éditeur :

(i) l'utilisabilité,
(ii) l'interface,
(iii) les fonctionnalités.

L'interface doit être la plus simple possible, car les traducteurs bénévoles ne sont pas souvent familiarisés avec la complexité d'éditeurs très riches (comme Word). Il vaut mieux enrichir un peu une interface minimale dédiée à la traduction Web. Par exemple, l'éditeur de traductions à la Wiki présenté dans la Figure 4 (p. 41) peut être étendu par l'ajout de la visualisation de tous les segments (la traduction ne se fait pas en une seule passe) et l'intégration de fonctionnalités linguistiques.

Les fonctionnalités de manipulation des ressources linguistiques doivent aussi être faciles à utiliser depuis cette interface.

L'éditeur doit donc présenter toutes les zones nécessaires à la visualisation de toutes les données : une pour le texte source et cible, une pour les dictionnaires, et enfin une pour les suggestions provenant de différentes sources.

Le deuxième point concerne les aides linguistiques automatiques (les suggestions linguistiques proposées par l'éditeur) et/ou manuelles (consultation par les traducteurs). Une qualité importante des suggestions automatiques est la « proactivité » : l'éditeur doit proposer toutes les données jugées pertinentes en relation avec le segment en cours de traduction, et ces données doivent avoir été calculées à l'avance. Les suggestions les plus importantes à offrir aux traducteurs sont :

- les suggestions de la MT, qui se font par calcul de la similarité entre un segment source à traduire et les segments déjà traduits précédemment ;
- les suggestions dictionnairiques, qui consistent à proposer des traductions personnalisées et spécialisées provenant de plusieurs

dictionnaires déjà établis par les traducteurs bénévoles (les traducteurs peuvent sélectionner un ou plusieurs dictionnaires) ;

- les suggestions provenant de la TA : dans ce cas, l'éditeur doit si possible, offrir aux traducteurs les résultats de plusieurs STA.

Un autre point intéressant mais moins important consiste à ajouter la possibilité de manipulation de données, en particulier des dictionnaires et des MT.

3.2.1 Le mode lecture/édition : le cas des Wikis

Une seule passe ne suffit pas pour avoir une traduction complète. Ainsi, notre méthode de traduction peut être un peu différente de celle des professionnels, où les traductions sont disséminées lorsqu'elles sont terminées, alors que, dans notre contexte, les traductions des documents seront disséminées au fur et à mesure, et ne seront pas nécessairement complètes.

C'est cette propriété qui permettra d'ouvrir notre environnement aux contributions et aux traductions *incrémentales*. Les traductions pourront êtres disséminées dans un état incomplet. Le passage à la post-édition se fera alors au moment de la lecture : si un paragraphe nécessite une amélioration, un traducteur pourra passer directement (sans changement d'interface) à la traduction. Dans les deux sens (lecture et édition), les traducteurs auront plus de liberté parce que toute modification sera visible instantanément par les autres traducteurs sur la Toile.

Les ressources linguistiques (dictionnaires, glossaires, etc.) ne seront pas figées non plus. Après avoir été compilées (prétraitement, import, codage, etc.), elles seront disponibles pour des améliorations à la volée en mode Wiki et de façon incrémentale, comme les traductions de documents.

3.2.2 Suggestions proactives

Les suggestions traductionnelles que nous tentons d'offrir aux traducteurs doivent être bien définies.

Que veut dire « proactivité » de façon précise dans notre contexte ?

Les deux points suivants permettent de clarifier cette notion.

- Les traductions de tous les « segments » devront être calculées d'avance par recherche dans la MT, appel à la TA et recherche sur le Web, suivis d'une évaluation. La zone de traduction sera initialisée à la « meilleure » traduction, et les 3 ou 4 premières suggestions seront montrées dans la zone des « suggestions ».

- Les informations dictionnairiques auront aussi été cherchées à l'avance, et seront présentées après fusion des parties identiques trouvées dans différents dictionnaires. On pourra aussi associer dynamiquement des raccourcis clavier aux différents équivalents pour faciliter l'insertion au cours de la frappe.

La proactivité sera couplée au processus de traduction dans l'éditeur. Au cours de la traduction, le segment actif (segment en édition) sera la base de la recherche dans les ressources.

3.2.3 Fonctions adaptées

Selon les étapes (traduction, révision, etc.), les traducteurs ont besoin de fonctions adaptées à chaque situation. Par exemple, la révision ou la post-édition nécessite des corrections qui peuvent porter sur un ensemble restreint de données ou sur des données à grande échelle.

Les fonctions de manipulation globale sont nécessaires. Par exemple, lors de la traduction du corpus BTEC en français (Boitet, 2004), en partant de la traduction par Systran-4, il faut remplacer partout « s.v.p » par « s'il vous plaît », c'est-à-dire dans plus de 8 000 phrases (5% de 163 000 phrases).

Il faut donc une fonction « chercher/remplacer » sur tout un document, voire sur un ensemble de documents en mode Wiki. On peut imaginer aussi de proposer dans un menu surgissant les actions d'édition déjà effectuées et applicables à l'endroit où est le curseur, et/ou sur la sélection en cours.

Outre l'édition globale, les fonctions dont nous parlons peuvent être : import et amélioration des dictionnaires, bascule entre les langues, ajout d'une nouvelle langue, sélection par descripteurs (structures permettant de filtrer les données à visualiser, cf. *Descripteurs riches*, p. 201), etc.

L'adaptation de ces fonctionnalités à la gestion d'une masse de données sera présentée un peu plus loin lorsqu'on abordera les problèmes de masse de données traductionnelles (cf. *Traduction de masses de données*, p. 64).

3.2.4 Personnalisation pré-traductrice

Les fonctions identifiées ci-dessus doivent être contrôlées par des paramétrages. Selon le contexte, les traducteurs auront besoin de choisir entre plusieurs dictionnaires. Les suggestions peuvent être proposées par combinaison de la TA et des MT ou séparément. Dans

le cas des MT, les traducteurs auront besoin de paramétrer les taux de similarité ou la méthode de calcul des coïncidences floues.

En ce qui concerne la TA, le paramétrage portera sur le choix des couples de langues, des systèmes de traduction automatique gratuits (exemple Google Translate), des formats, et des codages.

1.3.3 Traduction de masses de données

Travailler sur des MT ouvertes à usage libre en ligne peut leur faire atteindre une taille importante, ce qui peut générer des problèmes similaires à la gestion de corpus parallèles volumineux. C'est pourquoi nous avons élargi nos recherches à la gestion de corpus parallèles.

En effet, trouver des corpus parallèles et les préparer pour des applications en TAL reste un défi. Par exemple, en TA « experte », il suffit de quelques dizaines ou centaines de pages en langue source, et de ressources terminologiques bilingues ou multilingues, pour guider le développement d'un système. En TA « statistique », il faut typiquement 50M à 200M mots alignés (soit 200K à 800K pages). En TA « par l'exemple » utilisant des bi-textes « préparés », les tailles nécessaires (selon la méthode) sont moins grandes, par exemple 30 000 phrases, soit environ 2 000 pages (Boitet, 2007).

Les corpus restent des ressources linguistiques précieuses. Cependant, on ne trouve que peu d'outils permettant de gérer la masse des données issues des corpus. Ces chiffres reflètent les besoins de créer et de gérer des corpus de grande taille. Construire à partir de zéro des corpus en TA est très coûteux et nécessite des années d'investissement. Donc, nous pensons qu'on peut réduire l'impact de ces problèmes par la mutualisation des corpus.

D'une part, nous souhaitons traduire des données volumineuses en ayant un accès rapide à un environnement centralisé favorisant la mutualisation de la traduction et le partage de ressources linguistiques, et d'autre part permettre l'automatisation des aides linguistiques et la contribution des bénévoles.

L'environnement d'aide à la traduction bénévole sera étendu à la gestion de masses de données destinées à la TA.

3.3.1 Spécificité et diversité

Le traitement de masses de données est un sujet d'actualité. On utilise maintenant des corpus gigantesques en TA empirique directe, surtout en TA statistique, et de gros corpus

enrichis par des annotations plus ou moins complexes dans des variantes de TA « par les exemples ». De gros corpus multilingues contenant des représentations interlingues UNL (Universal Networking Language) sont aussi construits pour des systèmes de TA « experte », par règles et dictionnaires, pour construire des enconvertisseurs et des déconvertisseurs (Wang-Ju, 2004). Ces corpus sont variés, et différents des corpus collectés ou construits pour des études et recherches littéraires ou linguistiques. Seules les « mémoires de traductions » utilisées dans des outils d'aide aux traducteurs sont directement utilisables par la TA.

Par rapport aux corpus pour les études en littérature et en linguistique quantitative, ou pour la reconnaissance de la parole, et même par rapport aux MT (mémoires de traduction), on observe les différences suivantes (Boitet, 2007) :

- parallélisme : les corpus sont alignés au niveau des « segments » (phrases ou titres à l'écrit, phrases ou tours de parole à l'oral),
- nombre des langues et systèmes d'écriture : par exemple, le corpus EuroParl concerne 20 langues européennes (EuroParl, 2007), et presque autant de systèmes d'écriture différents.

D'autres corpus contiennent des annotations linguistiques lourdes qui sont soit internes au texte (balisage au fil du texte), soit externes (structures + correspondances), par exemple des arbres concrets ou abstraits, de constituants ou de dépendances, mononiveau ou multiniveau, ou encore des graphes UNL.

3.3.2 Taille

La diversité des corpus et leur hétérogénéité ne constituent pas les seuls problèmes de gestion d'une masse de données. La taille aussi nécessite d'être prise en considération (par exemple, la gestion instantanée en mémoire d'un grand corpus). Les applications en TAL sont souvent moins performantes lorsqu'il s'agit de traiter des données volumineuses (Kraif, 2001).

Le corpus JRC-Aquis est un corpus volumineux qui se trouve en version libre sur le Web. Il contient dans sa version complète 1 055 583 954 mots en 22 langues[34], ce qui est énorme ! Ce corpus peut nous donner une idée de la nécessité de développer des mécanismes spécifiques pour permettre une gestion efficaces de données volumineuses.

[34] D'autres informations précises existent sur les liens suivants : http://langtech.jrc.it/JRC-Acquis.html, http://europa.eu.int/celex/.

3.3.3 Visualisation parallèle

L'interface doit permettre la visualisation parallèle et l'édition de corpus et aider à la traduction et à la post-édition. Il est rare de trouver des environnements offrant des outils avancés d'aide à la traduction et à la génération de nouveaux énoncés en ligne.

Par exemple, le corpus OPUS est compilé dans un environnement en ligne qui regroupe plusieurs corpus multilingues en 60 langues. La plupart sont recyclés manuellement à partir des documents du système KDE et des traductions de la documentation PHP (Figure 11).

Bien qu'il offre une interface de navigation intéressante pour les corpus parallèles, l'environnement ne peut pas bénéficier des contributions des bénévoles, ni pour en augmenter la multilinguisation ou ni pour la mise à jour du contenu, car la documentation KDE et PHP est en évolution continue.

iso639	da	de	en_GB	es	et	fr	hu	it	ja	nl	nn	pt	pt_BR	ro	ru	sk	sl	sr	sv	tr	uk	wa	xh	zh_TW	iso639
da	-																	test	test	test	test	test	test	test	da
de		-	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	de
en_GB		all	-	test	test	test	test	test	test	test		test		test	test	test	test	test	test	test	test	test	test	test	en_GB
es		all	all	-	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	es
et		all	all	all	-	test	test			test		test			test	test	test								et
fr		all	all	all	all	-	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	fr
hu		all	all	all	all	all	-	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	test	hu
it		all	all	all		all	all	-	test	test		test	test	test	test	test	test	test	test	test	test	test	test	test	it
ja		all	all	all		all	all	all	-	test		test		test	test	test	test	test	test	test	test	test	test	test	ja
nl		all	all	all	all	all	all	all	all	-	test	test		test	test	test	test	test	test	test	test	test	test	test	nl
nn		all		all		all	all			all	-	test				test	test		test		test		test		nn
pt		all	all	all	all	all	all	all	all	all	all	-		test	test	test	test	test	test	test	test	test	test	test	pt
pt_BR		all		all		all	all	all					-		test	test	test		test				test		pt_BR
ro		all	all	all		all	all	all	all	all		all		-	test	test	test	test	test	test	test	test	test	test	ro
ru		all	all	all	all	all	all	all	all	all		all	all	all	-	test	test	test	test	test	test	test	test	test	ru
sk		all	all	all	all	all	all	all	all	all	all	all	all	all	all	-	test	test	test	test	test	test	test	test	sk
sl		all	all	all	all	all	all	all	all	all	all	all	all	all	all	all	-	test	test	test	test	test	test	test	sl
sr	all	all	all	all		all	all	all	all	all		all		all	all	all	all	-	test	test		test	test	test	sr
sv	all	all	all	all		all	all	all	all	all	all	all	all	all	all	all	all	all	-	test	test	test	test	test	sv
tr	all	all	all	all		all	all	all	all	all		all		all	all	all	all	all	all	-	test	test	test	test	tr
uk	all	all	all	all		all	all	all	all	all	all	all		all	all	all	all		all	all	-		test	test	uk
wa	all	all	all	all		all	all	all	all	all		all		all	all	all	all	all	all	all		-	test	test	wa
xh	all	all	all	all		all	all	all	all	all	all	all	all	all	all	all	all	all	all	all	all	all	-	test	xh
zh_TW	all	all	all	all		all	all	all	all	all		all		all	all	all	all	all	all	all	all	all	all	-	zh_TW
iso639	da	de	en_GB	es	et	fr	hu	it	ja	nl	nn	pt	pt_BR	ro	ru	sk	sl	sr	sv	tr	uk	wa	xh	zh_TW	iso639

Figure 11 : Présentation des bitextes « KDE » dans le corpus OPUS

Conclusion

L'état de la traduction professionnelle a été analysé et comparé avec celui de la traduction bénévole. Nous avons montré les avantages apportés par les outils des professionnels, et constaté que les bénévoles ne bénéficient actuellement d'aucun outil leur permettant d'organiser leurs projets et la traduction.

Nous avons identifié les problèmes que nous aborderons dans les chapitres suivants : le recyclage de données traductionnelles, la conception et le développement d'un environnement collaboratif d'aide à la traduction, et enfin l'extension au traitement de masses de données.

Nous abordons dans le chapitre suivant le système QRLex. Notre première expérience pour concrétiser un environnement en ligne destiné aux traducteurs bénévoles japonais a été développé en particulier pour aider la communauté de traducteurs de la *catégorie B* (cf. *Similarités et différences dans les pratiques observées*, p. 47).

Chapitre 2

Aides linguistiques dans l'environnement QRLex

Introduction

Comme nous l'avons montré dans le Chapitre 1 (cf. *Situation et problématique de la traduction bénévole*, p. 21), les communautés de traducteurs humanistes réalisent une grande partie de la traduction bénévole quotidienne sur le Web (W3C, PaxHumana, Traduct, etc.). Des milliers de documents sont traduits dans Wikipedia et ce dans des domaines variés (techniques, média, informations, droits de l'homme, etc.) (Bey, *et al.*, 2006).

Nous avons vu plus haut que le nombre de ces communautés est en augmentation continue. Cependant, la majorité de ces communautés ne dispose pas d'outils d'aide linguistique leur permettant à la fois de s'organiser et d'accélérer les traductions.

D'autre part, nous avons montré que les outils libres d'aide à la traduction sont incapables de répondre aux besoins spécifiques évoqués dans le premier chapitre. Par exemple, OmegaT[35] ne permet pas de gérer la traduction collaborative et ne peut pas être exploité en ligne.

Le projet QRLex a été lancé par le Pr. K. Kageura pour pallier ces manques. Il vise à fournir une plate-forme dédiée centralisant les ressources et partageant les fonctionnalités.[36]

Nous présentons dans ce chapitre l'architecture de l'environnement, et notre contribution aux autres modules ainsi qu'à la recherche et au recyclage de données traductionnelles sur le Web.

2.1 Architecture de l'environnement

L'architecture générale de l'environnement QRlex a été divisée en plusieurs modules[37]. Chacun est défini selon les besoins des traducteurs et est développé séparément pour être

[35] http://www.omegat.org/fr/omegat.html

[36] Il a été financé par la JSPS (société japonaise pour la promotion de la science) japonaise. Nous avons participé durant un an à sa réalisation dans le cadre d'un stage du JASSO (Japan Student Services Organization).

intégré par la suite, parce que plusieurs équipes de recherche sont impliquées dans le projet (Université de Tokyo, Université d'Okayama, Université Joseph Fourier et Université de Nantes).

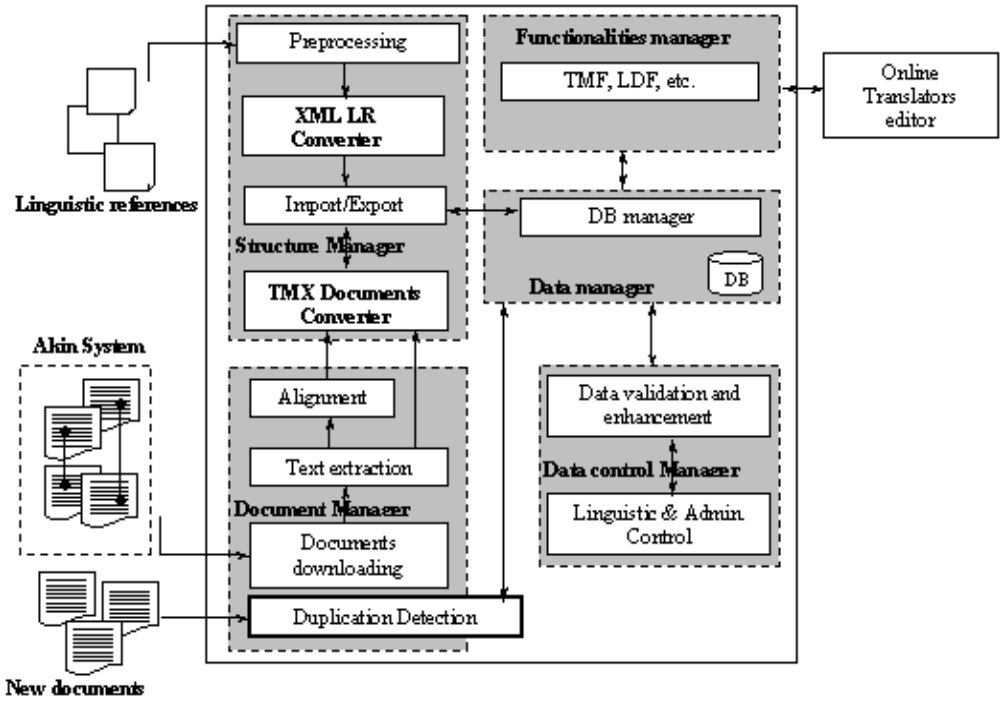


Figure 12 : Architecture générale du système QRLex

Le fonctionnement de l'environnement QRlex repose sur une architecture modulaire (Figure 12). Cinq modules y ont été intégrés pour gérer le flux de données et les interactions avec les utilisateurs. Les fonctionnalités de chaque module sont les suivantes.

- *Module de prétraitement et de structuration de données :* ce module prétraite les ressources linguistiques dans leurs formats originaux et construit des structures XML permettant leur gestion efficace (dictionnaires, glossaires, etc.).

- *Module de recyclage de données traductionnelles :* le recyclage de données traductionnelles se fait par ce module. Le noyau du module est un robot Web fonctionnant à la Google. En partant d'un ou plusieurs mot(s)-clé(s) dans une langue source, il renvoit un ensemble de paires de documents (source, cible).

[37] Au moment de la rédaction de cette thèse, le projet était toujours en cours, en collaboration avec plusieurs équipes, dont l'Université d'Okayama où le serveur sera installé. Les actions en cours étaient la compilation automatique d'une grande banque terminologique japonais-français et japonais-anglais et l'alignement de ressources textuelles telles que des bi-textes (le projet QRlex est financé par la JSPS - Japan Society for the Promotion of Science) et s'est achevé en 2008.

- *Module d'interaction avec les traducteurs « QRedit » :* ce module comprend l'éditeur et les fonctionnalités d'aide à la traduction. C'est grâce à ce module que l'appel aux fonctionnalités linguistiques se fait de façon systématique.

- *Module de gestion des données :* les données sont stockées dans une base de données centrale. Ce module permet la gestion centralisée des données (documents, ressources, utilisateurs) et répond aux requêtes des utilisateurs à travers l'éditeur.

- *Module de contrôle :* ce module permet aux administrateurs et aux utilisateurs de contrôler l'accès aux données. Par exemple, un administrateur a la possibilité d'associer certains privilèges aux utilisateurs ayant des compétences linguistiques élevées (exemple, la validation des révisions après une phase de prétraduction).

2.1.1 Module de prétraitement et de structuration des données

1.1.1 Prétraitements des données lexicales brutes

Le projet est destiné à aider une communauté spécifique, celle des traducteurs humanistes japonais qui produisent des documents en anglais et en japonais. Les ressources du système sont donc limitées à ce couple de langues. Les ressources libres compilées sont libres d'usage (nous n'avons utilisé aucune ressource commerciale). Elles se trouvent dans des formats et des codages différents (formats textuels non structurés). Il est donc naturel de penser à unifier le format et la gestion des données.

Ce module se charge du prétraitement de données diverses. Il inclut le prétraitement des ressources linguistiques (dictionnaires, banques terminologiques, etc.) et les données textuelles (les paires de documents et les segments multilingues alignés). Le module *recyclage de données* lui envoie des données brutes qui les transforme en format XML TMX (Lisa, 2008). En revanche, les ressources linguistiques sont compilées dans le format XLD (XML Linguistic Data) (Bey, 2006).

Le prétraitement consiste à détecter les articles dictionnairiques ainsi que les données pertinentes à partir des ressources textuelles originales et à les unifier sous un format adéquat. L'import dans des structures adaptées permet, entre autres, les fonctionnalités suivantes :

- unification de la gestion de données hétérogènes (Table 4),

- changement de direction de la traduction, et

- import facile de ressources linguistiques entières.

Dans les sections suivantes, nous détaillons les fonctions de chaque module. Nous expliquons le format XLD ainsi que ses propriétés qui nous ont permis de compiler des milliers d'entrées dictionnairiques.

1.1.2 Restructuration des données linguistiques

Pour faciliter l'import et la structuration de données hétérogènes, nous avons opté pour XML, une technologie qui a montré son succès dans plusieurs projets de gestion de données linguistiques, entre autres, le projet Papillon[38] et le dictionnaire japonais-anglais *Grand Concise*[39].

En effet, les ressources linguistiques du projet QRLex sont importées à partir de leurs formats textuels bruts en format XML. Toutes les ressources citées dans la Table 4 (pp. 80) sont compilées dans ce format.

Le module de gestion de structures permet de récupérer et de structurer en XML les articles présents dans chaque ressource.

2.1.2 Stockage centralisé

Ce module est un serveur de données qui, à travers les requêtes des utilisateurs, renvoie un ensemble d'entrées dictionnairiques ou textuelles, qui correspondent au contexte de la traduction en cours.

Figure 13 : Schéma du dictionnaire « Eijro » dans la BD

[38] http://www.papillon-dictionary.org/Home.po?lang=jpn
[39] http://www.gally.net/translation/GCJE.htm

Techniquement, cette centralisation est faite grâce à une architecture client/serveur : un conteneur de Servlet *Tomcat*, et un ensemble de fonctions implémentées en JSP et JAVA. L'interactivité est implémentée principalement en se basant sur *Javascript* et *Ajax*. Enfin, le stockage au niveau physique se fait dans une BD relationnelle pilotée par un SGBD de type *MySql*.

2.1.3 QRselect : module de recyclage des données traductionnelles

Le recyclage consiste à récupérer les données traductionnelles à partir de documents bilingues disséminés sur le Web. Ces documents peuvent êtres regroupés dans des « sites amis », ou aléatoirement dispersés sur le Web entier. Deux fonctionnalités doivent être détaillées, car elles constituent la base de ce module : le recyclage interne de documents, externes, et le couplage avec *l'éditeur bilingue* du système *QRLex*.

1.3.1 Recyclage externes

Le recyclage *externe* sert à la détection de documents déjà traduits sur le Web entier. Le but du recyclage est double. D'un côté, il permet aux traducteurs d'éviter une duplication de traduction d'un document déjà traduit. De l'autre côté, il permet de construire des MT par détection des UT bilingues.

Les algorithmes et des méthodes qui nous ont permis d'implémenter ce processus seront exposés de façon détaillée un peu plus loin.

1.3.2 Recyclage interne : le cas des sites amis

Le recyclage *interne* consiste à détecter les couples de documents qui constituent des traductions mutuelles sur un site Web spécifique, plutôt que sur le Web entier. Les paires de documents détectés sont alignées pour construire les MT (Ido, *et al.*, 1993).

L'idée du recyclage est motivée par le fait que la plupart des communautés de traducteurs n'ont pas à leur disposition une MT, alors qu'elles produisent des traductions qui peuvent être exploitées pour leur proposer des suggestions de traduction.

2.1.4 Le module QRedit : un éditeur couplé à des fonctionnalités traductionnelles

L'éditeur *QRedit* est une interface avec laquelle les traducteurs bénévoles communiquent avec QRLex. Il intègre des fonctionnalités linguistiques utiles telles que la consultation et la mise à jour des dictionnaires, et la traduction de documents proprement dit. Il est possible avec *QRedit* de lancer le recyclage pour détecter les documents existant sur le Web. Cette

fonctionnalité sera détaillée dans la deuxième partie lorsqu'on abordera le module de recyclage *QRselect*.

Les différentes fonctionnalités de l'éditeur sont présentées dans les paragraphes suivants.

1.4.1 Chargement des documents

Dans QRedit, un document source à traduire doit être en anglais, car il ne prend en charge par défaut que l'anglais comme langue source. Il ne peut être que du texte pur ou du HTML introduit par un lien Web. Le texte est donc extrait et segmenté en une suite d'unités de traduction (généralement des phrases) par des expressions régulières simples basées sur des symboles de ponctuation. Les UT sont synchronisées dans des zones source et cible pour garder l'alignement durant la traduction, lors de laquelle sont visualisées en mode document entier. La traduction se fait donc sur une seule UT à la fois, mais, pendant la traduction, les traducteurs ont une vue générale sur le document.

1.4.2 Aides dictionnairiques

Les aides linguistiques sont principalement intégrées dans l'éditeur au moment du chargement d'un document donné. Ces aides sont de quatre types :

(i) Pour les mots ordinaires : les traducteurs sont souvent satisfaits par les informations fournies par les dictionnaires existants. Le fait d'introduire ces informations dans un contexte traductionnel permet aux traducteurs de ne pas avoir besoin de consulter des références externes, ce qui réduit le temps de traduction.

(ii) Pour les expressions et les idiomes : les identifier est encore plus important. Les traducteurs sont souvent satisfaits par les propositions des dictionnaires existants, mais la recherche est souvent longue, pénible, avec un risque de « manquer » l'expression cherchée, ce qui conduit presque toujours à une mauvaise traduction.

(iii) Pour les termes techniques : il faut soit les rechercher et afficher leurs traductions, soit construire par les contributions de traducteurs bénévoles de nouvelles entrées terminologiques.

(iv) Pour les noms propres : il faut aussi intégrer un dictionnaire pour proposer des translittérations de noms propres. Les traducteurs ne sont pas satisfaits par les ressources de référence disponibles, et ont beaucoup de difficulté à identifier les références. Pour cela, ils consultent Internet mais ne trouvent pas toujours une traduction adéquate (Kageura, 2006).

1.4.3 Méthodes de suggestion dictionnairique

Les suggestions sont proposées selon deux façons :

- le contenu complet d'un article correspondant à la sélection en cours (texte source) ;
- les informations d'alerte (informations complémentaires d'orientation destinées aux traducteurs).

Si une traduction n'existe pas dans une ressource dictionnairique, les alertes servent à éviter aux traducteurs de lancer des recherches inutiles. Dans le cas où les articles existent, les alertes permettent aussi de prévenir les traducteurs de leur existence, ce qui les motive à les consulter.

Il est important de prévenir les traducteurs de l'existence ou de l'absence des articles ; dans les deux cas, les traducteurs ont un gain.

Le chargement d'un document dans QRlex active automatiquement la recherche des traductions de mots et d'expressions.

L'affichage de ces aides est fait de la façon suivante (Figure 24) :

- un affichage simplifié (information rapide) ;
- un affichage complet (tout le détail de la traduction).

La sélection d'une traduction à partir des informations affichées se fait de façon très conviviale. Une simple sélection par un clic de la souris insère la traduction dans la zone cible sans avoir perturbé le traducteur dans son rythme de traduction.

2.2 Aide à la traduction : couplage de références linguistiques

2.2.1 Approches existantes pour la récupération des dictionnaires

2.1.1 L'approche « RÉCUPDIC » et « PRODUCDIC »

Les deux outils RÉCUPDIC et PRODUCDIC ont été développés par H. Doan-Nguyen (Doan-Nguyen, 1998) dans le but de récupérer des dictionnaires de façon générique et de produire des « brouillons » de nouveaux dictionnaires en combinant des opérations de base (on peut par exemple inverser un dictionnaire ou combiner deux dictionnaires en parallèle ou en séquence).

La méthode à laquelle est arrivée H. Doan-Nguyen est en deux étapes : normalisation (*ad hoc*), puis compilation vers la forme interne souhaitée à l'aide d'un parseur écrit en H-grammar (Doan-Nguyen, 1998).

L'utilisateur écrit en H-grammar la grammaire du dictionnaire normalisé. Il ajoute ensuite les actions de construction d'objets et de détection d'erreurs.

La détection d'erreurs permet de corriger automatiquement les erreurs les plus fréquentes. Si un détail est faux dans un article, celui-ci n'est pas rejeté en bloc. Le « moteur » de RÉCUPDIC utilise ensuite la description compilée de la grammaire et des actions pour construire l'ensemble d'objets constituant la représentation structurée souhaitée du dictionnaire (Mangeot-Lerebours, 2001).

.COM Command (file name extension) + Commercial Business (Domain Name)	(BABEL (HWD . ".COM") (BODY LIST (SENSE (EXPS . "Command") (EXPL . "file name extension") (SUBJ . NIL)) (SENSE (EXPS . "Commercial Business") (EXPL . "Domain Name") (SUBJ . "Internet"))))

Figure 14 : Article de BABEL après récupération (objet LISP).

Le deuxième outil créé par H. Doan-Nguyen, PRODUCDIC, permet de produire de nouveaux ensembles lexicaux en utilisant des calculs ensemblistes. On peut par exemple produire un troisième dictionnaire bilingue (A-C) à partir de deux dictionnaires bilingues A-B et B-C. Plusieurs algorithmes ont été développés dans le cadre de sa thèse pour implémenter différentes techniques telles que la sélection, le chaînage, et la combinaison parallèle.

Pour ces deux systèmes, il manque actuellement une interface utilisable par un linguiste. Il se pose donc un problème de mise en œuvre pratique. Il faut donc continuer et améliorer cette technique pour la rendre utilisable par un non-informaticien.

L'outil RÉCUPDIC a permis de récupérer plus de 1 650 000 articles provenant de 20 sources dans 12 langues différentes, et PRODUCDIC a permis de produire 543 000 articles (le tout en à peu près un an à 40% de temps).

L'environnement dans lequel ont été implémentés ces outils est très puissant, mais il n'est pas adapté à un linguiste (lexicographe, lexicologue), donc il a été l'objet d'améliorations

dans le cadre du projet Papillon pour l'adaptation à un usage universel – minimiser les contraintes techniques pour les linguistes et lexicographes généralement non-informaticiens (Mangeot-Lerebours, 2001).

2.1.2 Approche Papillon

L'idée de créer une base lexicale multilingue générique et collaborative à grande couverture est apparue vers 1981 au GETA (GETALP depuis 2007) dans la cadre du projet TAO « ESOPE ». Après diverses études et prototypes (DBTAO, PARAX, SUBLIM) le projet Papillon a été lancé avec le NII en août 2000 et a conduit à la réalisation d'une plate-forme lexicale multilingue contributive en deux parties, Papillon-CDM et Papillon-NADIA.

La base de la partie « Papillon-CDM » contient environ 2M d'entrées dans 9 langues (anglais, français, japonais, lao, thaï, chinois, allemand, malais et vietnamien) provenant de dictionnaires « contribués » en licence GPL. Elle est libre d'usage et accessible en ligne. Son atout est de présenter les dictionnaires sous une forme unifiée, et pour différents usages. Les utilisateurs sont invités et encouragés à l'enrichir par des contributions qui sont temporairement stockées, pour être validées ensuite par les experts lexicographes.

Le noyau de l'environnement est fondé sur le système SUBLIM développé par G. Sérasset en 1994 (Sérasset, 1994) dont le but était le développement d'un système universel pour les bases lexicales multilingues. En SUBLIM, la description d'une base multilingue se fait par deux langages de haut niveau, LEXARD et LINGARD.

- LEXARD permet à l'utilisateur de définir la macrostructure de sa base, consistant en l'ensemble des dictionnaires de base, leurs types (monolingue, bilingue, interlingue) et leurs liens.
- LINGARD permet de définir la microstructure des dictionnaires. Pour chaque dictionnaire, on décrit la structure informatique des articles. Pour cela, on utilise des types de base : arbre, graphe, automate, structure de traits, liste, ensemble, énumération, etc.

Dans le projet Papillon, ces langages ont été traduits en XML, beaucoup plus verbeux, mais associé à beaucoup d'outils puissants et gratuits diffusés par le W3C.

Pour pouvoir manipuler et fusionner certaines parties des ressources, M. Mangeot a défini CDM (Mangeot-Lerebours, 2001), un formalisme commun de représentation de dictionnaires. Pour cela, il a identifié les types des informations contenues dans les dictionnaires accessibles

sous forme informatique et libres de droits, et les a nommés de façon unique dans l'espace de noms CDM (*Common Dictionary Markup*)[40].

éléments CDM	FeM	DHO	NODE
<entry>	<fem-entry>	<se>	<se>
<headword>	<entry>	<hw>	<hw>
<pronunciation>	<french_pron>	<pr><ph>	<pr><ph>
<etymology>			<etym>
<syntactic-cat>		<sense n=1>	<s1>
<pos>	<french_cat>	<pos>	<ps>
<lexie>		<sense n=2>	<s2>
<indicator>	<gloss>	<id>	
<label>	<label>		<la>
<example>	<french_sentence>	<ex>	<ex>
<definition>			<df>
<translation>	<english_equ><malay_equ>		<tr>
<collocate>		<co>	
<link>	<cross_ref_entry>	<xr>	<xg>/<vg>
<note>		<ann>	

Table 3 : Équivalents des éléments CDM dans le FeM, le DHO et le NODE

L'ensemble hiérarchisé CDM contient les éléments les plus courants trouvés dans ces ressources, à savoir le mot-vedette, la prononciation, la catégorie grammaticale, le vocable, la lexie, l'étymologie, les exemples, les étiquettes, les gloses, etc. Ces éléments ont toujours la même sémantique. Par exemple, <dml:entry> réfère toujours à un article et <dml:headword> au mot-vedette de l'article.

Pour certains éléments ayant des listes fermées de valeurs, CDM contient la définition d'une liste représentant l'ensemble des valeurs, et des règles de conversion pour chaque ressource et chaque langue concernée.

Un exemple est la liste des catégories grammaticales d'une langue. Lors de la récupération d'une ressource existante, les éléments originaux sont convertis vers des éléments de cet ensemble. Si toutefois certaines informations ne sont pas représentables avec cet ensemble, les éléments originaux sont conservés.

[40] CDM provient principalement de l'examen détaillé des dictionnaires FeM, DEC, DHO, OUPES, NODE, EDict, de la base ELRA-MÉMODATA, et du chapitre 12 de la structure TEI concernant les dictionnaires.

Dans notre travail, les éléments de l'ensemble CDM sont utilisés comme points de référence dans chaque dictionnaire converti (Table 3). La correspondance entre un élément de cet ensemble et un élément original lors de la récupération est définie au départ par un linguiste (Mangeot-Lerebours, 2002). Ces éléments ont été choisis sur la base de leur fréquence. L'ensemble lui-même peut évoluer si de nouveaux dictionnaires sont explorés et récupérés et font apparaître un nouveau type d'élément d'information.

L'état de l'art présenté ci-dessus n'a fait ressortir aucun nouveau problème qu'il faudrait résoudre pour nos besoins. Mais ce tour d'horizon sur ces méthodes de récupération nous a donné une idée sur le contenu des dictionnaires et la façon dont ils pourront être récupérés dans leurs formats originaux.

Les informations dictionnairiques et structurelles de CDM ont été adaptées pour construire notre format XLD.

2.2.2 Traitement des ressources dictionnairiques dans QRLex

2.2.1 Description de notre approche

L'approche de la compilation de dictionnaires dans *QRLex* est très simple. On étudie les dictionnaires manuellement, et ensuite on écrit des analyseurs pour les compiler dans un format *ad hoc* que nous avons appelé *XLD* (Bey, 2006). C'est la même méthode que celle de H. Doan-Nguyen (Doan-Nguyen, 1998), sauf que nous n'utilisons pas H-grammar, mais nos propres outils, basés sur XML et Java. Cette structure est basée sur les informations dictionnairiques de CDM (Common Dictionary Markup) du projet Papillon.

Il ne s'agit pas pour nous de fabriquer un dictionnaire unique composé de tous les dictionnaires. Les dictionnaires à importer restent autonomes. La base de données centrale de QRLex gère plusieurs dictionnaires différents.

Les dictionnaires de QRLex sont différents et spécialisés, mais compilés sous la même structure (c'est la spécialisation qui nous incite à ne pas les mettre ensemble dans le même volume).

Les traducteurs choisissent les dictionnaires à utiliser, et ce choix dépend effectivement de la nature du document à traduire. C'est exactement la même méthode que celle proposée par Jim Breen (Breen, 1995). Donc, on a le choix de spécifier des dictionnaires ou d'utiliser le tout lorsqu'on lance une requête. De plus, il y a une propriété très importante que les bénévoles ou tous les groupes en général veulent avoir : c'est de pouvoir de gérer leurs

ressources linguistiques séparément, et en plus, de créer un petit dictionnaire partagé pour gérer des traductions précises qui seront pas utiles aux prochaines traductions.

2.2.2 Sélection des dictionnaires

Les ressources linguistiques choisies et compilées dans le cadre du projet QRLex sont actuellement (Table 4) les suivants.

Données de référence	Description	Direction	Entrées
Eijiro	Dictionnaire général libre (EDP 2005)	EN-JP	1 640 000
Edict	Dictionnaire général libre	JP-EN	125 000
Nichigai	Guide de prononciation des mots étranger en Katakana	JP-EN	112 679
Termes scientifiques médicaux	Banque terminologique médicale	JP-EN	211 165
Total			**2 088 844**

Table 4 : Ressources linguistiques du projet QRLex[41]

- « Eijiro » : un dictionnaire bilingue unidirectionnel (anglais→japonais) de grande qualité qui est utilisé par beaucoup de traducteurs bénévoles japonais (1 576 138 entrées).
- « Nichigai » : un guide de prononciation des noms propres étrangers en Katakana (112 679 entrées).
- « Termes scientifiques médicaux » : nous les avons inclus pour tester la gestion de la structure des dictionnaires terminologiques (211 165 entrées).
- « Edict » : c'est un dictionnaire libre inclus pour renforcer le nombre des d'entrées traitées (112 898).

[41] L'utilisation de dictionnaires libres *Edict* et *Eijiro* permet d'avoir un total d'environ 1,7M entrées. Ce volume est 10 à 20 fois plus grand que celui des dictionnaires compilés dans le "denshi jisho" (http://www.japaneselanguagetools.com/index.html).

Figure 15 : Guide de prononciation des noms propres en katakanas

2.2.3 Quelques problèmes liés aux spécifiés structurelles

Les ressources de référence en format électronique sont compilées à partir d'un ensemble de ressources libres qui se présentent dans des formats hétérogènes.

Après une analyse des standards XML existants utilisés pour la gestion terminologique, tels que TBX (TermBase eXchange) et MARTIF (Machine-Readable Terminology Interchange Format), nous avons conclus que ces standards ne satisfaisaient pas les besoins cités dans les sections précédentes. Quant au format CDM, nous l'avons trouvé beaucoup trop détaillé pour nos besoins (Mangeot-Lerebours, 2002). D'ailleurs, le projet *Papillon* n'a jamais été expérimenté en collaboration par des bénévoles, car la structure des entrées est trop riche et demande des compétences linguistiques avancées.

Pour la satisfaction des besoins et critères cités ci-dessus, nous avons géré de façon unifiée les ressources linguistiques dans des structures XML que nous avons créées nous-mêmes ; cela nous permet de faciliter l'échange entre les différents modules et d'obtenir les propriétés suivantes.

- Les différents niveaux des unités recyclables sont traités dans un cadre unifié.

- Le contenu riche existant est accompagné par des fonctionnalités riches et à accès contextuel (par exemple, détection automatique des idiomes).

- Les informations non nécessaires sont exclues et les informations pertinentes pour les traducteurs sont incorporées.

- La gestion de données hétérogènes est possible (entrées phraséologiques, lexicographiques, terminologiques, etc.).

Dictionnaire	Format d'entrées
Eijiro	$__ annual membership : __ドルの年会費｛ねんかいひ｝
	$__ deposit required for making a bid on：～に入札｛にゅうさつ｝するため に必要｛ひつよう｝な__ドルの保証金｛ほしょうきん｝
Edict	仝/(n) "as above" mark/
	〆切 [しめきり] /(n) closing/cut-off/end/deadline/Closed/No Entrance/
Nichigai	アーガス　Ergas
	アーガルド Aagaard
Termes scientifiques médicaux	AST940035120　母惑星　parent planet
	AST940035130a　パリティ parity

Table 5 : Exemple d'entrées de référence dans QRLex

Il reste cependant des problèmes d'import, de gestion de données multilingues, et l'adaptation de diverses fonctionnalités : tri, suppression et fusion de duplications, etc.

2.2.3 Gestion unifiée des ressources linguistiques

2.3.1 Description de la méthode de normalisation

Les dictionnaires sélectionnés ont été étudiés séparément. Des scripts *ad hoc* ont été développés pour les importer et réaliser, de plus, les deux tâches suivantes :

(i) changement de direction de la traduction, car on souhaite traduire de l'anglais vers le japonais.

(ii) import automatique des dictionnaires dans le format XLD.

Les scripts ont servi à filtrer le contenu pour ne conserver que les informations utiles et changer la direction de certains dictionnaires (JP→EN en EN→JP), et enfin pour l'unification de l'encodage des données.

Le projet QRlex suppose que les traductions se font de l'anglais vers le japonais. Le changement de direction des dictionnaires est donc important du point de vue de l'utilisabilité pour les dictionnaires dont la langue source est le japonais.

Cependant, le passage d'une direction à une autre nécessite de réorganiser les données en les retriant et en éliminant les doublons des nouveaux dictionnaires, ainsi que les informations inutiles.

Le tri a été réalisé par un algorithme de tri rapide classique. En une seule passe, l'algorithme produit un dictionnaire trié, mais sans avoir éliminé les doublons. Le tri nous permet de les détecter facilement. Nous les éliminons automatiquement et ne gardons qu'un seul mot vedette, et en fusionant les traductions qui ont été séparées par une tabulation.

Le changement de direction des dictionnaires de QRLex n'est pas symétrique. Pour la base terminologique « Nichigai » et le dictionnaire des prononciations en katakanas, le changement de direction est à 100% symétrique car les traductions d'un article ne sont pas multiples. Par contre, le dictionnaire EDICT présente des informations différentes et variées : un mot vedette peut avoir plusieurs traductions, plusieurs catégories syntaxiques, et des exemples dans la partie traduction d'un article.

Bien que ces problèmes ne soient pas difficiles à résoudre, il existe beaucoup d'autres problèmes plus complexes liés directement au domaine de la lexicographie. Ce dernier n'est pas l'objet de notre thèse ; de plus, la majorité des problèmes sont déjà abordés dans le projet Papillon (Mangeot-Lerebours, 2001) et d'autres travaux tels que ceux de Jim Breen (Breen, 1995).

Notre méthode est *ad hoc*, elle ne vise aucune généricité, elle opte pour la simplicité et la spécificité, et se focalise surtout sur le traitement d'un volume dictionnairique bien déterminé d'avance.

Cette méthode nous a permis d'importer un total d'environ 1,7M entrées à partir des dictionnaires sélectionnés cités dans la Table 4.

Nous pourions en importer plus, mais la plus grande difficulté est de trouver des dictionnaires libres de droits, dans un format « propre », et adapté aux couples de langues et aux domaines visés.

2.3.2 Structuration en XML : le format XLD

Le format XLD (XML Linguistic Data) a été défini par une DTD[42] (Document Type Definition) dans laquelle trois parties ont été introduites : métadonnées, source, et cible.

- Métadonnées : c'est un ensemble d'informations décrivant le contenu d'une ressource linguistique donnée et nécessaires pour suivre son évolution et sa mise à jour (Figure 16).

<!ELEMENT	resource	(res-info, content)>
<!ELEMENT	res-info	(#PCDATA >
<!ATTLIST	res-info res-name	CDATA #REQUIRED>
<!ATTLIST	res-info author	CDATA #IMPLIED>
<!ATTLIST	res-info version	CDATA #IMPLIED>
<!ATTLIST	res-info dateCreation	CDATA #IMPLIED>
<!ATTLIST	res-info lastModification	CDATA #IMPLIED>
<!ATTLIST	res-info originalEncoding	CDATA #IMPLIED>
<!ATTLIST	res-info entriesNumber	CDATA #IMPLIED>
<!ATTLIST	res-info description	CDATA #IMPLIED>

Figure 16 : Les métadonnées dans XLD

- La source est est ensemble d'informations décrivant le mot vedette dans sa langue source (Figure 17).

<!ELEMENT	source	(headword, pronunciation, syntacticCat)>
<!ELEMENT	headword	(#PCDATA)>
<!ATTLIST	source xml:lang	CDATA #REQUIRED>
<!ELEMENT	pronunciation	(#PCDATA)>
<!ELEMENT	syntacticCat	(#PCDATA)>

Figure 17 : Définition de l'élément « source » dans XLD

- La cible est un ensemble d'informations décrivant et traduisant la source. Elle contient d'autres informations telles que les attributs : la prononciation, la langue cible, quelques exemples s'ils existaient dans la source, etc. D'autres informations lexicales sont aussi nécessaires pour rendre la structure des articles complète, telles que le domaine, la catégorie syntaxique, l'étymologie, etc. Notons

[42] http://www.w3schools.com/Xml/xml_dtd.asp

qu’une grande majorité de ces items ont été hérités du projet Papillon et du format CDM (Mangeot-Lerebours, 2001).

<!ELEMENT	target	(translation+, definition, etymology, syntacticCat, example*, collocate*, wikiLink, note)>
<!ATTLIST	target xml:lang	CDATA #REQUIRED>
<!ELEMENT	translation	CDATA #IMPLIED>
<!ATTLIST	id	CDATA #REQUIRED>
<!ELEMENT	definition	CDATA #IMPLIED>
<!ELEMENT	etymology	CDATA #IMPLIED>
<!ELEMENT	syntacticCat	CDATA #IMPLIED>
<!ELEMENT	example	CDATA #IMPLIED>
<!ELEMENT	collocate	CDATA #IMPLIED>
<!ELEMENT	wikiLink	CDATA #IMPLIED>
<!ELEMENT	note	CDATA #IMPLIED>

Figure 18 : Définition de l’élément « cible » dans XLD

Les représentations d’origine des entrées dictionnairiques de certaines ressources ne sont pas homogènes. On trouve des articles avec des mots vedettes simples ou sous forme de segments.

Par exemple, dans le dictionnaire Eijiro on trouve des mots vedettes tels que « detect », « tweak » et « semivocalic » traduits par « 見つける 発見する / 見破る / 感知する / かぎつける / 検出する », « 動揺 / 心痛 / ひねり / エラーや失敗ばっかりする選手 / ひねる » et « 半母音の［に関する］», respectivement.

Mais on peut aussi trouver des entrées sous forme de segments, comme « You will wait 24 hours before activating this service. » traduit en « このサービスが使用可能になるには 24 時間待ってください。 ».

À l’inverse des structures dictionnairiques usuelles, XLD peut contenir des bisegments. Sur certains aspects, elle ressemble à une MT en gardant sa forme dictionnairique. Ce mixage n’est en fait pas idéal, car les fonctionnalités de traitement des dictionnaires et des MT ne sont pas les mêmes (par exemple, TBX pour gérer les dictionnaires et TMX pour les MT). Mais nous avons été obligé de faire ce mixage car le dictionnaire Eijiro contenait des milliers d’entrées sous forme de bisegments. Le format XLD permet donc non seulement de représenter un ou plusieurs bisegments à l’intérieur d’un article, comme CDM (pour les idiomes et les exemples), mais aussi, ce que ne permet pas CDM, d’avoir des entrées réduites à un bisegment.

2.3.3 Avantages du format XLD dans le cadre de QRLex

Les représentations internes des articles sont dotées d'identificateurs uniques et d'autres qui servent à identifier les différentes traductions des articles, s'il y en a plusieurs.

Cependant, la réutilisation des éléments hérités du format CDM (projet Papillon) permet aux traducteurs bénévoles d'avoir des dictionnaires étendus et enrichis. Pour le travail collaboratif, il s'avère nécessaire d'inclure d'autres informations. Par exemple, l'introduction de liens Wiki « wikiLink » qui sont hérités directement des Wiki permet de dynamiser et d'enrichir le contenu des dictionnaires par les contributions des bénévoles. La présence de ces liens permet de créer des dictionnaires multilingues sans passer par des structures complexes.

La Figure 19 illustre ces liens dans le dictionnaire Wiktionary. On constate deux types de liens : *actifs* et *inactifs*. Les liens actifs pointent sur des pages existantes en relation avec l'article « ولد »[43], tandis que les liens inactifs pointent sur un contenu vide et doivent être complétés par les bénévoles, car dans ce Wiki l'édition est accessible sans aucune restriction.

Par l'introduction de ces liens à la Wiki, XLD permet de lier des articles d'un dictionnaire donné à d'autres articles provenant d'autres dictionnaires, suivant le même concept que celui de Papillon, mais avec une réalisation totalement différente. L'avantage de la nôtre réside dans l'ergonomie qu'apportent les Wiki à l'utilisateur et dans la structuration simple des données.

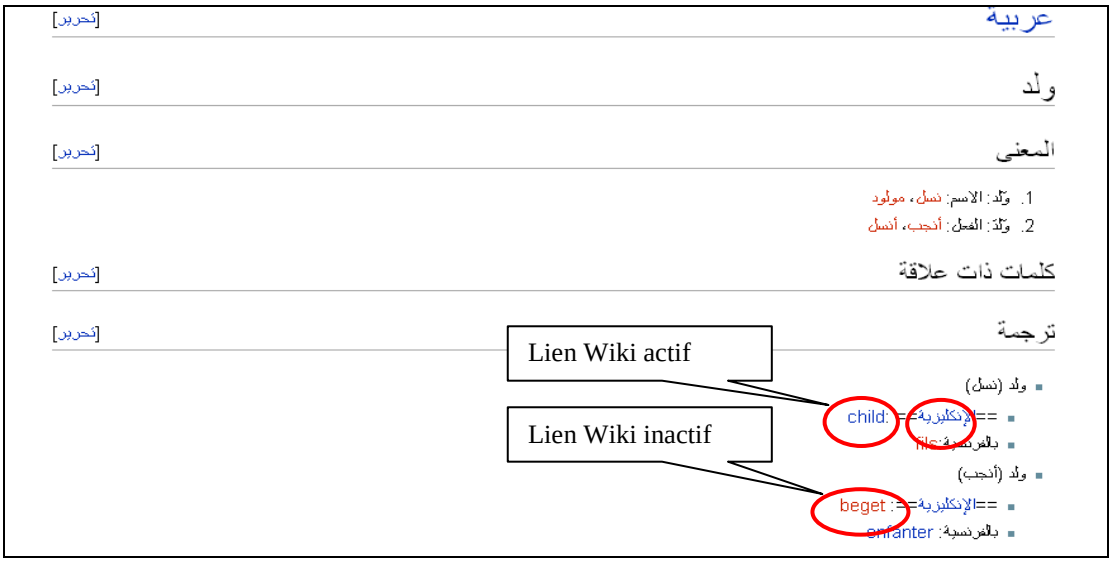


Figure 19 : Types des liens Wiki dans le Wiktionary

[43] « walada », « waladone », « wolida », etc. sont quelques translitérations du mot arabe «ولد ».

Nous verrons dans la partie II, consacrée à la conception de l'environnement BEYTrans, comment il est possible avec cette structure d'importer d'autres dictionnaires entiers en mode collaboratif et à la Wiki, chaque article devenant un document autonome enrichi par des liens Web.

2.3.4 Accès aux ressources

Les articles originaux importés dans XLD sont stockés dans une BD centrale qui fournit les données à tous les autres modules, en particulier l'éditeur avec lequel les interactions se font avec les utilisateurs. Les données sont un flux de données XML bilingues extraites, lors d'une requête, des dictionnaires stockés dans la BD.

La consultation de ressources peut être en mode « proactif » ou « manuel ».

Avec la consultation « proactive », les suggestions dictionnairiques sont proposées de façon systématique et sans intervention de l'utilisateur, en lui donnant le choix entre plusieurs traductions en provenance de plusieurs dictionnaires.

Par contre, avec la consultation manuelle, les ressources sont manipulées par l'intervention des utilisateurs.

Ces deux modes d'accès seront expliqués dans le chapitre suivant. À titre d'exemple, l'éditeur propose des suggestions de façon « proactive » à partir des dictionnaires, de la MT ou de la TA, etc. par la recherche de traductions les plus probables d'un segment source.

2.3 Recyclage de données traductionnelles : le cas du Web

La quantité d'informations sur le Web est en expansion continue et rapide, et devient une source précieuse où une grande quantité de données traductionnelles est disséminée quotidiennement en plusieurs langues par les bénévoles. Pour les aider à les détecter efficacement, il faut implémenter des méthodes de recyclage.

Les techniques de recyclage assimilent le Web à un « entrepôt de données » gigantesque et se basent sur des robots Web et des modules complémentaires pour construire des ressources linguistiques telles que les corpus unilingues, bilingues ou multilingues, et collecter d'autres informations linguistiques multilingues telles que les bases terminologiques spécialisées, les dictionnaires, les glossaires, etc.

Une grande partie des travaux sur le recyclage du Web se limite au recyclage de corpus parallèles (Resnik, *et al.*, 2003) (Chen, 2000), qui sont coûteux et rares. Mais les traducteurs

bénévoles sont peu dotés d'outils de recyclage et ne profitent que peu des traductions existantes sur le Web.

Il en résulte un besoin réel de recyclage de données traductionnelles disséminées sur la Toile. Le recyclage ne doit pas être fait pour prouver l'existence de ces traductions, mais plutôt pour augmenter la productivité des traducteurs bénévoles.

Les traducteurs bénévoles que nous avons consultés cherchent à identifier des documents constituant des traductions mutuelles pour sélectionner des données traductionnelles pertinentes telles que des paires de « bisegments ».

Nous présentons maintenant quelques systèmes de recyclage dans les sous-sections suivantes. Les limites de chacun seront identifiées et discutées, ce qui nous permettra de proposer un nouveau système permettant le recyclage et l'aide à la traduction.

2.3.1 Le Web comme « entrepôt traductionnel »

Plusieurs systèmes ont été développés récemment pour exploiter le Web comme « entrepôt traductionnel ».

Pour la collecte de corpus parallèles, Resnik a proposé le système STRAND (Resnik, *et al.*, 2003), intéressant et efficace. Il a permis de construire plusieurs corpus parallèles contenant des dizaines de millions de pages (collectées sous forme d'un ensemble de paires d'URL). Le système vise à construire avec un coût minimal des corpus parallèles bilingues.

La recherche de textes parallèles dans ce système se fait en trois étapes :

- détection des pages qui ont des traductions parallèles,
- génération des paires qui peuvent être des traductions mutuelles,
- élimination des paires erronées par filtrage structurel.

Dans la première étape, la recherche est réalisée par appel au moteur de recherche AltaVista.[44] Durant la première étape, les pages mères sont détectées dans le but de trouver des « ancres » pertinentes. Par exemple, pour effectuer la recherche de paires <anglais, français>, l'expression booléenne (ancre : « English » OU ancre : « anglais ») ET (ancre : « french » OU ancre : « français ») est utilisée avec un filtre à distance de « 10 lignes » permettant d'augmenter la pertinence des « ancres » pointant sur une page candidate.

[44]Le moteur de recherche d'Altavista est accessible à : http://www.av.com

En ce qui concerne la deuxième étape, le robot Web recherche des coïncidences entre les URL en exploitant le fait que les répertoires des pages traduites ont une organisation parallèle. Une liste de règles de substitution a été créée (par exemple, English → big5) pour générer toutes les URL qui peuvent apparaître dans la liste des URL dans l'autre langue. Si une URL apparaît effectivement dans la liste recyclée, elle est ajoutée à la liste des paires candidates.

Par exemple, si un site anglais-chinois contient l'URL http://mysite.com/english/home_en.html et si le jeu de règles de substitution génère une URL similaire à http://mysite.com/big5/home_ch.html, alors la page originale et l'URL générée sont fort probablement des traductions mutuelles. Elles sont donc ajoutées à la liste des documents candidats.

Un autre critère pour la recherche de correspondances est l'utilisation de la longueur des documents. Les textes qui sont des traductions mutuelles ont souvent des longueurs proches ; de ce fait, il est raisonnable de supposer, pour un texte *T* dans la langue L_1 et un texte *E* dans la langue L_2, que $longueur(T) \approx C x Longueur(E)$, où *C* est une constante dépendant de la paire de langues considérée. D'autres travaux ont montré qu'un filtre basé sur la longueur au niveau de la phrase réduit l'espace de la recherche des correspondances de façon exponentielle (Resnik, *et al.*, 2003).

Dans la troisième étape, le filtre structurel consiste à transformer les documents HTML candidats en un format linéaire dont le but est la comparaison entre les structures des paires pour renforcer la correspondance. Les deux documents dans les paires candidates détectées sont les entrées principales d'un analyseur de balises qui fonctionne comme un transducteur produisant trois types d'items :

[START:element_label] par ex : [START:A], [START:LI]
[END:element_label] par ex : [END:A]
[Chunk:length] par ex : [Chunk:174]

PTMiner[45] est un autre système basé sur un agent Web intelligent conçu pour la recherche de textes parallèles sur le Web. L'algorithme de recherche est totalement indépendant de la langue. Il est doté d'un module de reconnaissance de langue et adopte les techniques de calcul de longueur pour déterminer les paires <source,cible> candidates (Nie, *et al.*, 2001).

Avec PTMiner, le recyclage du Web se fait par les étapes suivantes :

[45] Le site officiel du projet PTMiner est : http://cfwb.otil.org

- Utilisation de moteurs de recherche (Google, Altavista, Yahoo, etc.) pour l'identification des sites susceptibles de contenir des paires parallèles.
- Comparaison des URL indexées par ces moteurs de recherche d'un site candidat pour décider si le site contient des traductions mutuelles.
- Exploration du site candidat, à la recherche de paires candidates.
- Examen des paires de documents à partir des URL collectées sur chaque site et identification des paires.
- Téléchargement des pages susceptibles d'être parallèles, comparaison de la taille du texte, de la langue et de l'encodage des pages.
- Sélection des paires candidates pertinentes et rejet des autres.

Dans PTMiner, ces étapes sont effectuées après avoir introduit dans le système quelques mots-clés permettant de délimiter la recherche des paires. Si la liste de paires de documents obtenue sert à construire des corpus alignés au niveau des phrases, alors il est capable (à l'inverse du système STRAND) d'affiner l'alignement en des constituants textuels plus fins. Cela dit, il est possible avec PTMiner d'aligner les documents au niveau des phrases. Cet alignement sert à construire des corpus destinés à la TA statistique.

3.1.1 Recyclage de données traductionnelles utiles à coût minimal

Les deux systèmes cités ci-dessus montrent qu'il est possible de collecter des ressources traductionnelles utiles à coût minimal. Plusieurs corpus bilingues ont été recyclés, contenant une dizaine de millions de documents.

La précision et le rappel de ces systèmes sont plus au moins acceptables. Par exemple, avec un paramétrage manuel pour une expérience destinée à identifier des pages Web anglais-français, STRAND a permis d'obtenir une précision de 100% et un rappel de 68,6%. L'expérience a été faite sur un échantillon de 326 paires candidates extraites d'un ensemble comportant 16.763 paires de documents candidats (recyclées par STRAND). Une autre expérience du même genre a été lancée pour l'identification de paires anglais-chinois et a donné des résultats aussi acceptables, avec une précision de 98% et un rappel de 61%.

Chen et Nie rapportent que 95% du texte d'un corpus anglais-français de 135MB obtenu par PTMiner est précis et que 90% du texte d'un autre corpus anglais-chinois de 137MB est précis (Chen, *et al.*, 2000).

La Table 6 montre les volumes gigantesques de données multilingues qui ont été collectées sur le Web[46].

Langue	Taille (paires d'URL)	Description
Japonais-anglais	non-cité	Des histoires de la nouvelle technologie
Russe monolingue	25 160 662 (1,8 Go)	Corpus général (octobre 2004)
Anglais-chinois	518 382	Corpus général (juillet 2003)
Anglais-arabe	2 190	Corpus général (juillet 2002)
Anglais-basque	59	Corpus général (octobre 2000)
Anglais-chinois	3 376	Corpus général (novembre 1999)
Anglais-français	2 491	Corpus général (novembre 1998)

Table 6 : Quelques corpus obtenus par recyclage

Ces résultats montrent clairement l'intérêt du recyclage de données et de l'assimilation du Web à un entrepôt gigantesque de données traductionnelles.

Cependant, les méthodes exposées ci-dessus nous montrent clairement que le recyclage n'est pas orienté vers une exploitation efficace par des traducteurs humains. De plus, les résultats produits ne peuvent être parcourus manuellement et une interface de recherche de traductions exactes parmi les données collectées devient primordiale.

3.1.2 Discussion

Les tentatives de recyclage actuelles ont comme but le développement ou la validation des systèmes de TA statistique ou la recherche d'information multilingues.

D'après les systèmes de recyclage existants, il s'avère que la majorité de ces derniers se focalis sur la collecte de corpus parallèles, car le coût de leur construction par recyclage est moins élevé.

[46] Chaque corpus est construit par recyclage du Web par des auteurs différents. P. Resnik a obtenu la permission de les exploiter pour détecter les paires de documents candidates et les a mis sur son site personnel où ils sont librement téléchargeables avec le système STRAND :
http://www.umiacs.umd.edu/~resnik/strand/

À travers cette étude, une question se pose : ces ressources sont-elles exploitables par les traducteurs bénévoles ?

Primo, les ressources recyclées et alignées au niveau des documents <source, cible> ne sont pas exploitables, car aucune détection automatique des unités linguistiques pertinentes n'est possible (noms propres, citations, collocations, etc.).

Secundo, les systèmes de recyclage existants partent de loin en visant ainsi beaucoup moins l'aide aux traducteurs. Les interfaces de suggestion et de recherche de données sont totalement absentes, ce qui rend les données inexploitables, car, comme déjà dit (cf. *Similarités et différences dans les pratiques observées*, p. 47), les traducteurs bénévoles ont souvent moins d'expertise en informatique.

Enfin, un examen attentif de plusieurs corpus montre qu'il existe énormément de bruit dans les paires obtenues (par exemple, ceux recyclés par STRAND). Ainsi, on trouve des paires qui sont des traductions des éléments d'interface tels que les boutons, titres, les onglets et les menus. Ces paires ne sont pas ou peu sollicitées par les traducteurs bénévoles, car ces derniers souhaitent avoir des aides linguistiques en relation avec le domaine dans lequel les unités source sont en train d'être traduites.

Il est donc important, voire urgent, de développer de meilleures méthodes de recyclage adaptées aux besoins des traducteurs bénévoles. Les spécificités de ce que nous proposons sont résumées dans les points suivants :

- construction d'une MT (à partir de corpus parallèles—voir méthodes ci-dessus) à partir de sites spécifiques (par exemple : Mozilla et Arabeyes).
- identification des bisegments dans un domaine donné. Il ne s'agit pas de l'alignement, mais plutôt de l'identification de traductions des segments sur des pages candidates, pour les proposer suivant les paradigmes d'une MT, avec un score de similarité.
- les bisegments doivent être proposés de façon « proactive » et conviviale.
- Les bisegments recyclés doivent correspondre aux domaines des documents à traduire (par exemple, les droits de l'homme).

Enfin, ces points importants doivent être unifiés et présentés dans un éditeur ergonomique, avec d'autres aides linguistiques (dictionnaires, terminologies, TA, etc.).

2.3.2 Recyclage des unités de traduction sur le Web

Notre analyse du recyclage dans les sections précédentes n'a été focalisée que sur la collecte de paires de documents. Bien que cela soit important comme phase primaire, cela n'est pas suffisant vis-à-vis des besoins des traducteurs bénévoles. Cela est dû à la nature des données que les traducteurs désirent utiliser durant la traduction.

Il faut donc trouver comment étendre le recyclage à la détection d'unités linguistiques bien précises telles que des expressions, des citations, etc. L'identification de telles traductions ne nécessite pas que la détection du document candidat, mais aussi la détection de la position du segment cible.

En effet, le recyclage peut aller de la détection de paires de documents jusqu'à l'extraction d'unités linguistiques de granularité plus fine.

Le problème en lui-même ressemble à un problème d'alignement. Mais on ne peut appliquer les algorithmes d'alignement (Gale, 1991) que si on a déjà des corpus préparés. Or, nous nous ne disposons que de segments en langue source !

Des solutions à ce problème seront proposées dans la deuxième partie (cf. *Identification de traductions mutuelles : le cas des bisegments*, p. 151).

3.2.1 Problèmes spécifiques

Les problèmes d'extraction de terminologie ou de lexiques bilingues à partir de traductions, ainsi que l'alignement à différents niveaux, ont été beaucoup étudiés et des solutions plus au moins acceptables ont été trouvées. Par exemple, il existe des algorithmes pour la détection des équivalences lexicales dans des documents comparables, et d'autres pour l'alignement à différents niveaux de granularité (par syntagmes constituants et mots) à partir de corpus parallèles (Daille, *et al.*, 1994). Jean Véronis (2000) a montré que les algorithmes d'alignement appliqués aux phrases dans le projet ARCADE [47] arrivaient à 98,5% de précision et ceux appliqués aux mots à 75% (Véronis, 2000).

En revanche, ces algorithmes nécessitent la *préparation* et/ou la *construction* de corpus à l'avance. Par exemple, dans le même projet (ARCADE), une année a été consacrée (la

[47] Le projet ARCADE est un projet collaboratif d'évaluation en traitement de la langue naturelle et de la parole. Il a été lancé (en 1996) et financé par le réseau des Universités Francophones, AUPELF-UREF. Deux phases on été planifiées : (i) la première phase (1996-1997) a été consacrée à la méthodologie, à la collecte de corpus et à la préparation, (i) la deuxième phase (1998-1999) a permis à d'autres équipes de participer et d'étendre l'alignement vers des mots et des expressions.

première phase du projet, 1996-1997) à la préparation de deux corpus par deux laboratoires. Le laboratoire Parole et Langage (LPL) a assuré la préparation du corpus JOC[48] (1,1M de mots par langue) et le laboratoire Recherche Appliquée en Linguistique Informatique (RALI) de l'Université de Montréal la préparation du corpus BAF[49] (400 000 mots par langue). En ce qui nous concerne, nous ne pouvons pas procéder ainsi, car la détection de données traductionnelles sur le Web par recyclage ne peut pas devoir être précédée d'une préparation de corpus par les traducteurs.

Trouver une méthode pour l'extraction de traductions mutuelles (un ou plusieurs mots contigus ou non) à partir de documents disséminés sur le Web serait très utile aux traducteurs et constitue un problème peu abordé et intéressant. La Figure 20 montre une paire de documents trouvée sur le site de traduction bénévole de la communauté W3C. Nous détecterons d'abord les paires de documents, puis, sur demande de l'utilisateur, la recherche sera affinée vers des bisegments.

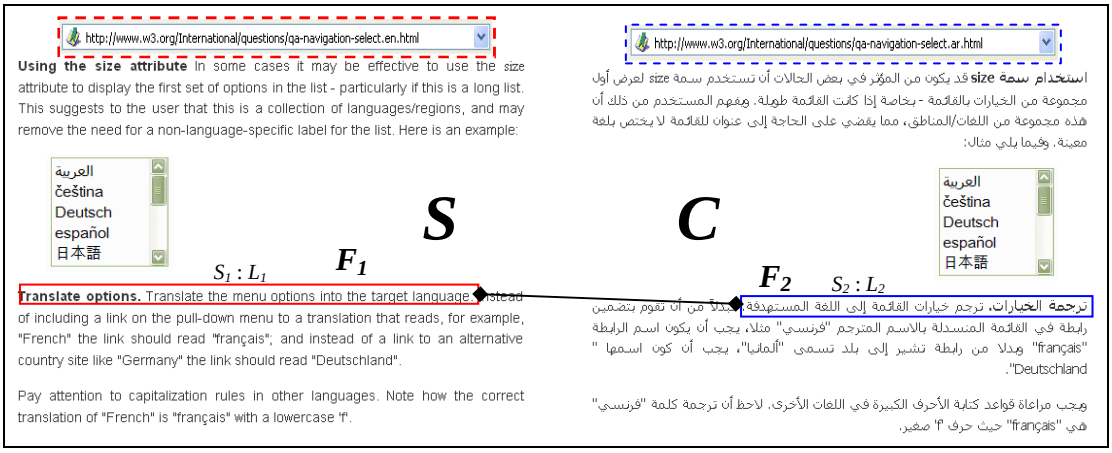


Figure 20 : Traduction anglais-arabe d'un document standard W3C

Le scénario est le suivant : durant la traduction, un traducteur peut lancer une requête dans le but de trouver une traduction sur la Toile d'un segment source (*S*). Le résultat à renvoyer doit être un ensemble de couples (désormais « bisegments ») constitués du texte source *S* et de ses traductions < (S_1,L_1), (S_2,L_2) >, où L_1 est la langue source du segment S_1 et L_2 est la langue cible du segment S_2. Par exemple, si la recherche se fait sur le fragment F_1 du document

[48] Le corpus « JOC » est composé de questions posées au Parlement Européen dans plusieurs domaines (santé, éducation, environnement, économie, etc.). Le corpus original a été collecté dans le cadre du projet MLCC-MULTEX et comprend 9 versions, chacune correspondant à une langue de la Communauté Européenne.

[49] Le corpus « BAF » est composé d'un ensemble de textes de genres différents en anglais-français, couvrant des textes institutionnels, des articles scientifiques, des manuels techniques et de la littérature. Quelques textes ont été importés du corpus multilingue ECI, de l'Association Bibliophile Universelle (ABU), et du projet Gutenberg.

source *S*, le robot doit pouvoir retrouver l'équivalence traductionnel F_2 dans le document *C* ou sinon, le segment complet contenant le fragment.

Nous développons cette idée en détail dans la sous-section suivante.

3.2.2 Construction automatique des mémoires de traduction

Au moment de la rédaction de notre thèse, le système QRselect était en train d'être étendu vers la construction de mémoires de traductions spécialisées, grâce au recyclage de sites « amis » (sites personnels auxquels les traducteurs bénévoles se réfèrent pour disséminer leurs traductions).

Dans ce système, la construction des MT est vue comme un double problème d'alignement. Il ne s'agit en effet pas de trouver directement des bisegments, mais de construire une ressource linguistique alignée au niveau des phrases à partir d'un ensemble de paires de documents recyclées sur le Web, ce qui est un premier problème d'alignement particulier, puis d'y rechercher des bisegments en appliquant des algorithmes d'alignement classiques tels que celui de Church et Gale (Gale, 1991) ou le modèle 4 d'IBM (Brown, *et al.*, 1991) qui sont déjà implémentés dans l'outil (Giza++, 2006).

En ce qui concerne la structuration des bisegments, l'ensemble des UT produites par l'aligneur sera compilé dans une structure TMX (Translation Memory eXchange) et stocké dans la BD centrale du système. L'exploitation des UT sera faite à travers l'éditeur QRedit que nous introduisions dans le chapitre suivant.

Dans les paragraphes suivants, nous introduisons le robot Web QRselect, développé spécialement pour résoudre le problème du recyclage de documents existants sur la Toile. Comme il a été développé dans le cadre du projet QRLex, il est limité aux paires de documents anglais-japonais.

2.3.3 QRselect : recyclage de documents déjà traduits sur le Web

3.3.1 Méthode de recyclage des paires EN-JP

L'outil QRselect a été explicitement développé pour le recyclage de paires de documents anglais-japonais en partant de mots-clés japonais. Il n'est donc possible de lancer la recherche que par une requête (à la Google) composée de mots japonais.

Le principe du recyclage dans QRselect se fait selon l'hypothèse suivante :

Beaucoup de documents source contiennent des « ancres » entourées par des mots spéciaux pointant sur l'URL d'un document cible.

Cette même hypothèse a été utilisée dans le développement de l'outil « STRAND » et est exploitée à nouveau dans QRselect.

La liste des mots-réservés entourant les ancres est établie manuellement. Dans le cas des documents japonais, une liste restreinte a été définie : « 原文 » (genbun = document original), « オリジナル » (originaru = original), « ソース » (sosu = source), « 英語 » (eigo = anglais), « 元記事 » (motokiji = document original), « 原著 » (gencho = article original) et « 原語 » (gengo = langue source).

L'une des propriétés importantes de QRselect est que :

La recherche de documents est limitée aux paires de documents anglais-japonais en se basant sur des mots-clés japonais à travers lesquels des ancres sont détectées dans les documents japonais et sur un calcul de proximité entre documents, en utilisant un dictionnaire bilingue.

Les mots-réservés doivent être des constituants de l'« ancre », en question au moment du recyclage d'une paire, autrement dit, au moins l'un des mots-réservés doit être une partie de chaîne de l'« ancre ». D'autres techniques sont aussi applicables telles que la définition d'une distance entourant les « ancres » et les mots-réservés (par exemple le système STRAND utilise une distance de 10 mots), mais elles n'ont pas été introduites, car QRselect vise plus de précision.

La méthode de recyclage employée par QRselect utilise les étapes suivantes (Figure 21) :

- Détection des pages japonaises correspondant aux mots-clés.
- Détection des ancres et récupération des pages cibles candidates.
- Calcul de scores par exploitation du dictionnaire Eijiro.
- Sélection d'une paire pertinente selon un seuil arbitraire choisi d'avance.
- Unification, et présentation des résultats aux traducteurs.

Pour juger si une paire est pertinente, un score est calculé pour déterminer la proximité traductionnelle entre le document source en japonais et la cible ((Équation 1). Le score est une fonction du nombre de couples de traduction des mots dans les deux documents <mot japonais, mots anglais> - l'ordre n'est plus important. Techniquement, QRselect utilise une

version ancienne du dictionnaire *Eijiro* contenant plus de 300 000 entrées pour déterminer ces couples.

Le score est défini plus précisément par le rapport entre le nombre de couples <mot source, mot cible> identifiés par le dictionnaire « Eijiro » et le nombre de mots dans le document source :

$$\text{Score}(A, B) = \frac{\text{Nombre de mots communs}(A, B)}{\text{Nombre de mots dans A}} \qquad \text{(Équation 1)}$$

où

- A : est un document en japonais, et B : est un document en anglais.
- le nombre de paires de mots communs est calculé après avoir traduit les mots présents dans le document japonais par le dictionnaire Eijiro.

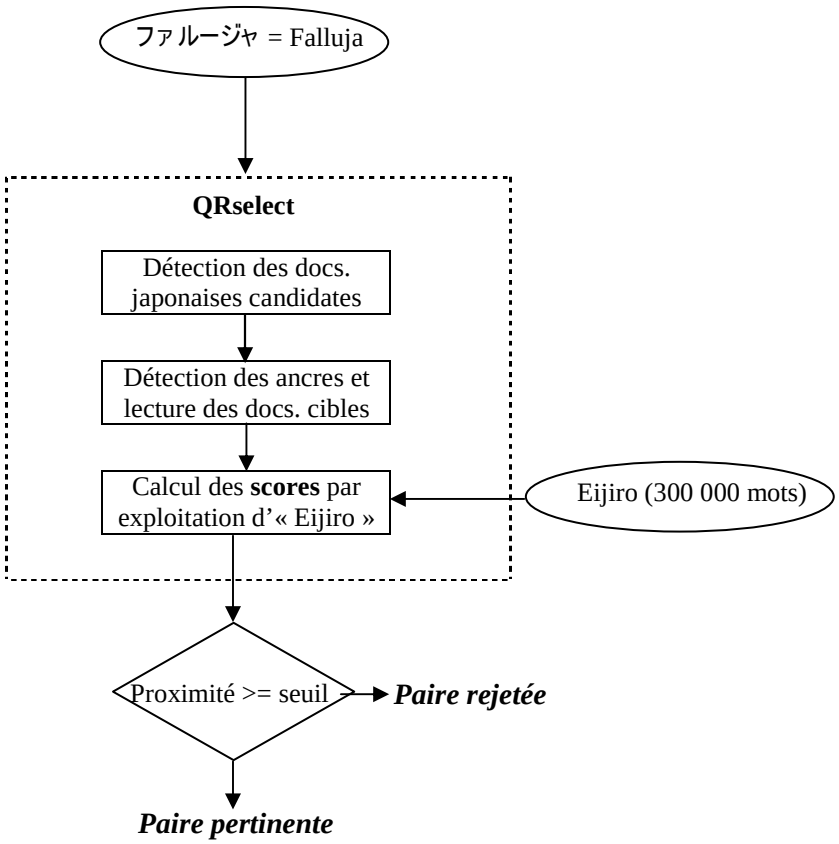


Figure 21 : Méthode de détection de paires pertinentes

Bien que cette méthode fonctionne, elle présente des inconvénients qui seront discutés plus loin (pp. 99).

Le dictionnaire « Eijiro » limite la recherche de documents durant le recyclage. Cela dit, le recyclage par le robot dépend surtout de la qualité du dictionnaire et des langues utilisées. Pour étendre la recherche vers d'autres langues, il est indispensable d'intégrer d'autres dictionnaires.

3.3.2 Interface de recyclage

Dans l'interface principale de QRselect, il est possible d'introduire des mots-clés et de paramétrer la recherche des paires de documents (Figure 22).

La recherche est effectuée à l'aide de plusieurs moteurs de recherche offrant des API pour les exploiter par programmation de la recherche Web. Les API les plus intéressantes sont celles proposées par « Google » et « YahooJapan ». Les deux sont exploitables par QRselect, mais celle de « YahooJapan » est plus utilisée, car elle permet un nombre de sessions supérieur, allant jusqu'à 5 000/jour, alors que celle de Google est limitée à 1 000/jour.

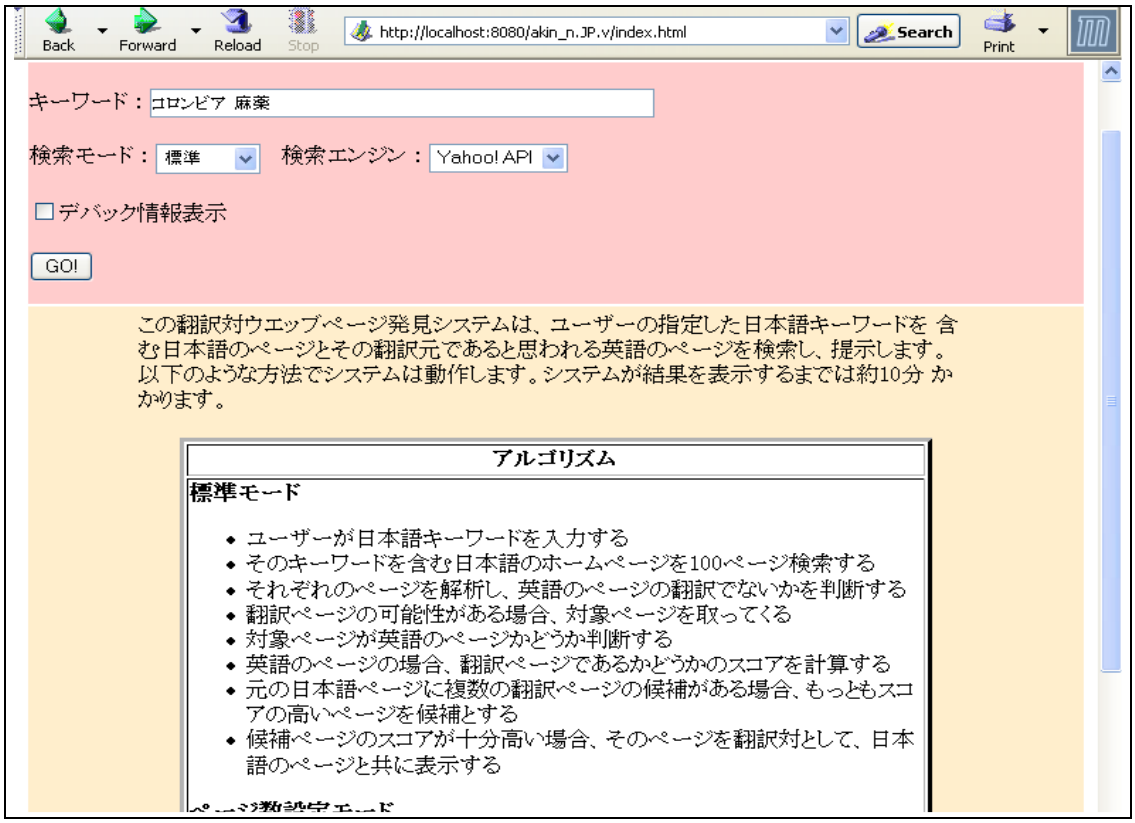


Figure 22 : Interface principale de configuration de QRselect

Le résultat est visualisé sous forme d'un ensemble de paires de documents jugées pertinentes et accompagné des titres des pages et d'un extrait décrivant brièvement le contenu des documents.

Sur la même interface, l'accès aux documents est possible par un simple clic de souris sur l'une des URL visualisées.

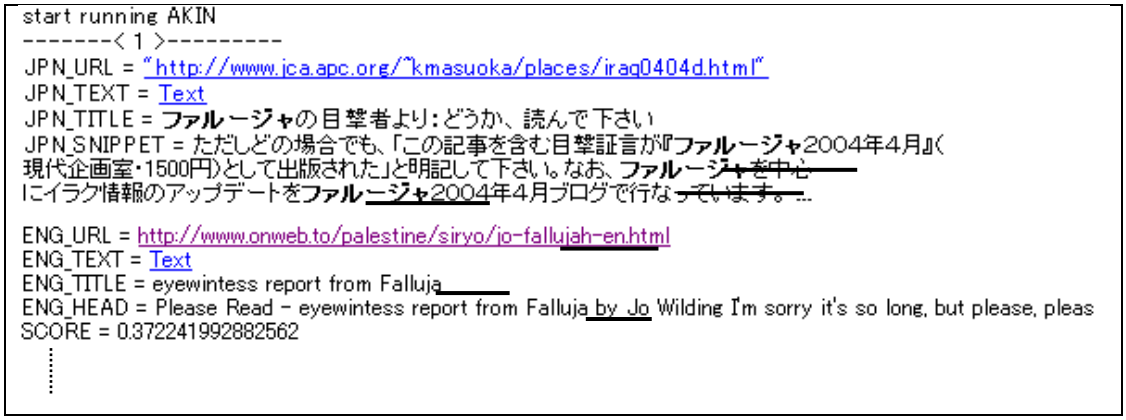


Figure 23 : Interface de détection de documents anglais-japonais avec QRselect

Les taux de proximité entre les documents sont visualisés après les informations descriptives de chaque paire. Par exemple, le mot « Falluja » génère les résultats illustrés sur la Figure 23 (cette figure n'illustre qu'une partie des résultats).

3.3.3 Les limitations du système QRselect

QRselect a été analysé sous différents angles, ce qui nous a permis de cerner les limitations suivantes.

- Bilinguisme du recyclage : les autres communautés de traducteurs bénévoles ne peuvent exploiter le système, car le dictionnaire sur lequel se base la recherche est limité à la paire de langues anglais-japonais.
- Correspondance « document-document » : les résultats du recyclage sont un ensemble de paires alignées au niveau des documents, or on sait que les traducteurs ont souvent besoin de traductions d'unités linguistiques beaucoup plus petites ou fines.
- Possible insuffisance du calcul de proximité : la méthode de calcul est basée sur les mots comme unités de comparaisons entre les documents sources et cibles. Cette méthode est figée, car il existe

des méthodes de calcul meilleures basées sur les caractères et pas sur les mots (Denoual, 2006).

- Insuffisance de l'analyse des « ancres » : certaines ancres sont référées par des graphiques qui ne présentent aucun texte, donc cela nécessite de penser à améliorer le rappel.

De plus, un seul dictionnaire est utilisé, et il ne peut être mis à jour. Cette fonctionnalité ne permet pas aux traducteurs de trouver des traductions précises pour des termes spécifiques, surtout lorsqu'il s'agit d'un domaine précis où le dictionnaire ne contient pas les traductions équivalentes.

Ces problèmes seront discutés de façon approfondie dans le deuxième chapitre de la deuxième partie, où QRselect sera étendu pour inclure de façon générique d'autres dictionnaires en d'autres langues, pour que l'outil puisse être utilisé par plusieurs communautés de traducteurs bénévoles.

Conclusion

Nous avons présenté dans ce chapitre les différents modules du système QRLex et avons discuté les différents problèmes que rencontrent les traducteurs bénévoles en ligne.

Nous avons d'abord étudié les différentes catégories de bénévoles à travers l'expérimentation directe de leurs sites ou par des interviews. Grâce à cette étude, nous avons catégorisé le travail traductionnel et identifié les besoins en aides linguistiques. Cela nous a permis de proposer une première maquette fonctionnelle correspondant aux besoins exacts des traducteurs bénévoles.

Les dictionnaires sélectionnés ont été étudiés chacun séparément. Cela nous a permis d'identifier des problèmes tels que l'hétérogénéité des données et la diversité des formats originaux de dictionnaires. Nous avons proposé le format XLD pour surmonter ces problèmes et compilé plus de 1,7M d'entrées bilingues anglais-japonais.

Nous avons proposé l'outil QRselect pour faciliter le recyclage de documents déjà traduits. En utilisant une méthode de calcul de proximité à l'aide d'un dictionnaire bilingue, le système QRselect détecte des documents qui constituent des traductions mutuelles. Nous avons proposé aux traducteurs bénévoles une interface conviviale de recherche dans laquelle il est possible d'introduire des mots-clés en relation avec le domaine de la traduction. Le système

renvoie un ensemble de paires de documents avec des liens pour l'accès direct aux pages candidates.

Dans le chapitre suivant, nous présentons l'éditeur bilingue QRedit. Nous discutons les différentes fonctionnalités d'aide à la traduction proposées par cet éditeur. Une grande importance sera donnée aux aides linguistiques que propose l'éditeur durant le processus de traduction.

Chapitre 3

QRedit : un éditeur Web bilingue d'aide à la traduction

Introduction

La plupart des outils d'aide à la traduction proposent aux traducteurs des interfaces bilingues ou multilingues sous forme d'éditeur, intégrant des fonctionnalités linguistiques. D'autres ne contiennent pas d'éditeur mais proposent des *plugins* exploitant les fonctionnalités d'édition d'un logiciel de traitement de texte (comme le plugin Trados MS Word).

Cependant, l'exploitation de ces outils est limitée aux traducteurs professionnels. Les traducteurs bénévoles, que nous avons consultés par des interviews directs et une étude de plusieurs sites de traduction bénévole (W3C, Traduct, FrenchMozilla, Arabeyes, etc.), ont confirmé que la majorité d'entre eux n'utilisent pas ces outils et refusent de les intégrer dans leurs travaux de traduction. Les raisons exprimées et ressenties par ces traducteurs sont :

1. le coût élevé, des outils existants, en particulier, en version réseau.
2. la conception orientée vers la traduction professionnelle.
3. le fonctionnement hors ligne : peu d'outils fonctionnent en ligne.
4. le manque d'un « éditeur en ligne ».

Les points (1), (2) et (3) ont été abordés dans la conception de l'environnement QRlex.

Nous attaquons ici le point (4).

Nous proposons et décrivons ici l'éditeur QRedit. Il a été conçu selon les besoins exprimés et ressentis par les traducteurs bénévoles et permet de traduire en ligne en exploitant gratuitement des outils d'aide, en particulier, les dictionnaires, les banques terminologies, etc. QRedit constitue la composante principale de communication et d'interaction avec les autres modules de QRLex (recyclage, exploitation de la BD, etc.) et avec les bénévoles.

Tout d’abord, nous expliquons comment coupler les aides linguistiques, ensuite nous présentons la méthode de traduction, et nous décrivons la méthode de traduction dans QRedit.

3.1 Couplage d’aides linguistiques dans QRedit

3.1.1 Traitement du document source

Dans QRedit, un document source (en anglais) est chargé en donnant son URL ou en copiant son texte directement dans la zone source (textarea en HTML). Le processus de chargement effectue une extraction textuelle basée sur les signes de ponctuation. Le texte est extrait des zones textuelles du document HTML (Figure 25).

3.1.2 Couplage de la traduction avec les mots source

Après le chargement d’un document, les informations dictionnairiques disponibles sont attachées aux mots et aux expressions identifiées dans les ressources linguistiques. Par contre, ceux et celles qui n’ont pas été détectés sont indiqués avec un affichage différent. Nous appelons cela le mode d’indication « positif ». Le mode négatif alerte les traducteurs sur la non-disponibilité d’une traduction de tel ou tel mot ou expression. Le mode positif indique au contraire la disponibilité d’une traduction, ce qui motive les traducteurs à consulter les ressources.

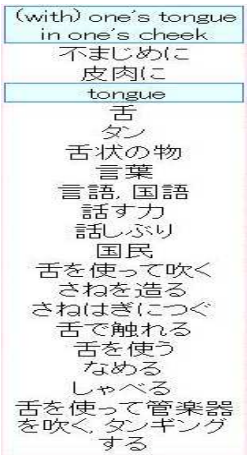


Figure 24 : Suggestions dictionnairiques dans QRedit

En ce qui concerne les informations attachées, elles sont cachées et présentées visuellement par des liens Web. Une fois un lien activé par le passage du curseur au-dessus de lui, il affiche les traductions compilées lors du chargement. L’affichage se fait de façon souple, c’est-à-dire que les traducteurs consultent des items dans une fenêtre déroulante à plusieurs choix, facilitant l’accès à la traduction recherchée.

3.1.3 Méthode d'aide durant la traduction

Avant de passer à une présentation exhaustive de la méthode d'aide à la traduction dans QRedit, il est nécessaire de souligner les besoins des traducteurs, non pas en termes de contenu des ressources linguistiques, mais plutôt de mode d'utilisation et d'accès aux traductions des segments lors du processus.

D'après les questionnaires et les interrogations des traducteurs, les points suivants se révèlent très importants dans la conception de l'éditeur.

- Dans le processus de traduction, les traducteurs vérifient le plus souvent de multiples ressources linguistiques et examinent plusieurs traductions et sens des segments source. Il est donc important de proposer aux traducteurs des informations pertinentes par l'intermédiaire d'un « système de suggestion » à plusieurs propositions.

- Le système doit pouvoir afficher les informations d'aide de la même façon que celle à laquelle ils sont habitués pour la consultation et la recherche des traductions. À part la diversité des propositions, ce système permet de plus aux traducteurs de trouver facilement le sens d'une traduction en cours, et cela d'une façon compréhensible et efficace.

Dans les paragraphes suivants, nous montrons comment ces besoins sont satisfaits dans l'outil QRedit.

3.2 Processus de traduction

3.2.1 Synchronisation de la visualisation source/cible

La disposition d'écran adoptée par la plupart des outils d'aide à la traduction est une interface visualisant les unités source et cible (Commercial-MT, 2006) (DéjàVu, 2007) (OmegaT, 2007) en parallèle. Une telle disposition des unités est intéressante car elle permet de visualiser ces paires de segments source/cible (UT) et l'intégration de nouvelles paires d'UT dans la MT. On adopte donc la même solution pour la présentation des UT aux traducteurs bénévoles dans QRedit.

Les traducteurs peuvent voir l'intégralité des documents source/cible avec une synchronisation cohérente des UT. Le passage d'une UT à une autre permet d'activer en arrière-plan des fonctionnalités spécifiques au segment actif (par exemple, extraction de coïncidences floues des MT).

Les UT sont détectées lors du chargement d'un document source et sont séparées par des balises <hr> (des lignes horizontales en HTML) et des identificateurs dans la zone d'édition source. Les UT dans la zone d'édition cible sont séparées de la même façon. Cela facilite la synchronisation et la gestion cohérente des UT (Figure 25).

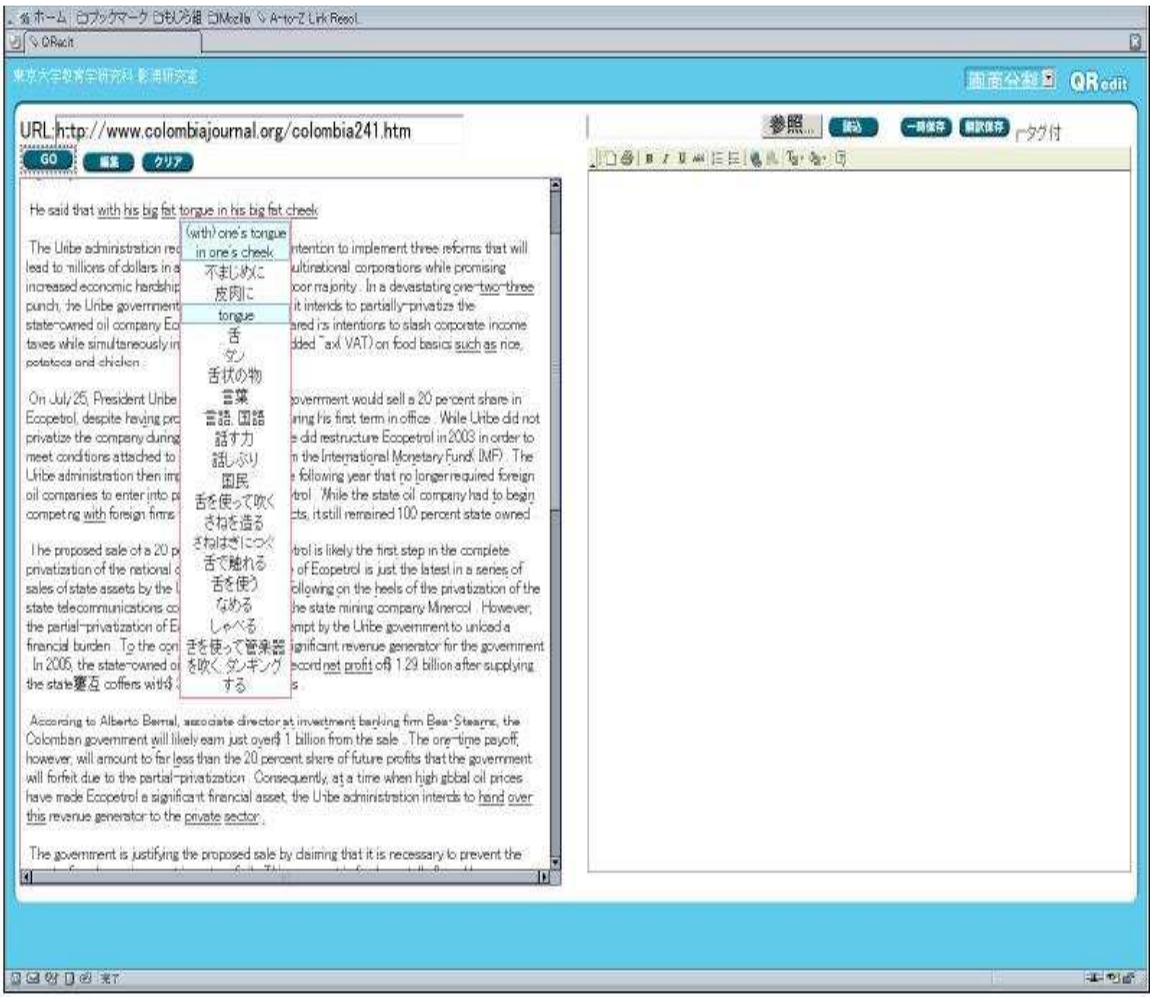


Figure 25 : Editeur de QRLex : interface de QRedit

À la fin de la traduction, les paires d'UT traduites sont intégrées dans la MT du système QRlex, qui sera exploitée à nouveau dans de nouvelles traductions par d'autres traducteurs.

La Figure 25 présente l'interface de l'éditeur QRedit. La zone de gauche est dédiée au contenu du document source et aux aides linguistiques. La zone droite est celle ou s'effectue la traduction.

La façon dont les aides linguistiques sont proposées aux traducteurs est présentée dans le paragraphe suivant.

3.2.2 Présentation des éléments sélectionnables

Dans la zone source, les unités linguistiques (mots, mots composés, phrases, fragments, etc.) ayant des traductions deviennent des ancres.

Les suggestions dictionnairiques sont sélectionnées par déplacement du curseur sur ces ancres. Lors d'une sélection pertinente, la traduction choisie de l'unité linguistique est insérée dans la zone cible (la zone de droite), à la position d'édition précédente (la zone cible).

3.3 Avantages et limitations de QRedit

3.3.1 Remarques liées à l'aide linguistique

Les ressources linguistiques sont déterminées et compilées d'avance ; elles sont prétraitées, et ensuite importées dans la BD de QRlex. L'utilisation de ces ressources présente plusieurs limitations importantes :

- la possibilité de mise à jour est totalement absente.
- la création de nouveaux dictionnaires (même temporaires) n'est pas possible.
- Il n'y a pas de possibilité de construction collaborative de dictionnaires.

La première possibilité est nécessaire parce que les traducteurs bénévoles cherchent à améliorer les ressources linguistiques. La deuxième et la troisième sont plus importantes lorsqu'il s'agit de créer des dictionnaires spécialisés et d'étendre le contenu vers d'autres données. La collaboration est primordiale, surtout lorsqu'il s'agit de la traduction en ligne. Un dictionnaire ne peut être construit par une seule personne. La collaboration peut donc favoriser la « collecte des efforts » et le partage des connaissances. On a vu que des dictionnaires comme le Wiktionary ont été construits de façon totalement collaborative (au début de l'année 2007, plus de 23 000 bénévoles sont inscrits à ce dictionnaire et participent activement pour l'améliorer en contenu).

Les ressources linguistiques compilées dans QRLex ne sont orientées que vers l'aide aux traducteurs humanistes (par exemple les droits de l'homme). Cette spécialisation ne couvre qu'une catégorie de traducteurs. De plus, les suggestions sont bilingues (EN-JP) et ne peuvent être appliquées à d'autres langues. Or, il arrive souvent qu'on veuille traduire un document dans plusieurs langues. Il faut donc donner la possibilité de gérer avec les mêmes fonctionnalités des ressources multilingues.

La construction automatique des MT nécessite une phase de segmentation. La segmentation s'effectue d'une manière automatique, mais nécessite généralement une intervention humaine pour améliorer les bornes des phrases. À part l'éditeur de traduction, il manque une interface pour éditer les UT pour corriger les erreurs de la segmentation automatique.

Nous nous attacherons à dépasser ces limites quand nous construirons un second environnement en ligne, encore plus complet, dans la partie II. Nous y résoudrons aussi le manque de « généralité multilingue » provenant de la limitation (voulue) de QRLex au couple anglais-japonais, dans ce seul sens.

3.3.2 Remarques liées à l'éditeur bilingue de traductions

L'éditeur QRedit est simple dans sa conception. Il permet d'importer facilement des documents en anglais. Par le processus de segmentation, les unités de traduction sont détectées et synchronisées. Mais les aides linguistiques sont limitées aux ressources linguistiques compilées d'avance. Il n'est donc pas possible de créer des dictionnaires temporaires qui sont souvent demandés par les traducteurs bénévoles.

Cependant, bien que cet éditeur soit utile pour certaines communautés de traducteurs, en particulier la *catégorie B*, il ne peut pas être utilisé dans toutes les situations de traduction et par toutes les communautés bénévoles de traduction, surtout les traducteurs appartenant à la *catégorie A*. Il y a plusieurs raisons pour expliuqer ces limitations.

Primo, l'éditeur ne permet pas la traduction incrémentale et la collaboration, ce qui empêche en partie son usage par la *catégorie A* (cf. *Similarités et différences dans les pratiques observées*, p. 47). En effet, la présence de ces deux fonctionnalités est importante dans le cas de projets de traduction de logiciels libres tels que Mozilla et Traduct. Mais il n'est pas possible avec QRedit de faire traduire un document simultanément par plusieurs traducteurs en ligne, alors les bénévoles le font dans l'environnement Translationwiki.net.

Secundo, la communication entre les traducteurs est quasiment absente, or elle est nécessaire pour l'échange de connaissances et pour l'organisation des projets de traduction. Les communautés que nous avons étudiées (par exemple Arabeyes) ont chacune une méthode de communication qui peut être le courriel, une liste de discussion, un forum, etc. Cette fonctionnalité n'a pas été envisagée dans QRLex ni durant la traduction dans QRedit.

Dans la deuxième partie, ces manques seront analysés de façon approfondie et un nouvel éditeur sera conçu et développé pour répondre aux besoins les plus importants et communs à toutes les communautés (diversité de langues, fonctionnalités plus avancées, etc.).

3.3.3 Remarques liées a la gestion collaborative des documents

Les traductions dans QRLex se font en ligne. Aller à l'encontre des concepts tels que ceux implémentés dans Wiktionary freine les traducteurs dans la mise à jour des ressources, parce que la construction de ressources volumineuses ne peut être faite avec un seul traducteur, et que par suite la collaboration devient indispensable.

Il faut donc aller au delà de QRLex et profiter des progrès des outils et de la technologie pour promouvoir la traduction collaborative à plusieurs niveaux : documents multilingues, dictionnaires, mémoires de traductions, etc.

Les traducteurs ont aussi besoin de partager des tâches et de suivre la progression des traductions, parce qu'ils traduisent dans un esprit communautaire. Un cas de figure est la localisation des logiciels libres (ArabicMozilla, 2007) (FrenchMozilla, 2005; W3C, 2007) où la traduction ne consiste pas seulement à traduire du texte, mais aussi à contrôler l'effet des traductions sur les interfaces et sur le fonctionnement général des logiciels.

3.3.4 Limites plus « dures »

3.4.1.a Impossibilité de traiter des documents volumineux

Le traitement et la dissémination de documents multiformat de grande taille est en demande croissante et est actuellement le thème de plusieurs appels d'offres. Mais QRLex/QRedit, en plus de ses limitations vis-à-vis du nombre de langues et de l'absence de collaboration, n'est pas capable de gérer des documents variés de taille énorme. Bien qu'il soit possible de transformer des documents divers en HTML, QRedit ne peut être appliqué à la traduction de documents de plus de quelques dizaines de pages.

La gestion de ressources et de documents volumineux nécessite d'utiliser des techniques modernes apparues avec le Web collaboratif (Web 2.0).

D'un côté, on dispose maintenant de dispositifs de stockage efficace de données gigantesques s'étendant sur des giga-octets, voire même des téraoctets, et une recherche performante avec la possibilité de mise à jour.

D’un autre côté, il faudrait aussi résoudre les problèmes de chargement de ces masses de données, proposer des solutions pour la visualisation, et trouver des solutions pour augmenter la vitesse du chargement, des manipulations globales, etc.

3.4.1.b Dictionnaire et TM

QRLex ne propose aucun processus d’automatisation permettant selon le contexte d’introduire des suggestions « proactives » et de sélectionner un dictionnaire spécifique à la tâche. La lemmatisation est également absente : l’utilisateur peut être dérouté en ne trouvant pas le mot-vedette correct dans les dictionnaires, et le temps de recherche dans les dictionnaires peut s’en trouver augmenté.

De plus, aucune modification en ligne n’est possible, ce qui rend les ressources linguistiques non « actives ».

3.4.1.c Contribution au dictionnaire par le Web seulement

L’un des clés de la réussite d’un système d’aide à la traduction est l’évolution du contenu des ressources linguistiques, l’autre étant la richesse initiale de ces ressources. Il faut donc développer des fonctionnalités permettant l’amélioration des ressources linguistiques de façon collaborative. La mise à jour se fera à la volée. A chaque amélioration, les nouvelles données devront être disponibles instantanément sur la Toile.

Conclusion

Le besoin d’un éditeur traductionnel utilisable par le Web a été noté lors de nos interviews avec les traducteurs et grâce à l’étude menée sur les environnements existants de traduction bénévole. On a constaté qu’il est nécessaire à l’heure actuelle de proposer aux traducteurs bénévoles un éditeur couplant des fonctionnalités linguistiques avec le texte source pour l’aide à la traduction et la dissémination des résultats dans la ou les langues cibles. Le développement de QRedit est une première tentative intéressante, mais il présente plusieurs limitations.

Dans la partie suivante, nous proposerons un nouvel éditeur, radicalement différent, conçu et intégré dans un environnement complètement collaboratif favorisant la traduction incrémentale, et offrant une grande couverture de langues et des fonctionnalités plus avancées que celles proposées par QRlex.

Partie II

BEYTrans : un service Web collaboratif libre d'aide à la traduction en ligne

Introduction

Comme nous l'avons vu dans la première partie (chapitre 1), la traduction bénévole est en augmentation continue et les traducteurs qui y sont impliqués se comptent par dizaines de milliers. Nous avons aussi montré les besoins et les manques apparus dans le cadre du projet QRLex.

Dans un temps où la demande de disposer des documentations en plusieurs langues devient plus importante, l'environnement QRLex ne peut répondre aux demandes multilingues de toutes les communautés.

Pour surmonter ces obstacles, nous proposons dans cette partie l'environnement BEYTrans (Better Environment for Your Translation), un environnement en ligne collaboratif, permettant la traduction incrémentale et non commerciale. Nous l'avons conçu de façon qu'il soit ouvert à toutes les langues et générique vis-à-vis de la gestion des ressources linguistiques. Un éditeur avancé de traduction, couplé aux différentes fonctionnalités d'aide à la traduction, a été développé et intégré dans l'environnement.

La totalité de l'environnement a été développée en se basant sur les concepts et la technologie Wiki. L'environnement est ouvert à tout usage pour tous les usagers intéressés à traduire ou à améliorer des ressources linguistiques.

L'environnement BEYTrans est mis en ligne sous la forme d'un Wiki. Il a été expérimenté de façon préliminaire dans le cadre du projet DEMGOL, dans lequel une centaine de notices italiennes ont été traduites de l'italien vers le français, et une expérience menée avec précision sur 90% d'entre elles par une spécialiste de domaine, italienne et non traductrice professionnelle, traduisant donc « à l'envers ». Une seconde procédure d'évaluation a ensuite été définie pour mettre en relief les avantages du Wiki (la localisation de plusieurs personnes sur de petits incréments et non des documents entiers) et des aides traductionnelles (TA, MT, dictionnaires). Une expérience a été lancé sur la localisation de Tikiwiki.

Cette partie est divisée en trois chapitres. Dans le premier, nous présentons la conception du nouvel environnement. Dans le deuxième, nous traitons le recyclage avec un nouveau point de vue. Dans le troisième, nous introduisons un nouvel éditeur en ligne.

Chapitre 4

Conception générale de l’environnement « BEYTrans »

Introduction

Nous nous attaquons dans cette partie au développement d’un environnement Wiki complet permettant la traduction collaborative incrémentale et favorisant la communication entre traducteurs bénévoles.

Les expérimentations que nous avons menées sur les environnements collaboratifs non destinés à la traduction, mais offrant l’édition collaborative et la gestion de documents multilingues (par exemple les environnements collaboratifs libres CMS « Content Management Systems »), nous ont guidé vers la technologie Wiki comme un choix intéressant pour le développement de notre environnement. Cela nous a été aussi confirmé par l’analyse du mode de fonctionnement du Wikipedia et du Wiktionary, deux environnements totalement construits de façon collaborative par les contributions de bénévoles.

En combinant cette technologie et les besoins des traducteurs bénévoles vus dans le chapitre précédent, nous proposons dans ce chapitre une nouvelle conception d’un environnement d’aide à la traduction, permettant non seulement la collaboration, mais la dissémination des traductions incrémentales et non commerciales. Selon le principe de la technologie Wiki, il sera ouvert à toutes les contributions des bénévoles sans aucune restriction. De plus, les composantes logicielles d’aide à la traduction devront aussi pouvoir être aussi exploitées par l’ensemble des bénévoles séparément ou en même temps.

4.1 Scenario et architecture

4.1.1 Scénario général

L’environnement est ouvert à tous les traducteurs bénévoles en ligne. Tout traducteur/utilisateur intéressé par la traduction peut donc charger et partager des documents avec d’autres traducteurs en mode collaboratif.

Avant de commencer une traduction, le document source doit être préparé par le système. Cela dit, le texte source est extrait et segmenté en plusieurs UT. En effet, l'import d'un document à traduire crée une structure interne TMX aidant à stocker efficacement les UT sources et cibles.

Pour réaliser les traductions, les traducteurs peuvent appliquer sur les UT les fonctionnalités suivantes :

- Appel de systèmes de TA : les traducteurs ont le choix entre plusieurs systèmes de TA gratuits pour avoir une première version (prétraduction) (Systran, Reverso, etc.). Les prétraductions des UT seront par la suite post-éditées et stockées dans une structure XML spéciale.

- Appel des MT : une fois trouvés, les segments détectés similaires à la source sont affichés, et leurs traductions sont proposées dans la zone cible. Cette méthode est plus efficace que la précédente si la similarité est élevée, parce que cela augmente la productivité et diminue le temps de traduction plus que si l'on part d'un résultat de TA « Web ».

Les résultats de traduction des UT sont synchronisés avec les UT source. Les traducteurs peuvent basculer du mode lecture au mode édition. L'éditeur de traductions multilingue offre une interface en ligne conçue à la « Excel », où toutes les UT source et cible sont synchronisées et visualisées en parallèle. Ainsi, il est possible d'exploiter le contenu des ressources linguistiques (dictionnaires, mémoires de traduction, etc.) par des fonctionnalités diverses de consultation.

Les fonctionnalités d'aide linguistique sont couplées aux événements déclenchés lors des manipulations des UT source via l'interface de traduction. Les fonctionnalités usuelles d'édition sont aussi disponibles : les suppressions, la fusion, l'ajout et la division des UT (des lignes dans l'interface) sont possibles. Un traducteur peut ajuster la segmentation par création de nouvelles UT ou par fusion de deux segments.

Durant la traduction, l'éditeur propose des suggestions de traduction par un calcul de taux de similarité entre les UT source et celles stockées dans la MT active[50]. Le seuil de correspondance est paramétrable. Les traducteurs peuvent lire le taux de similarité ainsi que les suggestions proposées par le système et les insérer à la bonne position. À la fin de la

[50] Il y'a plusieurs MT, chacune correspondant à une communauté précise.

traduction, les nouvelles versions des UT sont stockées à nouveau dans un « compagnon de traduction » (CT) en gardant les versions précédentes selon le principe des Wiki.

Dans le chapitre 3, on a vu que dans l'éditeur QRedit les UT source et cible sont synchronisées avec des ID qui sont inclus dans le code HTML, une méthode qui marche mais qui est moins efficace que la méthode basée sur XML utilisée ici. En effet, elle permet d'associer de la sémantique aux différents éléments structurant les données, et d'exploiter des API telles que DOM et SAX pour les manipulations des éléments (ajout, suppression, mise à jour, etc.).

Dans ce même scénario, à l'inverse de la traduction professionnelle, les traductions ne sont pas nécessairement disséminées complètes – à l'inverse des méthodes professionnelles où les traductions doivent être fournies complètes – mais disséminées partie par partie. Selon la progression, la traduction peut être incrémentale, et l'augmentation de la qualité aussi.

4.1.2 Architecture modulaire

De la même façon que pour QRlex, nous avons adopté une architecture modulaire pour permettre un développement séparé et facile. Ces modules sont les suivants.

1. *Module d'import et de segmentation multilingue :* le texte d'un document source est extrait et ensuite segmenté dans un CT. Les UT sont compilées dans cette structure (Figure 26). Selon le document, le texte source nécessite ou non un filtrage pour éliminer tout ce qui est incompatible avec la syntaxe Wiki (Leuf, *et al.*, 2001) (Cameron, *et al.*, 2006). L'outil LingPipe est exploité pour la segmentation (LingPipe, 2006).

2. *Gestion et accès aux ressources de référence :* ces ressources sont prétraitées et importées dans le format XLD présenté dans la partie précédente. L'échange interne entre les modules et l'export se fait à l'aide de cette structure. La structure XLD est facilement transformable en d'autres standards, ce qui permet à nos ressources de référence d'être échangées entre les experts et les logiciels. Il est possible en effet, grâce à ce module, d'importer des dictionnaires entiers qui deviennent actifs instantanément et sont immédiatement disponibles dans le processus de traduction. Les ressources de référence sont bilingues, mais ce module permet d'importer un nombre illimité de dictionnaires bilingues (dans plusieurs couples de langues). La mise à jour de ces ressources se fait en mode collaboratif, c'est-à-dire que les ressources évoluent selon les interventions collaboratives des traducteurs.

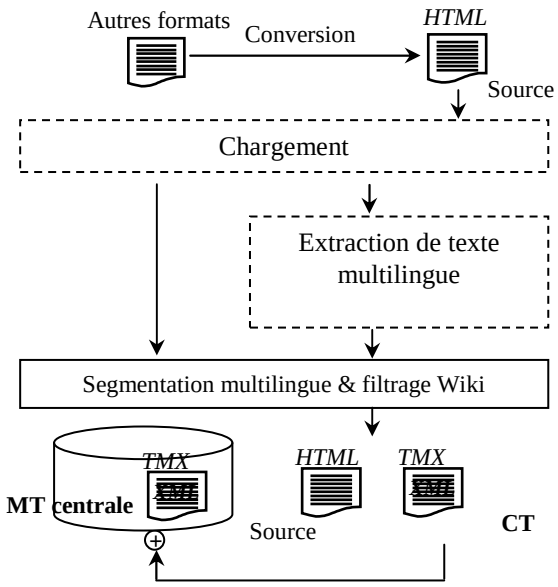


Figure 26 **:** Méthode d'import et prétraitement

3. *Module de gestion des MT :* l'environnement est conçu pour avoir plusieurs MT, une par domaine de traduction de chaque communauté. Tout au long du processus de traduction, un document à traduire est accompagné par une instance d'un CT multilingue gérant les UT. La division des MT selon les communautés permet de réduire le temps d'accès. L'accès aux MT est asynchrone et l'amélioration se fait durant le processus de traduction.

4. *Éditeur multilingue en ligne :* l'éditeur est le module central dans lequel se font les traductions et l'amélioration des ressources. Les aides linguistiques telles que les suggestions dictionnairiques et traductionnelles (par TA et MT) sont calculées de façon asynchrone et proposées de façon proactive.

   Lors de la sélection d'une UT, les fonctionnalités de recherche et de calcul sont déclenchées automatiquement pour servir le contexte en cours. Le traducteur consulte les suggestions ; si elles sont satisfaisantes, alors elles sont facilement sélectionnées et insérées dans la bonne position.

Voici quelques autres caractéristiques intéressantes de l'éditeur :

- les UT sont visualisées en parallèle, à la « Excel ».
- durant ce processus, une segmentation incorrecte peut être corrigée par ajout ou suppression de nouvelles UT.

- il est aussi possible de traduire une UT dans une nouvelle langue par ajout d'une colonne dans l'interface. Les unités source/cible sont synchronisées.

- il est possible aussi d'ajouter des annotations et des commentaires.

- les suggestions sont proposées de façon proactive.

- dans certaines situations, les traductions peuvent se faire en partant d'une version traduite. Un document peut être traduit à partir de la traduction produite par d'autres traducteurs dans une langue cible L_1 dans une nouvelle langue cible L_2 (Figure 27).

Figure 27 : La traduction transitive

5. *Recyclage avancé de données sur le Web :* ce module est une extension de QRselect (qui ne recycle que les documents bilingues japonais-anglais en utilisant un seul dictionnaire). Dans notre environnement, ce module est étendu à d'autres langues. Nous adoptons une solution générique grâce à laquelle plusieurs dictionnaires peuvent être exploités, éventuellement en même temps. Nous montrons comment il est possible de basculer systématiquement d'un dictionnaire à un autre pour un recyclage spécifique à la langue du dictionnaire choisi (Bey, *et al.*, 2006) (Resnik, *et al.*, 2003) .

6. *Gestion de masses de données :* ce module permet la gestion de données volumineuses, documents, dictionnaire ou corpus. Il s'occupe de la gestion interne et du chargement lors du traitement. Il permet la gestion et la récupération, la recherche et la visualisation de données dans le but de produire de nouvelles traductions. Sa construction nous a mené à résoudre les problèmes assez difficiles liés à la gestion de la mémoire et des données.

4.2 Gestion des ressources dans BEYTrans

2.1.1 Ressources dictionnairiques

Nous avons développé un processus d'import générique à travers lequel des dictionnaires en plusieurs langues peuvent être facilement intégrés aux fonctionnalités internes de l'environnement. Le processus d'import permet de transformer les entrées dictionnairiques en

un ensemble de documents Wiki, qui, une fois mis sur le Web, sont manipulés de façon collaborative (Leuf, *et al.*, 2001).

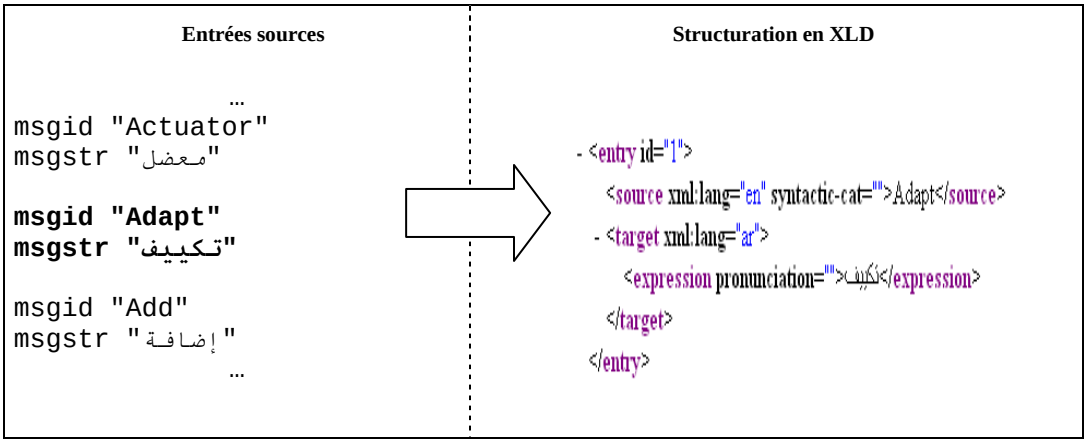


Figure 28 : Format XLD du dictionnaire technique arabe d'Arabeyes

Le principe est simple, mais efficace : le processus d'import récupère les articles dictionnairiques dans le format XLD et les transforme selon la transcription Wiki en un document Web présentant les mêmes informations lexicographiques que celles du format CDM du projet Papillon (Mangeot-Lerebours, 2002). L'injection des dictionnaires dans le Wiki de cette façon permet d'avoir un accès instantané aux nouveaux articles.

Cependant, pour avoir plus de cohérence et de contrôle de la qualité des ressources linguistiques, nous les regroupons par communautés sous des *espaces*. Un *espace* est introduit dans un Wiki pour regrouper certains types de documents selon certains critères. Cela évite le mélange de ressources et facilite l'accès et le suivi des modifications par chaque communauté.

En ce qui concerne l'évolution collaborative de ressources linguistiques, la mise à jour se fait dans deux situations différentes :

a. durant le processus de traduction ;

b. séparément.

Dans la situation (a), les articles absents ou mal traduits sont améliorés dans l'interface de traduction.

Par contre, dans la situation (b), les articles sont modifiés séparément (comme dans Papillon). Les modifications ne se font pas dans des formulaires HTML, mais dans des documents Wiki.

Nous avons développé plusieurs fonctionnalités permettant la manipulation des ressources dictionnairiques :

(i) la création de nouvelles ressources,

(ii) l'ajout de nouveaux articles à un dictionnaire donné,

(iii) la mise à jour des articles.

Les fonctionnalités (ii) et (iii) sont intégrées dans les situations (a) et (b) (Figure 29). Cependant, (i) ne peut être lancée que séparément. Cela est dû au fait que, durant la traduction, les dictionnaires sont déjà sélectionnés et que les traductions qui en dépendent sont déjà actives.

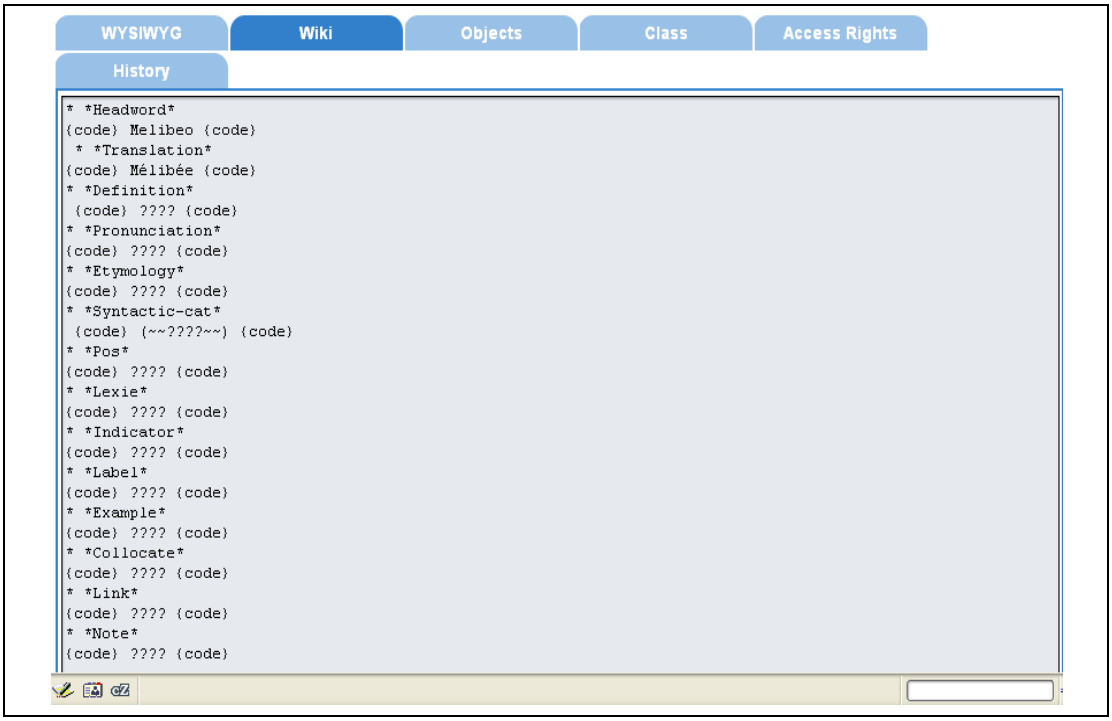


Figure 29 : Interface de manipulation des dictionnaires

L'environnement BEYTrans propose l'utilisation des dictionnaires sous certaines conditions : une ressource n'est manipulable que par la communauté de traducteurs à qui elle appartient, chacune dans un domaine spécifique. L'organisation des ressources est hiérarchique. Il existe des ressources à accès global et d'autres à accès limité, selon le contenu des dictionnaires (). Les droits d'accès sont définis dès l'enregistrement d'un traducteur dans la base des utilisateurs du système.

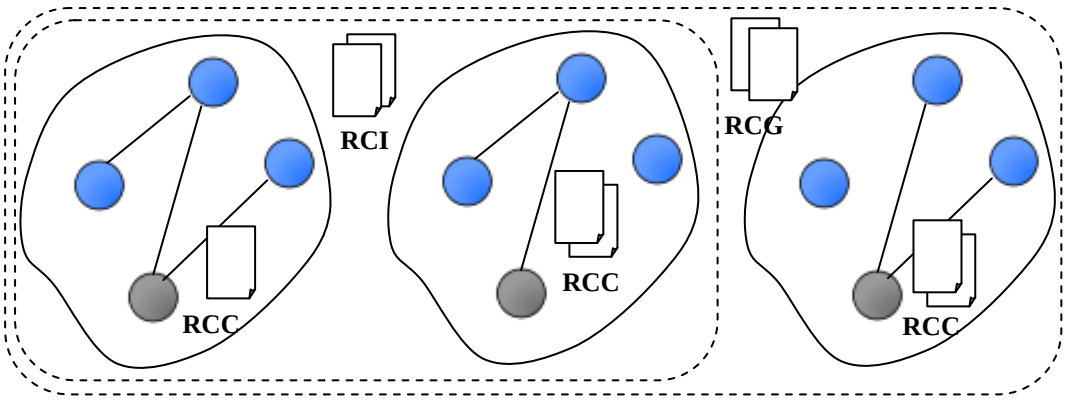


Figure 30 : Manipulation hiérarchique de ressources

Si un traducteur appartient à une communauté donnée et veut accéder aux ressources d'une autre communauté, il doit adresser une demande à l'administrateur de ladite communauté.

2.1.2 Les mémoires de traductions

Les MT sont organisées par domaine. Par exemple, les traducteurs dans le domaine des droits de l'homme disposent d'une MT spécifique à ce domaine. Chacune des communautés dispose donc d'une MT construite à partir des UT des documents qu'elle traduit.

La séparation en plusieurs MT permet une optimisation du processus de recherche et augmente la performance. De plus, les MT spécialisées par communautés permettent d'avoir une bonne cohérence de traduction et d'avoir une meilleure pertinence des suggestions automatiques.

Pour la gestion des UT dans une structure adéquate, nous avons introduit le compagnon de traduction (CT) (cette structure sera décrite en détail dans la sous-section 5.2). L'ensemble des UT contenues dans tous les « CT » constitue la MT d'une communauté. Autrement dit, le nombre des documents d'une communauté donnée correspond au nombre des « CT » dans lesquels toutes les UT sources et traduites sont stockées, et l'ensemble des UT présentes dans les « CT » constitue la MT d'une communauté.

L'utilisation d'un « CT » permet d'avoir une gestion souple des UT et de perfectionner la traduction de la façon suivante :

(i) changement de la direction de traduction sans changement d'interface,

(ii) gestion de plusieurs versions multilingues des UT,

(iii) mise à jour aisée de la MT.

Grace à (ii), les UT peuvent être améliorées progressivement par différents traducteurs. Chaque UT a un ID unique composé du nom de la communauté, du nom du document et d'un numéro d'ordre. Les langues et les autres informations sont facilement identifiées par leurs balises et ces ID uniques.

4.3 Amélioration de la traduction

4.3.1 Amélioration incrémentale

Les traductions incrémentales favorisent le travail communautaire. Les bénévoles peuvent donc traduire une ou plusieurs UT d'un document donné, et les disséminent progressivement. Au début, les documents importés sont disséminés dans leurs états initiaux. Ensuite, il est possible de créer d'autres traductions dans de nouvelles langues.

Figure 31 : Mode « lecture »

On adopte cette stratégie pour permettre la collaboration et la traduction progressive. L'effort des bénévoles peut donc être limité à la traduction de petites unités linguistiques.

4.3.2 Amélioration par des composants de traduction

Les suggestions proactives constituent les principales aides linguistiques de l'éditeur multilingue. Ces suggestions sont les suivantes.

1. *Appel à la TA :* l'exploitation de la TA du Web est peu coûteuse. Il est possible maintenant de proposer des prétraductions par appel aux services gratuits offerts par plusieurs sites tels que Systran et Reverso.

2. *Suggestion par la MT :* plus les fragments de texte sont redondants, plus la MT devient utile. Par exemple, on trouve beaucoup de redondances dans la documentation technique (Chenon, 2005) (Planas, 2000). On a souvent recours au calcul de similarité entre chaînes pour proposer des suggestions. L'algorithme « EDIT » est utilisé pour le calcul de la similarité entre chaîne source et chaînes stockées dans la MT (Levenshtein, 1966).

Les aides de ces trois types, de même que les suggestions dictionnairiques, doivent être proactives, i.e préparées à l'avance et proposées dès qu'on arrive sur un segment, comme le suggérait déjà Ph. Langlais (Langlais, *et al.*, 2000) (Langlais, *et al.*, 2002).

Appeler des systèmes de TA et chercher dans la TM au moment où le traducteur arrive sur un segment courant conduit à des attentes beaucoup trop ou longues (plusieurs secondes), alors que ½ seconde est ressenti comme un peu trop long.

3. *Suggestion externes :* les suggestions externes sont proposées par recyclage de bisegments à partir du Web. Ce problème a été brièvement introduit plus haut (cf. *Recyclage des unités de traduction sur le Web*, p. 93) et sera l'objet d'une étude approfondie dans le chapitre suivant.

4.3.3 Amélioration par communication entre traducteurs

À l'intérieur des communautés organisées existantes, en particulier celles orientées vers une mission précise, la communication entre traducteurs est considérable. L'échange de connaissances se fait par l'envoi de messages via une liste de discussion, par des forums, etc. Les traducteurs demandent souvent des aides pour la traduction de mots techniques, des expressions, etc.

Un moyen intéressant serait de proposer des canaux de communication tels que les IRC où chaque canal serait spécifique à un projet de traduction.

Les mécanismes de suivi de « versions » successives de documents et l'annotation des traductions permettent aussi d'avoir une communication indirecte et transparente à travers laquelle les traducteurs peuvent avoir des idées sur les interventions d'autres traducteurs sur une partie des données.

Conclusion

Nous avons présenté une nouvelle architecture d'un environnement collaboratif avec lequel il serait possible d'améliorer les traductions de documents multilingues de façon incrémentale.

Au début, nous avons exposé un scénario selon lequel la traduction est incrémentale et les traductions sont disséminées partiellement à la volée. Cette stratégie est nouvelle par rapport à celle suivie par les traducteurs professionnels, dont les travaux ne sont disséminés qu'à la fin des traductions.

Le suivi de versions et la communication par l'annotation des traductions permettent aux membres de communautés de se connecter entre eux en fonction de thèmes précis et d'échanger des connaissances pour l'organisation de projets et l'amélioration efficace des traductions.

La technologie Wiki a été largement utilisée dans le prototypage de BEYTrans. L'environnement libre XWiki a été remodelé par l'introduction de nouvelles composantes traductionnelles. D'un côté, nous avons compilé des dictionnaires libres (par exemple 180.000 entrées du dictionnaire Arabeyes) dont l'amélioration et la mise à jour sont faites à la Wiki. D'un autre côté, l'éditeur initial de documents Wiki a été remplacé par un éditeur traductionnel dans lequel des fonctionnalités linguistiques ont été intégrées de façon systématique.

Le premier prototype de BEYTrans permet aux traducteurs une meilleure organisation de leurs projets traductionnels et une gestion efficace des sessions des bénévoles.

Nous nous sommes concentré sur le prototypage de l'environnement et sur les aides linguistiques.

Chapitre 5

Recyclage de documents multilingues sur le Web

Introduction

Nous avons montré dans les chapitres précédents que le Web peut être exploité comme une ressource linguistique à part entière pour identifier les documents et les bisegments déjà traduits.

En ce qui concerne les documents, nous avons déjà proposé et réalisé (en collaboration) dans le cadre de QRLex le prototype QRselect pour le recyclage de paires de documents. Nous l'avons multilinguisé pour le rendre beaucoup plus accessible aux autres communautés. Dans le cadre de BEYTrans, nous l'avons aussi amélioré sur le fond, en proposant et implémentant des solutions au problème de la détection de correspondances traductionnelles plus fines, au niveau des bisegments.

5.1 Unification du codage des caractères en Unicode/UTF-8

5.1.1 Gestion unifiée de l'encodage

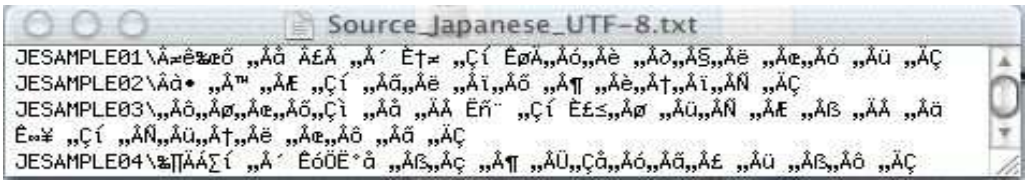
La détection de l'encodage permet d'unifier et de traiter correctement les textes multilingues lors du recyclage. Ce traitement se fait presque par toutes les applications informatiques. Pour une représentation interne et un traitement correct, plusieurs standards ont été développés, dont le but est de pouvoir représenter le jeu de caractères d'une ou plusieurs langues. Par exemple, les standards « ISO 8859-1 » et « CP1252 » sont utilisés pour coder le jeu de caractères français, et les standards « ISO 8859-5 », « CP1251 » et « CP 886 » pour coder le jeu de caractères cyrilliques.

Cependant, ces systèmes de codage n'ont pas cessé de poser de sérieux problèmes. Certains ont été constatés lors de l'échange de données sur les réseaux.

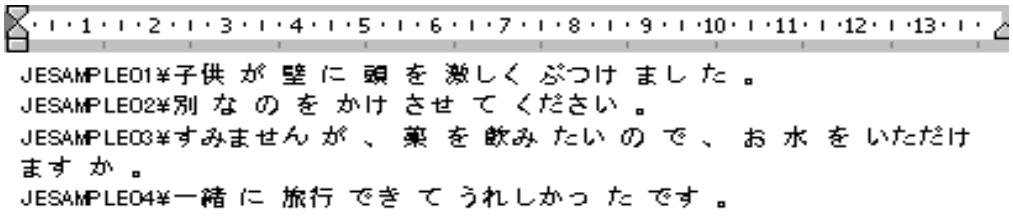
D'autre part, les données nécessitent parfois une conversion pour être correctement visualisées sur les machines non dotées du système d'encodage approprié. Les mêmes problèmes sont apparus à propos de l'échange de paquets de données à travers les réseaux. De

plus, des standards comme « ISO 8859-1 » ne gèrent qu'un jeu de caractères sur 8 bits, ce qui permet de représenter 256 caractères, or, il existe des milliers de caractères qui doivent être représentés. C'est l'une des raisons pour lesquelles le codage UNICODE a été développé. Nous utiliserons l'encodage UTF-8 d'Unicode.

Un système de codage correct permet de gérer et de visualiser correctement des données multilingues (Figure 32), en particulier lorsqu'il s'agit de les importer et de les gérer dans un seul volume.



(a) : Affichage incorrect



(b) : affichage correct

Figure 32 : Problème de visualisation dû à une mauvaise détection du codage

Il est donc nécessaire de détecter automatiquement le codage du jeu de caractères des documents source pour avoir un traitement correct et unifié des textes durant le recyclage. Nous avons unifié nos données en UTF-8 car ce standard est largement utilisé pour gérer les données multilingues, et peut être produit à l'aide de multiples outils disponibles sur le Web.

D'autres problèmes se posent aussi lors de l'exploration du Web. Des documents dans une langue donnée peuvent être codés différemment. Par exemple, les documents japonais peuvent être codés en EUC-JP ou en SHIFT-JS. Il est donc primordial de détecter les langues et les codages pour pouvoir traiter le texte correctement et l'importer dans la BD.

1.1.1 Détection de la langue source : documents homogènes

Nous avons utilisé le système SANDOH (Vo-Trung, 2004). Ce système identifie dans un document quelconque les zones « homogènes » de même langue et de même codage. Il

transforme le document en UTF-8 et encadre chaque zone homogène par des balises indiquant le couple <langue, codage> de cette zone.

Pour les textes hétérogènes, l'interface Web est la suivante :

Figure 33 : Exemple de l'interface d'analyse d'un texte hétérogène par SANDOH

SANDOH analyse le texte et renvoie un texte structuré avec des balises donnant le couple <L, C> pour chaque zone homogène :

<Fr-cp1252>Au commencement, Dieu créa les cieux et la terre. <\Fr-cp1252>
<Ge-cp1252> Im Anfang schuf Gott die Himmel und die Erde. <\Ge-cp1252>
<It-cp1252> In principio Dio creò il cielo e la terra. <\It-cp1252>
<No-cp1252>I begynnelsen skapte Gud himmelen og jorden. <\No-cp1252>
<Po-cp1252>Registro da genealogia de Jesus Cristo, filho de Davi, filho de Abraão: <\Po-cp1252>
<Sp-cp1252> En el principio creó Dios los cielos y la tierra. <\Sp-cp1252>
<Ru-cp1251>Родословие Иисуса Христа, потомка Давида, потомка Авраама<Ru-cp1251>

Ce résultat est visualisé après avoir converti le résultat en UTF-8.

1.1.2 Détection de la langue source : documents hétérogènes

SANDOH est aussi capable de détecter les zones hétérogènes dans les documents multilingues (Vo-Trung, 2004). Il a été évalué par une expérience sur un document contenant 11.000 phrases en 11 langues. Les résultats obtenus sont encourageants (Table 7).

Langue	Codage	Nb phrases testée	Nb phrases exactes	Taux exact
Anglais	CP 1252	1000	998	99,80
Français	UTF-8	1000	1000	100,00
Espagnol	CP 1252	1000	990	99,00
Allemand	CP 1252	1000	993	99,30
Portugais	CP 1252	1000	995	99,50
Italien	CP 1252	1000	990	99,00
Russe	KOI-8	1000	1000	100,00
Vietnamien	TCVN3	1000	900	90,00
Vietnamien	UTF-8	1000	900	90,00
Vietnamien	VNI	1000	850	85,00
Vietnamien	VPS	1000	890	89,00

Table 7 : Evaluation de la performance de SANDOH sur un texte hétérogène

Ce système a été intégré dans nos outils pour gérer efficacement la détection des UT (unités de traduction) et les MT (mémoires de traductions).

5.1.2 Gestion des correspondances sources/cibles

La synchronisation entre les UT source et cible est gérée dans des documents XML TMX correspondant à un CT (compagnon de traduction). Un segment source et ses traductions sont enveloppés dans un élément XML TMX <tu>. Un <tu> contient le segment source et ses traductions, qui sont gérées par des éléments XML TMX <tuv>.

La synchronisation se fait par l'attribution d'un identificateur unique à chaque UT, donc à chaque élément <tu>. Ces identificateurs permettent donc de synchroniser les segments (<tuv>) dans différentes langues correspondant à une même UT (<tu>) .

La construction des bisegments <S,C> se fait par traduction incrémentale collaborative. Au début, il n'y a que les UT source dans le CT. L'ajout d'une nouvelle traduction (un segment cible) insère un nouvel élément <tuv> dans l'élément <tu>. Un élément <tu> peut contenir plusieurs éléments <tuv> (Figure 34).

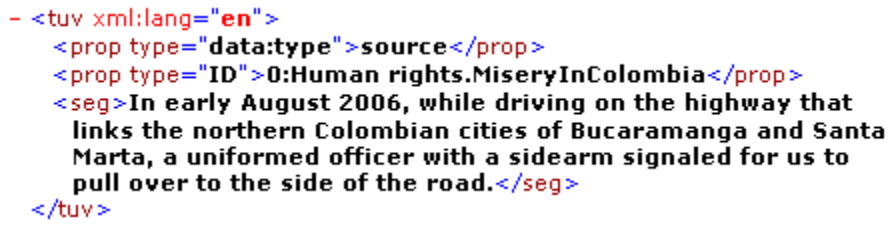


Figure 34 : Structure d'un <tuv> dans le compagnon de traduction (CT)

Pour récupérer une UT, le système utilise son identificateur (ID) qui se compose du nom de la communauté (ex : W3C), du nom du document, et du numéro d'ordre de l'UT dans le document.

5.2 Prétraitement de documents déjà traduitts

5.2.1 Segmentation multilingue

La segmentation est un problème à part entière, difficile, et non résolu complètement pour toutes les langues (Kraif, 2001). Nous avons donc dû traiter cette question, bien qu'elle soit éloignée de notre thème principal de recherche, parce qu'elle est nécessaire pour la construction des MT et que les outils que nous avons essayés n'étaient pas du tout satisfaisants.

Comme nous voulons traiter beaucoup de langues, nous ne pouvons pas envisager d'utiliser une segmentation « experte » (à règles et dictionnaires) pour chaque langue. Nous nous sommes donc limité à l'exploitation de méthodes statistiques basées sur les n-grammes en utilisant l'outil « Linpipe », qui nous donne la possibilité d'apprentissage à base de corpus. Les modèles de langues pour les langues européennes et le chinois existent déjà. Dans nos futurs travaux, nous souhaitons l'étendre vers d'autres langues telles que l'arabe et le japonais. Il faut en effet construire ou trouver des corpus convenables pour cette tâche. Dans notre environnement, deux niveaux de segmentation sont nécessaires pour le traitement de données dans notre environnement : segmentation en phrases et segmentation en mots.

2.1.1 Segmentation en mots

La plupart des méthodes existantes se basent sur la détection des mots en utilisant des symboles ou séparateurs. Les langues à écriture alphabétique sont plus au moins faciles à segmenter en mots : c'est facile pour toutes les langues de l'Europe mais difficile des langues asiatiques comme le thai, le lao, le khmer, et même le vietnamien (où les blancs séparant les sylabes et pas les mots). Certaines (thai, lao) n'ont pas non plus de séparateurs de phrases. Les écritures utilisant des idéogrammes (chinois, japonais) posent le même problème (absence de

séparateurs de mots), avec une difficulté supplémentaire due au grand nombre des idéogrammes.

La tâche n'est pas triviale, notamment à cause des rôles ambigus des séparateurs (comme le blanc, l'apostrophe, le tiret, etc.). Leur présence est parfois arbitraire (par exemple, « parce que » et « lorsque ») et, malheureusement, n'implique pas une frontière entre mots « linguistiques » (ex : « aujourd'hui », « rendez-vous », « pomme de terre », etc.). La segmentation en mots dépend du dictionnaire morphologique utilisé et des stratégies choisies pour le traitement des mots composés (Aït-Mokhtar, *et al.*, 2003).

Nous exploitons des modèles de langue existants pour la segmentation des langues écrites avec des séparateurs de mots typographiques.

Les langues écrites sans séparateurs de mots seront traitées par des outils spécifiques. Il est toutefois important de dire qu'il est impossible de gérer toutes les langues par des modèles de langue, parce que ce processus nécessite la construction pour chaque langue d'un corpus annoté. Nous nous sommes limité au japonais et au chinois jusqu'à présent.

2.1.2 Segmentation en phrases

Avant d'exposer les problèmes liés à la segmentation en « phrases », il faut donner une définition de ce terme.

Selon (Aït-Mokhtar, *et al.*, 2003) :

« Une phrase est l'unité minimale de communication linguistique. Si on considère la phrase comme une unité linguistique (suivie d'une pause importante), le problème est alors de déterminer ces pauses dans les textes. Elles sont généralement représentées par un point, mais aussi par les points de suspension, le point d'interrogation, le point d'exclamation, le point-virgule, le double point.»

Il écrit encore en faisant référence au Grévisse (1993) que :

« La virgule peut même séparer des phrases, que nous appelons sous-phrases dans ce cas. L'absence du point n'implique donc pas la continuité de ce qui est intuitivement appelé *"phrase". »*

De plus, ces signes de ponctuation peuvent marquer une simple coupure, ou « discontinuité syntaxique », comme le montrent les exemples suivants :

(1) Je vais à la pêche avec toi ! cria-t-il. (Colette)
(2) Le prochain rendez-vous de la section est prévu le 5 mai. Pour voter. (Le Monde)

Cette analyse montre la difficulté que pose la segmentation par les signes de ponctuation qui peuvent ne donner qu'une segmentation partiellement correcte.

Le problème est loin d'être résolu complètement, surtout pour la segmentation multilingue. Les applications informatiques utilisent toujours une segmentation basée sur ces signes de ponctuation. (Kraif, 2001) a proposé une segmentation basée sur des règles syntaxiques construites à la main, mais il s'est limité aux cas de figure les plus fréquents. Il a noté qu'il peut être nécessaire (pour certains textes) de fournir un dictionnaire d'abréviations *ad hoc*. En se basant sur les signes de ponctuation, il a ainsi défini les règles suivantes :

- Phrases = (Mots Séparateurs (Séparateurs)*)* DerniersMots point espace Maj | (Mots Séparateurs (Séparateurs)*)* SymbolesTerminaux.

- Séparateurs = {espace, virgule, point d'exclamation, point d'interrogation, tiret, guillemets, parenthèses}.

(Un mot = toute suite maximale de caractères ne contenant pas de séparateur.)

- DerniersMots = Mots à l'exception des abréviations (etc., cf., pp., MM.) et des mots d'une lettre.

Notons qu'il existe beaucoup d'algorithmes implémentés dans des applications diverses offrant la segmentation par les signes de ponctuation.

Par exemple, nous avons expérimenté l'outil « OmegaT » (outil gratuit d'aide à la traduction) pour tester l'efficacité de la segmentation dans un contexte traductionnel. L'outil produit beaucoup de segments incorrects et échoue devant des abréviations comme « Mr. » ou même les adresses de courriel. Par exemple, une adresse « prenom.nom@domaine.fr » est segmentée en « prenom » « nom@domaine » « fr » parce que « OmegaT » traite de façon statique les phrases en se basant sur des signes définis à l'avance en ne prenant pas en compte les items précédents et suivants pour savoir si c'est la fin d'un segment.

2.1.3 Méthode de segmentation (heuristique et semi-automatique)

Il y a une autre méthode de segmentation proposée par (Carpenter, 2006) et implémentée dans l'outil LingPipe (LingPipe, 2006).

Cette méthode identifie les frontières des segments (ou des phrases) en se basant sur la détection des items suivants et précédents d'un item candidat à partir d'un ensemble de séquences extraites a priori d'une phrase. Si un item est jugé être une frontière dans une séquence, alors la frontière de la phrase est l'index de l'avant-dernier caractère de l'item. Pour

qu'un item délimite la « fin » d'une phrase, il doit être un élément de l'ensemble des items de ponctuation tels que le point « . », le point d'interrogation « ? », etc. Il doit aussi être suivi par une espace et l'item suivant doit être un « item » de début d'une nouvelle phrase (par exemple, un « item » avec une majuscule initiale).

Les phrases contenant des abréviations telles que « Mr. Smith » sont problématiques parce qu'un modèle de phrase simpliste traitera le point dans « Mr. » comme un item délimitant. En effet, il est nécessaire alors de vérifier l'item pénultième et de rejeter les abréviations communes. Le modèle de phrases adopté est basé sur trois types d'item :

Pour illustrer ce modèle, nous prenons l'exemple suivant :

I saw Mr. Smith.

Où les items dans cette pharese sont : (1) « I » (2) « saw » (3) « Mr » (4) « . » (5) « Smith » (6) « . »

- Les arrêts possibles : ce sont les items qui peuvent être pris comme fin de phrase (par exemple « . », « ? », « ! », etc.), dans l'exemple, les items 4 et 6

- Les « pénultièmes impossibles » : ce sont les items qui ne sont pas des items « pénultièmes » dans les phrases. Cet ensemble contient typiquement des abréviations ou des acronymes. On arrive à l'item 4 (« . »), appelé le « dernier ». Le « pénultième » est l'item 3 (« Mr ») et l'antépénultième est l'item 2 (« saw »).

- Les débuts impossibles : ce sont les items qui ne peuvent pas être en début de phrase. Cet ensemble inclut typiquement les caractères de ponctuation qui doivent être attachés à ce qui précède, comme les guillemets (par exemple « » », « ' », « } », etc.)

Selon (Carpenter, 2006), il y a aussi deux critères qui déterminent les aspects de la détection des frontières d'une phrase :

- Frontière finale forcée : il est possible de prendre un item final dans n'importe quelle entrée comme un délimiteur d'une phrase, dans le cas ou non d'un possible item d'arrêt.

- Parenthèses équilibrées : si les parenthèses sont généralement équilibrées, alors tant qu'il y a des parenthèses ouvrantes qui ne sont pas fermées, la phrase en cours ne doit pas être terminée. Les

crochets carrés (« [», «] ») et les parenthèses (« (», «) ») sont équilibrés séparément, autrement dit, un couple de crochets ne doit pas fermer une parenthèse ouvrante et vice versa. Le modèle à construire ne prend volontairement pas en charge les parenthèses imbriquées. C'est-à-dire que la première parenthèse fermante suivant n'importe quel nombre de parenthèses ouvrantes ferme toutes les parenthèses et n'importe quelle parenthèse fermante suivante «)» ou crochet «]» est tout simplement rejetée. Cette approche évite au segmenteur d'être piégé si les extrémités d'une séquence de parenthèses (crochets ou accolades) sont mal équilibrées.

Cette méthode de segmentation a été implémentée dans l'outil « LingPipe » qui intègre aussi des méthodes d'apprentissage basées sur des n-grammes et la création de modèles de langage. Il est possible de l'utiliser pour l'apprentissage de la segmentation de toutes les langues que nous traitons. Pour cela, il faut préparer des corpus d'apprentissage pour chaque langue.

Dans notre cas, nous limitons la segmentation au japonais, au chinois et à l'arabe, car le paquetage de l'outil offre des corpus d'apprentissage déjà préparés. En ce qui concerne les langues européennes principales (anglais, français, italien, etc.), l'outil est déjà entraîné et effectue la segmentation des textes dans ces langues. Notons cependant qu'il ne s'applique qu'à du texte brut, pas à du html par exemple.

Le choix de cet outil est justifié par son extensibilité à d'autres langues. Nous combinons dans notre environnement cette approche avec l'éditeur de traductions qui, selon les besoins, offre la possibilité de produire facilement en mode collaboratif des corpus annotés pour l'apprentissage de nouvelles langues (la création collaborative de corpus est détaillée dans la troisième partie).

Les segments générés par l'approche ci-dessus nécessitent dans certains cas des améliorations par intervention humaine. C'est pourquoi notre environnement intègre dans l'éditeur des possibilités pour améliorer la segmentation (division, fusion, suppression de cellules, etc.). Les segments correctement segmentés sont stockés dans le compagnon de traduction (CT).

2.1.4 Gestion des segments : le concept de compagnon de traduction

Le compagnon de traduction (CT) est la structure la plus importante dans la gestion des documents en UT et des MT dans BEYTrans.

Un « CT » est un document en format XML « TMX » (Translation Memory eXchange) dans lequel sont stockées les UT extraites d'un document source. Le « CT » est le compagnon du document source tout au long de sa traduction ; il évolue au fur et à mesure que la traduction progresse. Une UT peut être mise à jour, ajoutée ou supprimée. Cette structure est le support dans lequel les paires de segments <source, cible> sont gérées. À chaque document source est donc associé un CT.

Les nouvelles traductions sont ajoutées à cette structure dans une position permettant leur synchronisation avec la source.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <tmx version="1.0">
    <header creationtoolversion="1.0.0" datatype="text" segtype="sentence"
     adminlang="EN-US" srclang="en" o-tmf="unknow" creationtool="BEYTrans" />
  - <body>
    - <tu>
      - <tuv xml:lang="en">
          <prop type="data:type">source</prop>
          <prop type="ID">0:Human rights.MiseryInColombia</prop>
          <seg>In early August 2006, while driving on the highway that
            links the northern Colombian cities of Bucaramanga and Santa
            Marta, a uniformed officer with a sidearm signaled for us to
            pull over to the side of the road.</seg>
        </tuv>
      - <tuv xml:lang="ar">
          <prop type="data:type">target</prop>
          <prop type="ID">0:Human rights.MiseryInColombia</prop>
          <seg />
        </tuv>
      - <tuv xml:lang="fr">
          <prop type="data:type">target</prop>
          <prop type="ID">0:Human rights.MiseryInColombia</prop>
          <seg />
        </tuv>
      - <tuv xml:lang="jp">
          <prop type="data:type">target</prop>
          <prop type="ID">0:Human rights.MiseryInColombia</prop>
          <seg>2006年8月上旬、コロンビア北部の都市ブカラマンガとサンタ・マルタを結
            ぶ高速道路を車で走っていたとき、制服を着たオフィサが私たちに道路脇によって
            停車するよう手で信号を送ってきた。</seg>
        </tuv>
      - <tuv xml:lang="sp">
          <prop type="data:type">target</prop>
          <prop type="ID">0:Human rights.MiseryInColombia</prop>
          <seg />
        </tuv>
      </tu>
```

Figure 35 : Structuration des segments dans un compagnon de traduction

Le standard TMX a été complètement utilisé (Lisa, 2008). De plus, il a été étendu pour ajouter d'autres informations concernant les utilisateurs, les droits d'accès aux données, etc.

Ces nouvelles modifications seront détaillées dans la troisième partie où nous traiterons des données beaucoup plus complexes et volumineuses concernant les corpus destinés à la TA.

5.2.2 Construction de la mémoire de traductions

2.2.1 Adaptation du standard TMX

Le standard TMX (Translation Memory Exchange) est un standard développé par le consortium LISA dans le but d'offrir une structure compatible à tous les outils d'aide à la traduction et pour éviter que les traducteurs ne créent plusieurs formats incompatibles. Il permet entre autres de gérer des traductions multiples des segments source. Le standard TMX est utilisé par la majorité des outils d'aide à la traduction (Trados, 2005) (Similis, 2005) (DéjàVu, 2007). Il présente les avantages suivants :

- Échange des MT : l'avantage le plus évident de TMX est de pouvoir échanger des MT entre outils et experts humains, sans devoir à chaque fois développer de nouveaux « filtres » (transformations de formats).

```
<p>The big<b>black</b>cat</p>
<p>Le gros chat<b>noir</b></p>
```

Figure 36 : Un code HTML source – à transformer en TMX

- Liberté de choix : les experts ont plus de liberté pour choisir leurs systèmes, cela leur assure de ne pas être dépendants ou bloqués sur un outil particulier.

- Portabilité : ce standard est bien défini. Il permet aux développeurs d'ajouter des fonctionnalités en les implémentant de façon portable et indépendante des formats originaux.

La Figure 36 montre deux énoncés dans leur format HTML. La transformation en TMX met les segments textuels entre les balises *tuv* (translation unit version), elles-mêmes placées entre les balises *tu* (translation unit) (Figure 37).

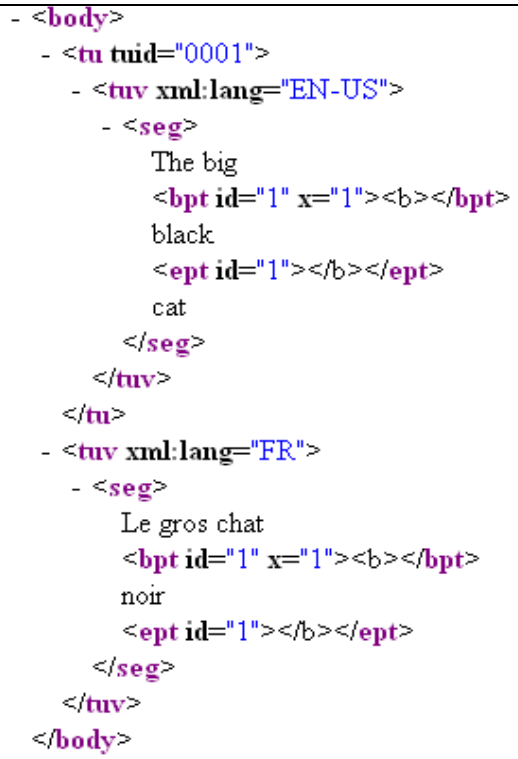


Figure 37 : Transformation du code HTML ci-dessus en TMX

2.2.2 Import des UT de communautés spécifiques

Dans BEYTrans, le processus de recyclage se fait de deux façons différentes :

1. par recyclage de « tout » le Web,
2. par le recyclage de sites « amis ».

Dans cette section, nous ne nous intéressons qu'au deuxième type de recyclage, c'est-à-dire au recyclage de « sites amis ».

Le système QRselect a été amélioré pour permettre le recyclage d'un « domaine Web » bien déterminé (par exemple, PaxHumana). La méthode de recyclage est la suivante.

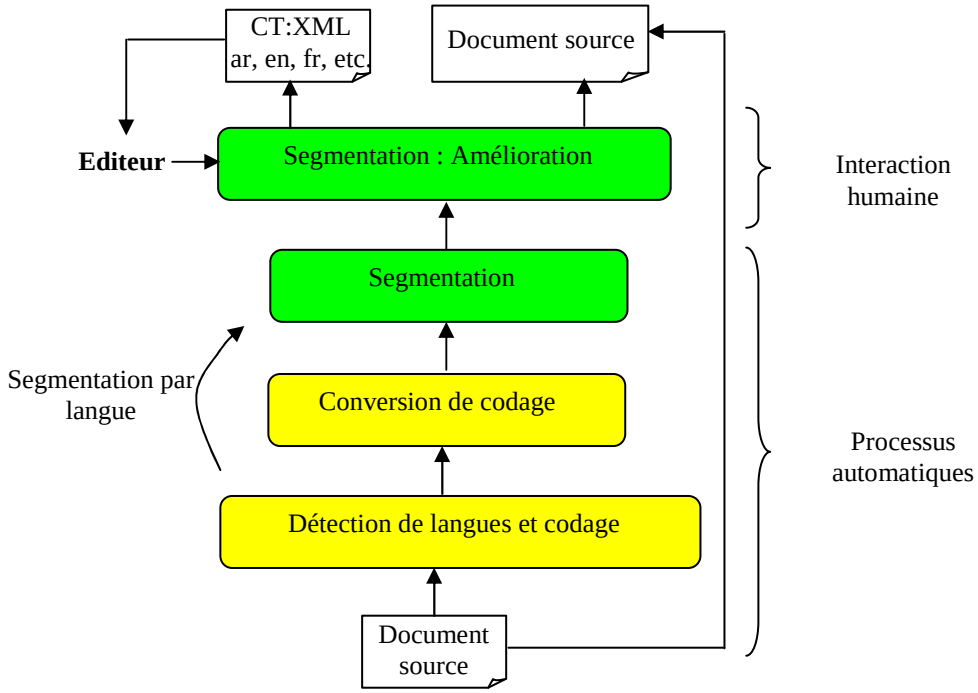


Figure 38 : Processus d'import et de segmentation

Le système reçoit en entrée un domaine Web. Il parcourt toutes les pages Web accessibles de ce domaine pour y chercher des paires de documents (source/cible). Une fois une paire de documents identifiée, elle est passée à l'aligneur « GIZA++ » pour construire des paires d'UT (unités de traduction). À la fin, les paires d'UT sont compilées dans la mémoire de traductions dans un CT.

2.2.3 Fonctionnalités utiles

Les fonctionnalités de la construction de ressources linguistiques par recyclage ainsi que l'exploitation d'autres ressources dictionnairiques doivent être complétées par des fonctionnalités additionnelles telles que la visualisation des données et le réglage de l'interface. Dans BEYTrans, nous avons développé les fonctionnalités additionnelles suivantes :

- Visualisation : les UT sont présentées aux traducteurs sous un format « parallèle ». Les traductions se font donc segment par segment. Cela favorise la traduction incrémentale où les traducteurs peuvent travailler sur des segments différents en même temps.
- Correction de la segmentation : la correction des erreurs détectées à ce niveau est possible. Sous l'éditeur, on peut :

- diviser des cellules,
- fusionner des cellules,
- multi-sélectionner plusieurs cellules (UT) pour appliquer une fonctionnalité commune (suppression, fusion, etc.).
- voir les contextes : les UT recyclées sont présentées dans le contexte de leur apparition dans le document.

Ces fonctionnalités sont utiles et pas très difficiles à réaliser. Dans la section suivante, nous aborderons le problème bien plus difficile du recyclage de données traductionnelles.

5.3 Recherche de données traductionnelles dans BEYTrans

Le recyclage de traductions existant sur le Web permet de compenser le fait que la MT d'une communauté est initialement vide, et d'accélérer sa croissance, puisqu'on y met aussi bien les résultats des traductions faites sur le site que des traductions faites ailleurs.

Notre méthode de recyclage adopte le principe des systèmes STRAND et QRselect : nous recherchons d'abord des paires de documents par un robot et par un calcul de taux de proximité. Ensuite, nous déterminons les bisegments par une méthode d'alignement (utilisation de GIZA++).

Nous « multilinguisons » QRselect en intégrant plusieurs dictionnaires. Le résultat est assez bon, comme le montrent nos expériences et les évaluations que nous avons faites.

5.3.1 Multilinguisation de QRselect

Pour le calcul des équivalences lexicales entre les paires de documents, QRselect utilise un dictionnaire bilingue. La première version de l'outil exploite le dictionnaire « Eijiro » (30 000 entrées) pour détecter les paires de documents anglais-japonais. L'extension à d'autres langues nécessite donc l'introduction de nouveaux dictionnaires libres et la localisation de l'interface utilisateur.

Comme il s'agit du recyclage pour l'aide aux traducteurs bénévoles, nous souhaitons sélectionner des dictionnaires disponibles librement sur certains sites de traduction bénévole.

Bien que nous visions à terme une multilinguisation très générale de QRselect, nous nous sommes jusqu'ici limité aux couples de langues anglais-arabe et anglais-français, faute de

moyens et de temps (cela nécessite des personnes connaissant ces couples de langues). Mais l'outil reste ouvert à l'intégration d'autres langues.

La multilinguisation de QRselect est faite à trois niveaux : import de dictionnaire, transformation en format Eijiro, et définition des mots réservés.

3.1.1 Sélection et import de nouveaux dictionnaires

Pour adapter QRselect à un couple L_1-L_2, il faut disposer d'un dictionnaire L_1-L_2.

Est-il vraiment possible de multilinguïser QRselect dans toutes les langues ? En théorie oui, en pratique pas encore.

Primo, il est difficile de trouver des dictionnaires dans toutes les langues, en particulier ceux qui correspondent aux besoins des traducteurs bénévoles.

Secundo, il est possible d'ouvrir l'environnement aux traducteurs pour leur permettre d'importer ou de créer de nouveaux dictionnaires. Mais une phase de prétraitement et de préparation est nécessaire. Si elle pouvait être faite correctement, alors il serait possible théoriquement de faire le recyclage dans toutes les langues. Cela dit, une fois un dictionnaire importé dans la BD du système, il ne reste alors qu'à définir si nécessaire des balises ou étiquettes spécifiques pour les nouvelles langues.

Le premier dictionnaire que nous avons traité est celui de la communauté « Arabeyes ». Le deuxième est un dictionnaire de la communauté « FrenchMozilla » qui travaille sur la traduction de documents techniques et la localisation des logiciels de la suite Mozilla.

(a) **msgid** "Annotation"
msgstr "تحشية، تعليقات ، تعليق ،حاشية"
(b) **msgid** "Antialiasing"
msgstr "إزالة التسنين"

Figure 39 : Structure des entrées du dictionnaire « Arabeyes »

- Dictionnaire technique d'Arabeyes : ce dictionnaire contient 18 000 entrées anglais-arabe. C'est un des plus grands dictionnaires, comparé avec les dictionnaires d'autres communautés de localisation/traduction de logiciels et de systèmes libres (par exemple, FrenchMozilla). Il est disponible en format « po », ce qui le rend facilement manipulable par les outils libres tels que

« poedit »51, et il est maintenu sous un Wiki en mode collaboratif par plusieurs dizaines de bénévoles.

Les mots vedette sont identifiés par les balises « msgid » qui sont propres au format « po ». Les traductions sont récupérables par les balises « msgstr ». Dans le cas de traductions multiples, elles sont séparées par des virgules, comme dans « تحشية، تعليقات ، تعليق ،حاشية » associé au mot vedette « Annotation ».

- Dictionnaire technique FrenchMozilla : ce dictionnaire a été récupéré à partir du site de francisation de logiciels libres du progiciel Mozilla. Il contient 465 entrées anglais-français consacrées totalement au domaine technique informatique. Bien que sa taille soit restreinte, nous l'avons choisi pour aider la communauté francophone de localisation à identifier les traductions techniques existantes. Ce dictionnaire existe sous forme de document HTML (cf. Les documents à traduire sont organisés dans un arbre CVS (Concurrent Versions System) où les documents sont répertoriés avec les informations qu'ils contiennent. Les traducteurs doivent savoir comment exploiter ce type de serveur de version (récupération, mise à jour, etc.) pour envoyer et récupérer les documents à traduire.

- Ressources linguistiques, p. 13 : Figure 2). Nous avons récupéré toutes les pages, les avons prétraitées, puis avons alors facilement compilé leur contenu en XLD.

Mots vedettes	Traduction	Note
Access key	Clé d'accès	usage
Account	Compte	tacite
Active character codings	Jeux de caractères	actifs
Address bar	Barre d'adresse	revoir

Table 8 : Quelques entrées du dictionnaire « FrenchMozilla »

Le prétraitement de FrenchMozilla a été fait manuellement car la structure des entrées n'était pas homogène.

Les « notes » donnent aux traducteurs une idée sur l'état de traduction de termes (FrenchMozilla, 2005) :

- *Figé* : un choix a été fait et approuvé.

[51] L'un des outils libres les plus utilisés par les traducteurs/localiseurs de logiciels et systèmes libres. Les messages et chaînes à traduire sont généralement compilés dans des fichiers « po ».

- *Ouvert* : aucun choix n'a été fait et le débat est ouvert.
- *Débattu* : la traduction a été discutée, mais aucun choix ne s'est dégagé.
- *Tacite* : le terme n'a été l'objet d'aucune discussion. Tout le monde semble d'accord.
- *Revoir* : le terme n'a été l'objet d'aucune discussion, mais à la relecture, le choix semble incertain, voire douteux.
- *Usage* : la traduction semble être liée à un contexte particulier, ou il s'agit d'une erreur d'interprétation.

Tous les termes du dictionnaire ont été gardés, aucun terme n'a été privilégié par rapport aux notes des traducteurs. D'autres prétraitements sont expliqués dans la sous-section suivante.

3.1.2 Transformation en format d'« Eijiro »

Tous les dictionnaires ont été importés dans le même schéma que celui du dictionnaire « Eijiro » dans la BD. L'une des raisons principales consiste à éviter de reprogrammer ou de modifier des modules de QRselect. Par exemple, après la phase de prétraitement, l'entrée (a) (Figure 39) aura le format :

« Annotation تحشية تعليقات تعليق حاشية »

Cette phase est suivie d'une phase de transformation de dictionnaires en format XLD.

3.1.3 Définitions de mots réservés

Chaque nouveau dictionnaire doit être accompagné d'un ensemble de nouveaux mots réservés (étiquettes ou balises). Cet ensemble est utile pour la détection des « ancres » qui orientent la recherche vers la source de la traduction.

Deux ensembles ont été définis « manuellement » après avoir analysé les sites de traduction des communautés correspondant aux couples de langues anglais-arabe et anglais-français : W3C (anglais-arabe) et (anglais-français), Paxhumana (anglais-français), Arabeyes (anglais-arabe), etc.[52].

[52] Notons que certaines communautés ne mettent pas d'ancre mais indiquent la source par des liens HREF et IMG (HTML). Cela dit que les documents sources ne sont pas référenciés par des textes.

Pour avoir plus de souplesse et de robustesse durant le recyclage et la recherche des ancres, nous avons fait en sorte que ces deux ensembles soient extensibles directement par les traducteurs et que l'interface de QRselect soit multilingue. Pour rendre cela possible, nous avons amélioré l'interface de QRselect.

MRAA (anglais-arabe) = {(traduction) ترجمة, (anglais) انجليزيه, (source) مصدر, (source) الأصل, (source) المرجع, (informations) اخبار, (anglais) english}[53].

MRAF (anglais-français) = {anglais, source, original, destination, news, translation, référence}.[54]

Nous avons multilingualisé l'interface en arabe, français, anglais et japonais (une tâche nécessaire pour les communautés bilingues) (Figure 40). Ensuite, nous avons ajouté un module d'import de nouveaux dictionnaires.

D'un autre côté, nous avons ajouté une liste avec des fonctionnalités dynamiques : les traducteurs peuvent introduire de nouveaux mots réservés par des ajouts, des mises à jour ou des suppressions. La Figure 41 illustre une liste dynamique de mots réservés pour la langue arabe.

QRselect devient alors beaucoup plus souple et générique. L'ajout dynamique de nouveaux dictionnaires permet théoriquement d'ouvrir l'environnement à toutes les langues. Les traducteurs n'ont pas non plus besoin de paramétrer le recyclage.

[53] La source en arabe a plusieurs traductions possibles.
[54] MRAA : mots réservés anglais-arabe, MRAF : mots réservés anglais-français.

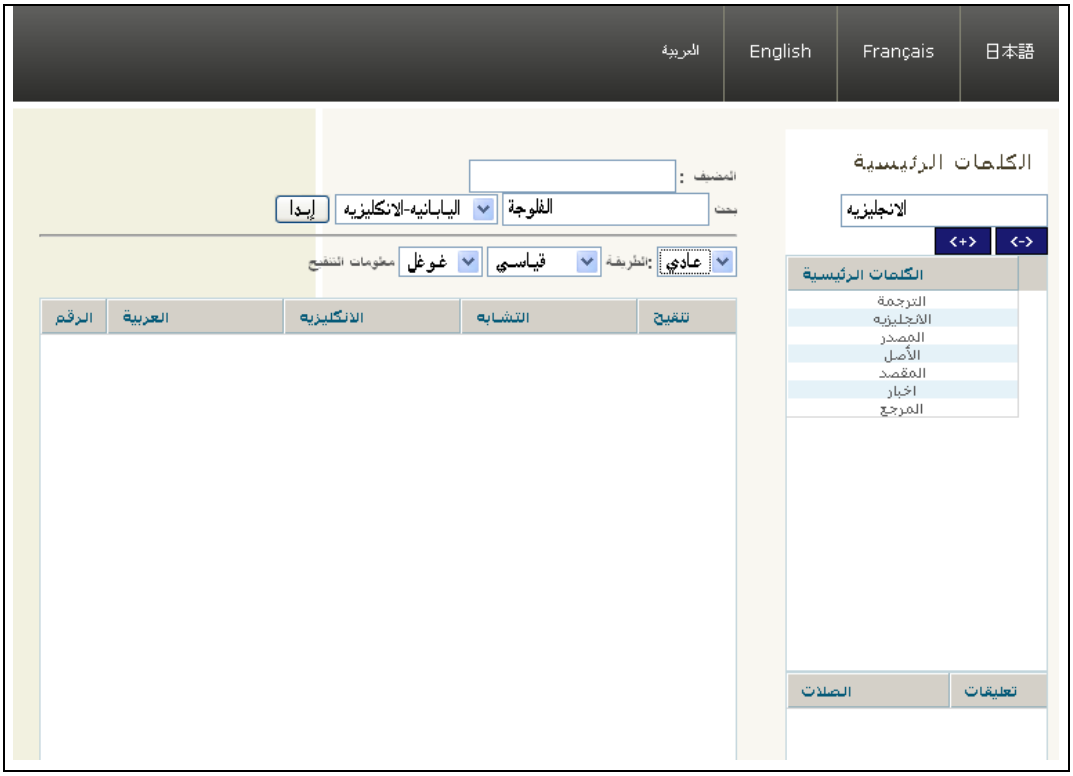


Figure 40 : Localisation de QRselect en Arabe (QRselect-2)

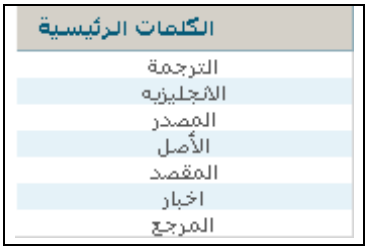


Figure 41 : Liste extensible de mots réservés.

En effet, le choix d'une langue disponible dans l'interface ou d'un dictionnaire bilingue permet de sélectionner automatiquement l'ensemble de mots réservés et les paramètres internes : encodage des pages, méthode de segmentation, etc.

La Figure 42 illustre ces modifications.

La partie supérieure (A) de l'interface permet le changement de la langue de l'interface.

Les mots réservés sont modifiables dans la zone (B) où il est possible d'éditer chaque mot dans sa propre cellule.

La zone (C) contient les informations de configuration de recyclage (introduction de mots-clés « コロンビア 麻薬 » et sélection manuelle d'un dictionnaire).

Figure 42 : Nouvelle interface de QRselect

Enfin, la zone (D) affiche les résultats de recyclage sous forme d'une grille : ID, URL (+information : titre, encodage, etc.) de la cible, URL de la source, information de débogage (affichage des paires non pertinentes), taux de proximité (l'algorithme complet et la méthode de calcul sont présentés dans la sous-section suivante).

Techniquement, les paramètres de recyclage sont envoyés en UTF-8 au serveur à travers une requête *HTTPRequest* à l'aide d'une API *Ajax*. Le serveur les extrait à l'aide d'une « Servlet » qui, a son tour, appelle le robot de recyclage qui détecte les paires candidates et renvoie une liste d'URL de documents.

3.1.4 Nouvel algorithme de recyclage de documents

L'algorithme implémenté dans la première version de QRselect (cf. *Méthode de recyclage*, p. 95) ne permet de recycler que les paires de documents du couple anglais-japonais, et les mots-réservés sont codés en dur dans le code.

La nouvelle version de QRselect implémente un nouvel algorithme qui prend en charge les nouvelles données de multilinguisation telles que les dictionnaires, les mots réservés, etc. (Figure 43). L'algorithme reçoit en entrée une liste de mots réservés, une liste de mots-clés, et

un nom d'un dictionnaire. Pour décider si une paire est pertinente, la fonction « Prox » calcule une valeur de proximité en prenant les mots comme des items de base pour la comparaison. Cette proximité est aussi utilisée pour calculer la « distance de Jaccard » qui permet d'avoir une idée sur la similarité entre les documents (Lee, 1999). Les formules de calcul de proximité et de distance de Jaccard sont :

$$\text{Proximité}(cs_1, cs_2) = \frac{\text{Taille}(\text{Items}(cs_1) \cap \text{Items}(cs_2))}{\text{Taille}(\text{Items}(cs_1) \cup \text{Items}(cs_2))} \quad \text{(Équation 2)}$$

$$\text{Distance}(cs_1, cs_2) = 1 - \text{Proximité}(cs_1, cs_2) \quad \text{(Équation 3)}$$

(si la distance = 0 alors la coïncidence est parfaite, sinon elle prend la valeur 1)

Algorithme

```
// mc : liste de mots-clés, mr : liste des mots réservés, nameDico : nom du dictionnaire.
Fonction pairesDetection( cw, mr, nomDico, seuil )
// Détection de documents source correspondant à la liste « mc ».
  listeURL_1 ← webCollection( mc );
  pour i allant de 0 à Taille( listeURL_1 ) faire
     C1  ← html2text( listeURL_1[i] );
     // Détection des ancres dans les pages cible.
     listeURL_2 ← ancresDetection( mr,listeURL_1[i] );
      pour j allant de 0 à Taille( listeURL_2 ) faire
        // Extraction de contenu textuel.
        C2  ←  html2text ( listeURL_2[j] );
        // Calcul de la proximité avec la méthode de Jaccard.
        P ← Prox( nameDico, C1, C2 );
         // si la proximité est suffisante, la paire est sélectionnée.
          si P > = seuil alors
           // Ajout à la liste de paires candidates.
           liste_ Paires ← liste_Paires + paire( listeURL_1[i], listeURL_2[j]);
          fsi
      fpour
  fpour
  return list_candidate;
ffonction
```

Figure 43 : Nouvel algorithme de QRselect

Si cette valeur est acceptable, c'est-à-dire supérieure à un certain seuil, les URL des deux documents sont sauvegardées dans une liste à part appelé « list_candidate ». Techniquement, cette liste est structurée en format XML et est facilement interprétable par la grille de l'interface de QRselect (Figure 42, zone D) (DHTMLxGrid, 2007).

Exemple :

<?xml ... ?> <rows> <row id="0"> <cell> id </cell> <cell> URL Cible </cell> <cell> URL source </cell> <cell> taux de proximité </cell> </row> ... </rows>.

5.3.2 Expérience et évaluation

3.2.1 Le cas des paires anglais-japonais

Une première expérience du système a été faite pour son évaluation sur une liste de 33 mots-clés japonais. Cette liste est fournie par deux traducteurs et deux évaluateurs professionnels regroupant cinq catégories : la zone géographique ou les pays, affaires, droit, culture et sport, informatique, autres.

Le nombre de documents cibles à extraire a été fixé à 100 pages, ce qui signifie que, pour chaque mot-clé, le système ne détecte qu'un nombre restreint de documents dépendant en premier de la capacité de vérification par les traducteurs des URL fournies[55]. Pour la recherche, le moteur « YahooJapan » a été utilisé parce qu'il offre plus d'instances de recherche par jour que « Google ».

La pertinence d'une paire de documents est définie à l'aide d'un seuil égal à 0,1 (cf. *(Équation* 1, p. 97).

Les résultats d'évaluation sont donnés en termes de précision et de rappel.

La Table 9 montre les résultats d'évaluation ; les lignes correspondent aux 33 mots-clés utilisés durant la recherche des paires. En ce qui concerne les métriques, la notation suivante est utilisée pour évaluer les différentes catégories des paires détectées.

- A : nombre total de pages produit par QRselect.
- T : nombre de paires formées d'un document japonais et du document anglais source.
- C : paires candidates produites par QRselect. Elles sont divisées en :
- CY : paires correctes.
- CE : paires incorrectes.

[55] Ce nombre a été fixé après avoir consulté une quinzaine de traducteurs bénévoles. La plupart d'entre eux confirme qu'ils vérifient moins de 100 URL lors d'une recherche sur le Web.

- M : les paires manquantes (silence). C'est le cas où QRselect ne détecte pas un document japonais traduction d'un document anglais, alors que les deux existent sur le Web. Cet ensemble est aussi divisé en deux sous-ensembles :

- MH : la page source existe en HTML ou en format balisé.

- MI : la page cible ne contient pas de lien vers la page source, ou la page source n'est plus accessible par ce lien.

- N : le nombre de pages japonaises détectées comme des traductions incorrectes

- P : précision = CY/C

- R : rappel = CY/T

Keyword set	A	T	C	CY	CE	M	MH	MI	N	P	R
(a) Colombia	92	1	1	0	1	2	1	1	89	0	0
(a) Colombia, drug	58	25	17	17	0	11	8	3	30	1	0,68
(a) Colombia, Uribe	32	20	19	19	0	1	1	0	12	1	0,95
(a) Venezuela	95	3	1	1	0	3	2	1	91	1	0,33
(a) Venezuela, Chavez	81	3	3	3	0	0	0	0	78	1	1
(a) Falluja	97	14	3	3	0	19	11	8	75	1	0,21
(a) Falluja, Aljazeera	96	31	7	4	3	29	27	2	60	0 ,57	0,13
(a) Baghdad, resistance	98	36	17	17	0	20	19	1	61	1	0,47
(b) Abu Graib, human right	97	26	14	7	7	19	19	0	64	0,5	0,27
(b) Separation wall	93	11	6	3	3	12	8	4	75	0,5	0,27
(b) Chomsky, Iraq, invasion	96	14	9	9	0	7	5	2	80	1	0,64
(b) Katrina, Hispanic	93	4	21	1	20	3	3	0	69	0,05	0,25
(b) China, censorship	98	30	12	12	0	20	18	2	66	1	0,4
(b) Sellafield, BNG	93	8	1	1	0	12	7	5	80	1	0,125
(b) Said, Arafat, Zionism	51	9	8	7	1	5	2	3	38	0,88	0,78
(b) Catholic, contraception	94	9	6	4	2	6	5	1	82	0,67	0,44
(b) Veteran, suicide	91	3	4	1	3	4	2	2	83	0,25	0,33
(c) Torvalds	97	40	4	4	0	36	36	0	57	1	01
(c) Stallman	94	17	13	10	3	7	7	0	74	0,77	0,59
(c) Napster, file exchange	97	52	31	31	0	22	21	1	44	1	0,60
(c) Halloween document	79	2	5	0	5	7	2	5	67	0	0
(c) Linux, developing countries	93	13	15	10	5	4	3	0	1	0,74	0,67
(c) Google, library, scan	96	16	2	1	1	15	15	0	79	05	0,06
(d) Free culture	94	25	6	4	2	21	21	0	67	0,67	0,16
(d) Krugman, column	97	15	22	7	15	10	8	2	65	0,32	0,47
(d) Seattle Post, Mariners	94	36	1	1	0	40	35	5	53	1	0,028

(d)	China, Football	99	11	1	1	0	17	10	7	81	1	0,09
(d)	Shunsuke, local, media	87	0	0	0	0	0	0	0	87	-	-
(d)	F1, interview, driver	99	62	1	1	0	64	61	3	34	1	0,02
(d)	Ghilbi, export	63	1	0	0	0	2	1	1	61	-	0
(d)	Hollywood, star, article	97	1	1	1	0	0	0	0	96	1	1
(e)	Nablus report	100	23	11	10	1	23	13	10	66	0,91	0,43
(e)	John Pilger	95	26	19	18	1	10	8	2	66	0,95	0,69
		2936	587	281	208	73	451	379	72	2204	**0 ,74**	**0,35**

Table 9 : Résultats d'évaluation du système QRselect avec 33 mots clés

Cette expérience montre des résultats avec une précision de 0,74 et un rappel de 0,35. Bien que les paires obtenues soient utiles aux traducteurs et permettent d'économiser énormément d'effort par rapport aux moteurs de recherche tels que « Google » ou « YahooJapan », la performance du système reste modeste. Les causes principales sont les paires obtenues dans CE (incorrectes) et MH.

En effet, les paires de documents incorrectes peuvent être divisées en deux catégories :

- les pages japonaises ne sont pas des traductions mais indiquent des liens vers des pages anglaises contenant des informations sur le contenu recherché. Le nombre de paires causant cette situation a été identifié dans 67 cas.

- les pages japonaises sont des traductions, mais le système traverse des liens erronés pour obtenir la source. 6 cas de ce genre ont été identifiés au cours de l'analyse des résultats obtenus.

Les paires existantes non trouvées peuvent aussi être à leur tour divisées en deux catégories :

- le lien n'a pas été détecté par le système parce que l'ancre du document original ne contient pas l'un des mots réservés. QRselect a produit 187 cas dans cette catégorie.

- Le système a détecté une page originale par une ancre correcte mais, dans la phase de calcul, le document a été rejeté car le score n'a pas été suffisant. Ce manque a été identifié dans 192 cas. Cela est dû principalement aux mauvaises transformations faites par le dictionnaire « Eijiro ».

En résumé, les erreurs et les manques sont causés par trois facteurs :

(i) une mauvaise détection des ancres,

(ii) une mauvaise extraction du texte des zones balisées, et

(iii) une mauvaise transformation par le dictionnaire.

Si les deux premiers facteurs peuvent être surmontés techniquement, le dernier peut être surmonté socialement en promouvant l'implication et les contributions des traducteurs bénévoles. Si les traducteurs se rendent compte du mérite du recyclage, alors ils fourniront des « liens » correcte vers les pages anglaises source.

Pour confirmer cela, nous avons consulté 9 traducteurs bénévoles. 8 d'entre eux ont accepté de changer leur méthode de dissémination des traductions en incluant dans chaque page traduite un lien vers la page anglaise source. 5 traducteurs ont confirmé qu'ils voulaient changer leurs traductions déjà disséminées en les rendant conformes au mode de fonctionnement de QRselect. Dans le futur, si les contributions des traducteurs évoluent dans ce sens, alors les 187 paires manquantes (à cause du manque des ancres) seront détectées comme des paires pertinentes.

Ce système est en amélioration continue, dans le but de surmonter les problèmes confrontés et d'augmenter la performance générale du recyclage.

5.3.3 Identification de traductions mutuelles : le cas des bisegments

Les traducteurs peuvent recevoir des propositions à partir de l'ensemble des segments déjà stockés. Bien que ces suggestions soient utiles, la MT ne peut répondre que par des segments existants. Pour surmonter l'insuffisance des ressources linguistiques, nous voulions étendre les suggestions vers des ressources plus volumineuses, comme le Web. Ce besoin a déjà été ressenti à travers les besoins des traducteurs bénévoles qui ont souvent recours au Web pour trouver des traductions des expressions déjà traduites.

Nous proposons l'hypothèse que la traduction d'un segment source (S_1,L_1) peut avoir une traduction de bonne qualité et cohérente sur le Web.

Si la méthode de recyclage proposée dans QRselect (cf. *QRselect : recyclage de documents déjà traduits*, p. 95) utilise un dictionnaire bilingue pour détecter les correspondances dictionnairiques (traduction mot à mot) et une formule de calcul de proximité, alors le recyclage de bisegments doit détecter des correspondances entre les

composants des segments (mots) et une similarité entre la chaîne source (à traduire) et une chaîne cible.

Mais cette fois-ci, nous souhaitons proposer une alternative, car la méthode basée sur l'utilisation des dictionnaires est figée. Pour permettre la recherche de bisegments dans plusieurs langues, et exploiter les services de TA gratuits, nous adaptons l'utilisation de la TA. Cela nous permet de surmonter deux problèmes par rapport à la méthode de recyclage de documents :

- Éviter les tâches de prétraitement, compilation et import de dictionnaires.
- Si le nombre de langues est restreint dans les dictionnaires, alors la TA peut offrir plus de couples de langue.

La TA peut nous aider à avoir des traductions approximatives qu'on peut utiliser pour le calcul des similarités. On sait aussi que, pour certains couples de langues, la TA produit plus de 50% de mots corrects dans la langue cible. On peut donc essayer de partir des prétraductions pour avoir des approximations et ensuite trouver les traductions correctes sur le Web.

Voici un exemple. Prenons la citation suivante (trouvée sur le site des traducteurs bénévoles PaxHumana) : SS_1 = « **not a single person made a public complaint. Fear prevented them from doing so.** ». Généralement, les traducteurs bénévoles cherchent à avoir une cohérence dans les traductions qu'ils produisent. Ils cherchent une traduction précise que d'autres traducteurs ont déjà produite sur le Web.

Sur le même site, nous trouvons aussi la traduction correcte en français de la citation : SC_1 = « **pas une seule personne n'a porté plainte. La peur les en a empêchés.** ». En utilisant la TA, la traduction de la citation est la suivante : ST_1 = «**pas une seule personne a fait une plainte du public. CRAINTES les ont empêchés de le faire.** ». En supposant que les mots {**une**, **ne**, **la**, **les**, **en**, **du**, **de**, **le**} extraits des segments S_2 et S_3 sont des mots vides, le nombre de mots dans S_2 et S_3 (sans les mots vides) est de 9 et 11 respectivement, et les mots communs aux deux phrases sont : {**pas**, **seule**, **personne**, **a (avoir)**, **porté**, **plainte**, **empêchés**}. Le nombre de mots communs est donc plus grand que 50% dans deux segments S_2 et S_3 (6 mots dans S_2 et 7 mots dans S_3). Pour plus de clarté, prenons deux extraits du document source et cible :

... In a rare public speech, Arar addressed this fear directly. He told the audience that an independent commissioner has been trying to gather evidence of law-enforcement officials breaking the rules when investigating Muslim Canadians. The commissioner has heard dozens of stories of threats, harassment and inappropriate home visits. But, Arar said, "**not a single person made a public complaint. Fear prevented them from doing so**." Fear of being the next Maher Arar ...	.... À l'occasion d'une rare allocution publique, Arar a abordé cette peur sans détour. Il a déclaré à l'assemblée qu'un commissaire indépendant a essayé de rassembler des preuves montrant que des fonctionnaires censés faire appliquer la loi agissaient dans l'illégalité lors d'enquêtes concernant des musulmans canadiens. Le commissaire a entendu des douzaines de récits de menaces, de harcèlement et de visites inopportunes au domicile. Mais, a dit Arar, « **pas une seule personne n'a porté plainte. La peur les en a empêchés.** » La peur d'être le prochain Maher Arar ...

La méthode de détection de bisegments est résumée dans le pseudo-algorithme suivant :

1. Sélectionner un segment SS_1 en L_1 (langue source).
2. Choisir un système de TA et traduire SS_1 en L_2 (langue cible). On obtient SS_2'
3. Lancer le robot Web pour chercher les chaines SC_i (en L_2) similaires à SS'_1.
4. Si la similarité entre SS'_1 et SC_i est supérieure à un certain seuil (fixé de façon empirique), alors ajouter SC_i à la liste des chaînes candidates.

La segmentation est faite par l'outil Lingpipe grâce au modèle de phrase qu'on peut construire dynamiquement à partir de corpus préparés à l'avance (code source : annexe B).

Exemple : application de l'API Lingpipe pour la segmentation.

Texte en entrée :

The induction of immediate-early (IE) response genes, such as egr-1, c-fos, and c-jun, occurs rapidly after the activation of T lymphocytes. The process of activation involves calcium mobilization, activation of protein kinase C (PKC), and phosphorylation of tyrosine kinases. p21(ras), a guanine nucleotide binding factor, mediates T-cell signal transduction through PKC-dependent and PKC-independent pathways. The involvement of p21(ras) in the regulation of calcium-dependent signals has been suggested through analysis of its role in the activation of NF-AT. We have investigated the inductions of the IE genes in response to calcium signals in Jurkat cells (in the presence of activated p21(ras)) and their correlated consequences.

Lingpipe produit le résultat suivant :

Résultat de la segmentation :

150 TOKENS
151 WHITESPACES
5 SENTENCE END TOKEN OFFSETS
SENTENCE 1:
The induction of immediate-early (IE) response genes, such as egr-1, c-fos, and c-jun, occurs rapidly after the activation of T lymphocytes.
SENTENCE 2:
The process of activation involves calcium mobilization, activation of protein kinase C (PKC), and phosphorylation of tyrosine kinases.
SENTENCE 3:
p21(ras), a guanine nucleotide binding factor, mediates T-cell signal transduction through PKC-dependent and PKC-independent pathways.
SENTENCE 4:
The involvement of p21(ras) in the regulation of calcium-dependent signals has been suggested through analysis of its role in the activation of NF-AT.
SENTENCE 5:
We have investigated the inductions of the IE genes in response to calcium signals in Jurkat cells (in the presence of activated p21(ras)) and their correlated consequences.

Lingpipe exploite le corpus « MEDLINE » comme exemple de données pour la construction du modèle de phrase. Le corpus contient plus de 13M citations dans le domaine du biomédical et est actuellement maintenu par la « United States National Library of Medicine (NLM) ».

Conclusion

Nous avons présenté dans ce chapitre un tour d'horizon sur les méthodes existantes de recyclage de données sur le Web. Nous avons aussi montré la différence entre les méthodes de collecte existantes, qui butent principalement sur la construction de corpus destinés au TAL. En combinant les principes de recyclage de ces méthodes et les spécificités des données traductionnelles, nous avons proposé une nouvelle méthode de recyclage.

Primo, l'outil QRselect a été augmenté pour être utilisé par plusieurs communautés de traducteurs bénévoles. Cela a permis d'avoir plus de généricité et d'augmenter la couverture de langues.

Secundo, nous avons étendu le recyclage vers des unités traductionnelles plus fines. Par l'exploitation de la TA et de méthodes statistiques, nous avons montré comment il est maintenant possible de recycler des correspondances traductionnelles et des bisegments.

Chapitre 6

Editeur multilingue intégrant des fonctionnalités d'aide

Introduction

Nous avons décrit dans la partie I notre contribution à la spécification et à la réalisation de QRedit, utilisable sur le Web et couplé à des aides linguistiques basées sur des ressources partagées sur un site Web dédié à une communauté de traducteurs bénévoles (anglais-japonais).

Certaines de ces communautés préfèrent toujours travailler en local, et QRedit leur convient. Cependant, d'autres communautés, de plus en plus nombreuses, ressentent le besoin de travailler en ligne de bout en bout. Il s'agit non seulement des projets de multilinguisation de logiciels libres, mais aussi de traducteurs motivés par des « causes », plus jeunes, ayant toujours connu le Web et en attendant tout. Ce sont eux qui traduisent déjà sur des sites comme translationwiki.net. Le pas suivant consiste à leur offrir un éditeur comme QRedit utilisable en ligne, et offrant un support beaucoup plus complet. C'est ce que nous avons fait en réalisant BTedit, qui permet aux traducteurs de traduire en collaboration avec d'autres traducteurs dans une interface à la Excel.

Les deux avantages de BTedit par rapport aux outils commerciaux sont :

- Utilisabilité sur le Web.
- Intégration des fonctionnalités d'aide aux traducteurs.

Les progrès en technologie nous ont permis de rendre BTedit assez souple et convivial (XML, AJAX, Wiki, etc.). Il est intégré dans un Wiki, qui, par construction, offre des fonctionnalités collaboratives que nous exploitons de façon générale dans BEYTrans.

Dans ce chapitre, nous présentons la conception générale de cet éditeur, le mode de lecture/édition à la Wiki, et les différentes étapes à suivre pour produire une traduction complète d'un document source vers plusieurs langues. Nous expliquerons ensuite les fonctionnalités linguistiques que nous avons développées.

Pour montrer l'efficacité de notre environnent et l'éditeur dans des situations traductionnelles concrètes, nous présenterons une expérimentation faite sur la traduction de presque une centaine de documents du projet en cours DEMGOL.

6.1 Visualisation et édition des documents

6.1.1 Visualisation en mode lecture

Le mode « lecture » consiste à visualiser des documents source partiellement traduits. Dans ce mode, les traducteurs peuvent suivre la progression de la traduction et contribuer en passant directement en mode « édition ». Les traductions peuvent être disséminées complètes ou partielles. De façon incrémentale, les traductions partielles progressent vers des versions finales.

Notons que les traductions en mode édition se font segment par segment. Au passage en mode lecture, les segments sont rassemblés pour produire une version lisible en format Wiki.

6.1.2 Visualisation en mode édition

1.2.1 Édition pour la traduction

L'éditeur est conçu pour permettre l'édition de plusieurs types de données.

Primo, deux zones d'interface d'édition permettent l'édition des segments source et cible (ces segment sont synchronisés). L'édition de la cible se fait pour produire une traduction d'un segment cible ou post-éditer une prétraduction (MT ou TA).

Secundo, l'éditeur permet de détecter les termes ou les mots manquant dans les dictionnaires, et de créer de nouvelles entrées.

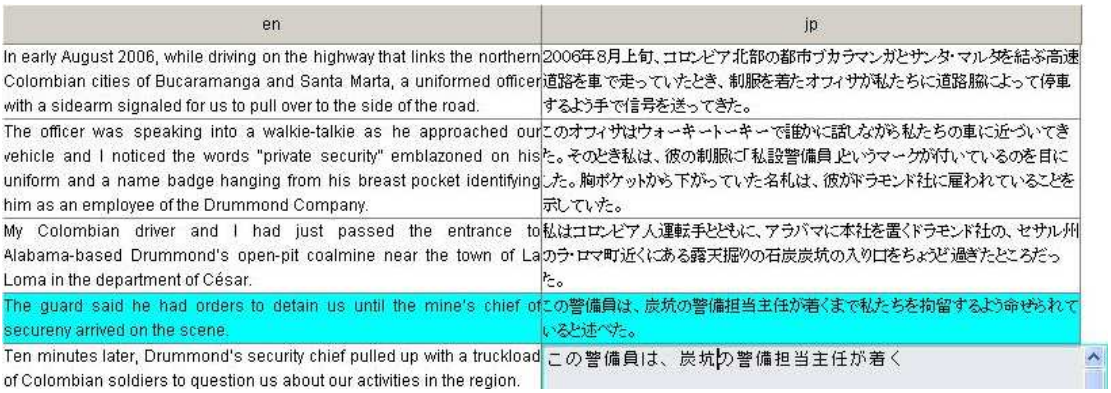
Le passage en mode édition se fait à partir du mode lecture. À partir de ce dernier mode et lors de la sélection de l'édition, l'éditeur charge le CT correspondant au document sélectionné. Les UT sont détectées une par une et affichées dans une grille à la Excel. La grille n'accepte pas le format TMX du CT, ce qui implique que l'éditeur fait appel à une fonction de transformation du format de CT vers le format acceptable par la grille.

Il est possible d'inverser les couples de langues. Les traductions peuvent se faire à partir d'une traduction cible. Cette fonctionnalité permet d'avoir une traduction « associative ». Les traducteurs ne maîtrisant pas la langue source peuvent ainsi démarrer des traductions à partir d'une cible déjà traduite pour traduire dans une nouvelle langue.

1.2.2 Édition des ressources linguistiques

1.2.2.a Édition parallèle de la MT

L'édition des segments des MT se fait dans une interface semblable à celle de la traduction. Les UT sont présentées dans une grille à la Excel, chacune sur une ligne, où chaque cellule est dans une langue.



en	jp
In early August 2006, while driving on the highway that links the northern Colombian cities of Bucaramanga and Santa Marta, a uniformed officer with a sidearm signaled for us to pull over to the side of the road.	2006年8月上旬、コロンビア北部の都市ブカラマンガとサンタ・マルタを結ぶ高速道路を車で走っていたとき、制服を着たオフィサが私たちに道路脇によって停車するよう手で信号を送ってきた。
The officer was speaking into a walkie-talkie as he approached our vehicle and I noticed the words "private security" emblazoned on his uniform and a name badge hanging from his breast pocket identifying him as an employee of the Drummond Company.	このオフィサはウォーキートーキーで誰かに話しながら私たちの車に近づいてきた。そのとき私は、彼の制服に「私設警備員」というマークが付いているのを目にした。胸ポケットから下がっていた名札は、彼がドラモンド社に雇われていることを示していた。
My Colombian driver and I had just passed the entrance to Alabama-based Drummond's open-pit coalmine near the town of La Loma in the department of César.	私はコロンビア人運転手とともに、アラバマに本社を置くドラモンド社の、セサル州のラ・ロマ町近くにある露天掘りの石炭炭坑の入り口をちょうど過ぎたところだった。
The guard said he had orders to detain us until the mine's chief of secureny arrived on the scene.	この警備員は、炭坑の警備担当主任が着くまで私たちを拘留するよう命ぜられていると述べた。
Ten minutes later, Drummond's security chief pulled up with a truckload of Colombian soldiers to question us about our activities in the region.	この警備員は、炭坑の警備担当主任が着く

Figure 44 : Édition des mémoires de traduction dans BTedit

Les traducteurs ont la possibilité d'édition et de mise à jour des segments source et cible. La figure ci-dessus montre un exemple d'une mémoire de traduction en cours d'édition (édition du dernier segment en langue japonaise).

1.2.2.b Édition des dictionnaires

La mise à jour des ressources linguistiques se fait dans une interface différente de celle de la traduction. La figure ci-dessous montre la façon dont une entrée est ajoutée dans le dictionnaire. Le traducteur n'a pas besoin de formater lui-même les entrées. Il reçoit une interface avec une zone d'édition contenant la plupart des informations utiles à la saisie d'une entrée dictionnairique : mot-vedette, traduction, exemples, catégorie syntaxique, ...

L'ajout de nouvelles entrées engendre la création d'une page Wiki. Cette page peut être créée directement au moment de la traduction. Si un mot est détecté comme absent dans un dictionnaire, le traducteur peut directement ajouter sa traduction sans quitter l'interface de traduction BTedit.

Figure 45 : Augmentation de dictionnaires dans BEYTrans

1.2.3 Édition générale

On entend par *édition générale* l'édition de toutes les données dans une seule interface : durant le processus de traduction, les dictionnaires, les MT et les documents peuvent être mis à jour sans changer d'interface. Un dictionnaire peut être amélioré par une nouvelle entrée si un mot non traduit existe dans le segment en cours de traduction.

Dans BEYTrans, le passage à la mise à jour est possible dans toutes les directions sans changer d'interface. Cela a un grand intérêt, surtout pour minimiser le temps de la recherche et de basculement vers les ressources pour chercher des traductions.

6.2 Fonctionnalités d'aide linguistique durant la traduction

6.2.1 Fonctionnalités combinées des MT et des TA

Dans les outils d'aide à la traduction, les suggestions de la MT se font après avoir calculé la similarité entre un segment actif et les segments stockés dans la MT.

Il existe plusieurs méthodes pour le calcul de la similarité. La plus simple consiste à calculer la distance d'édition (DE) en mots entre deux segments, puis à déterminer les mots qui sont communs aux deux segments.

La distance d'édition est le nombre minimal d'opérations d'édition nécessaires pour transformer la première chaîne en la seconde. Les coûts attribués aux opérations sont typiquement les suivants (distance de Levenshtein) : 0 pour la conservation et 1 pour la substitution, l'insertion et la suppression (Levenshtein, 1966) (Wagner, *et al.*, 1974).

L'algorithme

L'algorithme de programmation dynamique pour calculer la distance d'édition de Levenshtein a été proposé par Wagner et Fischer (1974). L'algorithme a une complexité de $O(n \times m)$, où n et m sont les longueurs des deux chaînes.

Algorithme *Wagner-Fisher*

```
Entier LevenshteinDistance(char s[1..m], char t[1..n])
  // d est un tableau de m+1 lignes et n+1 colonnes
  Entier D[0..m, 0..n] ;

  Pour i allant de 0 à m faire
     D[i, 0] := i ;
  Pour j allant de 0 à n faire
     D[0, j] := j ;

  Pour i allant de 1 à m faire
     Pour j allant de 1 à n faire
        si s[i-1] = t[j-1] alors C := 0
                          sinon C := 1
        D[i, j] := minimum(
                   D[i-1, j] + 1,  // suppression
                   D[i, j-1] + 1,  // insertion
                   D[i-1, j-1] + C // substitution
                        )
   return D[m, n]
```

Exemple :

On utilise les notations suivantes :

- Correspondance (par exemple « a » & « a ») : m(a)
- Insertion (par exemple « » & « a ») : i(a)
- Suppression (par exemple « a » & « ») : d(a)
- Substitution (par exemple « a » & « b ») : s(b,a)
- Transposition (par exemple « ab » & « ba ») : t(ab)

(i) « gage » & « gauge ». Opérations = m(g) m(a) i(u) m(g) m(e) avec DE =1.

(ii) « tenser » & « tensor ». Opérations = m(t) m(e) m(n) m(s) s(o,e) m(r) avec DE=1.

(iii) « hte » & « the ». Opérations = d(h) m(t) i(h) m(e) [t(ht) m(e)] avec DE=2[1]

Nous avons utilisé la distance d'édition pour le calcul de la similarité basé sur les mots ((Équation 1). D'autres techniques sont aussi applicables, comme celles basées sur les n-grammes où les comparaisons se font un à niveau plus fin, sur les séquences de caractères, sur les séquences d'octets, ou sur les séquences de bits (Denoual, 2006).

$$Similarité = 1 - D(s,t) \Big/ |s| + |t| \qquad \text{(Équation 4)}$$

Où *s* et *t* sont les chaînes source et cible, respectivement. La similarité varie entre la valeur 0 (dissimilarité totale) et la valeur 1 (similarité parfaite).

Les autres suggestions sont des prétraductions qui ne se font pas par des recherches dans la MT, mais par des appels à la TA gratuite sur le Web (Vo-Trung, 2004).

Dans le futur, nous voudrions remplacer cette distance, simple à calculer mais très peu performante en pratique dans les outils d'aide aux traducteurs, par une distance « composite », comme celle utilisée dans Similis (Planas, 2000). Une telle distance fait intervenir plusieurs « étages » de représentation des segments : flot d'entrée usuel, balises, texte normalisé sans balises, lemmes et termes, fragments (chunks).

6.2.2 Fonctionnalité d'aide liée aux dictionnaires

Les aides dictionnairiques se résument aux choix et aux suggestions proposés durant le processus de traduction de façon proactive. Les suggestions dictionnairiques se présentent de façon asynchrone (sans bloquer la traduction en cours).

Une alerte pour chaque mot vedette permet de savoir si sa traduction existe ou non dans le dictionnaire choisi. Grâce à ces alertes, les traducteurs définissent l'action appropriée. Donc, en accédant à un mot, il est préalablement possible de savoir s'il s'agira d'une action d'ajout, de mise à jour ou de simple consultation.

Dans la même interface, les traducteurs peuvent consulter une traduction provenant des références et l'insérer à la position du curseur ou, dans le cas contraire, de mettre à jour par

ajout d'une nouvelle traduction dans le dictionnaire, qui sera disponible instantanément pour tous les traducteurs connectés sur la même référence (Figure 49).

L'édition nécessite des fonctionnalités multiples pour mener la traduction à bien :

- recherche et manipulation globale des ressources linguistiques durant la traduction (dictionnaires et les MT).
- consultation et mise à jour de documents.

Les lexicographes (bénévoles ou professionnels) qui s'intéressent aux ressources compilées peuvent directement participer à la vérification et à l'amélioration du contenu sans passer par l'interface de traduction.

2.2.1 Fonctionnalités d'aide avancées

Les fonctionnalités les plus importantes sont les suggestions qui sont fournies par la TA ou les MT. Le déclenchement des suggestions est basé sur la détection d'événements (par exemple, déplacement d'une cellule à une autre) lors des manipulations de la grille de l'éditeur. Le tout se fait en arrière-plan. À part celles proposées par la TA et la MT, ces suggestions peuvent être des paires de bisegments recyclées du Web.

2.2.2 Normalisation de données traductionnelles

Avant de proposer les suggestions traductionnelles, quelques traitements préliminaires sont nécessaires. Le plus important est la normalisation des mots, qui consiste à réduire chaque mot à une forme unique radicale en enlevant les désinences ou les préfixes pour l'accès dictionnairique. Cela est fait dans notre environnement par la nouvelle version de l'algorithme de Porter (Porter, 1980) implémentée dans Pystemmer (Pystemmer, 2008).

La normalisation des mots apporte une grande efficacité dans les traitements de texte et la récupération des lemmes pour l'accès aux dictionnaires. Cela est particulièrement vrai pour les langues européennes qui ont une morphologie flexionnelle plus riche que l'anglais.

L'une des formes de normalisation consiste à identifier le lemme d'un mot. La lemmatisation consiste à identifier la « forme canonique » d'un mot, qui permet d'accéder aux dictionnaires. Les formes d'un lemme, dans une langue donnée, varient en fonction de leur genre (masculin ou féminin), leur nombre (un ou plusieurs), leur personne (venons, venez, viennent, etc.), leur mode (indicatif, impératif, etc.). En français, par exemple, la forme canonique est identifiée par :

- l'infinitif pour les verbes.
- le masculin singulier pour les noms et les adjectifs « réguliers ».
- les mots eux-mêmes pour les catégories invariables (préposition, adverbe, conjonction,...).
- toues les formes « nécessaires » pour les mots irréguliers (ex : empereur/impératrice).

Par exemple, l'adjectif *petit* existe sous quatre formes : *petit*, *petite*, *petits* et *petites*. La « forme canonique » de tous ces mots est le mot *petit*. Pour certains verbes, il existe beaucoup plus de formes. Par exemple, pour le verbe avoir, on trouve : *ai*, *as*, *a*, *avons*, *ais*, *avons eu*, *ayez eu*, *eussions eu*, *aurions eu*, etc.[56]

La normalisation est nécessaire dans notre environnement car elle facilite la détection des entrées dans les dictionnaires. Même si la tâche n'est pas triviale, surtout lorsqu'on traite des documents multilingues, nous avons essayé de montrer l'efficacité de la lemmatisation pour les langues européennes.

La lemmatisation en elle-même constitue un problème à part. Les langues riches morphologiquement telles que l'arabe sont beaucoup plus difficiles à lemmatiser que la langue anglaise. Ce problème ne constitue pas l'un de nos objectifs, nous limitons donc la lemmatisation à l'intégration d'algorithmes existants, et pour un nombre de langues restreint, tout en laissant la possibilité d'intégrer de nouveaux lemmatiseurs.

6.3 Premier protocole d'évaluation : le projet DEMGOL

Une première phase d'expérimentation et de développement a été faite conjointement avec le sous-projet de traduction du DEMGOL[57] qui a pour but la traduction en trois langues (français, espagnol, anglais) de presque 1200 notices rédigées en italien. Cette expérimentation nous a permis de montrer l'efficacité de notre environnement pour accélérer la traduction en français des notices non encore traduites, et de le tester dans une situation réelle, avec une traductrice.

[56] Une partie de ces définitions a été extraite de l'encyclopédie Wikipedia : http://fr.wikipedia.org/wiki/Lemmatisation

[57] Dictionnaire Etymologique de la Mythologie Crecque Online, http://demgol.units.it.

6.3.1 Le projet DEMGOL

3.1.1 But du projet

Le Groupe de Recherche sur le Mythe et la Mythographie de l'Université de Trieste (GRIMM, Département des Sciences de l'Antiquité « Leonardo Ferrero ») a élaboré un projet pour la construction d'un Dictionnaire Étymologique de la Mythologie Grecque en ligne (DEMGOL) de grande ampleur.

3.1.2 Participants

Grâce à une bourse accordée à Mme Francesca Marzari (traductrice et spécialiste de la mythographie grecque, Université de Sienne) dans le cadre d'un projet Émergence obtenu à Grenoble par le groupe de Mme le Pr. Létoublon, la collaboration entre Trieste, Sienne et Grenoble a consisté en une vérification par F. Létoublon de toutes les notices au fur et à mesure que la traduction progressait.

3.1.3 Conditions techniques

Le serveur du projet DEMGOL est géré à Trieste (Italie) par un responsable informaticien (Mr. Giovanni Zorzetti) sous la direction du Prof. Ezio Pellizer. L'accès au site est libre et ne nécessite aucune authentification, sauf pour les administrateurs, pour lesquels, en particulier durant la traduction, une authentification est nécessaire. Les administrateurs qui font la traduction peuvent mettre en ligne les notices dans la langue de traduction au fur et à mesure de leur validation.

Une instance de traduction (BTdemgol) a été installée au Japon pour l'expérience. Il est actuellement accessible à L'URL suivant : http://87.98.171.238:8080/beytrans.

6.3.2 Préparation d'une instance « BTdemgol »

Une instance personnalisée a été réalisée pour la communauté alors potentielle des traducteurs souhaitant contribuer à la traduction des notices du projet DEMGOL. L'instance BTdemgol permet de lancer rapidement les traductions, cela en disposant de toutes les notices et d'un dictionnaire spécialisé. Les nouvelles notices doivent donc être transférées dans la BD. Les notices traduites doivent à leur tour être présentées aux traducteurs de façon alignée.

3.2.1 Préparation et import de notices

Toutes les notices constituant le projet DEMGOL ont été récupérées à partir d'une BD envoyée par l'administrateur du site original. Les 1200 notices ont été tout d'abord converties

en UTF-8 et injectées dans la BD du système BTdemgol pour fonctionner en mode collaboratif à la « Wiki » (**Erreur ! Source du renvoi introuvable.**).

Figure 46 : Mode lecture à la Wiki de notices traduites dans BTdemgol

Au moment où nous avons commencé l'expérience, la moitié des notices avaient déjà été traduites de l'italien vers le français, et environ 300 notices avaient été traduites de l'italien vers l'espagnol.

BTdemgol permettait dès sa première version de traduire vers le français, l'espagnol et l'anglais. Cependant, l'expérimentation n'a pu être faite que pour l'italien-français.

3.2.2 Segmentation

En arrière-plan, les notices (en italien) ont été segmentées par le segmenteur multilingue en un ensemble d'UT et stockées dans les CT (compagnons de traduction). Toutes les notices sont transférées dans la langue d'origine (italien), avec les traductions disponibles (en français et/ou espagnol).

Un processus d'alignement a été indispensable pour aligner les notices déjà traduites et du coup construire une MT préliminaire pour ce projet. Les notices traduites dans au moins une

langue sont alignées au niveau segment (un TU dans TMX pour inclure toutes les langues) dans les CT. Les notices qui n'ont pas de correspondance en langue cible doivent être traduites.

6.3.3 Utilisation par DEMGOL

3.3.1 Conditions de développement

Durant l'été 2006, nous avons fait quelques réunions avec les membres du projet DEMGOL à Grenoble et nous avons conclu que la meilleure méthode consisterait à adapter une instance de l'environnement BEYTrans. Le développement a eu lieu au Japon en 2007. Les membres du projet DEMGOL ont suivi le développement, le test et la traduction, à Trieste, Grenoble et Sienne.

Le projet BEYTrans était trop important pour être réalisé par une seule personne dans le cadre d'une thèse. Nous avons bénéficié de la participation de plusieurs développeurs. Nous avons associé en particulier 4 étudiants de Polytech Grenoble (RICM-2) durant leur stage d'été (2007) au laboratoire DATIC (Danang, Vietnam). L'un d'eux a participé à l'amélioration de l'éditeur de traduction BTedit. L'encadrement a été fait à distance à partir de l'Université de Tokyo (Japon) avec l'aide sur place du Dr. Vo-Trung Hung (ancien thésard du GETA, directeur du laboratoire DATIC, et responsable des études doctorales et de la coopération internationale à l'UTDN, Université de Technologie de Da Nang) et de Cong-Phap HUYNH (arrivé fin 2006 au GETALP).

Les différentes fonctionnalités ont été intégrées avec succès. De leur côté, les membres de notre équipe (GETALP) intéressés par BEYTrans ont testé et fait des retours constructifs pour l'amélioration des différents modules. Enfin, Mme Letoublon à Grenoble a participé à l'amélioration de la qualité des traductions des notices en français.

3.3.2 Aspects techniques

3.3.2.a Appel à la TA et post-édition

Nous avons traduit toutes les notices non traduites en français à l'aide du traducteur gratuit de Babel Fish (basé sur Systran) sur le Web. Cet appel à la TA a été fait en une passe. Les UT des notices traduites sont inclus dans les CT de chaque notice.

Les traducteurs ont trouvé les prétraductions assez bonnes, parce que la plupart des mots en italien sont bien traduits en français. C'est grâce à cela que le temps de traduction a été réduit, comme le montre l'évaluation (cf. infra).

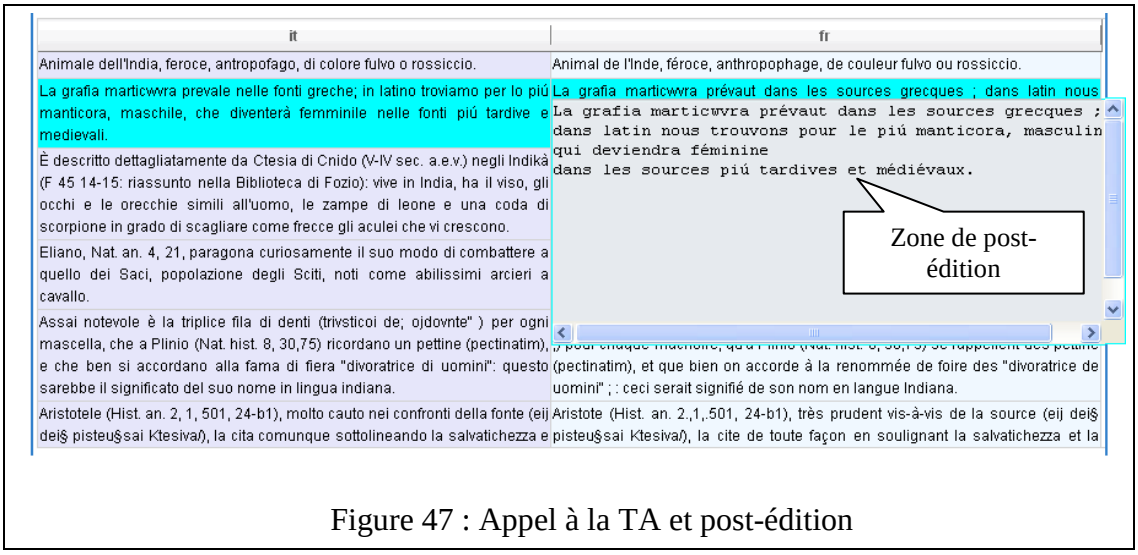


Figure 47 : Appel à la TA et post-édition

La post-édition est le processus d'amélioration humaine qui consiste à exploiter l'éditeur pour améliorer les prétraductions produites par les traducteurs automatiques (Figure 47). L'état de la traduction est modifié si les post-éditions sont finies correctement et si le texte est révisé et de bonne qualité. C'est cet état qui indiquera aux autres traducteurs d'éviter de faire des traductions multiples de la notice.

3.3.2.b Lexique de traduction et mémoire de traductions

Un dictionnaire (au sens de BEYTrans) a été créé explicitement pour le projet DEMGOL, en utilisant la syntaxe Wiki et un contenu linguistique structuré hérité de l'environnement Papillon pour le contenu des articles. Nous l'appelons ici « lexique de traduction », pour éviter toute confusion avec le dictionnaire DEMGOL, objet de la traduction.

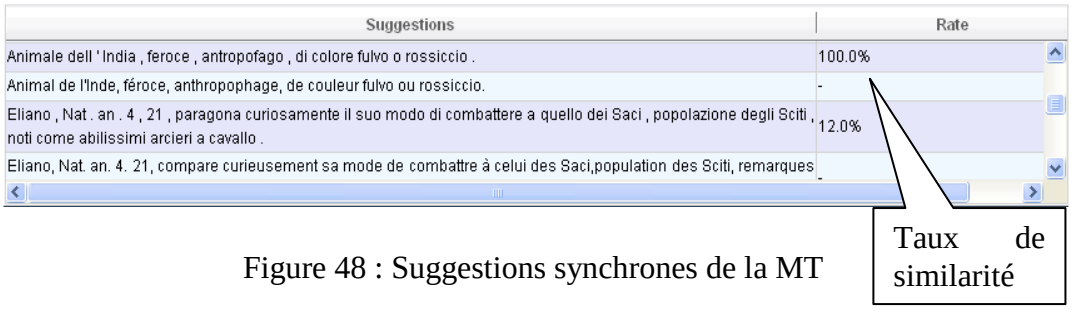


Figure 48 : Suggestions synchrones de la MT

Les noms de personne (dieux, héros, …) en grec ainsi que leurs translittérations, ou les mots liés à la langue, sont traduits vers le français durant la traduction des notices. De plus, les segments sont insérés dans la mémoire de traductions de façon alignée à chaque nouvelle traduction. Les traducteurs du projet peuvent mettre à jour le dictionnaire séparément, ou durant le processus de traduction.

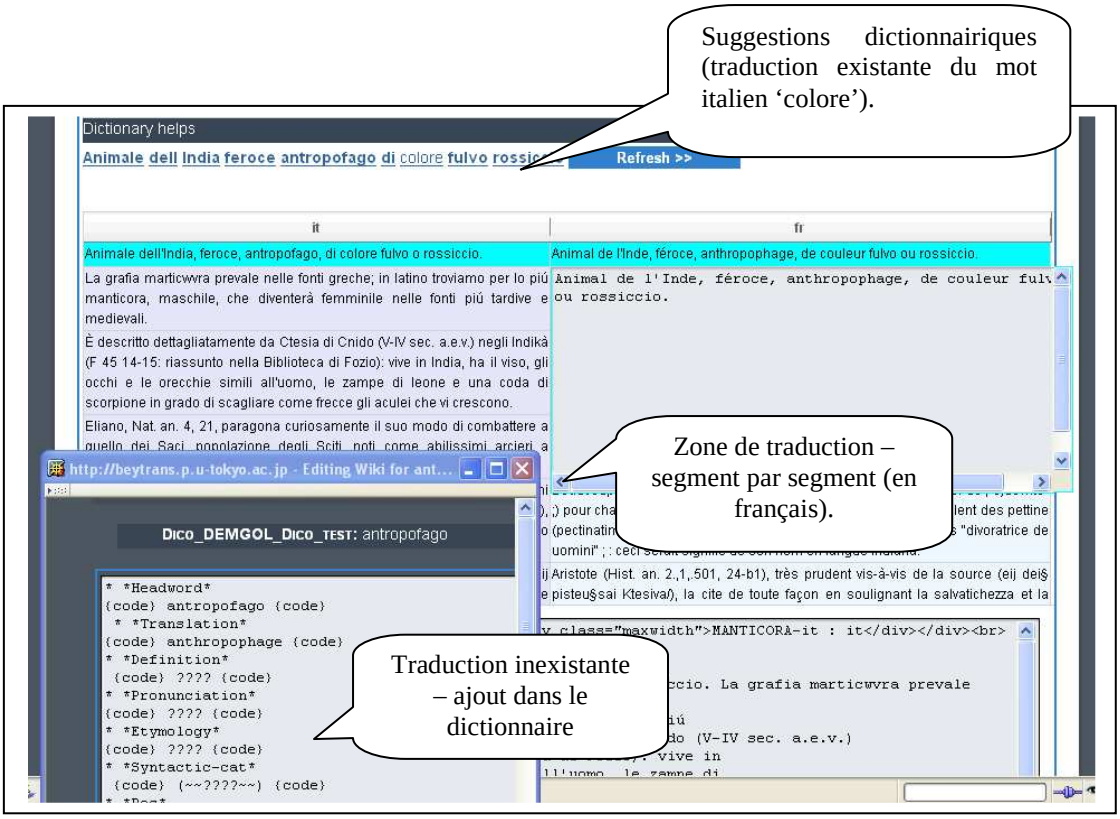


Figure 49 : Suggestions dictionnairiques avec mise à jour directe

À partir du segment source donné, les mots sont extraits, lemmatisés et ensuite des correspondances traductionnelles dans les dictionnaires sont recherchées. L’aide dictionnairique est présentée sous forme d’un ensemble de liens Web (un pour chaque mot) vers les entrées dictionnairiques (Figure 49).

Le lexique de traduction ne contenait au début aucune entrée. Après des expérimentations de traduction qui ont duré deux semaines, environ 147 entrées y ont été ajoutées.

3.3.2.c Autres points marquants

L’encodage initial des caractères était UNICODE /UTF-16 (16 bits). Les caractères accentués sont représentés dans la BD de DEMGOL (à Trieste) par leurs codes hexadécimaux de la table UNICODE. Par exemple, le caractère « é » a été codé par « \u00e9 ». L’interprétation du code hexadécimal se fait au moment de la visualisation par les navigateurs Web.

De plus, les machines non dotées de fontes grecques rencontrent des problèmes de visualisation des caractères grecs. Sans fonte grecque, il est donc impossible d'afficher la chaîne suivante : Ἄβας (Abas)[58].

Pour remédier à ces problèmes, il a fallu convertir le codage des caractères en un codage unique permettant la visualisation de données multilingues. UTF-8 s'avère une solution adéquate, car il est facilement interprétable par tous les navigateurs Web. Il a été proposé pour sumonter les problèmes de compatiblité entre des systèmes qui ne gèrent pas la représentation sur 16 bits. De plus, une représentation sur 8 bits réduit la taille des données et évite d'avoir des octets sperflus. Il a aussi montré son efficacité dans plusieurs projets gérant des données multilingues (par exemple, la base lexicale multilingue Papillon). Les 1200 notices du projet DEMGOL ont donc été converties en UTF-8.

3.3.3 Evaluation

Une expérience a été faite sur deux ensembles de notices non traduites en suivant un protocole bien défini à l'avance. Le premier ensemble (E_1) contient 60 notices dont les noms commencent par la lettre « L » (par exemple « LABDACOS », « LACEDEMONE », « LAKIOS », etc.). L'autre ensemble (E2) contient 30 notices dont les noms commencent par la lettre « M » (par exemple « MACEDONE », « MACHEREO », « MACISTO », etc.).

3.3.3.a Protocole d'évaluation et intérêt d'utilisation de BEYTrans

Le protocole de l'expérience a été défini dans l'ordre des tâches à réaliser pour produire des documents traduits de bonne qualité.

Ses étapes sont les suivantes :

- Re-segmentation : cette étape a consisté à corriger les erreurs de segmentation produites par LingPipe.
- L'appel de la TA : pour toutes les notices de la lettre « M », les prétraductions sont produites par la TA en ligne, appelée automatiquement par notre éditeur sur l'ensemble des documents.
- Suggestions par la MT : au début de l'expérience, la MT est vide. Au fur et à mesure de la progression dans les traductions, la MT doit s'enrichir car les segments traduits seront exploités par la MT. Bien qu'il eût été possible de démarrer un processus d'alignement sur les notices déjà traduites, nous nous sommes contenté de fournir

[58] Extrait de la notice « Abas » du projet DEMGOL : http://demgol.units.it.

notre environnement vide, comme cela se fait dans la plupart des outils d'aide à la traduction.

- Suggestions dictionnairiques : le dictionnaire d'aide à la traduction est aussi vide au début. La traductrice doit enrichir le dictionnaire à chaque fois que l'éditeur de traductions lui signale qu'un mot est absent du lexique de traduction.[59]

Dans un premier temps, la traductrice a traduit l'ensemble E_1 sans l'environnement BEYTrans (suivant la méthode classique). Ensuite, l'ensemble E_2 a été traduit avec l'environnement BEYTrans, en exploitant les aides linguistiques, et en enrichissant le lexique de traduction et la MT.

3.3.3.b Evaluation du temps

L'un des objectifs principaux des aides à la traduction est d'économiser l'effort et le temps de traduction. Selon le protocole proposé à la traductrice, le temps (minutes et secondes) de traduction doit être soigneusement noté après la fin de la traduction de chaque notice, pour pouvoir juger l'efficacité de l'environnement.

La durée de la traduction entre le début de la traduction et la soumission des résultats est d'environ de 2 semaines (incluant les temps de rupture). Les traductions n'ont pas été faites en continu, car cela a dépendu de la disponibilité de la traductrice[60]. Une meilleure condition expérimentale aurait été d'avoir aux moins deux traducteurs avec des compétences plus ou moins équivalentes, pour traduire avec et sans BEYTrans le même ensemble de documents.

30 notices	BEYTrans	Lexiques traduits (Lt)	Temps total (Tt)	Temps réel (Lr)
TH classique	-	0	3h33	3h33
BEYTrans (TAO)	+	2h54	5h08	2h13

Table 10 : Temps de traduction avec et sans BEYTrans

Avant de commencer l'expérience, nous avons voulu construire pour la communauté DEMGOL un lexique de traduction auquel les traducteurs pourraient ajouter de nouvelles entrées.

[59] Notons que cela se fait après avoir segmenté et normalisé les mots, et qu'une recherche systématique dans le lexique de traduction est lancée en arrière-plan.

[60] L'expérience a été faite par une seule traductrice employée avec un contrat à durée limitée. Son contrat a expiré durant l'expérience. Nous la remercions vivement d'avoir continué, à titre bénévole, ce qui a permis de faire des mesures sur une centaine de notices.

Cela n’a pas été possible, car les traducteurs du projet DEMGOL n’avaient pas de dictionnaire électronique (ils ne consultent que des dictionnaires papier). À l’inverse de ce que nous avons fait pour la communauté Arabeyes et celle des droits de l’homme, il ne nous a été possible d’importer aucune ressource linguistique de cette communauté.

Ce résultat ne reflète pas la performance de l’environnement, car les traductions ont été faites par une seule personne. En effet, l’aspect collaboratif implémenté n’a pas été totalement exploité. Pour une meilleure utilisation par des communautés de bénévoles, nous proposons à la fin de ce chapitre un deuxième proprocole d’évaluation collaborative, qui nous permettra montrer les gains de la traduction collaborative basée sur un Wiki.

Le temps de traduction des notices avec BEYTrans comprend aussi le temps d’ajout des entrées dictionnairiques. Une opération d’ajout inclut les étapes suivantes :

- vérification de l'absence de traduction (affichage différent sur l’interface des mots normalisés).
- clic sur le lien de mot absent.
- affichage de la fenêtre de l’ajout.
- positionnement sur la source.
- saisie de la source (mot vedette).
- saisie de la traduction.
- sauvegarde.

À la fin de la traduction des 30 notices (lettre M) par la traductrice, le nombre total d'ajouts dans le dictionnaire était de 174 entrées. Pour clarifier la différence de temps de traduction avec et sans l’utilisation de BEYTrans, nous prenons comme base de calcul 30 notices. Si un ajout dure 1 minute, alors le temps total de traduction avec BEYTrans, en prenant en compte les 174 entrées, devient 2,23h.

Nous avons supposé, selon nos contraintes, que la taille et la complexité de toutes les notices sont presque les mêmes. La différence dans le temps de traduction est alors 1,31h (le temps d’ajout d’un lexique de traduction est 1mn).

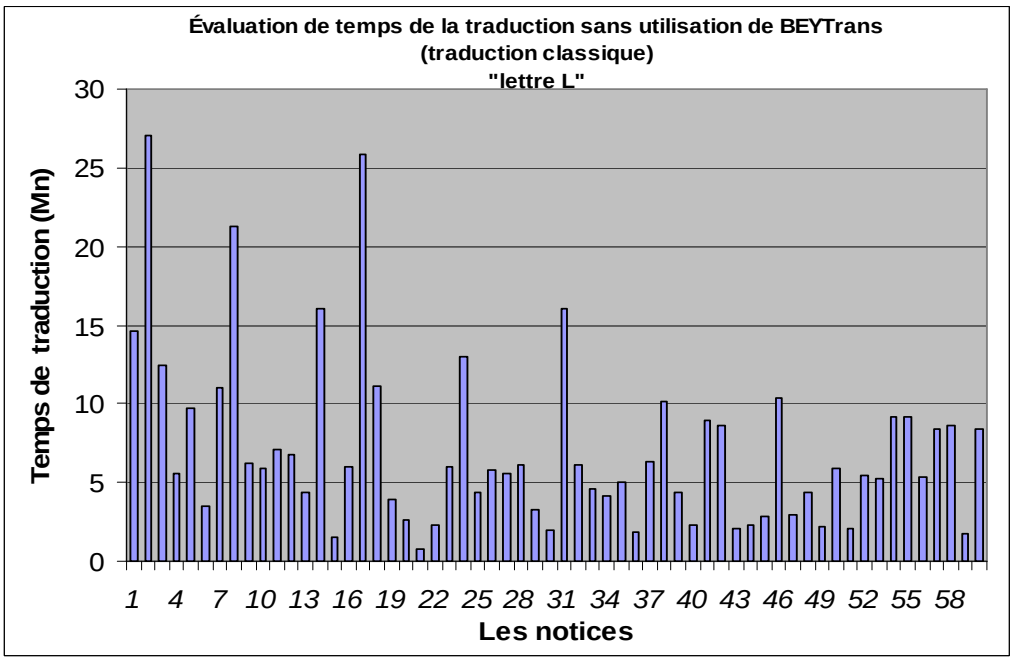


Figure 50 : Temps dépensé pour les notices de la lettre L de DEMGOL

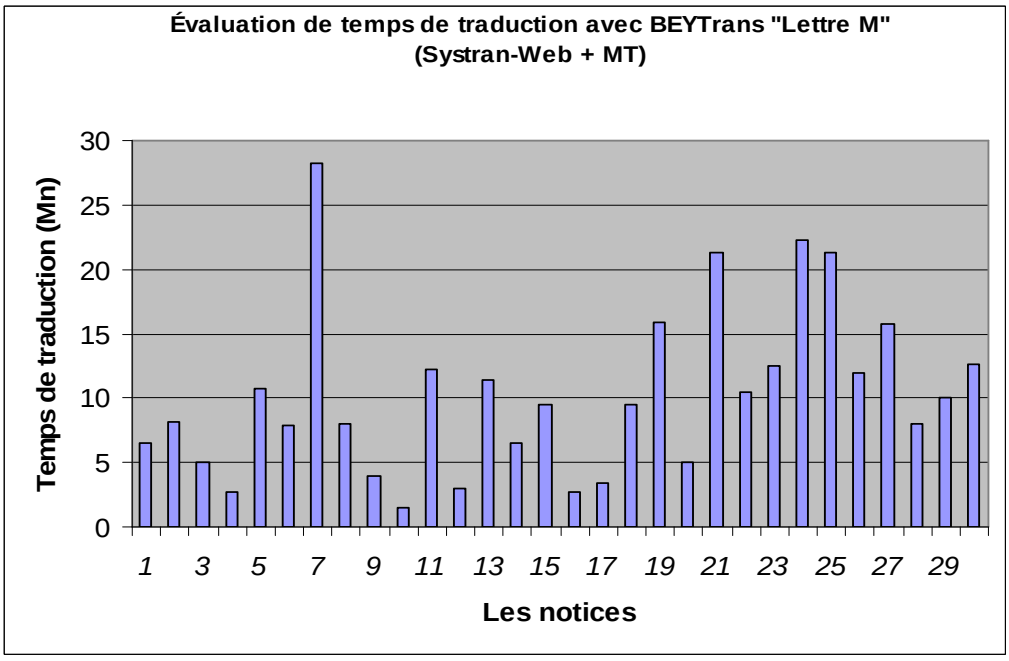


Figure 51 : Temps dépensé pour les notices de la lettre M avec BEYTrans

Bien que notre environnement soit fourni avec une MT et un dictionnaire vide, il a été en avance par rapport à la méthode classique. Ces résultats, bien que modestes, nous laissent optimiste, car la traduction collaborative à la Wiki peut réduire encore le temps de traduction. La Figure 50 montre les temps de post-édition de chaque notice de l'ensemble E_1 sans

l’exploitation de l’environnement BEYTrans. La traductrice n’a consulté qu’un seul dictionnaire (version papier) pour avoir des traductions et d’autres informations utiles sur les personnalités historiques grecques. Le temps total de la traduction inclut le temps de post-édition des prétraductions automatiques (appel de la TA gratuite proposé par Systran) et la consultation manuelle du dictionnaire.

La Figure 51 montre les temps de post-édition de chaque notice de l’ensemble E_2 dans l’environnement BEYTrans. Le traducteur ajoute au dictionnaire de nouvelles entrées à chaque fois qu’une traduction d’un terme est absente, et vérifie en même temps s’il existe des coïncidences floues suggérées par la MT.

En ce qui concerne la MT, environ 100 UT ont été ajoutées, dont quelques-unes ont été exploitées durant la traduction des nouvelles notices. Nous pensons aussi que la MT augmentera si le nombre de traducteurs bénévoles pouvant contribuer aux traductions augmente.

En résumé, cette expérience a été spécifiquement faite pour avoir des retours de la part du traducteur et pour améliorer l’utilisabilité de l’environnement. Les résultats ont été modestes, car nous n’avons pas eu la possibilité de faire intervenir plusieurs bénévoles.

Dans les pragraphes suivants, un nouveau protocole est présenté afin de surmonter les limitations de l’expérience précédente.

6.4 Deuxième protocole d'évaluation : le projet Tikiwiki

Nous abordons ici les points suivants :

1. Analyse et critique du premier protocole d'évaluation
2. Préparation de l'environnement et améliorations fonctionnelles
3. Proposition du protocole
4. Quelques expériences préliminaires et préparatoires
5. Premiers résultats

Analyse et critique du premier protocole d'évaluation

Le premier protocole appliqué au projet DEMGOL nous a permis de vérifier la faisabilité du concept de traduction collaborative basée sur un Wiki. Nous avions sélectionné deux ensembles de notices à traduire afin de mesurer le temps de la traduction. Un premier ensemble a été traduit avec la méthode classique, tandis que le deuxième a été traduit à l'aide de BEYTrans (mais sans lexique de traduction initiale). Le temps a été mesuré pour calculer la différence entre les traductions.

Bien que cette première expérience ait montré la faisabilité de notre concept, la preuve de l'avantage apporté par l'utilisation de BEYTrans n'avait pas pu être vraiment apportée, à cause de limites venant des conditions de l'expérience :

- Absence de bénévoles : le projet DEMGOL ne disposait pas d'une communauté de bénévoles pour lancer une expérience réellement collaborative. C'est pourquoi nous avions limité le test à l'usage des fonctionnalités de BEYTrans, avec la participation d'une seule traductrice[61].

- Éxpériences et prototypage : les expériences ont été réalisées pendant le processus d'implémentation du prototype. C'est grâce aux retours de la part de la traductrice que nous avons pu améliorer progressivement notre environnement.

- Limitation du nombre de langues : la langue originale était l'italien. Les notices on été donc traduites de l'italien vers le français et l'espagnol.

[61] La traductrice Francesca Marzari avait été recrutée en post-doc pour réaliser les traductions IT-FR.

- Le nombre restreint de notices[62] sur lesquelles a porté l'expérience (90 + 50 = 140 sur 1200 au total) ne permettait pas de juger correctement la performance du système.

Nous avons donc défini un second protocole expérimental, dans lequel notre objectif global est de montrer le gain en termes du temps passé aux différentes opérations (post-édition, construction et mise à jour des dictionnaires, terminologies, mémoires de traduction, etc.), lorsque plusieurs personnes sont impliquées collaborativement.

Ce nouveau protocole permet de montrer le gain des traductions et les atouts de la technologie Wiki.

6.4.1 Préparation de l'environnement et améliorations fonctionnelles

Avant d'aborder la définition précise de ce nouveau protocole, il nous a été nécessaire d'améliorer plusieurs points dans l'environnement, que nous résumons comme suit[63] :

- amélioration de plusieurs aspects ergonomiques de l'éditeur en ligne de BEYTrans ;
- élimination de manipulations superflues qui engendraient des pertes de temps inutiles ;
- gestion par défaut d'un grand nombre de langues (démarrage du tamoul) ;
- intégration de l'appel de ressources linguistiques en ligne ;
- ajout et amélioration de plusieurs fonctionnalités : import, extension du traitement de formats variés, accessibilité ;
- intégration d'une fonction de mesure de la durée de la traduction ;
- appel proactif de la TA ;
- traduction associative possible dans la même interface.

6.4.2 Définition du protocole

Le protocole met en œuvre deux méthodes complémentaires :

[62] Une notice ne correspond pas à une page standard. La taille d'une notice varie entre 5 et 10 segments (1 segment fait en moyenne 10 à 15 mots), alors qu'une page standard fait en moyenne 250 mots, soit 15-16 segments (1 segment fait en moyenne 15-17 mots).
[63] Guide d'utilisateur BEYTRans version 2.0 : http://panflute.p.u-tokyo.ac.jp/~bey/pdf/documentation.pdf

(1) *Première méthode :* traduction par des bénévoles dans des conditions similaires à la méthode classique. On suppose qu'un traducteur prend une page standard (250 mots, 1400 caractères) sur un site Web, au temps T_1, et la traduit avec son éditeur favori (vi, vim, TextWrangler, etc.) dans une langue cible. Il peut utiliser les dictionnaires en ligne ou en version papier. Quand il a fini, au temps T_2, il charge la page traduite sur le site. Par définition, la durée de la traduction de cette page par cette méthode est $D_h = T_1 - T_2$. Par exemple, on fait l'expérience sur 50 pages, et on obtient un délai de $(50/N + 2cN)$ heure(s) si on a N traductions et si c est le temps de téléchargement.

(2) *Deuxième méthode :* les traductions se font dans BEYTrans. Dans un premier temps, avant le début du travail humain de post-édition, les segments de chaque page à traduire sont « pré-traduits » par TA. Ensuite, au temps T_3, les traducteurs commencent à post-éditer les 50 pages en question. Une différence essentielle est que plusieurs personnes peuvent travailler en même temps sur la même page. Une autre différence est que les traducteurs n'ont pas à télécharger de fichiers. Supposons que les 50 pages soient toutes post-éditées au temps T_4. Le temps humain passé à la post-édition n'est évidemment pas T_4-T_3 (qui est le délai total), c'est la somme des temps passés par l'un ou l'autre des traducteurs sur chaque segment, et ces temps sont automatiquement mesurés par BEYTrans.

Dans la première méthode, chaque traducteur peut utiliser les outils et les ressources dont il dispose sans aucune restriction. Il est dans un contexte « classique » où on « prend » un travail à faire sur un site Web et on « rend » le travail fini sur le même site.

Dans la deuxième méthode, les traducteurs utilisent BEYTrans, en suivant les étapes suivantes :

a. *Choix d'une communauté :* le choix d'une communauté à participation forte de bénévoles joue un rôle important dans la réalisation du protocole. La communauté doit être active et préexistante. L'import dans BEYTrans des données d'un projet existant motivera la communauté à traduire, d'autant plus que ses membres maîtrisent déjà le domaine.

b. *Choix de données à traduire :* la majorité des communautés développant des environnements en ligne a aujourd'hui tendance à traduire la documentation et les messages d'interface en plusieurs langues. Les données à importer se présentent sous des formats divers et nécessitent l'écriture de parseurs pour les importer dans la base de données centrale de BEYTrans.

c. *Préparation de dictionnaires :* dans le cas où la communauté dispose d'un ou de plusieurs dictionnaires, un parseur d'import devra être développé pour que les traducteurs soient guidés systématiquement par des suggestions linguistiques.

En revanche, de nouveaux dictionnaires seront créés et l'appel des dictionnaires en ligne remplacera le premier manque (absence de dictionnaires préexistants) signalé à propos du premier protocole.

d. Prétraduction : la pré-traduction de chaînes non traduites se fait en parallèle avec la post-édition. Une innovation très intéressante est que le processus de la TA se lance en arrière-plan pendant que les traducteurs font la révision. Cette méthode permet d'avoir un double parallélisme : (1) pré-traduction, (2) post-édition collaborative.

e. Post-édition : les contributeurs font la post-édition en mode collaboratif pendant que les pré-traductions se font en arrière-plan.

f. *Recherche et construction de la terminologie :* les termes absents de la terminologie sont insérés automatiquement dans le lexique de traduction de la communauté.

g. *Calcul du temps de traduction :* l'éditeur de BEYTrans permet de calculer le temps de pré-traduction, de post-édition, d'ajout dictionnairique, etc. Dans l'évaluation, on pourra compter dans le temps humain seulement le temps de post-édition, ou y ajouter le temps de travail sur les dictionnaires.

Le choix des données d'une communauté existante est primordial pour motiver les bénévoles à utiliser notre outil. Certaines communautés de bénévoles développent des logiciels et font elles-même les traductions. Par exemple, les développeurs de Mozilla, Tikiwiki et Arabeyes participent à la traduction d'interfaces, de messages et de documentations. Les bénévoles ont donc des connaissances sur le domaine de la traduction.

"All Fields must be non empty" => "**يجب الا تترك حقول فارغة**",

"You do not have permission to write the mapfile" => "**ليس لديك صلاحية كتابة ملف التوجيه**",

"pageviews" => "**استعراضات الصفحة**",

"Invalid password. You current password is required to change your email address." => "**كلمة السر خاطئة. كلمة سرك الحالية ضرورية لتغيير عنوان بريدك الإلكتروني**.",

"Twi" => "**صيني تايواني**",

"note: those parameters are exclusive" => "**ملاحظة: هذه المعاملات خاصة**",

"Split a page into columns" => "**تقيسم صفحة إلى أعمدة**",

```
"column" => "عمود",
"Automatically creates a link to the appropriate SourceForge? object" => " يخلق آلياً رابط لكائن SourceForge? المناسب",
"Reply to parent comment" => "إجابة عن التعليق السابق",
"compose message tpl" => "حرر قالب لرسالة",
"messages tpl" => "قوالب الرسائل",
"Fields to display:" => "حقول للعرض",
"View articles" => "تصفح المقالات",
"admin HtmlPages" => "صفحات ال HTML الخاصة بالمدير",
"Admin Direrctory Sites" => "دليل المدير للمواقع",
"change password" => "تغيير الرمز السري",
…
```

Figure 52 **:** Fichier de messages anglais-arabe du projet Tikiwiki (en PHP)

De manière générale, les bénévoles en traduction ne maîtrisent pas toujours les interfaces d'aide à la traduction. Un effort supplémentaire de développement a été fourni dans ce sens dans BEYTrans 2.0 pour améliorer l'ergonomie de son interface, afin qu'elle soit encore plus conviviale. Par exemple, les suggestions linguistiques sont présentées conjointement dans une seule fenêtre pour éviter les navigations inutiles.

6.4.3 Expérimentation : le projet Tikiwiki

Nous avons choisi un projet libre pour montrer comment démarrer ce protocole. Le projet Tikiwiki (Tiki) existe en 20 langues, dont les éléments d'interface sont partiellement (arabe, coréen, espagnol, etc.) ou complètement (français, catalan, etc.) traduits.

Plusieurs types de données à traduire coexistent dans ce projet. Notre choix s'est porté sur les messages de l'interface Tiki qui sont facilement récupérables et pré-segmentés (Figure 52).

Parmi les 20 langues existantes, nous avons importé 11 langues (anglais, arabe, catalan, danois, espagnol, farsi, français, japonais, italien, russe, taiwanais, etc.). Les messages sont stockés dans des fichiers bilingues EN-XX (où XX=langue cible). La phase d'import sans duplication a permis d'avoir 14.929 messages en langue source (anglais). Dans l'éditeur, cet ensemble est divisé en 149 pages virtuelles accessibles par une fenêtre coulissante affichant 100 messages.

Le projet Tiki se distingue par le grand nombre des langues traitées (environ 25 langues) et de messages (environ 300 pages standard de 50 messages, où un message contient en

moyenne 5 mots). La présence d'une communauté active de bénévoles rend donc le choix du Tiki pertinent[64].

Quelques problèmes linguistiques dans Tiki

(1) La communauté Tiki ne dispose pas d'une base terminologique pour guider les traductions.

(2) Il n'y a pas d'interface pour la manipulation de messages.

(3) L'anglais en tant que langue source nécessite d'être révisé, car il comporte beaucoup de fautes d'orthographe (exemple dispay → dispaly, version → vesion, etc.), comme on en voit dans la Figure 52.

(4) Tiki évolue continuellement et de nouvelles chaînes sont introduites quotidiennement.

(5) La localisation en certaines langues (exemple espagnol, arabe, japonais, coréen, farsi, etc.) nécessite d'être complétée par l'intervention de bénévoles.

(6) L'environnement BEYTrans propose une solution au problème (1) par l'appel de ressources linguistiques en ligne. Il propose un éditeur convivial, ce qui résout le problème (2). Les problèmes (3), (4) et (5) seront résolus par l'intervention de bénévoles et l'exploitation des ressources proposées par BEYTrans.

Nombre de sujets	3 sujets	Langues maîtrisées : EN-FR
Nombre de messages	100 messages	
Nombre de page(s) virtuelle(s)	1 page	
Temps d'apprentissage	20 mn	100 messages/page (page virtuelle numéro 147)
Nombre de messages attribués aux sujets	33, 33 et 34 messages	
Mode de traduction	Collaboratif, interactif et proactif	

Table 11 : Conditions de l'expérience du deuxième protocole

[64] Les modérateurs Marc Laporte, Alain Désilet et Louis-Philippe Huberdeau ont montré un intérêt particulier à la localisation en plusieurs langues du projet Tiki.

Ce nouveau protocole a déjà été appliqué à un ensemble de messages anglais-français du projet Tiki, car les 3 personnes impliquées dans le processus maîtrisaient la traduction technique dans ces deux langues (Table 11).

[02:26] <bey> je vais voir avec Jean-Claude s'il peut nous aider
[02:54] == JeanClaude [i=8158408e@gateway/web/freenode/x-dqdtgsychyrwvehj] has joined #tikiwikimessage
[02:56] <bey> on est connecté les trois
[02:56] <@tomokiyo> je suis prête
[02:57] <bey> très bien
[02:57] <bey> et toi Jean-Claude?
[02:57] <@tomokiyo> toujours anglais français?
[02:57] <bey> oui Mutsuko
[02:57] <JeanClaude> Ok, prêt !
[02:57] <bey> en attend la confirmation je Jean-Claude
[02:57] <bey> très bien
[02:57] <bey> ok
[02:58] <bey> je vais commencer à traduire à partir du ID 14602 au 14030
[02:59] <@tomokiyo> moi, 14031
[02:59] <bey> Mutsuko 146030 ay 14662
[02:59] <@tomokiyo> ok
[02:59] <bey> oui pardon, à parti de 31 oui effectivement
[03:00] <JeanClaude> et moi du 14663 au 14693
[03:00] <bey> Jean-Claude: a partir de 14662 jusquà la fin = 14701
[03:00] <bey> il faut me dire exactement l'heure
[03:00] <bey> car on doit commencer ensemble
[03:00] <bey> chez moi est 15:00
[03:01] <JeanClaude> ici aussi
[03:01] <bey> ok on commence alors
[03:01] <bey> merci Jean-Claude
[03:01] <@tomokiyo> je ne trouve pas 146030
[03:02] <@tomokiyo> ce n'est pas 14630?
[03:03] == JeanClaude [i=8158408e@gateway/web/freenode/x-dqdtgsychyrwvehj] has quit ["Page closed"]
[03:04] <@tomokiyo> dis moi quelle page, stp
[03:04] == JeanClaude [i=8158408e@gateway/web/freenode/x-czdqpghpwpysbocd] has joined #tikiwikimessage
[03:05] <bey> tu commence du suivant
[03:11] <bey> je suis au moitié! ça avance :-)
[03:16] <bey> j'ai fini
[03:16] <bey> 16 minutes pour moi
[03:16] <bey> 10 segments n'ont pas été post-édité car bien traduits!
[03:20] <JeanClaude> J'ai fini !
[03:26] <bey> on a tous fini
[03:26] <bey> on a passé en paralle 24 minutes

[03:26] <bey> Plus 4 minutes de TA
[03:26] <bey> 28 minute au total
[03:27] <bey> On a gagné à trois 15 minutes au lieu de 50 minutes à deux par rapport à ce matin...
[03:27] <bey> Merci pour Mutsuko et Jean-Claudes, je suis reconnaissant à vos aides
[03:27] <bey> A bientôt …

Figure 53 : Contenu de communications entre les trois sujets

Les bénévoles ont été invités à faire la post-édition après avoir lancé la TA (en arrière-plan) d'une manière purement collaborative, et simultanément sur la même page virtuelle[65].

Déroulement de l'expérience

Un système de tchat intégré à BEYTrans a été lancé afin d'accompagner l'expérience. La Figure 53 illustre la communication entre les contributeurs au cours de cette expérience.

Il faut souligner que l'expérience s'est déroulée sans aucun problème, car les bénévoles étaient satisfaits par les améliorations ergonomiques de l'interface, et pourtant c'était leur première utilisation. L'accès direct aux données (la page virtuelle) ainsi que l'outil de tchat ont permis aussi une amélioration considérable de la qualité de la communication, du partage de tâches et du pilotage de l'expérience.

La Figure 54 illustre un extrait des traductions dans cette expérience.

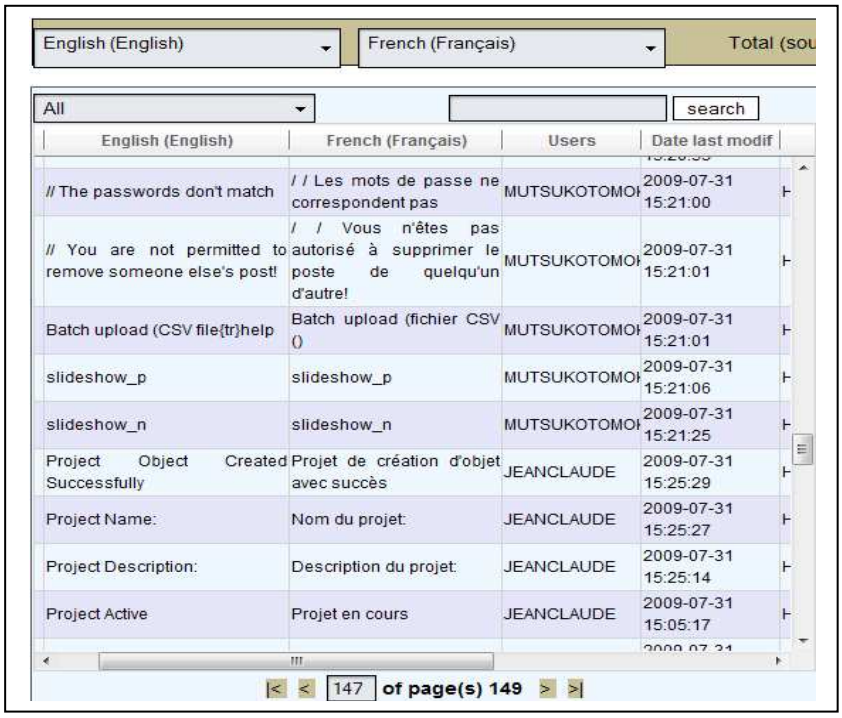


Figure 54 : Extrait des traductions collaboratives (3 bénévoles)

On peut d'ores et déjà déduire de ce qui précède un ordre de grandeur de l'amélioration apportée par BEYTrans dans ces conditions (avec 3 traducteurs), par rapport à la situation classique, où les 3 traducteurs auraient téléchargé 3 fichiers de 33, 33, 34 segments.

[65] Une page virtuelle est une page Web contenant plusieurs messages qui sert à faciliter la navigation. Elle peut correspondre selon la configuration à une ou plusieurs pages standard.

	Temps humain	Délai
Traduction humaine	> 2 heures	40 minutes
BEYTrans	> 1 heure	24 minutes
Gain	> 50%	> 40%

Table 12 : Comparaison entre méthode classique et la méthode collaborative

Au total, on avait environ 500 mots soit deux pages standard. Le temps de la traduction humaine aurait donc été de 2 heures, et le délai d'au moins 40 minutes, auxquelles aurait fallu rajouter 1 à 2 minutes pour les téléchargements (en les supposant effectués en parallèle).

Conclusion

L'éditeur a été développé en respectant les pratiques des traducteurs bénévoles. Tout d'abord, nous l'avons conçu de façon à permettre la visualisation et l'édition de documents multilingues en ligne, en mode collaboratif intégré à notre Wiki. D'autre part, les traductions se font avec l'aide de dictionnaires divers et spécialisés, que les traducteurs ont importés ou créés. La mémoire de traduction est activée et enrichie durant chaque traduction, et cela pour plusieurs sessions.

Nous avons montré la faisabilité d'un éditeur en ligne d'aide à la traduction comparable à ceux des professionnels. Notre éditeur, ainsi que l'environnement BEYTrans, a été testé sur le projet DEMGOL avec un premier protocole qui a permis de démontrer la faisabilité technique de l'approche, mais pas les avantages liés au travail collaboratif (réduction espérée des délais, partage immédiat de nouvelles ressources lexicales), ni ceux liés à l'utilisation de la TA (problèmes de la segmentation et du traitement de formats divers).

Les améliorations considérables apportées à l'environnement BEYTrans 2.0 nous ont permis de définir et de commencer à mettre en œuvre un second protocole d'évaluation permettant la comparaison des temps et des délais de traduction associés aux deux méthodes suivantes : (1) la méthode classique avec téléchargement de N fichiers par N traducteurs, traduction en local et remise des résultats sur le site Web, et (2) la méthode collaborative à petits grains (segments) fondée sur la post-édition de résultats de TA, et sur la contribution à une ressource lexicale partagée.

Les avantages de la méthode (2) ressortent déjà clairement de cette première expérimentation.

Partie III

Traduction et évaluation de gros corpus multilingues pour la TA

Introduction

Les mémoires de traductions construites au fur et à mesure de l'utilisation de BEYTrans constituent elles-mêmes des documents bilingues de structure particulière. Une idée naturelle a été d'utiliser BEYTrans, non pas pour les améliorer, puisqu'elles sont le résultat de traductions « finalisées », mais pour les étendre à d'autres langues. Leur taille est plus importante que celle de chaque document traduit, sans toutefois être très grande, puisqu'elles correspondent pour l'instant à quelques dizaines ou centaines d'ouvrages, soit au plus 5 000 à 10 000 pages (ou 125 000 à 250 000 mots) pour des communautés comme Paxhumana.

C'est à cause de notre implication dans deux campagnes d'évaluation de TA (IWSLT-04, IWSLT-06) que nous avons eu l'idée d'utiliser notre environnement d'aide à la traduction pour construire, améliorer, évaluer et étendre (en taille et en nombre de langues) des corpus multilingues parallèles pour la TA. À titre d'exemple, le BTEC (Basic Travel Expression Corpus) du consortium C-STAR faisait 163 000 segments (de 6,5 mots en moyenne), soit environ 1,1 M mots ou 4 500 pages standard, en anglais, japonais, chinois, et coréen. Depuis, il a atteint 1M segments de même longueur moyenne dans ces langues, et a été étendu partiellement à l'italien, à l'espagnol, à l'arabe, et au français. Mais il n'est toujours disponible (pour les partenaires uniquement) que sous forme de fichiers, on ne peut ni le visualiser ni a fortiori l'améliorer et l'étendre sur le Web.

Le but de cette partie est donc de voir comment adapter BEYTrans au traitement de ces énormes documents multilingues parallèles que sont les corpus pour la TA empirique et pour l'évaluation de la TA. Les problèmes ne concernent pas seulement la masse de données (EuroParl, c'est 20M mots en 11 langues, soit 11M segments de 20 mots en moyenne, ou 880 000 pages), mais aussi la structure (il faut ajouter de multiples traductions de référence, et pour certains sous-ensembles les résultats des STA en compétition), et les opérations à effectuer (évaluations objectives et subjectives) par de nombreux « juges » et/ou contributeurs.

Dans cette perspective, nous étudions dans cette partie les différents problèmes de gestion de masses de données multilingues. Ensuite, nous proposons des extensions à l'environnement « BEYTrans » pour les surmonter. Enfin, dans le dernier chapitre, nous discutons les résultats obtenus par quelques expérimentations sur des corpus volumineux.

Chapitre 7

Problèmes spécifiques des corpus multilingues pour la TA

Introduction

Nous nous concentrons dans ce chapitre sur des problèmes précis qui concernent la gestion de données multilingues volumineuses, et nous nous limitons aux données utilisées en TA. À l'inverse des parties précédentes, les documents sur lesquels nous travaillons ne sont pas des articles, des notices ou des livres, mais la mémoire de traductions elle-même. L'objet de la traduction et la mémoire de traductions coïncident.

La taille et le nombre de langues posent de nouveaux problèmes : il est évidemment impossible de charger un gros corpus de 80 000 pages en 11 langues comme EuroParl comme une page Web, alors que cela est possible pour quelques pages ou dizaines de pages en 2 langues.

D'autre part, l'utilisation de ces corpus pour l'évaluation nécessite de complexifier la structure (on veut souvent 4, 10 ou 16 traductions de référence, et N traductions automatiques), et de prévoir des interfaces spécifiques pour l'évaluation et la visualisation des résultats.

Nous commençons par étendre BEYTrans (avec ses fonctionnalités de départ) au traitement et à l'import de corpus bilingues ou multilingues volumineux dans un environnement collaboratif. Nous présentons ensuite les structures XML adaptées à une gestion efficace de gros corpus parallèles, et abordons les problèmes liés à la complexité de quelques corpus (transcriptions, structures intermédiaires, arbres décorés, structures originales, etc.).

Enfin, nous présentons une méthode pour faciliter l'extension rapide à d'autres langues en exploitant les fonctionnalités développées dans le cadre de BEYTrans (appel de la TA, MT, aide dictionnairique, etc.).

7.1 Problèmes liés à la complexité structurelle et à la taille

7.1.1 Complexité et taille

1.1.1 Ordre de grandeur de la taille d'un corpus « brut »

La collecte des corpus nécessite de prendre en considération les problèmes posés par la taille qui varie d'un corpus à un autre ou dans l'ensemble. La taille peut être énorme dans chaque langue, et il peut y avoir des données de taille proportionnelle au carré du nombre de segments (comme les distances). L'une des premières tâches à résoudre est l'unification du codage et du format d'accès. Pour cette raison, nous avons procédé à l'étude de plusieurs corpus.

Par exemple, le corpus multilingue parallèle « JRC-Acquis » de l'Union Européenne concerne le domaine du droit, avec les comptages suivants :

Language ISO code	Nb of texts	Text body			Signatures	Annexes	Total Nb words (text + signatures + annexes):
		Total Nb words	Total Nb characters	Average nb words	Total Nb words	Total Nb words	
bg	11384	16140819	104522671	1417.85	2170075	14114612	32425506
cs	21438	22843279	148972981	1065.55	7225300	16763733	46832312
da	23624	31459627	213468135	1331.68	2629786	16855213	50944626
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	**239583543**	1572.37	2973574	16459680	55887003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	**3490204**	19716243	**62132608**
et	23541	24621625	192700704	1045.9	1336051	14995748	40953424
fi	23284	24883012	212178964	1068.67	2677798	12547171	40107981
fr	**23627**	**39100499**	234758290	1654.91	3021013	**19978920**	62100432
hu	22801	28602380	213804614	1254.44	2529488	15056496	46188364
it	23472	35764670	230677013	1523.72	3120797	18331535	57217002
lt	23379	26937773	199438258	1152.22	2436585	15018484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44703607
mt	10545	20926909	126906748	**1984.53**	1336042	15620611	37883562
nl	23564	35265161	231963539	1496.57	3039580	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513141	17027393	49253537
pt	23505	37221668	227499418	1583.56	3034308	19350227	59606203
ro	6573	9186947	60537301	1397.68	514296	11185842	20887085
sk	21943	26792637	179920434	1221.01	3227852	16190546	46211035
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14965384	46974192
Total	**463792**	**636216050**	**4288962348**	**1387.23**	**60368893**	**358999011**	**1055583954**

Table 13 : Quelques chiffres à propos du corpus JRC-Aquis

Un autre exemple de corpus volumineux est le corpus OPUS. C'est une tentative de construire le corpus parallèle informatisé le plus grand possible par collecte de documents libres et de leur traductions. Dans sa version (v 0.2), le corpus OPUS contient environ 30 millions de mots dans 60 langues, collectés à partir de trois sources :

- la documentation OpenOffice.org (http://www.openoffice.org),
- les messages du système KDE (http://i18n.kde.org),

- le manuel du langage PHP (http://www.php.net/download-docs.php).

Le sous-corpus d'OpenOffice contient environ 2,6 millions de mots en 6 langues. Le corpus est complètement parallèle. Tous les documents en anglais ont été traduits en 5 autres langues. Le sous-corpus du manuel KDE (KDEdoc) inclut 24 langues avec 3,8 millions de mots. Enfin, le sous-corpus de la documentation PHP contient environ 3,5 millions de mots en 21 langues.

Un dernier exemple d'un corpus parallèle volumineux intéressant est celui des procès-verbaux du parlement européen, EuroParl. C'est un ensemble d'énoncés alignés dans 11 langues.[66] Le but de Ph. Koen en construisant ce corpus par alignement des corpus monolingues disponibles était de proposer un corpus complet et aligné pour la construction de 110 traducteurs automatiques statistiques (pour tous les couples de langues). C'est la raison pour laquelle les composants linguistiques du corpus sont alignés et référencés par des identificateurs uniques (ID). L'alignement de ce corpus a été fait en utilisant l'algorithme de Church et Gale (Gale, 1991).

Les trois exemples ci-dessus montrent la taille de quelques corpus libres existants. On les trouve librement téléchargeables sur le Web, mais sans aucune interface de navigation et de manipulation. Les autres corpus que nous avons explorés sont disponibles sous forme textuelle ou sont structurés (souvent en XML).

1.1.2 Ordre de grandeur de la taille d'un corpus « travaillé »

Pour la validation des systèmes de TA, ou plus généralement de TAL multilingue, on a besoin de construire des corpus qui varient selon la taille et la complexité des structures (annotations diverses, arbres, représentations sémantiques, etc.). Ces corpus sont généralement orientés vers le développement interne des outils dans les laboratoires de recherche et ne sont pas disponibles pour une utilisation libre.

Il existe plusieurs exemples de ces corpus. Par exemple, le corpus construit à l'USM (Penang) par l'UTMK a permis de développer un système de TA anglais-malais de bonne qualité. Le corpus contient 25 286 exemples de traductions construits à partir de textes généraux. Des structures de correspondances S-SSTC (Synchronized Structured String-Tree Correspondences) ont été développées pour synchroniser les exemples sources et cibles (Al-Adhaileh, *et al.*, 1999) (Boitet, *et al.*, 1988).

[66] Voir aussi le lien : http://www.statmt.org/europarl/.

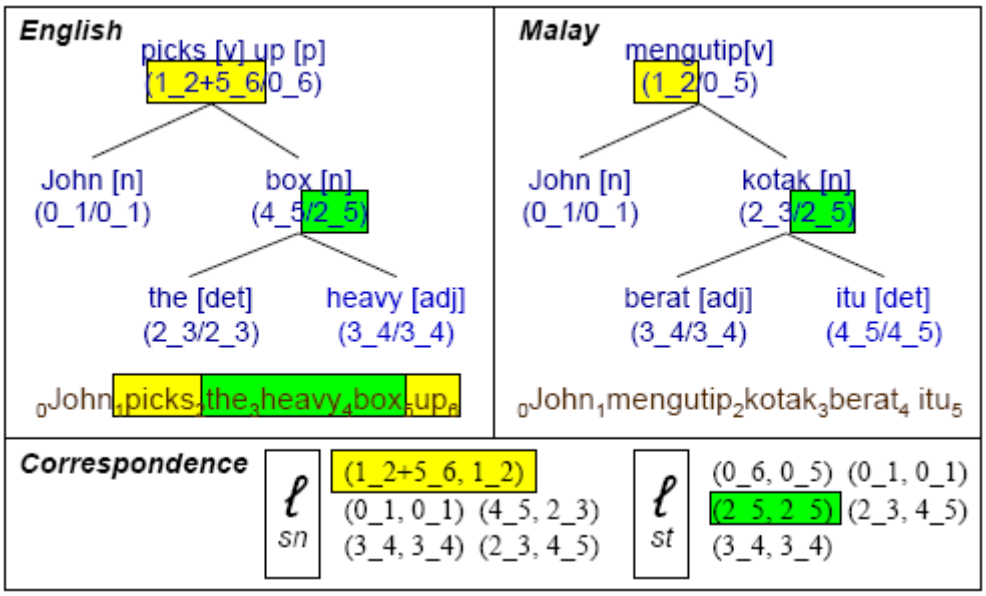


Figure 55 : Exemple complet d'une S-SSTC

Une SSTC est une correspondance entre une chaîne (énoncé en LN) et un arbre « linguistique » qui la représente. Il peut s'agir d'un arbre de constituants ou d'un arbre de dépendances, et la correspondance peut être concrète ou abstraite. La SSTC est codée par deux applications, SNODE et STREE, associant à chaque nœud de l'arbre une sous-chaîne, éventuellement non connexe. Pour un nœud N, SNODE(N) correspond à N seul, et STREE à tout le sous-arbre de racine N.

Certains corpus ont comme annotations des structures d'hypergraphe. Par exemple, les corpus UNL (Universal Networking Language) sont construits pour valider des systèmes de TA à base d'enconvertisseurs et de déconvertisseurs. Ils contiennent des hypergraphes UNL servant de « pivot » pour passer d'une langue à une autre (Figure 55). Un (hyper-)graphe UNL est une structure sémantique abstraite d'un énoncé anglais équivalent à l'énoncé à traduire (Tsai, 2004).

```
<?xml version="1.0" encoding="Unicode"?>
<!--<?xml-stylesheet type="text/xsl" href="unifem.xsl"?> -->
<unl:D unl:dn="FB2004" unl:on="Symposium 2001 Geneva " unl:dt="2001"
xmlns:unl="http://www.undl.org/2002/schema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.undl.org/2002/schema
UNL-XML.xsd">
<unl:P number="1">
<unl:S number="1">
<unl:org lang="el">
The Universal Forum of Cultures - Barcelona 2004
</unl:org>
<unl:unl>
```

```
mod:01(forum.@def.@entry,universal(mod&lt;thing))
mod:01(forum.@def.@entry,culture(icl>abstract thing).@def.@pl)
cnt(:01.@entry.@title,Barcelona_2004)
</unl:unl>
<unl:GS lang="es">
el foro universal de las culturas, Barcelona_2004.
</unl:GS>
<unl:GS lang="ru">
Универсальный форум культур - Barcelona 2004. </unl:GS>
<unl:GS lang="it">
Il forum universale delle culture , Barcelona_2004 , .
</unl:GS>
<unl:GS lang="hd">
saMskQwiyoM kI sArvaBOmika saMgoRTI bArsilonA 2004 .
</unl:GS>
<unl:GS lang="fr">
Forum universel des cultures , Barcelone_2004 , .
</unl:GS>
</unl:S>
```

Figure 56 : Exemple d'un document UNL XML-isé

D'autres corpus sont proposés spécifiquement pour des tâches en TA. Ce sont essentiellement des corpus post-édités utilisables dans les campagnes d'évaluation (par exemple, BTEC). C'est à ce type de corpus que nous donnons le plus d'importance, car nous souhaitons par la suite adapter BEYTrans à l'évaluation des systèmes de TA.

1.1.2.a Traductions

1.1.2.a.i Traductions candidates

L'un des problèmes que nous rencontrons actuellement est l'extension de corpus volumineux existants à d'autres langues. Par exemple, la traduction du BTEC de l'anglais (langue source) en français nécessite la traduction, l'édition et la révision manuelle par des humains. Cette tâche peut s'étendre sur plusieurs mois ou plusieurs années si les moyens humains sont minimaux. L'appel à la TA est utile pour réduire le temps humain nécessaire.

Premièrement, cela permet de produire des prétraductions, dans lesquelles des parties (mots propres ou techniques, expressions, etc.) sont correctes et n'auront donc pas à être saisies. Selon la qualité des systèmes produisant ces prétraductions, une certaine proportion de segments cibles peut être acceptée telle quelle. Par exemple, dans le cas de la traduction par Systran de l'italien vers le français, les prétraductions font gagner presque 2/3 du temps. C'est ce que nous avons remarqué lors des expérimentations de traduction des notices du projet DEMGOL.

De plus, ces prétraductions peuvent être considérées comme des traductions candidates avec lesquelles il est possible d'évaluer la qualité de traduction des systèmes de TA en compétition.

1.1.2.a.ii Traductions de référence

Les traductions de référence sont toujours produites par des humains. Elles servent à plusieurs applications en TA, dont la plus connue est l'évaluation « objective » dans laquelle les prétraductions sont comparées avec des énoncés de référence à l'aide de calculs produisant des scores (similarité dans le cas de BLEU, distance d'édition dans le cas de WER, score « ouvert » dans le cas de NIST).

Par exemple, ces calculs ont été utilisés dans la campagne d'évaluation IWSLT-06 à laquelle nous avons participé. Les résultats de ces métriques ont été combinés avec ceux des jugements humains pour classer les systèmes de TA des participants (Denoual, 2006).

1.1.2.b Stockage des évaluations « subjectives » et « objectives »

À part notre désir de construire un système permettant la gestion efficace de corpus multilingues volumineux, nous voulions aussi avoir des applications aux campagnes d'évaluation pour la génération de nouvelles références, et appliquer des métriques dans des expérimentations internes pour valider nos hypothèses.

Les participants aux campagnes d'évaluation telles que IWSLT doivent suivre un protocole bien précis exigé par les organisateurs. Au début, les participants reçoivent des corpus d'entrainement volumineux (20K énoncés en 2006) pour construire ou adapter des systèmes de TA (comme nous l'avons fait dans notre équipe avec le système Systran). Ensuite, des corpus restreints (100 à 500 énoncés) sont envoyés aux participants pour produire des traductions candidates. Les traductions produites par chaque système de TA servent à calculer divers scores (BLEU, NIST, mWER, METEOR, etc.). De plus, les participants s'organisent pour participer à l'évaluation « subjective » (jugements humains). Les deux résultats obtenus sont combinés pour classer les systèmes de TA.

L'évaluation subjective consiste à produire des jugements humains de 2 types : fluidité et adéquation.

1. Fluidité (Fluency) : elle indique comment le segment à évaluer est perçu par un locuteur natif (Figure 57). L'évaluateur doit classer le niveau de langue de la

traduction candidate en choisissant une valeur parmi : Flawless English, Good English, Non-native English, Disfluent English, Incomprehensible.

2. Adéquation (Adequation) : la traduction de référence (gold standard) ainsi que la traduction candidate sont présentées à l'évaluateur humain, qui doit juger quelle proportion de l''information est transmise par la phrase candidate en la comparant avec une traduction de référence (Figure 58). L'évaluateur humain doit choisir une valeur parmi : All of the information, Most of the information, Much of the information, Little information et None of it.

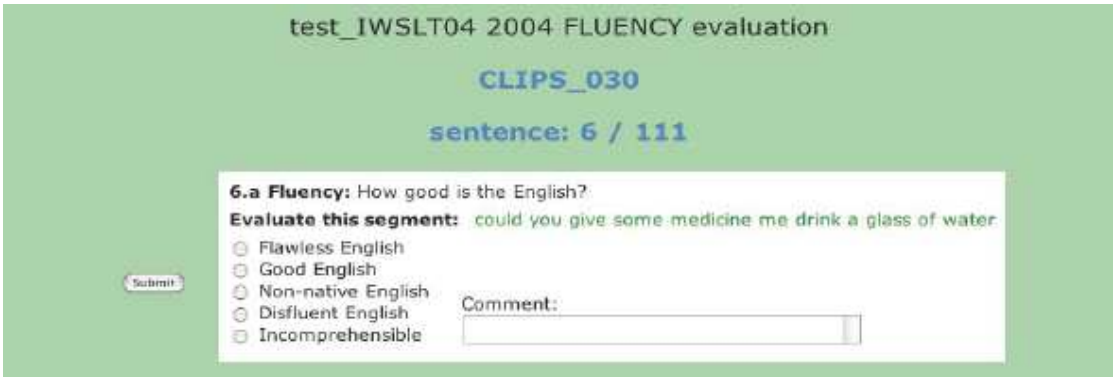


Figure 57 : Interface d'évaluation de la fluidité

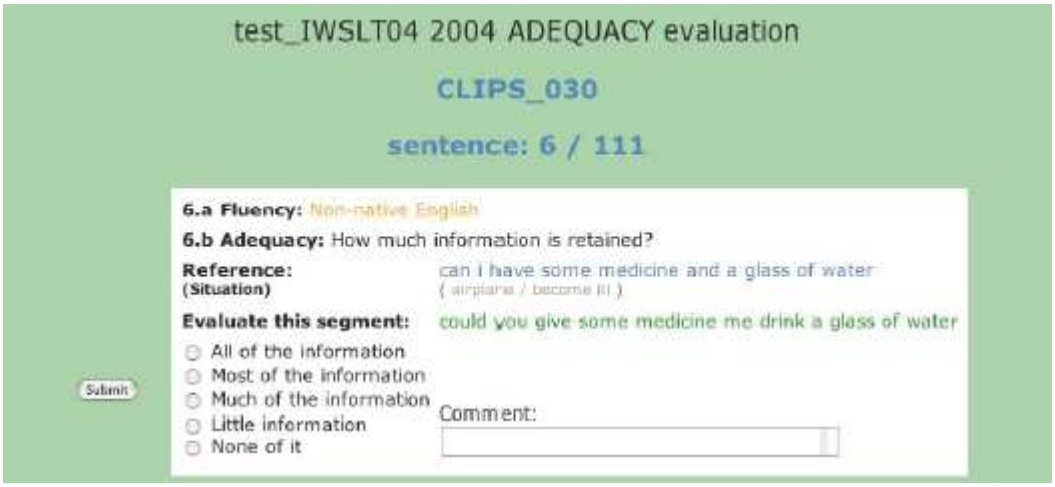


Figure 58 : Interface de l'évaluation d'adéquation (phrase de référence visualisée)

Pour lancer ces évaluations, les campagnes d'évaluation proposent généralement une interface. La Figure 57 montre l'interface d'évaluation de la fluidité proposée par IWSLT-06 : on voit un lot de phrases affecté à un évaluateur (111 phrases) et le numéro de la phrase courante (6). L'interface ne présente que la traduction candidate (phrases de référence

cachées) pour éviter que l'évaluateur (spontanément) ne se réfère à la référence, ce qui diminuerait l'impartialité des jugements subjectifs.

La Figure 58 montre l'interface d'évaluation de l'adéquation, qui présente les deux traductions (référence et candidate). L'évaluateur compare les deux phrases et sélectionne l'une des notes proposées par le protocole d'évaluation, représentées par des boutons radio.

Cette interface, bien qu'elle soit fonctionnelle, présente quelques limitations. L'évaluation subjective est faite par un groupe d'évaluateurs humains coordonnés par un administrateur (ou coordinateur) humain. Ce dernier gère plusieurs groupes dispersés sur plusieurs pays (France, Allemagne, USA, Japon, Corée, Chine, Italie, etc.) et attribue aux participants un ensemble d'informations d'accès et un lot de phrases à évaluer.

Un problème se pose lorsqu'un évaluateur manque de temps (par exemple en période de vacances), et qu'aucune autre personne ne peut suivre l'état de progression de ses phrases. Dans l'environnement proposé aux évaluateurs, une session d'évaluation subjective sur un ensemble de phrases ne peut être ouverte par plusieurs personnes. Un autre problème est que l'interface présente à un évaluateur une phrase à la fois, ce qui rend la tâche d'évaluation très pénible. Une interface présentant un nombre paramétrable de phrases pourrait accélérer l'évaluation.

Pour adapter BEYTrans aux protocoles d'évaluation, on a donc besoin de gérer les résultats d'évaluations subjectives et objectives de traductions automatiques supposées être faites dans le cadre de tâches précises. Nos structures internes doivent aussi permettre la sélection des échantillons de corpus, et le calcul des scores par les protocoles utiliés par les campagnes d'évaluation.

1.1.3 Structure logique envisagée

Pour arriver à unifier la gestion de corpus destinés à la TA, il suffit de représenter toutes les structures utilisées, car chaque corpus est structuré selon les besoins de l'application. Généralement, un corpus se compose des énoncés source et cible, dans un ou plusieurs langues cible avec des annotations diverses (morphosyntaxiques, sémantiques, etc.). Certains corpus comme le BTEC et EuroParl sont bruts, sans annotations spécifiques, mais les énoncés sont alignés au niveau des phrases. D'autres corpus parallèles contiennent plusieurs propositions de traduction pour le même énoncé. Nous prenons le BTEC et le JRC-Aquis comme cas de figure.

Les segments du BTEC sont stockés dans des fichiers textuels monolingues (les énoncés n'ont aucune structure), chacun étant repéré par un identificateur (ID). C'est grâce à cet « ID »

que les énoncés sont alignés au niveau des phrases. On trouve aussi dans ce même corpus des variantes ou des propositions pour chaque énoncé. La Table 14 illustre un cas de figure où la structure d'un énoncé est représentée par un ID (par exemple 000001\1\, 000001\2\, etc.) qui est un numéro d'ordre par rapport aux autres segments, et d'une ou de plusieurs propositions de traduction du même segment (un segment peut contenir un ou plusieurs énoncés), chacune avec un numéro d'ordre.

ID (source)	Segments sources	ID (cible)	Segments cibles
000001\1\	Hamburger and stew on the right side and salad, please.	000001\1\	햄버거랑 오른쪽에 있는 스튜랑 샐러드로 할게요.
		000001\2\	햄버거하고 오른쪽에 있는 스튜하고 샐러드 주세요.
000341\1\	Regular size, please.	000341\1\	Tamaño normal, por favor.
		000341\2\	Tamaño mediano, por favor.
000354\1\	Can I eat this without cooking?	000354\1\	Posso mangiare questo senza cuocerlo?
		000354\2\	Posso mangiare questa senza cuocerla?

Table 14 : Forme des énoncés et des propositions dans le BTEC

Le corpus JRC-Aquis est une compilation de plusieurs documents XML structurés sous le format « TEI.2 » et alignés au niveau des phrases. La récupération d'une paire de phrases se fait par des attribut XML « xtargets » (par exemple xtargets="X:Y Z" où X, Y et Z sont des ID de phrases). Le corpus existe en plusieurs langues, et les liens (par exemple FR-EN, IT-EN, SP-FR, etc.) représentent les bisegments (bilingues).

L'attribut « type » dans l'exemple ci-dessus désigne le type d'alignement qui peut être : 1:1, 1:2, 2:1, 2:2. Ensuite, les ID des énoncés alignés sont regroupés sous l'attribut lien «xtargets». Par exemple, <link type="1:2" xtargets="217;212 213"/> représente un alignement du couple de langues EN-FR de type 1:2. Le premier énoncé correspond à l'entrée 217 (EN) et les deux autres énoncés sont identifiés par les deux identificateurs 217 et 212 (FR).

```
<link type="1:2"  xtargets="217;212 213"/>

<link type="1:1"  xtargets="218;214"/>

<link type="1:1"  xtargets="219;215"/>

<link type="1:1"  xtargets="220;216"/>

<link type="1:1"  xtargets="221;217"/>
```

La complexité des corpus multilingues parallèles dépend de leurs caractéristiques et usages. Par exemple, dans les corpus destinés aux études littéraires et de linguistique

quantitative, ou pour la reconnaissance de parole, et même dans certaines MT (mémoires de traductions), on observe plusieurs différences. Selon (Boitet, *et al.*, 2006), ces différences se résument par les points suivants :

- parallélisme : les corpus sont alignés au niveau des « segments » (phrases ou titres à l'écrit, phrases ou tours de parole à l'oral), et, dans chaque paire, l'un est traduction de l'autre — cela peut varier, comme dans les transcriptions de débats parlementaires.

- nombre des langues et systèmes d'écriture : par exemple, le corpus EuroParl concerne 20 langues européennes, et presque autant de systèmes d'écriture différents (différences sur le jeu de caractères de base, ou sur les diacritiques, ou sur la notion même de « signe » : ch est une lettre bi-gramme en espagnol, mais pas dans les autres langues).

- annotations linguistiques lourdes : elles sont soit internes au texte (balisage au fil du texte), soit externes (structures + correspondances), par exemple des arbres concrets ou abstraits, de constituants ou de dépendances, mononiveau ou multiniveau, ou encore des graphes UNL. Ces annotations peuvent comprendre des correspondances plus ou moins fines entre texte et structure, de forme (txt, str), et entre correspondances de deux langues, de forme (txt, str)—(txt', str') — ex: S-SSTC à l'USM (Malaisie). D'où la différence par rapport aux MT.

En partant de ces propriétés et des deux cas de figure cités précédemment, nous constatons qu'un corpus parallèle, quelle que soit sa forme, peut être réduit à la structure logique plus simple suivante :

- Chaque corpus parallèle est constitué d'un ensemble de paires (dans le cas le plus simple, un énoncé source et cible). Ces paires sont compilées dans des fichiers séparés monolingues ou dans des fichiers multilingues. Chaque énoncé a un identificateur(ID). Dans le cas des propositions de traduction, d'autres identificateurs doivent être ajoutés, désignant par ce fait, une succession allant de 1 jusqu'à N (N est le nombre de propositions).

- Les fichiers monolingues séparés sont codés différemment. Par exemple, dans le cas du BTEC, les codages EUC-JP et le EUC-KR sont utilisés pour coder les énoncés en japonais et en coréen respectivement. Il faudrait recoder les corpus en codage unique pour permettre une gestion efficace de données multilingue et éliminer les problèmes liés aux mauvaises visualisations par les navigateurs Web.

- Les segments séparés dans des fichiers monolingues doivent être à un moment donné compilés sous leur forme multilingue, surtout lorsqu'il s'agit de visualiser en parallèle dans un éditeur les énoncés alignés.

Nous avons rappelé dans la partie précédente que le format TMX est efficace pour la gestion des mémoires de traduction. En partant du principe qu'un corpus parallèle est une sorte de MT, nous prenons la même structure et l'étendons pour inclure d'autres informations spécifiques aux corpus. Cette extension englobera l'ajout des propositions de différentes sources. Une proposition peut être en effet en provenance du corpus original (travail humain), générée automatiquement par la TA, ou produite par post-édition.

```
<tu tuid="074005" datatype="text" usagecount="bey" lastusagedate="20071023">
    <tuv xml:lang="en">
      <seg>
        I'm sorry, but we don't take reservations.
      </seg>
    </tuv>
    <tuv xml:lang="jp">
      <seg id="1" type="source" mt="" creationdate="20071023" lastusagedate= "20071023"
usagecount="bey">
        すみません、予約を取っていないのですが。
      </seg>
    </tuv>
    <tuv xml:lang="cn">
      <seg id="1" type="source" mt="" creationdate="20071023" lastusagedate= "20071023"
usagecount="bey">
        对不起，我们不接受预订。
      </seg>
    </tuv>
    <tuv xml:lang="kr">
      <seg id="1" type="source" mt="" creationdate="20071023" lastusagedate= "20071023"
usagecount="bey">
        죄송합니다만, 저희 식당에서는 예약을 받지 않습니다.
      </seg>
    </tuv>
</tu>
```

Figure 59 : Une instance du BTEC en TMX étendu

Chaque énoncé est représenté par une balise tuv (translation unit), qui enveloppe les propositions dans des éléments seg. Une tuv renvoie à une version monolingue d'un énoncé, et on lui associe l'attribut xml:lang pour désigner sa langue (kr=coréen, cn=chinois, ar=arabe, etc.). L'attribut type donne à un énoncé la nature du segment, qui peut être l'énoncé « source » issu de la compilation initiale à partir du corpus d'origine ou en provenance de sources diverses, qui pour le moment sont limitées aux « sources » indiquant qu'il s'agit d'un segment non modifié et importé du corpus original. Sinon, l'attribut mt indique que le

segment (énoncé) a été produit par TA, et contient alors le nom du système de TA qui l'a produit.

Il se peut aussi que le contenu textuel d'un segment seg soit généré par création ou post-édition humaine. L'attribut type prendra alors la valeur « human » ou « postedition », respectivement.

Important : nous éliminons toutes les formes de données en provenance des calculs, en particulier celles de l'évaluation. Nous incluons ces données dans une autre structure que nous définissons ultérieurement.

Lorsqu'il s'agit de corpus « voraces » en place comme ceux qui contiennent de la voix et de la vidéo, il faut prendre en compte certaines restrictions. Par exemple, les corpus de la campagne ESTER (Phase 2) se compose des éléments suivants (Le Meu, 2004) (ESTER, 2007):

- un corpus audio manuellement transcrit (environ 100 heures),
- un corpus audio non transcrit (environ 2.000 heures),
- un corpus de textes provenant de transcriptions.

S'il s'agit de ce type de corpus, nous limitons le traitement et la gestion à la transcription textuelle déjà faite. Si nous prenons les corpus ESTER, les corpus audio transcrits se composent d'enregistrements d'informations à la radio, enregistrés sur plusieurs chaînes de radio. Voici un exemple de ces transcriptions portant sur les entités nommées :

<**Event** **desc**= "??" **type**="entities" **extent**="begin"> *entity* <**Event** **desc**="??" **type**="entities" **extent**="end">

Les transcriptions dans cet exemple ainsi que d'autres peuvent être stockées dans les entrées de la structure proposée dans la sous-section précédente. Il suffit de stocker ces transcriptions dans des entités CDATA où il est possible d'insérer des portions en XML. Cependant, les annotations de surface restent inchangeables. Quel que soit le niveau de granularité de la transcription, notre structure peut garder les transcriptions profondes intactes et n'adapter que les annotations et les métadonnées nécessaires pour la compilation et l'import dans BEYTrans.

1.1.4 Représentations intermédiaires

Les représentations intermédiaires permettent de coder le sens d'un énoncé en LN. Elles ont des spécifications, généralement liées à un domaine précis. Le format « interchange format » (IF) du projet international C-STAR[67] en est un exemple. Il a été défini spécialement pour représenter le sens de dialogues oraux finalisés dans le domaine du tourisme.

L'autre langage pivot qui nous intéresse est UNL (Universal Networking Language), un langage considéré comme le « HTML des contenus multilingues ». Ce langage est orienté vers le texte écrit, et son utilisation nécessite deux composants essentiels : un enconvertisseur et déconvertisseur (un pour chaque langue naturelle).

Il existe actuellement des déconvertisseurs plus au moins complets pour 13 langues : arabe, arménien, chinois, anglais, espagnol, français, hindi, italien, indonésien, japonais, portugais, russe et thaï (Boitet, 2002). Il existe aussi des enconvertisseurs semi-automatiques ou automatiques pour 10 langues. Le développement de l'arménien est aussi en cours. Enfin, l'UNDL (la fondation UNL) a créé des corpus de taille importante annotés par des graphes UNL (environ 25 articles de l'encyclopédie EOLSS, soit 880 000 mots et 13675 segments, avec en moyenne un peu plus d'un graphe UNL par segments).

Le paragraphe suivant présente la spécification d'UNL mais ne constitue pas une étude exhaustive.

Les énoncés en langue naturelle sont représentés sous forme d'hypergraphes. Un hypergraphe UNL peut être décrit comme un ensemble de relations sémantiques binaires entre des nœuds. Chaque nœud porte un UW (Universal Word « mot universel », ou « lexème interlingue ») et des attributs. Une relation correspond à peu près à un cas profond (sémantique). Il y a 41 relations de ce type, et aussi 4 relations de « thésaurus » réservées à la KB (Knowledge Base) et à la définition de la hiérarchie des UW (icl, iof, equ, pof). Les attributs expriment des informations subjectives (modalité, aspect, nombre abstrait, négation, temps abstrait, acte de parole, etc.).

```
{org:es}
Los ríos dejaron prácticamente de llegar, taponados por presas.
{/org}
{unl}
```

[67] Ce projet est achevé depuis quelques années. Une plate-forme expérimentale a été développée dans le domaine du tourisme : plusieurs agents touristiques dispersés sur plusieurs sites ont pu communiquer dans leur langue maternelle avec des clients parlant une autre langue.

obj(flow(icl>occur).@past.@entry.@not, river.@def.@pl)
man(flow(icl>occur).@past.@entry.@not, almost)
rsn(flow(icl>occur).@past.@entry.@not, :01)
obj:01(block(icl>do).@past.@entry, river.@def.@pl)
agt:01(block(icl>do).@past.@entry, dam.@pl)
{/unl}
{fr} Bloquées par des barrages, les rivières ne coulaient presque plus.{/fr}
{el} Blocked by dams, the rivers almost stopped flowing.{/el}

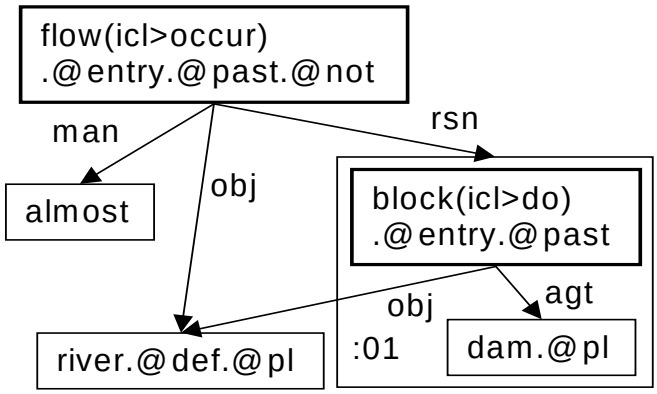


Figure 60 : Graphe UNL d'un énoncé espagnol déconverti en français et anglais

Un graphe UNL est généralement écrit en format textuel, et peut être visualisé à l'aide de certain outils come UNL-viewer ou le générateur de dessins de UNLDECO[68] (Sérasset, 2003), comme le montre la Figure 60.

Comme notre équipe ne s'intéresse qu'à la partie concernant le français, nous limiterons l'étude aux corpus UNL-FR. Nous montrerons toutefois comment il est possible, par appel des serveurs de déconversion d'autres langues (dans d'autres pays), de générer des versions multilingues en partant du même corpus.

Nous ne nous intéressons pas ici au développement de ce pivot, mais à stocker, visualiser et éventuellement éditer des graphes UNL se présentant comme des annotations des énoncés (en langue naturelle) d'un corpus. Il devrait aussi être possible d'appeler les enconvertisseurs et déconvertisseurs disponibles pour créer les graphes UNL associés aux énoncés, puis d'obtenir leurs traductions dans diverses langues.

[68] http://gohan.imag.fr/unldeco/

7.1.2 Descripteurs riches

La masse de données que nous souhaitons mettre en œuvre n'est pas homogène, et doit pouvoir contenir plusieurs corpus de plusieurs centaines de millions de mots. À part les contenus des énoncés, on doit stocker des informations additionnelles sous forme de métadonnées liées à chaque corpus (ou source de données).

Ces informations doivent être gérées dans notre environnement en gardant toutes les descriptions, et on doit pouvoir en ajouter d'autres.

Nous rencontrons le problème suivant : comment gérer efficacement une masse de données de sources différentes.

Il ne semble pas judicieux de penser à une manipulation globale dans une interface unique. Cette situation nous conduit à diviser les données et à leur donner des descripteurs qui serviront à la fois au système pour récupérer et manipuler efficacement ces données, et à renseigner les linguistes/traducteurs en leur présentant une série d'informations linguistiques héritées de la version d'origine de chaque corpus.

Nous définissons cinq descripteurs. Certains sont définis de façon permanente et d'autres temporairement, pour une tâche précise, par exemple une expérience d'évaluation sur un ensemble de données. Ces descripteurs sont les suivants :

1. Descripteurs permanents : ils concernent les données, les corpus, et les unités de traduction.

   a. *Descripteur des données.* Il décrit le nombre de mots et segment du corpus compilé dans le système, ainsi que le nombre de langues et leurs codages. Il contient aussi des liens XML de type « xtarget » pointant sur des parties textuelles d'autres corpus.

   b. *Descripteurs de corpus.* Un tel descripteur précise le nom d'un corpus, les droits d'auteur, la langue source, les langues cible, le genre, le domaine, le nombre d'énoncés, l'auteur et les dates de création ainsi que de dernière modification. Il contient l'ID maximum et minimum des énoncés. Le nombre de langues et les ID peuvent être modifiés lors de la manipulation (exemple, ajout d'un nouvel énoncé ou de nouvelles langues). Les auteurs peuvent être des noms de compte d'utilisateurs

s’ils sont membres, ou une adresse IP s’il s’agit d’un invité (accès temporaire).

c. Descripteurs d’unité de traduction. Un tel descripteur précise la date de création et de dernière modification, un nom d’auteur, le nombre des propositions, et un identificateur de référence.

2. Descripteurs de navigation : ce sont des informations décrivant une partie des données générée à la demande d’un utilisateur (par exemple, sélection de toutes les paires anglais-français du corpus BTEC modifiées avant une date donnée). Ce sont les suivants :

   a. *Descripteur de navigation.* Avant de lancer une opération de visualisation sur un ensemble de données, il faut spécifier les besoins sous forme de paramètres. Ces derniers seront combinés avec les descripteurs des données sélectionnées pour permettre une navigation aisée. Par exemple, les données retenues provenir de sources différentes avec un intervalle d’ID spécifique. Ce descripteur servira à faciliter l’acquisition de données lors du passage d’un ensemble de données à un autre, parce que les données ne peuvent être visualisées dans leur totalité.

   b. *Descripteur d’évaluation.* Une évaluation de TA se fait toujours sur un sous-ensemble de données sélectionnées manuellement ou au hasard. Une évaluation a donc certaines spécificités (corpus, langue, calcul, intervenants humains, etc.) et est faite que dans une session précise (pour chaque groupe). Chaque session contient ses propres données et paramètres à savoir le nombre de paires, les langues, et les données supplémentaires (colonnes dans l’éditeur) telles que le temps de post-édition, les distances d’édition, et les scores calculés par les métriques statistiques (BLEU, NIST, etc). C’est grâce à ce dernier descripteur qu’une session est maintenue et peut être restaurée par la suite.

Le premier ensemble de descripteurs pourra être compilé en TMX étendu. Par contre, le deuxième ensemble ne sera que temporaire et sera supprimé du système après la fin d’une session donnée, qui correspondra à la fin d’une expérience.

7.2 Problèmes liés aux traitements globaux

7.2.1 Lenteur

Le temps de réponse est un facteur important. Les réponses aux requêtes doivent être effectuées par notre environnement dans un temps de réponse minimal. Nos premières expériences nous ont montré qu'il faudrait créer des index pour accélérer l'accès aux données, et que l'ajustement du cache (« heap ») améliore la performance.

Nous adoptons une architecture « 3 tiers » avec division de chaque corpus en un ensemble de petits documents de taille variant entre 10 000 à 50 000 segments, selon la taille des segments. Un tel « document virtuel » peut contenir environ 1M mots. Cela correspond par exemple à 40 000 segments quadrilingues du BTEC (anglais, japonais, chinois, coréen) soit ¼ du total. Cette quantité de données à visualiser est devenue possible grâce à la technologie actuellement employée pour le développement des éditeurs Web (exemple gestion de cache, Ajax, HTTPRequest, etc.).

Ces documents virtuels sont accompagnés d'informations supplémentaires telles que les métadonnées (par exemple début et fin d'énoncés, langues, etc.). On peut donc visualiser et manipuler les 163 000 énoncés du BTEC parallèlement sur 4 interfaces Web ! Nous verrons qu'avec l'éditeur implémenté dans un Wiki, il est possible de gérer efficacement de gros corpus.

Si l'on note par C_i un corpus i (supposons le BTEC) alors C_i est stocké physiquement dans un ensemble de documents : $\{E_1, E_2, E_3, \ldots, E_n\}$. Chaque ensemble E_i est représenté par un descripteur qui servira pendant la navigation et la visualisation.

Pour avoir plus de souplesse, la taille d'un ensemble E_i est spécifiée au moment de l'import du corpus. Par exemple, il est beaucoup plus commode de visualiser et de manipuler les énoncés du BTEC dans des documents de taille acceptable (de 10 000 à 50 000 énoncés) que par blocs de 1 000 énoncés, comme proposé dans sa version origine. Nous pouvons avoir plus d'efficacité si nous utilisons un mécanisme semblable à une « fenêtre coulissante » où nous pouvons avoir sur une même fenêtre (ou interface) des données provenant de plusieurs descripteurs différents, contigus ou non.

7.2.2 Visualisation et navigation

La visualisation dépend de la nature des données et de leurs structures. Par exemple, une meilleure visualisation d'un corpus parallèle est obtenue par une interface où les énoncés sont

aussi présentés en parallèle. Il faut aussi prendre en considération le passage à l'échelle vers une masse de données, car d'autres critères interviennent. Tout d'abord, il faut concevoir une interface qui permet aux utilisateurs de récupérer des données à partir d'un ou de plusieurs corpus par des requêtes sur tout ou partie des données et de la configurer selon les expériences et les tâches. D'autre part, certaines données structurées comme les graphes UNL nécessitent une présentation particulière car, à part la transcription textuelle des énoncés, il faut visualiser les graphes.

La Figure 61 illustre une interface avancée d'un éditeur de visualisation et de navigation qui peut être appliqué à l'évaluation de la TA. La figure montre une conception possible qui peut être paramétrable selon les besoins et les tâches.

Cette conception se divise en deux parties :

1. Une partie persistante. Elle consiste à gérer toutes les opérations nécessaires à la visualisation et à la navigation et est composée des identificateurs (ID), des énoncés source et cible, et des propositions de traduction.

(1) Édition et navigation								**(2) Édition et évaluation**							
Enoncés			**Propositions**					**Traductions candidates**			**Post-éditions**			**Calculs**	
Id	source	Cible	Id_1	Id_{p2}	Id_{p3}	...	Id_{pn}	STA_1	...	STA_k	$PostE_1$	...	$PostE_p$	NIST	BLEU
Id_1															
Id_2															
...		...		...	...	...	...	...	...	...	...	...	...	...	...

Figure 61 : Visualisation de corpus parallèles

2. Une partie propre aux tâches d'évaluation. Un ensemble de traductions candidates est produit par TA. Les énoncés produits par la post-édition peuvent être à leur tour pris comme des références pour d'autres expériences d'évaluation. Enfin, les résultats des métriques appliquées sont visualisés dans cette partie.

Cette interface permet de représenter et de gérer une expérience d'évaluation où il est possible d'avoir plusieurs propositions de traduction pour un même énoncé source. Les métriques peuvent être configurées de sorte qu'elles ne soient applicables que sur un sous-ensemble des données sélectionnable directement à partir de l'interface. Par exemple, nous pouvons choisir une « colonne de post-édition » comme cible pour les calculs au lieu de la colonne « cible ».

Avant de lancer une expérience, le nombre et le type des colonnes doivent être configurés à l'avance. On peut ne prendre qu'une seule proposition ou ne pas avoir la colonne de la post-édition. La configuration de l'interface dépend de la nature de la tâche d'évaluation.

Notons aussi que l'interface est compatible avec la manipulation d'autres corpus annotés tels qu'UNL. Les énoncés et les graphes peuvent être visualisés et édités tant qu'ils sont en format textuel. Des services externes tels que les déconvertisseurs et enconvertisseurs peuvent être appelés pour avoir des prétraductions au lieu des STA (par exemple, un déconvertisseur UNL-russe).

L'originalité de cette interface (éditeur) réside dans sa dynamicité. C'est aux utilisateurs de définir sa forme en partant de leur besoins. Si l'on ne s'intéresse qu'à UNL, les graphes sous leur forme textuelle seront présentés dans des colonnes séparées et synchronisés avec une langue source et cible ou même plusieurs langues à la fois. L'appel aux composants UNL se fait de la même façon que l'appel des traducteurs automatiques disponibles en libre service sur la Toile.

7.3 Problèmes liés aux autres spécificités

7.3.1 Traitements informatiques externes

3.1.1 Appel de mesures statistiques comme BLEU, NIST, mWER

Des métriques telles que BLEU, NIST, mWER ou METEOR sont souvent utilisées pour remplacer les jugements qualitatifs humains, car ces derniers sont très coûteux. La plupart des approches développées pour automatiser l'évaluation calculent la similarité entre des traductions de référence et des traductions candidates produites par TA en se basant sur des idées intuitives (Denoual, 2006) :

« ...l'idée que le score de similarité doit être proportionnel au nombre de mots en commun, ou encore celle que des mots correspondants présents dans le même ordre devraient produire un score supérieur à celui de mots présents dans le désordre. »

Ou,

« ...plus la somme des longueurs des sous-chaînes contiguës, communes à la chaîne candidate et à la chaîne référence, est grande, plus le score devrait être élevé. »

Les comparaisons sont effectuées en termes de comptage de courtes séquences de mots entre les traductions de référence et les traductions candidates. Par exemple, la méthode BLEU repose sur cette idée et est formulée comme suit.[69]

Le score BLEU est désigné par BLEU_{wN} où *N* est la longueur maximale des n-grammes considérés, appelée aussi « ordre » du score.[70] Un score BLEU_{wN} est le produit d'une pénalité *BP* et de la moyenne géométrique des "précisions modifiées" à l'ordre *n*, notées p_n, calculées pour toutes les longueurs de chaînes jusqu'à *N*.

$$\text{score BLEU}_{\text{wN}} = BP \times \sqrt[N]{\prod_{n=1}^{N} p_n}$$

La mesure BLEU est fondée sur le comptage d'occurrences de n-grammes de mots, et la méthode est en quelque sorte « aveugle » : il est donné autant d'importance à un mot qu'à un autre. Une simple permutation, ou un simple ajout, ne pénalisent souvent pas une phrase candidate longue. Cette limitation fait l'objet de l'un des reproches les plus courants adressés à la méthode BLEU, mais ce n'est pas la seule.

Les méthodes d'évaluation en elles-mêmes n'entrent pas dans le cadre de cette thèse. Nous ne nous focalisons en effet que sur leur application dans l'évaluation pour produire des scores durant nos expérimentations.

3.1.2 Appel de systèmes de TA

L'appel à la TA n'est effectué que dans les deux situations suivantes :

1. produire des prétraductions pour l'extension de corpus vers d'autres langues.

2. produire des traductions candidates lors des tâches associées à la TA.

Ces dernières seront produites et utilisées pour l'évaluation subjective et objective.

7.3.2 Mode d'exécution des tâches

A part la visualisation et la navigation, l'éditeur dédié offre des fonctionnalités linguistiques internes et externes avancées (segmentation, synchronisation, appel à la TA et aux calculs statistiques, etc.). Il doit donc permettre d'exécuter plusieurs tâches, pour certaines

[69] La méthode BLEU a été proposée par IBM en 2001.
[70] Ces notations ont été prises de la thèse d'Etienne Denoual selon une référence à la notation de Babych.

en parallèle. Comme la liste est longue, nous ne présentons que les tâches les plus utiles et difficiles à résoudre.

3.2.1 Mode d'exécution asynchrone

C'est un mode pseudo-parallèle où l'exécution s'effectue sur le même serveur. Il est possible de lancer plusieurs requêtes asynchrones et de continuer à travailler en parallèle sur le traitement. Une fois les résultats obtenus, ils sont affichés aux traducteurs. Un traducteur peut lancer une opération de remplacement sur un gros volume de données (par exemple, remplacement de « s.v.p » par « s'il vous plaît» sur 163.000 phrases du BTEC), tandis qu'en parallèle, d'autres requêtes d'appel à la TA et de recherche dans la MT sont en cours.

3.2.2 Mode d'exécution parallèle

Les corpus volumineux que nous souhaitons gérer et traduire nécessitent de prendre en compte les contraintes imposées par les systèmes de TA gratuits. Par exemple, une requête Web de traduction sur Systran (Web) ne peut dépasser le nombre de mots autorisé. Le système rejette toute requête dépassant ce nombre. Il n'est donc pas possible, par exemple, de traduire tout le BTEC avec une seule requête (163 000 segments).

Ce problème nécessite d'adopter la même méthode que celle utilisée dans le système TRADOH, c'est-à-dire, diviser les documents en petits morceaux acceptables par les STA libres et ensuite regrouper les traductions partielles obtenues pour avoir la traduction complète d'un document. En revanche, cela nécessite une préparation des données à traduire (division, codage, gestion des résultats obtenus). Pour arriver à réaliser une tâche de traduction par TA « libre », il faut un certain nombre de requêtes (*R*) de traduction, en fonction de la division possible d'un corpus donné.

$$R = \frac{T \times N}{C} \qquad \text{(Équation 5)}$$

où *N* est le nombre de documents $\{E_1, E_2, ..., E_i, ..., E_n\}$.

Chaque document correspond à un descripteur D_i. Il y a donc autant de descripteurs à lire que de documents à traduire. Si l'on suppose que les documents sont équilibrés et contiennent un nombre *P* d'énoncés, alors le nombre de mots (*T*) par document (E_i) est : $T = P \times (M_1 + M_2 + ... + M_p)$ (*M* est la taille en mots). *C* est le nombre de mots à traduire.

Certains systèmes de TA limitent la traduction à un certain nombre de caractères. Dans ce cas, une approximation est prise en fonction du nombre de caractères autorisés pour définir le nombre de mots acceptés. Par exemple, nous avons déterminé la moyenne du nombre de mots présents dans des échantillons d'énoncés pris du BTEC et avons trouvé que le nombre moyen de mots *C* est 90 en prenant 500 pour le nombre de caractères autorisés par un système de TA gratuits. Cette limite de 500 caractères est souvent utilisée par les systèmes gratuits. D'autres STA sont moins restrictifs, comme « Babelfish » qui exploite la technologie Systran mais permet de traduire des textes plus longs.

Note : nous avons pris les mots comme unités de calcul pour avoir un nombre de requêtes exact, mais en réalité les unités de traduction sont le plus souvent des énoncés et ne doivent pas être divisées. Pour la traduction des énoncés, nous suivons donc le principe de *tout ou rien*, c'est-à-dire qu'on traduit tout l'énoncé ou qu'on le fait passer à la requête suivante. Des précautions doivent aussi être prises concernant la taille en mots du dernier document, et le dernier énoncé. Nos expérimentations ont montré que la traduction de l'anglais en français d'un énoncé du BTEC contenant entre 15 et 20 mots à l'aide d'une requête « HTTPRequest » sous BEYTrans à l'aide de « Babelfish » prenait entre 3 et 5 secondes[71].

La traduction de tout le BTEC vers le français nécessite : $R = 1\,059\,500/C$ requêtes (163 000 énoncés ≅ 6 500x163 mots)[72] où $C \cong 150$ mots qui correspondent (si nous prenons le minimum) à 10 phrases. Le nombre de requêtes *R* est donc 7063 et le temps total de traduction du BTEC est 59 heures (≅ 2,5 jours).

Cette durée peut être réduite si plusieurs traductions sont lancées sur plusieurs postes en parallèle. Cela est possible grâce à la technologie Wiki où plusieurs sessions peuvent être ouvertes pour la manipulation d'un document donné.

Une autre solution au problème a été proposée par l'architecture de Yakushite.net (Kitamura, *et al.*, 2003), où les traductions sont réparties sur plusieurs serveurs grâce à une architecture parallèle. Sur chacun d'entre eux fonctionne un traducteur automatique. Cela n'est pas possible dans notre environnement, car nous ne disposons pas de notre propre système japonais-anglais ou chinois-anglais. Nous pensons aussi que le nombre de requêtes à lancer ne diminue pas la performance, parce qu'elles sont exécutées au niveau serveur.

[71] Ces traductions ont été faites par appel à Systran via Babelfish. L'interface de ce dernier permet de traduire 150 mots au lieu de 500 caractères proposés par le site officiel de Systran.
721 fichiers = 1 000 segments ≅ 6 500 mots ≅ 26 pages (une page standard contient 250 mots).

Techniquement, chaque requête de traduction correspond à une requête « HTTPRequest » en « Ajax ». Il faut lancer *R* requêtes HTTPRequest pour obtenir la traduction d'un corpus complet.

En résumé, la traduction d'une masse de données nécessite de lancer un nombre considérable de requêtes et ce nombre dépend des paramètres autorisés par les systèmes de traduction gratuits sur le Web. Pour avoir des traductions rapides, il faut choisir des systèmes proposant un *C* élevé. Cela peut minimiser le nombre de requêtes, et en conséquence le temps de traduction.

3.2.3 Mode d'exécution bloquant

On perçoit le mode bloquant lors de l'exécution d'une requête Web classique où les utilisateurs sont bloqués face à leur navigateur. L'exécution bloquante est utile lorsque deux tâches peuvent modifier les mêmes données. Par exemple, un usager doit attendre la fin d'une opération de remplacement global avant de passer à la traduction globale d'une quantité de données.

Cependant, un autre problème apparaît, pour l'instant incontrôlable, dû au principe Wiki : les modifications en arrière-plan des données en cours de manipulation par d'autres usagers ne peuvent bloquer une opération faite sur les mêmes données par un autre utilisateur. Nous ne proposons aucune solution dans le cadre de cette thèse, mais ce sera un sujet intéressant à développer dans le futur.

Conclusion

Les corpus destinés à la construction ou à l'évaluation des systèmes de TA sont nombreux. D'un corpus à l'autre, les structures et la taille changent. Nous avons étendu notre environnement à la gestion d'une masse de données, constituée principalement de corpus parallèles destinés à la TA. Nous avons étudié une dizaine de corpus et identifié quelques problèmes.

L'une des applications que nous avons particulièrement étudiées est l'évaluation, où l'on a besoin de gérer des données spécifiques pour charger, visualiser, manipuler et calculer des scores selon plusieurs métriques. Nous avons aussi défini des structures logiques qui permettent de gérer non seulement les énoncés textuels, mais des structures complexes telles que les transcriptions et les graphes (par exemple UNL). D'autres propositions ont permis de concevoir et réaliser une interface de navigation, qui grâce aux descripteurs définis, offre aux utilisateurs une configuration adéquate aux tâches.

Chapitre 8

Extension de BEYTrans aux gros corpus multilingues

Introduction

Dans le chapitre précédent, nous avons exposé les problèmes de gestion et de multilinguisation d'une masse de donnée, en partant des problèmes de gestion que posent la taille et la complexité de structure, tout en concentrant l'analyse sur les représentations internes de données.

Cependant, ces données doivent d'être gérées dans des interfaces adéquates pour permettre aux usagers une interaction souple et efficace avec des corpus parallèles et volumineux, et leur donner la possibilité de lancer des opérations globales telles que la recherche, le remplacement, etc. Autrement dit, la navigation et la manipulation de données de masse doivent se faire dans des éditeurs Web qui permettent non seulement la navigation, mais aussi la modification et l'interaction, en prenant en compte la présentation et la persistance des données durant la post-édition.

Nous proposons dans ce chapitre des méthodes de visualisation, de navigation et de manipulation des données. Nous expliquons comment nous avons étendu BEYTrans, en particulier l'éditeur Web BTedit, pour gérer de très gros corpus parallèles.

8.1 Visualisation et navigation

8.1.1 Navigation

1.1.1 Navigation parallèle en mode lecture

Les corpus parallèles sont présentés dans des pages HTML où il est possible de naviguer par des liens Web sous une transcription Wiki. Les descripteurs sont des structures à sélectionner avant toute opération de navigation. Le nombre de descripteurs correspond au nombre de documents constituant un corpus. Dans notre configuration, chaque document correspond à un descripteur sélectionné d'avance. La visualisation se fait en sélectionnant un

ou plusieurs descripteurs. Cela permet de ne charger en mémoire qu'un nombre restreint de descripteurs et les données qui leur correspondent.

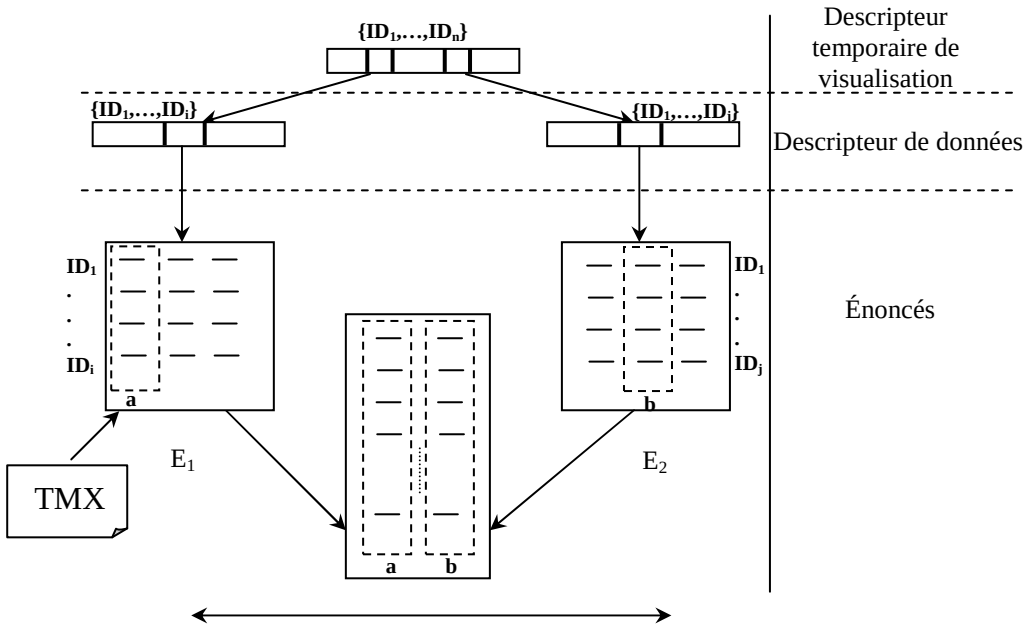


Figure 62 : Visualisation d'une masse de données dans BEYTrans

A travers le descripteur temporaire de visualisation, d'autres descripteurs de données sont identifiés, chacun correspondant à un ensemble précis de données.

1.1.2 Navigation parallèle en mode édition

Nous avons développé une interface d'édition similaire à notre éditeur développé dans la deuxième partie, mais avec une structure modifiée. L'unité de base reste l'UT.

Un ensemble d'énoncés est considéré comme un ensemble de bisegments d'une mémoire de traduction et peut être modifié énoncé par énoncé. Dans cette nouvelle interface, il est possible, selon la configuration montrée dans la Figure 61, d'ajouter de nouvelles propositions de traduction, une nouvelle référence, et des calculs en cas d'évaluation de la TA. La navigation en édition se fait comme précédemment dans une interface à la Excel.

Il existe un mode « lecture », notamment utilisé lors des évaluations subjectives, qui ne permet pas la modification des données.

8.1.2 Visualisation

Les corpus sont visualisés en les divisant en documents de taille réduite, affichables dans un Wiki. Ces documents sont lus dans une interface où il est possible de passer en mode d'édition collaborative.

1.2.1 Visualisation dans une page html « active »

La forme de visualisation idéale pour les corpus parallèles compilés est une table où les colonnes représentent les langues et les lignes les énoncés. Mais cela nécessite une transformation du mode de représentation XML (TMX) en mode HTML. De plus, il important de savoir que, bien que les Wiki acceptent le code HTML, il est préférable de générer des formes selon les transcriptions Wiki usuelles (Figure 64). Pour cela, nous avons écrit une feuille de style XSLT pour extraire le contenu et générer un modèle (template) de représentation en « Velocity » (Figure 63).

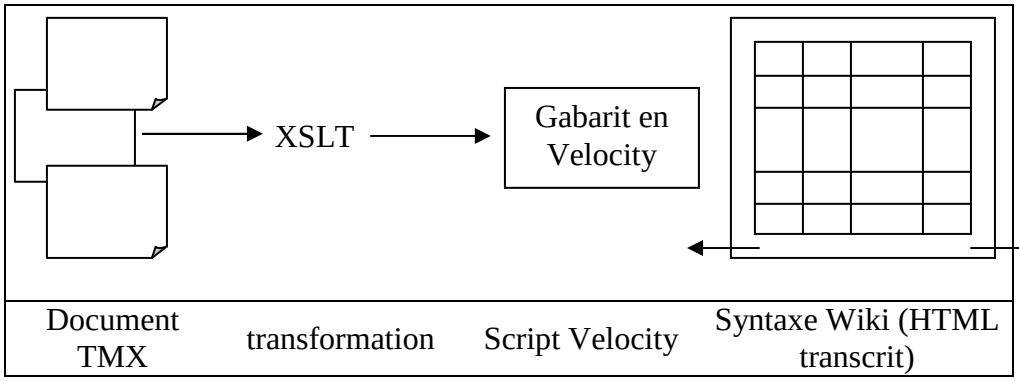


Figure 63 : Visualisation parallèle de corpus en HTML transcrit (Wiki)

La visualisation peut être directe ou par le biais d'une configuration utilisateur déterminée.

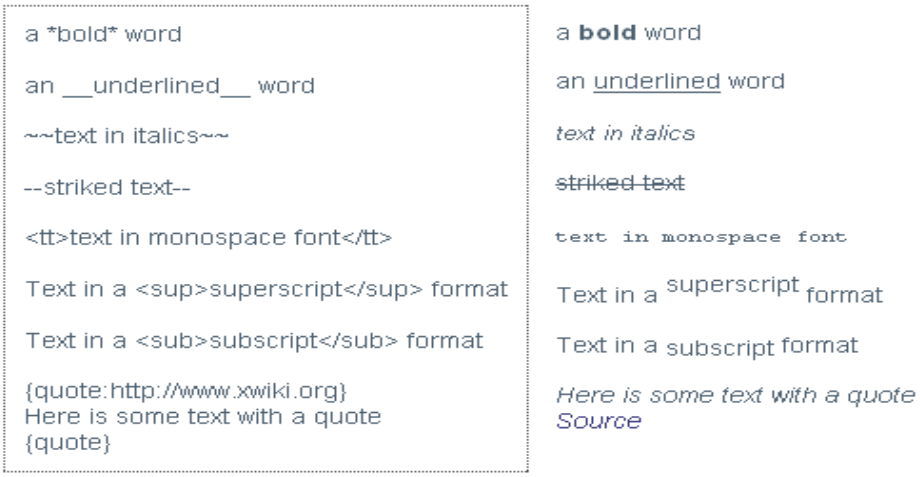


Figure 64 : Transcription du HTML dans les Wiki

L’accès aux données est censé se faire de la même façon que celui en consultation ordinaire pour un document court en Wiki.

1.2.2 Mode de passage en contexte d'édition

1.2.3 Fenêtre logique sur les données

Une fenêtre logique sur un sous-ensemble de données est définie avant la navigation. Le passage entre deux descripteurs nécessite la récupération de données de documents différents : dans ce cas, les documents XML correspondant au sous-ensemble présenté dans une « fenêtre coulissante » sont récupérés et traités à l’aide d’un parseur qui permet aussi de produire un format XML facilement interprétable par l’éditeur Web implémenté.

Le passage de la lecture à l’édition se fait par des liens Wiki « edit » où l’on a remplacé le modèle (template) d’édition initial par un nouveau modèle qui présente les données sous forme d’un éditeur à la Excel au lieu d’un éditeur Wysiwyg.

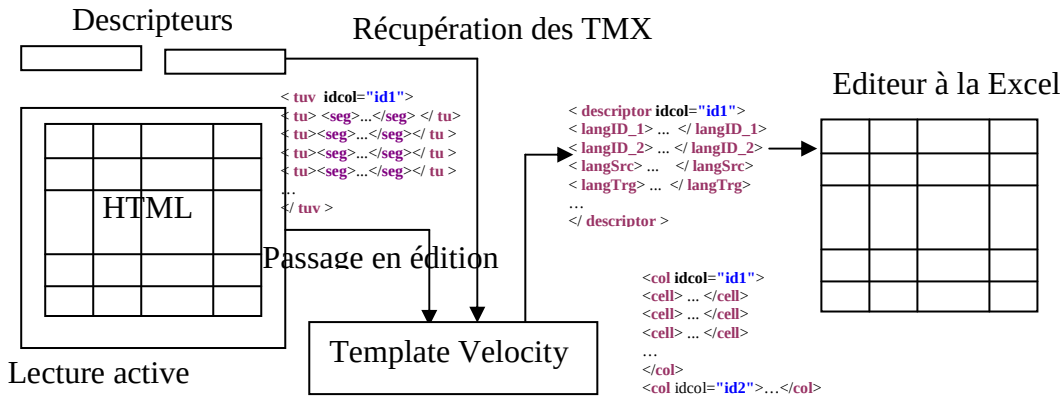


Figure 65 : Passage en mode lecture au mode édition

Dans un contexte de lecture, une session contient un ensemble de descripteurs sur lesquels l’utilisateur est positionné. Si un passage en mode édition est déclenché, alors l’ensemble des descripteurs actifs est utilisé pour créer le format de l’éditeur ainsi que pour récupérer les données à présenter.

1.2.4 Parcours

1.2.4.a Dispositifs d'interface

L’un des critères que nous prenons le plus en compte est la convivialité. L’interface doit être aussi simple que possible pour le lancement des requêtes ou de fonctionnalités. Ces dernières couvrent autant les calculs et les opérations de manipulation de corpus (par exemple

manipulations globales, tris, etc.) que les manipulations de l'interface elle-même (ajout de ligne, fusion, division, etc.).

D'un côté, l'interface permet la manipulation de corpus (post-édition, ajout, fusion, annotations, etc.) et d'un autre côté, elle permet le lancement des expériences d'évaluation de TA (appel des métriques, appel de systèmes de TA, etc.).

1.2.4.b Tris

Toutes les opérations de manipulation peuvent porter sur un segment isolé, sur un sous-ensemble, ou sur tous les corpus stockés dans la BD.

En particulier, il faut pouvoir trier un ensemble d'énoncés selon divers critères (appliqués aux métadonnées), et éventuellement selon le contenu (ordre lexicographique, longueur, etc.).

Le temps de réponse de BTedit reste acceptable, jusqu'à environ 50 000 énoncés. Au-delà, il faut attendre la fin du traitement. Si le tri est lancé sur l'ensemble des données, cela ne suffit plus, l'attente est trop longue.

Nous n'avons pas vraiment abordé cette question, mais une solution serait de répartir le calcul sur plusieurs machines, puis de faire en temps linéaire une fusion des monotonies obtenues. Si par exemple on a un tri en $O(n \log_2 n)$ qui prend 1 seconde pour 10.000 segments (ou lignes dans la grille « à la Excel » de l'éditeur), cela correspond à un peu moins de 140.000 comparaisons par seconde.

Si l'on avait 1M segments à trier, cela demanderait un peu moins de 150s sur un seul processeur. On pourrait alors lancer 100 tris en parallèle, ce qui coûterait 1s, et fusionner les 100 monotonies obtenues, ce qui coûterait $10^6 \log_2 100$ comparaisons, soit moins de 50s, d'où environ 50s au total, ce qui donnerait une attente 3 fois moins longue, mais encore trop longue.

Pour obtenir un tri presque en « temps réel » sur l'ensemble des données, la seule possibilité semble, non pas d'aller vers une architecture parallèle, mais, beaucoup plus simplement, de maintenir en permanence l'ensemble trié par rapport aux critères de tri utilisés (jamais plus de 2 ou 3), en utilisant des structures comme les arbres de recherche équilibrés. Ainsi, dans un « AVL »[73], le coût d'une modification est au maximum $(\log_2 n)$ comparaisons et $(1 + \log_2 n)$ « rotations » (moins coûteuses), soit si n=10^6 moins de 1,5x10^{-4} seconde.

[73] http://fr.wikipedia.org/wiki/Arbre_AVL

Par contre, d'autres opérations telles que les remplacements globaux de chaînes (mots, segments, etc.) sont faisables parce qu'elles s'effectuent en temps linéaire, et non pas en $O(n \log_2 n)$ comme le tri. Si l'on fait par exemple 20 000 remplacements par seconde, cela fait seulement 5 secondes pour 1M segments sur 1 processeur, donc c'est très supportable.

Nous limitons donc pour l'instant les tris aux données visualisées sur une instance de l'éditeur. L'idée de maintenir les données triées sera expérimentée dans le futur.

1.2.5 Filtrage et sélection de sous-corpus

La division d'un corpus volumineux en un ensemble de « documents virtuels » de taille limitée (à 10 000 – 50 000 segments) est une technique que nous avons adoptée dès le début. La même idée peut être utilisée pour offrir d'autres fonctionnalités telles que le filtrage, et la sélection de données précises sera nécessairement contenue dans le même document virtuel.

Le filtrage peut porter sur l'ensemble des données d'un corpus ou sur les données sélectionnées par des descripteurs et est réalisé par la spécification d'un ensemble de paramètres (par exemple, langue source et cible, intervalle des ID, auteur, date, un corpus ou plusieurs, etc.). Ces paramètres permettent de construire (et éventuellement de nommer) un ensemble de segments parmi d'autres pour lui appliquer des fonctionnalités précises (par exemple, appel à la TA). Dans certaines situations, on a aussi besoin de récupérer des données aléatoirement à partir d'un ou de plusieurs corpus (par exemple, dans la campagne IWSLT, les organisateurs sélectionnent aléatoirement des énoncés du BTEC pour construire des corpus d'évaluation (dev sets, test sets).

1.2.6 Edition

1.2.6.a Sélection d'un sous-ensemble à éditer

La sélection d'un sous-ensemble d'énoncés d'un corpus donné (ou de plusieurs) dépend des tâches, qui sont de diverses natures. L'édition peut être faite pour améliorer le contenu des énoncés du point de vue du niveau de la langue, si ce contenu est produit par des intervenants en ligne moins compétents.

Dans le cas de l'évaluation, l'édition est beaucoup plus complexe, et dépend fortement du but poursuivi. Par exemple, dans le cas de l'extension vers une nouvelle langue, les intervenants peuvent appeler la TA pour avoir des prétraductions, qui à leur tour peuvent être post-éditées et annotées par des évaluateurs pour leur associer un niveau de qualité.

L'un des problèmes que nous avons rencontrés est la définition de la forme de la tâche, qui diffère d'une expérience à une autre. Il peut s'agir dans le cas d'une évaluation humaine « subjective » ou « objective » par calcul statistique, d'un format précis de l'interface d'édition. En tout, l'édition offre la manipulation avec une performance raisonnable d'environ 50 000 énoncés si l'on considère les colonnes des propositions, de la post-édition, et des prétraductions comme des énoncés indépendants du sous-ensemble sélectionné.

Les fonctionnalités offertes par l'interface d'édition sont actuellement :

1. Ajout d'une nouvelle langue.
2. Appel à la TA.
3. Annotation multiniveau des prétraductions.
4. Evaluation statistique BLEU, NIST.
5. Gestion de propositions multiples.
6. Suivi des versions en mode collaboratif à la Wiki.
7. Post-édition.
8. Tri s'il se fait dans un temps acceptable : pour l'instant on se limite à 50 000 énoncés.
9. Recherche/remplacement avec possibilité d'annulation.
10. Gestion d'annotations XML au cœur du code HTML.

Comme notre équipe participe activement au projet international UNL, nous avons étendu l'édition aux graphes sémantiques UNL. La transcription des graphes UNL stockés dans des fichiers UNL-XML a été importée dans notre environnement pour l'expérimentation. Il est maintenant possible d'éditer ces représentations complexes, tout en ayant la possibilité d'appeler les déconvertisseurs et les enconvertisseurs disponibles gratuitement sur la Toile[74] (Figure 66).

Les résultats produits par les déconvertisseurs peuvent être post-édités par des intervenants.

[74] Exemple : le déconvertisseur russe est disponible à : http://www.unl.ru/server.html

Il est à noter que le langage UNL est une représentation sémantique qui ne peut être manipulée ou générée par des non-spécialistes en mode textuel. Les graphes UNL peuvent être modifiés en mode graphique, mais nous n'avons pas encore intégré d'éditeur graphique dans BTedit[75].

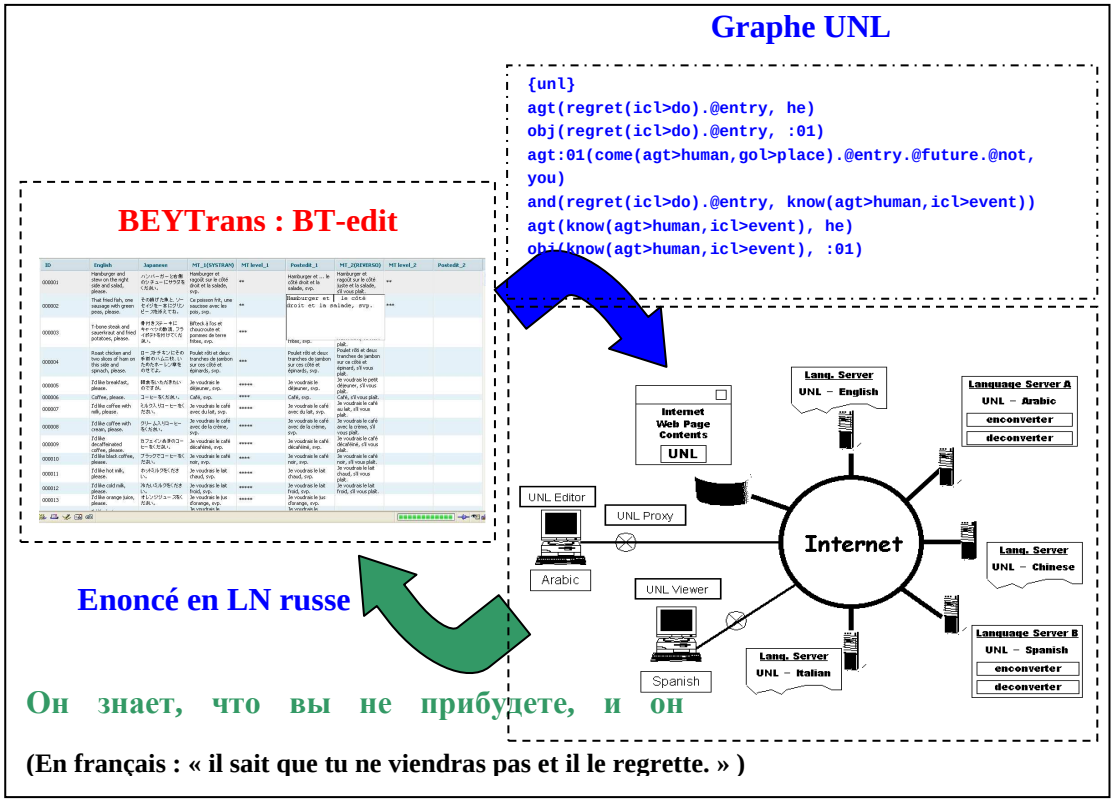


Figure 66 : Communication de BEYTrans avec le déconvertisseur UNL-russe

L'éditeur ainsi que les fonctions collaboratives de suivi de la progression des travaux en cours seront, nous l'espérons, très intéressants pour la communauté scientifique travaillant sur UNL. En effet, il est maintenant possible avec BEYTrans de générer des corpus UNL, de les évaluer directement sans changement d'interface, et d'accéder aux déconvertisseurs, enconvertisseurs et autres composants UNL (vérificateurs, visualiseurs).

Notons enfin que cette idée a été reprise dans une thèse commencée en mars 2007 par HUYNH Phap Cong, qui a développé, également en Wiki, un système spécifiquement dédié à la gestion de corpus de TA, et en particulier à ceux du projet EOLSS/UNL-UnescoL de traduction de l'*Encyclopedia Of Life Support Systems* de l'anglais vers les autres langues de l'Unesco. Dans ce cas, un document ou « article » est fourni sous forme de deux fichiers, l'un

[75] Il en existe un à l'UPM (Madrid), et à Jakarta (Indonésie), et nous en avons fait plusieurs prototypes à Grenoble. Mais aucun n'est utilisable sur le Web, et en adapter un au Web dépasserait le cadre de cette thèse.

.aspx de type html et donnant la présentation, et l'autre, .unl, donnant le format unl-html et les graphes UNL.

1.2.6.b Présentation personnalisée

1.2.6.b.i Cacher ou montrer des « colonnes »

La visualisation d'un grand volume de données peut surcharger les utilisateurs. Il est donc important de donner aux utilisateurs de l'éditeur la possibilité de limiter la visualisation et l'application de fonctionnalités à un ensemble restreint de données, et cela en les gardant en mémoire (éviter un rechargement). Pour arriver à limiter ces données dans une interface à un moment donné, il faut permettre à l'utilisateur de définir des filtres qui dépendent des critères et des besoins à un moment donné.

```
function filter_BLEU(Min_BLEU){
   for(var i=0; i< mygrid.getRowsNum();i++){
    var eval_BLEU = parseFloat (grid_ instance.cells2(i,1).getValue().toInteger());
    if((eval_BLEU > Min_BLEU))
       instance_editor.setRowHidden(instance_editor.getRowId(i),false)
    else
       instance_ editor.setRowHidden(instance_editor.getRowId(i),true)
  }
}
```

Figure 67 : Exemple de filtrage de données avec masquage de ligne

Le script Javascript de l'exemple ci-dessus montre un filtre simple permettant de sélectionner les énoncés d'une expérience d'évaluation ayant des valeurs d'évaluation BLEU supérieures à un minimum donné (Min_BLEU). La fonction setRowHidden d'une instance de l'éditeur permet de masquer la ligne correspondant à un segment ayant l'ID instance_editor.getRowId(i) et dont la valeur BLEU est inférieure à Min_BLEU.

1.2.6.b.ii Montrer N états successifs d'une évolution Wiki

Les modifications se font sur des positions précises selon le sous-ensemble sélectionné. La complexité du traitement apparaît, lorsque, lors d'une visualisation, d'autres utilisateurs modifient le contenu, alors que le présent utilisateur, passant d'un énoncé à un autre dans l'ensemble, ne peut se rendre compte des modifications.

Le problème est que l'état de visualisation peut ne pas être fidèle.

Ce type de problème est important et nécessite d'être pris en compte. Une solution telle que la copie de données pour chaque session ne peut résoudre le problème, parce qu'elle dupliquerait le contenu et que la taille deviendrait incontrôlable.

La solution que nous adoptons est de ne garder que le sous-ensemble sur lequel on travaille (actuellement visualisé). On ne prend pas en compte les modifications faites sur les autres ensembles de données.

Dans ce cas, la visualisation se fait par des flux HTML, sous forme de tableaux visualisés successivement, et le sous-ensemble courant est fidèle au contenu sélectionné. Le passage à un autre sous-ensemble présentera une deuxième interface de données parallèle, sans garantie sur le contenu. L'utilisateur peut avoir des données modifiées, mais peut obtenir un état des modifications par recours aux techniques Wiki : il peut alors restituer le contenu ou accepter de continuer ses manipulations en gardant le même contenu.

1.2.6.b.iii Montrer plusieurs sorties de TA

Les traductions automatiques diffèrent d'un système à un autre, et selon les couples de langue. C'est pourquoi il est a priori utile de demander plusieurs « prétraductions » à plusieurs systèmes de TA, si l'on veut traduire, et bien sûr aussi si l'on veut comparer des systèmes de TA. L'utilisation de prétraductions en provenance de différents systèmes, dans notre cas, sera faite dans le cadre de deux tâches :

1. La multilinguisation de corpus (extension à d'autres langues) : les traductions automatiques sur un couple de langues sont différentes d'un système à un autre au niveau de la qualité ou de la grammaticalité de la langue cible. Par exemple, le système Systran produit de bonnes traductions lorsqu'il s'agit de certains couples (par exemple français-italien ou anglais-français). (Boitet, 2004) relate par exemple une expérience dans laquelle il a fait traduire tout le BTEC par Systran-5, puis a testé le résultat sur un sous-ensemble de 510 segments utilisé lors de la campagne IWSLT-04 (environ 13 pages standard). En post-éditant (sous Excel et sans aide de la MT ni de dictionnaires), il a mis 12 mn par page standard de 250 mots, tandis que 3 chercheurs et ingénieurs français mettaient chacun 59 mn en moyenne, dans le même environnement, mais sans disposer des résultats de TA.

2. L'évaluation de la TA : l'évaluation de la TA est souvent faite sur plusieurs systèmes ; la qualité des traductions est jugée grâce à des jugements humains et à des métriques automatiques. Ensuite, des corrélations entre les jugements humains et ces métriques (BLEU, NIST, mWER, etc.) sont aussi calculées (par exemple

corrélation de Pearson) pour en déduire la qualité finale. Par exemple, les campagnes d'évaluation telles que IWSLT ou NIST considèrent la participation d'une bonne dizaine de systèmes chaque année pour l'évaluation de systèmes ayant l'anglais comme langue cible (voir l'annexe C : participation à IWSLT 2006). D'autres se focalisent sur d'autres langues, comme la campagne CESTA qui est destinée à tester et évaluer des STA ayant le français comme langue cible.

Les différentes sorties des systèmes de traduction automatique seront gérées dans des instances liées à une tâche précise. Dans les deux cas, et pour chaque expérience de multilinguisation ou d'évaluation, une structure XML spécifique à chaque expérience sera créée.

En effet, c'est à partir de l'éditeur d'évaluation que les services externes Web sont appelés pour produire des traductions multicible en provenance des systèmes de traduction automatique.

1.2.6.c Modifications globales

Nous avons vu comment il est possible de gérer une masse de données avec la possibilité d'appel de services externes (TA, métriques, évaluations humaines, etc.) et de post-édition des prétraductions. La génération de ces dernières peut être effectuée sur un sous-ensemble en visualisation parallèle dans l'éditeur.

Cependant, nous sommes confrontés à un problème en relation avec la manipulation à grande échelle d'une masse de données. Nous nous posons alors la question suivante : quelle serait la performance du système/éditeur dans un environnement « collaboratif » face aux requêtes de modification lancées, d'un côté, sur les sous-ensembles en visualisation dans l'éditeur (chacun dans une session utilisateur), et de l'autre côté, sur toutes les données dans la BD centrale suivant une requête d'un utilisateur, ou d'un programme dans le cas d'une expérience ?

L'éditeur permet de charger une quantité de données allant jusqu'à 50 000 segments, sans compter les colonnes qui peuvent à leur tour multiplier le volume, mais la vitesse de chargement ne reflète pas la vitesse des opérations. Par exemple, l'appel à la TA libre sur le Web est parfois lourd et prend plusieurs dizaines de secondes. Cela ne dépend pas de notre environnement ni de notre éditeur, mais de la surcharge du service de TA libre sur le Web. Le temps d'attente dépendra donc de deux facteurs, interne et externe (interne : lié au temps de traitement de notre environnement ; externe : lié aux temps de réponse des services de TA libre).

Cependant, il semble nécessaire de permettre des modifications globales, par exemple pour remplacer « s.v.p » par « s'il vous plaît » dans au moins 20% des énoncés du BTEC traduit en français, puisqu'il s'agit de dialogues oraux. Nous avons donc cherché à identifier et à résoudre les problèmes en relation avec le lancement des opérations globales, telles que « rechercher/remplacer ».

1.2.6.c.i Vue de l'utilisateur

Les opérations globales évoquées ci-dessus doivent être exécutées sous le contrôle des utilisateurs, qui, chacun dans une session, voient les modifications s'effectuer tout en ayant la possibilité de les annuler. Ces opérations peuvent être de 3 sortes : recherche, recherche et remplacement contrôlé, recherche et remplacement automatique. À tout moment, au cours des manipulations globales, il est souhaitable de pouvoir récupérer les positions de modification globale et exécuter des opérations inverses. Ces opérations peuvent être temporaires (après la fin d'une session) ou persistantes : dans les sessions suivantes, les utilisateurs peuvent ou non annuler les opérations globales exécutées dans des sessions différentes.

Ces opérations ne montrent pas les modifications effectuées en arrière-plan sur l'ensemble des données parce que l'interface n'en ne montre qu'un sous-ensemble. L'utilisateur ne voit en effet que les données sur lesquelles il travaille dans une session donnée, les données non chargées dans l'interface sont inaccessibles. Cela rend difficile le suivi des modifications, surtout lorsqu'il s'agit de modifications des mêmes données modifiées en collaboration par plusieurs utilisateurs dans plusieurs sessions différentes.

Nous supposons que les modifications incrémentales sont gérées par les mécanismes Wiki où les différentes versions sont stockées par sessions et par utilisateurs. Il ne reste alors qu'à trouver des solutions pour exécuter les opérations globales et pouvoir les restituer à nouveau pour retrouver l'état initial.

Pour cela on considère deux cas :

1. les opérations limitées aux données de l'interface en cours,
2. les manipulations des données non visibles.

Dans les deux cas, les opérations sont lancées en collaboration en plusieurs sessions.

1.2.6.c.ii Performance

Des opérations telles que « rechercher/remplacer » sont des opérations de base facilement intégrables dans n'importe quel éditeur. Mais elles posent des problèmes de performance lorsqu'il s'agit de les appliquer sur une masse de données, dans des sessions différentes en mode collaboratif où les opérations et les requêtes sont lancées par des scripts côté client et des Servlets/Portlets côté serveur sur des données allant jusqu'à plusieurs centaines de millions de mots.

Il ne faut pas non plus oublier de mentionner les problèmes en relation avec les Wikis, où toutes ces opérations sont exécutées dans des sessions différentes (une par utilisateur) et où les versions modifiées sont enregistrées dans chaque session. Il est à noter que dans ce cas le système sera surchargé par la lourdeur des opérations « globales » et les flux de données transactionnels venant de différentes sessions.

8.2 Solutions aux problèmes liés à la taille et à la complexité

Le chargement et la disposition des données à un moment donné doivent être maintenus, que ce soit en mémoire ou sur disque. Lors du chargement, l'éditeur utilise des techniques de pagination et un chargement dynamique, ce qui augmente considérablement sa puissance : c'est pendant la navigation que le chargement s'effectue, progressivement à la demande de l'éditeur qui envoie des requêtes en arrière-plan au serveur pour récupérer à la volée les données restant à visualiser.[76]

8.3 Aspect générique pour les traitements externes

8.3.1 Interfaces d'évaluation « subjective »

Ces évaluations nécessitent une interface dans laquelle toutes les informations sont présentées (source ou/et référence) de façon que les évaluateurs humains émettent leurs jugements avec la méthode la plus correcte et la plus sûre.

Il y a des controverses sur la conception de cette interface. Par exemple, dans IWSLT, les énoncés candidats sont proposés par groupes de 5 énoncés (voir Figure 57 et Figure 58, pp.193).

[76] http://b.hatena.ne.jp/entry/1541312

Nous avons constaté lors de notre participation à IWSLT-06 que les évaluateurs du couple (EN-CN) effectuent des comparaisons de façon spontanée entre les traductions candidates (déjà annotées) et les énoncés en cours d'évaluation.

Pour cette raison, nous proposons à l'organisateur d'une évaluation « subjective » de paramétrer l'interface, de façon à produire soit l'interface usuelle (mauvaise à notre avis), soit une interface dans laquelle on ne juge qu'un système à la fois sur un segment donné.

8.3.2 Interfaces d'évaluation « objective »

A l'inverse l'évaluation subjective, l'objectif de l'évaluation objective traditionnelle est de lancer des métriques basées sur des méthodes statistiques pour évaluer automatiquement la qualité des prétraductions par comparaison de sous-chaînes avec les énoncés de référence. L'interface d'évaluation, dans ce cas, doit offrir la possibilité d'appeler ces métriques et de les gérer avec l'ensemble de données d'une expérience.

Conclusion

Nous avons montré dans ce chapitre comment nous avons étendu l'environnement BEYTrans à la gestion d'une masse de données, résolvant pour cela les problèmes et les difficultés exposés dans le chapitre précédent. L'interface de navigation a été améliorée de façon à permettre le chargement de plus de 50 000 énoncés à la fois, avec la possibilité de lancer des opérations à grande échelle sur l'ensemble des données. C'est grâce aux descripteurs et à la division des corpus en sous-ensembles au format TMX que cela a été rendu possible. Enfin, nous avons pu adapter notre éditeur à la visualisation parallèle (source, post-édition, plusieurs prétraductions, et calculs divers liés à l'évaluation) tout en gardant la possibilité d'édition et d'appel de différentes méthodes externes.

Notons que Blanchon et Boitet (Blanchon, *et al.*, 2004) proposent depuis longtemps d'ajouter des mesures objectives liées à la tâche, et en particulier le temps de post-édition, ou une distance entre la post-édition et la (pré)traduction à juger, si la tâche est de produire des traductions de qualité professionnelle. Cette idée a été adoptée par le gros projet GALE aux USA, qui a introduit la procédure d'évaluation dite « HER » (Human Error Rate) et HTER (Human Translation Error Rate) (Roukos, 2006).

BEYTrans intègre par construction la possibilité d'implémenter cela très facilement, puisqu'on peut post-éditer toute « suggestion de traduction » venant de la TA ou de la MT, et même des traductions a priori bonnes car humaines ! Il suffit d'ajouter une fonction de distance comme celle de Levenshtein et de visualiser (et stocker) les résultats.

Chapitre 9

Expérimentations

Introduction

L'extension de BEYTrans au traitement de grands corpus multilingues a donné lieu à quelques expérimentations que nous présentons dans ce chapitre. Dans un premier temps, nous présentons les corpus parallèles compilés dans notre environnement, et donnons quelques détails sur la méthodologie d'import, la conversion en codage unique et le format interne permettant l'unification du traitement et de la gestion interne. Dans un deuxième temps, les deux campagnes d'évaluation IWSLT (avec l'anglais en langue cible) et CESTA (avec le français en langue cible) seront introduites brièvement. Dans la dernière section, nous présentons les expériences que nous avons menées sur le corpus BTEC du consortium international C-STAR. En particulier, nous montrerons les avantages de notre environnement en ce qui concerne la création des énoncés de référence et l'extension vers d'autres langues grâce à la TA et à la post-édition collaborative à la Wiki.

9.1 Import de divers corpus parallèles bilingues et multilingues

La liste des corpus libres disponibles et téléchargeables en ligne est longue, mais il y a peu de corpus volumineux et fortement multilingues. La Table 14 présente la liste de corpus qui ont été récupérés et transformés pour servir à nos expériences.

Ces corpus ont été analysés séparément pour en comprendre la structure. Ces structures sont diverses et variées dans la complexité d'annotation et d'alignement. Par exemple, les entrées du corpus JRC-Aquis sont structurées et alignées en XML par des liens (par exemple <link type="1:2" xtargets="217;212 213"/>). Un lien a un type (par exemple, "1:2" pour un segment source correspondant à deux segments cibles) et les positions des segments reliés par ce lien, notées dans l'attribut « xtargets » (voir aussi : cf. *Structure logique envisagée*, p. 194).

Pour avoir une gestion homogène des énoncés, il nous a donc fallu récupérer les positions des indices et réaligner les énoncés pour obtenir une structure de MT en format TMX. Ils ont alors pu être traités correctement dans BTedit.

Corpus	Taille
Acquis corpus (JRC)	6 300 000 mots
Corpus BAF	400 000 mots
Corpus CARMEL	10 000 000 mots
CRATER Multilingual Aligned Annotated Corpus	1 000 000 mots
the IJS ELAN (Slovene-English Aligned Corpus)	1 000 000 mots
OPUS (open source parallel corpus)	30 000 000 mots (en 60 langues)
Tanaka corpus	180 000 énoncés
Swedish political texts corpus (Uppsala Universitet)	11 000 mots

Table 15 : Liste de corpus libres[77]

L'analyse et le prétraitement préliminaire sont faits par des programmes (écrits en Java) que nous avons développés séparément pour chaque corpus et qui ont aussi été utilisés pour unifier le codage en UTF-8 et importer les données en TMX.

Des corpus présentés en texte brut et séparés dans plusieurs fichiers (par exemple BTEC) ont été transformés en documents TMX monolingues et ensuite en versions multilingues.

Les versions monolingues ne contiennent qu'une seule occurrence *tuv* enveloppée dans un élément *tu*. En revanche, les versions multilingues contiennent plusieurs *tuv* (une par langue). De plus, les programmes développés ont permis la division des corpus sur plusieurs documents de taille différente.

Enfin, les documents produits sont injectés par un script dans la BD de BEYTrans[78], et deviennent instantanément disponibles et consultables sur la Toile.

Lors de la phase de prétraitement, des descripteurs contenant des informations telles que le nombre d'énoncés, les dates de création et de mise à jour, l'auteur, les droits d'accès, des informations sur le corpus origine, et les langues, sont associés aux documents créés.

La navigation et la visualisation parallèle passent par la sélection des descripteurs en fonction des requêtes des utilisateurs. Le résultat d'une requête est un ensemble d'énoncés en provenance d'un ou de plusieurs descripteurs structurés dans un document XML directement interprétable par BTedit. Le format de cette structure est présenté dans l'exemple suivant (exemple pris du BTEC).

[77] Voir aussi l'annexe C pour plus d'information.
[78] Tous nos scripts ont été développés en Velocity ou en Groovy (deux langages de scriptage imposés par XWiki).

```
<rows>
   <row id="00001">
      <cell> Hamburger and stew on the right side and salad, please.</cell>
      <cell>ハンバーガーと右側のシチューにサラダをください。</cell>
      <cell >请来个汉堡，右边的那个炖菜和色拉。</cell>
      <cell >햄버거랑 오른쪽에 있는 스튜랑 샐러드로 할게요..</cell>
      <cell > Hamburger e stufato dalla parte destra e insalata, per favore.</cell>
   </row>
   <row>
      …
   </row>
      …
</rows>
```

Figure 68 : Structure XML interprétable par BTedit

Notons que si le mode « pagination » (buffering) est actif, alors les énoncés qui ne tiennent pas sur une interface sont chargés page par page lors de la navigation, ce qui réduit le temps de chargement et l'attente des données du serveur.

9.2 Visualisation et édition de corpus

L'adaptation du mode de navigation et de visualisation a été déjà abordée dans le chapitre 7, où nous avons montré l'intérêt d'une présentation parallèle. Par exemple, dans le cas du projet DEMGOL, cette présentation nous a permis de récupérer les UT et de les visualiser lors de la traduction des notices.

D'autre part, certaines fonctionnalités sont déclenchées par des événements lors de la navigation (par exemple les suggestions du dictionnaire et de la MT, l'appel à la TA, etc.). Dans le cas des corpus, nous avons adopté une représentation parallèle analogue pour manipuler et visualiser les corpus convenablement. Ce que nous avons présenté dans la Figure 61 (cf. *Visualisation et navigation*, p. 203) est concrétisé dans la Figure 71 (cf. *Interface d'une expérience*, p. 232) où les énoncés sont visualisés en parallèle avec la possibilité d'édition et d'appel de fonctionnalités diverses.

9.2.1 Implémentation

La synchronisation des énoncés visualisés dans l'éditeur est l'image de la synchronisation interne des énoncés. Cette synchronisation est conservée pendant la transformation des données vers le format XML de l'éditeur par des lignes annotées en XML par *row* contenant plusieurs *cell* dont les valeurs sont les énoncés (voir exemple ci-dessus). L'éditeur permettant la visualisation, la navigation et l'édition de corpus multilingues est illustré dans la Figure 69.

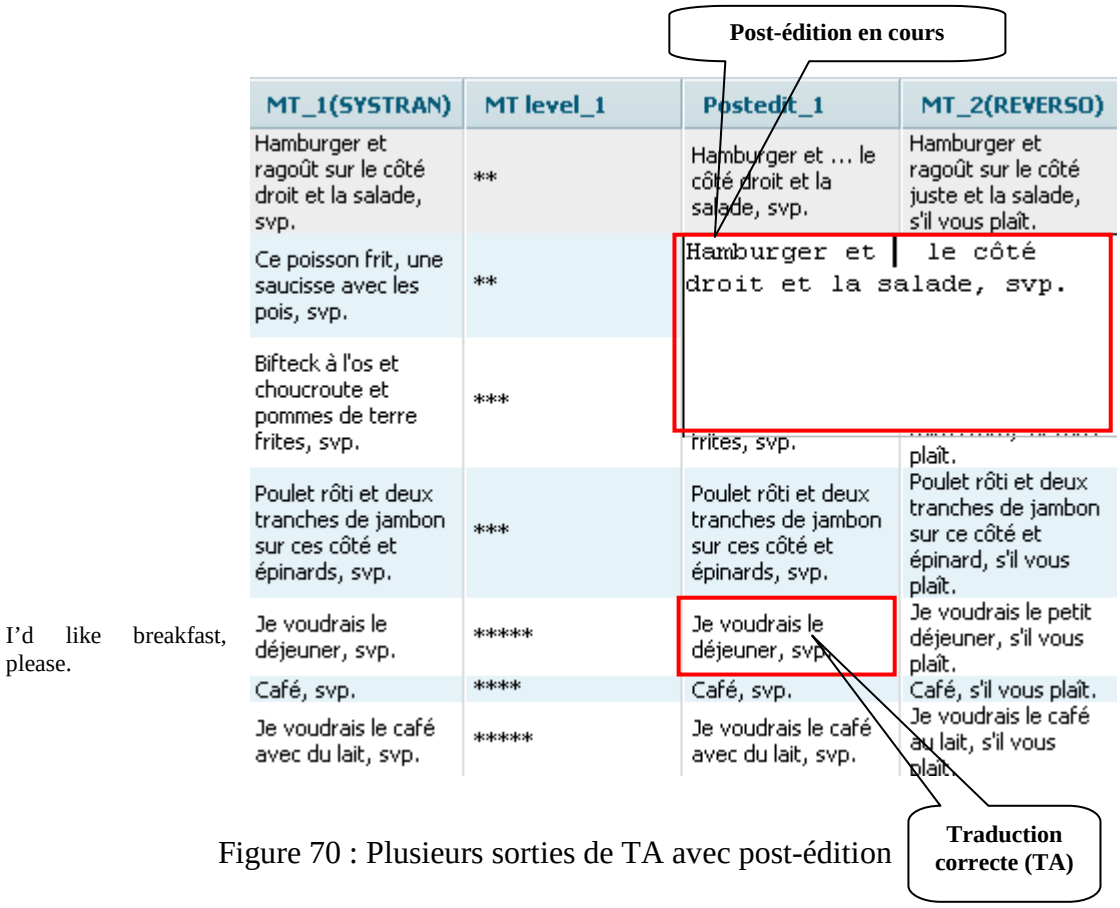
Figure 69 : Navigation et visualisation parallèle dans un corpus multilingue

L'éditeur a été implémenté dans BEYTrans suivant l'architecture et le principe de fonctionnement des Wiki, c'est-à-dire que les modifications sont faites en mode collaboratif et incrémental. Toutes les versions générées par les bénévoles sont enregistrées et restaurées à la demande. Cela permet d'avoir plus d'efficacité et de souplesse, car les données modifiées ne sont jamais perdues et la traçabilité est aussi maintenue.

9.2.2 Appel à la TA et post-édition

L'appel à la TA se fait dans l'éditeur à la demande, énoncé par énoncé, ou sur tout l'ensemble des énoncés en cours de visualisation.

Les utilisateurs peuvent post-éditer les prétraductions pour en tirer de nouvelles références ou pour effectuer une évaluation objective liée à la tâche (calcul de distance d'édition, temps de post-édition, etc.).



I'd like breakfast, please.

Figure 70 : Plusieurs sorties de TA avec post-édition

Il est possible d'avoir plusieurs suggestions ou prétraductions d'un ou de plusieurs systèmes de traduction automatique. Par exploitation de la TA Web libre, la Figure 70 montre les traductions produites par Systran et Reverso. L'énoncé « *I'd like breakfast, please.* » est traduit par « *Je voudrais le déjeuner, svp.* » par Systran, et par « *Je voudrais le petit déjeuner, s'il vous plaît.* » par Reverso ce qui est plus correct. Dans un cas de figure comme celui-là, en post-édition, la distance d'édition est 0, mais le temps de fabrication de la nouvelle référence est 1, car le champ « post-édition » a été initialisé à la traduction Systran : il faut une opération de copier-coller pour le remplacer par la traduction Reverso.

9.3 Les campagnes d'évaluation de la TA

Les campagnes d'évaluation de la TA qui nous intéressent sont celles du projet CESTA qui a pour langue cible le français, et celle de l'atelier (workshop) annuel IWSLT, qui a pour langue cible l'anglais.

9.3.1 CESTA

Le projet CESTA a proposé une série de campagnes d'évaluation de systèmes de traduction automatique pour diverses paires de langues allant vers le français. Ce projet, qui a débuté en janvier 2003, est financé par le Ministère de la Recherche français dans le cadre du programme Technolangue.

La campagne CESTA constitue une avancée par rapport aux campagnes d'évaluation de la traduction automatique, jusqu'ici menées par le NIST (aux Etats-Unis) en se basant sur les métriques d'évaluation statistique NIST/BLEU (IBM). En effet, à ces mêmes métriques ont été ajoutées d'autres méthodes plus fiables basées sur les notions de score grammatical et de score sémantique. C'est au cours d'une phase de méta-évaluation par rapport à l'évaluation de juges humains que ces métriques ont été elles-mêmes comparées entre elles et évaluées.

9.3.2 IWSLT

La campagne d'évaluation IWSLT-06 à laquelle nous nous référons ici a été réalisée en se basant sur un corpus de parole multilingue. Ce corpus contient des énoncés dans le domaine touristique similaires à ceux trouvés souvent dans les livres de phrases destinés aux touristes voyageant dans d'autres pays. Les détails concernant le Basic Travel Expression Corpus (BTEC), l'ensemble des différentes conditions pour chaque groupe de données et les spécifications de participation dépendent de chaque campagne d'évaluation. Chaque année, les organisateurs proposent une évaluation concernant des phénomènes précis concernant l'impact de la traduction sur la qualité de traduction.

Notre laboratoire a été impliqué officiellement dans IWSLT-06 pour expérimenter des traducteurs commerciaux, Systran pour le couple italien-anglais, japonais-anglais, et ATLAS-II pour le couple japonais-anglais.

9.4 Application à IWSLT-06

La campagne est gérée par les organisateurs du workshop annuel du consortium international C-STAR, et porte sur un sujet en relation avec la traduction d'énoncés apparaissant dans des dialogues spontanés finalisés. En 2006, les phénomènes liés aux dialogues oraux tels que les hésitations... ont été l'objet d'évaluation des systèmes de TA des participants. Ne travaillant pas sur des couples de langues ayant l'anglais comme cible, nous avons participé en mettant en œuvre deux systèmes, ATLAS-II et Systran. Nous étions les seuls à participer avec des systèmes commerciaux.

Nous avons eu quelques difficultés à reconstruire les résultats pour faire des évaluations internes et les comparer avec celles publiées par les organisateurs d'IWSLT. Notre objectif était en effet de produire de nouvelles références par post-édition des prétraductions et d'appliquer les mêmes métriques, et de recalculer de nouveaux scores selon le protocole proposé par les organisateurs.

Le but était de prouver (encore une fois) l'inadéquation des mesures BLEU, NIST, qui utilisent un petit nombre de traductions de référence pour chaque segment (alors que l'ensemble des « bonnes traductions » d'un énoncé, même de 6 à 10 mots, peut contenir des centaines ou des milliers d'énoncés, assez « loin » les uns des autres pour les mesures basées sur les n-grammes de mots).

Nous avons constaté les besoins suivants :

- Environnement d'évaluation pour lancer les expériences.
- Construction de nouvelles références.
- Paramétrage de la méthode d'évaluation subjective.
- Division des tâches de façon collaborative sur plusieurs évaluateurs humains (pas nécessairement au laboratoire) en déplacement ou depuis chez eux.

Tous ces besoins ont été pris en compte pour l'adaptation de BEYTrans aux évaluations internes.

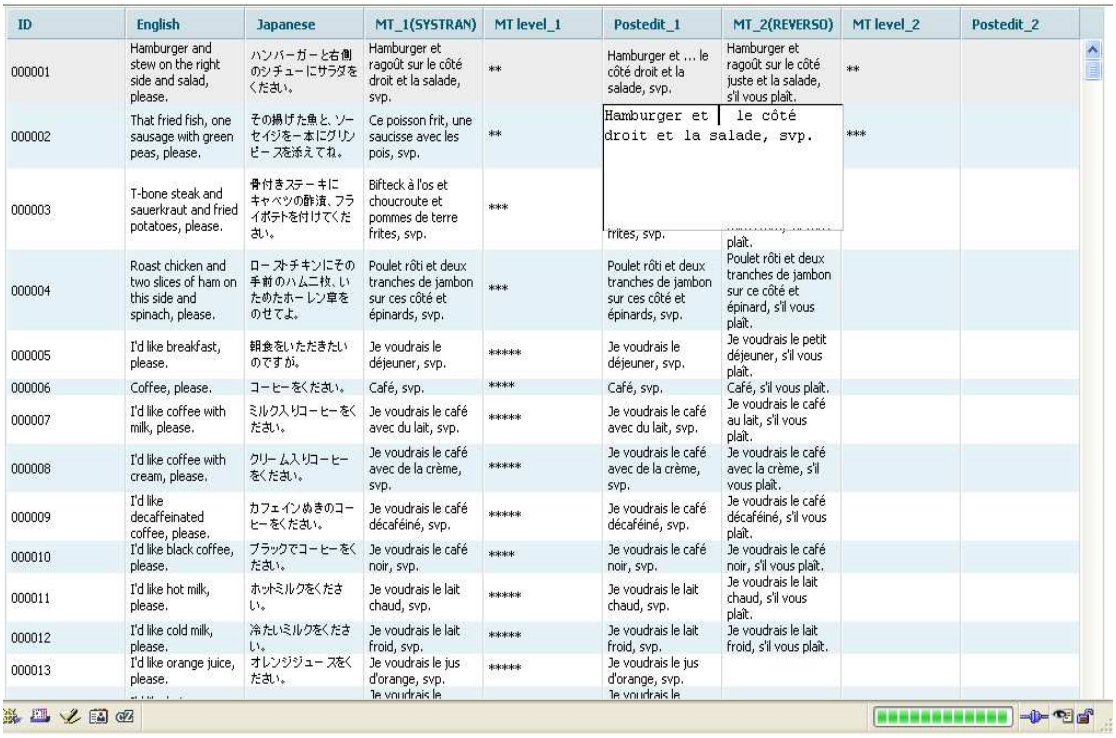
9.4.1 Instance d'une expérience d'évaluation

Dans une évaluation, on a besoin des énoncés de référence, des énoncés sources et des traductions candidates. La configuration des paramètres d'une expérience se fait avant son lancement : on choisit les données à impliquer, les métriques à calculer, et les systèmes de la TA à appeler. Cette configuration permet de générer les colonnes et les lignes de l'interface et d'appeler la TA.

Par exemple, une expérience peut être configurée pour contenir dans l'interface une colonne source, une colonne de référence, des prétraductions (N prétraductions), des colonnes pour la post-édition et enfin des colonnes pour les métriques NIST et BLEU.

9.4.2 Interface d'une expérience

Une fois une expérience créée, l'interface de présentation parallèle des énoncés est construite automatiquement. Cette interface n'est pas figée, on peut la modifier à sa guise par des ajouts de nouvelles colonnes ou lignes.



ID	English	Japanese	MT_1(SYSTRAN)	MT level_1	Postedit_1	MT_2(REVERSO)	MT level_2	Postedit_2
000001	Hamburger and stew on the right side and salad, please.	ハンバーガーと右側のシチューにサラダをください。	Hamburger et ragoût sur le côté droit et la salade, svp.	**	Hamburger et ... le côté droit et la salade, svp.	Hamburger et ragoût sur le côté juste et la salade, s'il vous plaît.	**	
000002	That fried fish, one sausage with green peas, please.	その揚げた魚と、ソーセイジを一本にグリンピースを添えてね。	Ce poisson frit, une saucisse avec les pois, svp.	**	Hamburger et le côté droit et la salade, svp.		***	
000003	T-bone steak and sauerkraut and fried potatoes, please.	骨付きステーキにキャベツの酢漬、フライポテトを付けてください。	Bifteck à l'os et choucroute et pommes de terre frites, svp.	***	frites, svp.	plaît.		
000004	Roast chicken and two slices of ham on this side and spinach, please.	ローストチキンにその手前のハムニ枚、いたのホーレン草をのせてよ。	Poulet rôti et deux tranches de jambon sur ces côté et épinards, svp.	***	Poulet rôti et deux tranches de jambon sur ces côté et épinards, svp.	Poulet rôti et deux tranches de jambon sur ce côté et épinard, s'il vous plaît.		
000005	I'd like breakfast, please.	朝食をいただきたいのですが。	Je voudrais le déjeuner, svp.	*****	Je voudrais le déjeuner, svp.	Je voudrais le petit déjeuner, s'il vous plaît.		
000006	Coffee, please.	コーヒーをください。	Café, svp.	****	Café, svp.	Café, s'il vous plaît.		
000007	I'd like coffee with milk, please.	ミルク入りコーヒーをください。	Je voudrais le café avec du lait, svp.	*****	Je voudrais le café avec du lait, svp.	Je voudrais le café au lait, s'il vous plaît.		
000008	I'd like coffee with cream, please.	クリーム入りコーヒーをください。	Je voudrais le café avec de la crème, svp.	*****	Je voudrais le café avec de la crème, svp.	Je voudrais le café avec la crème, s'il vous plaît.		
000009	I'd like decaffeinated coffee, please.	カフェインぬきのコーヒーをください。	Je voudrais le café décaféiné, svp.	*****	Je voudrais le café décaféiné, svp.	Je voudrais le café décaféiné, s'il vous plaît.		
000010	I'd like black coffee, please.	ブラックでコーヒーをください。	Je voudrais le café noir, svp.	****	Je voudrais le café noir, svp.	Je voudrais le café noir, s'il vous plaît.		
000011	I'd like hot milk, please.	ホットミルクをください。	Je voudrais le lait chaud, svp.	*****	Je voudrais le lait chaud, svp.	Je voudrais le lait chaud, s'il vous plaît.		
000012	I'd like cold milk, please.	冷たいミルクをください。	Je voudrais le lait froid, svp.	*****	Je voudrais le lait froid, svp.	Je voudrais le lait froid, s'il vous plaît.		
000013	I'd like orange juice, please.	オレンジジュースをください。	Je voudrais le jus d'orange, svp.	*****	Je voudrais le jus d'orange, svp.			

Figure 71 : Instance de l'éditeur d'évaluation

Des opérations de manipulation globale comme les tris et les « chercher/remplacer » sont appelées selon les besoins. On peut trier les résultats des métriques BLEU ou lancer des remplacements sur des segments (nécessaire pour la post-édition).

Nous illustrons cela par l'exemple présenté dans la Figure 71 où on voulait lancer les métriques statistiques pour l'évaluation automatique du couple anglais-japonais. La colonne source ainsi que la référence, et leurs identificateurs, sont présents, et chaque ligne présente un énoncé. Deux systèmes de TA sont appelés pour produire des prétraductions en anglais.

9.4.3 Résultats d'expérience en IWSLT-06

L'éditeur nous a permis de détecter des erreurs dans les références employées par IWSLT-06, et ainsi de montrer les limitations de la méthode d'évaluation subjective proposée par les organisateurs.

Problèmes de qualité des segments sources : la Table 16 illustre ces problèmes sur la langue japonaise. Ces erreurs sont passées inaperçues parce que les données du BTEC n'ont jamais été présentées sous un format parallèle, alors que notre éditeur montre ces problèmes facilement.

Source (japonais)	ATLAS translation	Correct
トイレ は 機内(kinai) **高校 (koukou=high school)** です. ご 案内 (annai=guidance) 致 (itta)し ます.	It will be a guide of the rest room that an in-flight high school has (*O).	**後方 (backward)**
はい クレジットカード を ご 利用 頂け ます し **帰る** カード は ビザ マスター アメリカンエクスプレス です.	The yes credit card can be had to be used and the card where (*S) returns is visa	**使える()**

Table 16 : Une fausse conversion Kana-Kanji

Problèmes d'évaluation humaine en « adéquation » : du point de vue de l'évaluateur humain, la décision d'évaluation doit être basée sur le degré où une personne peut comprendre le sens d'origine en utilisant la traduction anglaise. Ce qui suit concerne le chinois en langue source (car une de nos étudiantes en thèse était chinoise et a accepté de participer à l'évaluation). Durant le processus d'évaluation, nous avons essayé de ne pas donner d'importance à la forme de sortie : même si elle est mauvaise, on donne un bon score d'adéquation.

Concernant l'interface d'évaluation subjective, nous avons alerté les organisateurs durant nos premières rencontres avec eux, car nous étions contre la présentation de plusieurs traductions (par plusieurs systèmes de TA) d'une même entrée dans l'interface.

En effet, nous avions prévu la tendance spontanée des évaluateurs humains à faire des comparaisons inutiles entre les différentes sorties pour les trier, ce qui fait perdre beaucoup de temps.

Nous avons argumenté (Boitet, *et al.,* 2006) que, si la vérification d'une sortie a comme coût une unité μ, si une comparaison coûte ν, et si on dispose de 20 sorties, alors environ $20 \log_2 20 \approx 100$ comparaisons sont nécessaires pour trier lesdites 20 sorties présentées dans un écran de l'interface. Le temps T d'évaluation de 20 énoncés passe alors de $T = 20\mu$ à $T' = 20\mu + 100\nu$. Comme on a probablement $1{,}5\mu \le \nu \le 2\mu$, on a $170\mu \le T' \le 200\mu$, d'où une multiplication du temps par 8 ou par 10 (si on fait toutes les comparaisons) et en pratique par 5 à 6 (mesure effectuée sur le temps mis par notre évaluatrice Wenjie CAO).

Notre expérience confirme cela. Elle ne prend que 3 minutes pour vérifier et évaluer les segments d'un seul « paquet », sans classification. Dans ce cas, $\mu \approx 9$ secondes et une comparaison prenait environ v = 20 secondes. Le temps aurait été augmenté de 12 fois si les évaluateurs (des autres participants, utilisant l'interface « groupée ») avaient essayé d'établir une classification complète. Dans la pratique, l'évaluation « groupée » (avec tentative de classement) a pris en moyenne de 20 à 40 minutes. Nous en déduisons qu'en moyenne un évaluateur ne fait pas toutes les comparaisons, en particulier, les comparaisons entre les énoncés placés en haut et ceux placés en bas.

Nous avons donc conseillé à nos évaluateurs de ne plus faire de comparaisons entre les différentes sorties en anglais. Notre principal évaluateur pour le chinois-anglais a évalué suivant notre méthode environ 5.400 énoncés en une journée et demi (270 interfaces ; 13,5 heures ; μ =3 sec.). Or, l'estimation des organisateurs (basée sur l'interface d'évaluation d'IWSLT-05) était de 4 à 5 jours.

Nous avons donc prouvé l'intérêt de notre méthode d'évaluation subjective, par rapport à celle proposée par les organisateurs. Il nous reste à montrer que les évaluations subjectives combinées avec les métriques automatiques sont valides.

On a constaté en effet plusieurs erreurs dans les énoncés source, ainsi notre intuition nous a orienté vers la construction de nouvelles références pour voir si les évaluations de nos systèmes dans cette campagne étaient valides ou non.

9.4.4 Nouvelle interface : nouvelles références par post-édition

Les besoins révélés dans les sous-sections précédentes nous ont poussé à transformer notre éditeur pour la génération de nouvelles références et à appliquer les métriques suivant le protocole exigé par les organisateurs, mais en produisant cette fois-ci de nouvelles références par post-édition des prétraductions produites par la TA, et en refaisant ensuite les mêmes expériences pour obtenir de nouveaux scores, et montrer si possible que l'insertion de références obtenues par post-édition donne des scores BLEU et NIST plus en accord avec les scores des évaluations subjectives que les scores BLEU et NIST basés sur des références n'ayant aucun rapport avec le « style » et le vocabulaire des sorties de TA.

La Figure 72 montre l'interface de BTedit. Les colonnes contiennent des données d'une tâche de traduction comportant les prétraductions de deux systèmes de traduction automatique, Systran et Reverso. Les deux colonnes de post-édition permettent de corriger les prétraductions produites par les deux systèmes. Les calculs BLEU et NIST sont appliqués sur

la colonne source, cible, et sur les prétraductions produites par Systran. Ces calculs peuvent aussi être paramétrés de façon à produire un ou plusieurs calculs, BLEU ou NIST ou les deux à la fois. Les scores obtenus par ces métriques sont présentés dans la partie haute de la figure.

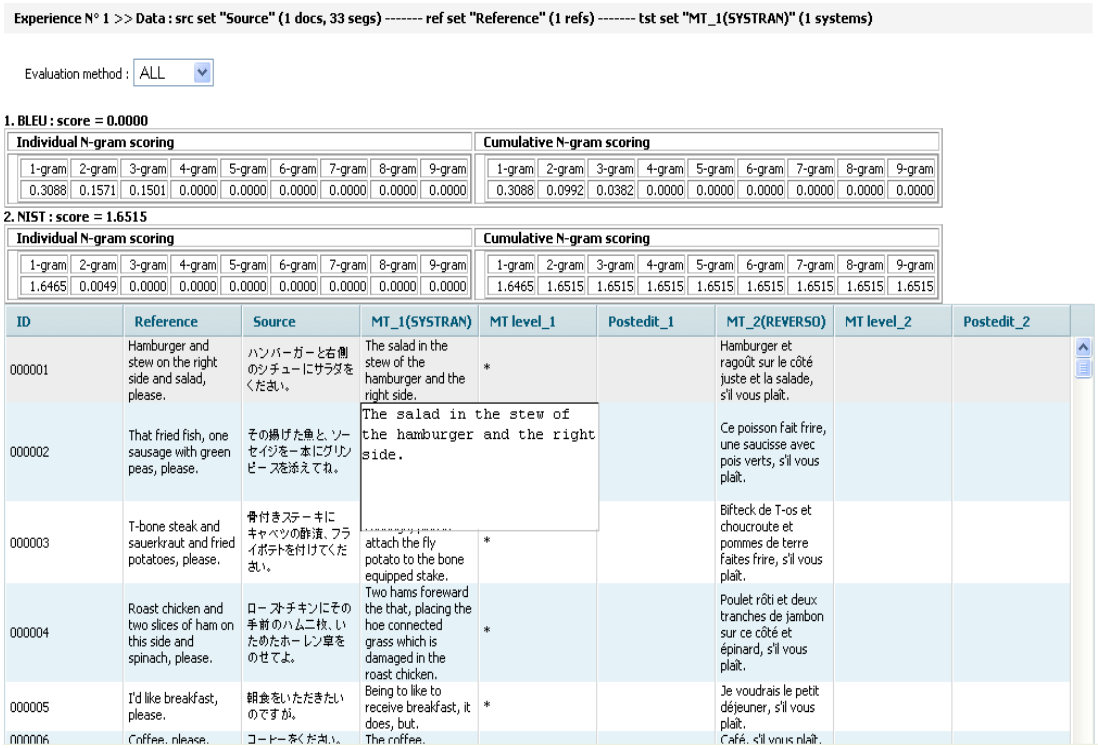


Figure 72 : Interface d'évaluation objective

Les prétraductions sont post-éditées et les bonnes traductions obtenues peuvent servir de nouvelles références. On peut maintenant calculer des scores NIST et BLEU pour avoir des scores en prenant comme références :

1. les références initiales (1,2,..., jusqu'à 16 par segment).
2. les traductions obtenues par post-édition.
3. l'union des deux.

Cette interface nous permettra dans le futur d'effectuer nos évaluations en interne, sans passer par les sites d'organisateurs, et de comparer les classements des systèmes avant et après l'ajout de références obtenues par post-édition des résultats de TA. Idéalement, il faut d'ailleurs ajouter de telles références non seulement pour les systèmes commerciaux, construits avec des méthodes « expertes » (dictionnaires, automates, grammaires), mais aussi pour les systèmes statistiques « réglés » pour « imiter » les références.

Conclusion

L'évaluation des systèmes de traduction automatique a été présentée dans ce chapitre suivant le protocole proposé par les organisateurs des campagnes d'évaluation. Nos expériences en 2004 et 2006 nous ont montré que les classements fournis par les évaluations « objectives » sont souvent en désaccord avec les évaluations « subjectives ». Ainsi, Systran, avec lequel nous avons participé 2 fois à IWSLT, était presque toujours dernier en évaluation objective et souvent premier en évaluation « subjective ».

L'adaptation de BEYTrans aux corpus pour la TAO nous permet non seulement de présenter les corpus et leurs évaluations de façon « sensible », et pas seulement par des tableaux de chiffres, mais aussi d'expérimenter et de valider une nouvelle méthode « objective » d'évaluation, fondée sur l'insertion de références obtenues par post-édition des résultats de TA.

Conclusion générale

Les traductions disséminées par les bénévoles sont variées et faites dans plusieurs domaines (littérature, santé, informatique, etc.). La plupart des sites que nous avons étudiés mettent à la disposition des internautes des traductions variées de plus ou moins de bonne qualité, et en grande quantité (exemple W3C, Arabeyes, etc.).

Cette thèse a été consacrée à l'identification et à la résolution des problèmes informatiques qui apparaissent quand on veut permettre aux traducteurs bénévoles d'avoir plus de productivité et d'efficacité durant le processus de traduction.

Trois principaux axes de recherche ont été identifiés, correspondant à trois problèmes intéressants :

1. la réalisation d'un environnement collaboratif complet pour l'aide à la traduction incrémentale en ligne,

2. le développement d'un module de recyclage multilingue de traductions existantes sur le Web,

3. la gestion de masses de données, en particulier de très gros corpus parallèles destinés à la TA.

Environnement collaboratif libre d'aide à la traduction collaborative

Dans un premier temps, nous avons participé à la réalisation de l'environnement QRLex et au développement de l'éditeur QRedit. Ce faisant, nous avons proposé une solution pour la gestion de données hétérogènes par la définition de structures XML *ad hoc* qui nous ont permis d'importer plus de 1,7M d'entrées. Cependant, les limitations de QRLex nous ont mené à étendre nos recherches vers des solutions plus efficaces.

Nous avons conçu un nouvel environnement, BEYTrans, qui permet à la fois une grande couverture de langues et une diversité de pratiques, et l'avons développé en suivant l'architecture Wiki. Il permet la traduction incrémentale et collaborative, tout en offrant des aides linguistiques et la gestion de documents multilingues en ligne. Son éditeur multilingue, BTedit, est plus adapté que QRedit, et offre des fonctionnalités traductionnelles avancées (suggestions proactives par les MT et dictionnaires, par TA, etc.). Pour montrer l'utilisabilité et l'utilité de cet environnement, nous l'avons expérimenté sur la traduction d'une centaine de notices du projet DEMGOL. Plusieurs retours et échanges ont été entrepris avec la traductrice

de ce projet. Cela nous a permis d'améliorer l'environnement et de le tester dans une situation concrète.

Les résultats obtenus par cette expérience sont encourageants, bien que les ressources linguistiques (MT, Dictionnaires, etc.) aient été vides au départ. Nous avons ensuite amélioré notre système et défini un protocole d'évaluation susceptible de démontrer clairement les avantages de la post-édition collaborative de résultats de TA dans un environnement Wiki. Une expérience préliminaire sur la localisation de messages de Tikiwiki a donné des résultats très prometeuses : avec seulement 3 contributeurs en parallèle, le délai a été divisé par un peu plus de 3, et le temps humain total par un peu plus de 2, prouvant que l'aide par la TA est plus efficace que l'aide par MT (mémoire de traductions).

Outil de recyclage de données traductionnelles

Le développement de l'outil QRselect (dans QRLex) nous a permis de montrer la faisabilité et l'intérêt du recyclage de traductions existantes pour l'aide aux traducteurs bénévoles. Nous avons développé cet outil en deux étapes. Dans un premier temps, nous avons limité le recyclage de documents aux paires anglais-japonais, car la communauté visée était celle des traducteurs bénévoles japonais.

Comme BEYTrans supporte plusieurs langues, nous avons transformé QRselect pour qu'il supporte plusieurs langues. Cela permet aux traducteurs de différentes communautés de faire le recyclage dans plusieurs couples de langues. Pour montrer l'utilité du recyclage, nous l'avons expérimenté sur la détection de paires de documents en anglais et japonais. Les résultats sont à présent satisfaisants. De plus, cette expérimentation nous a permis d'évaluer l'outil et d'identifier d'autres problèmes que nous souhaitons aborder dans nos futures recherches.

Le traitement de masse de données multilingues destinée à la TA

Enfin, dans la troisième partie, nous avons étendu le problème de la traduction incrémentale et collaborative à la gestion de corpus parallèles volumineux destinés à la TA. Nous avons montré quels sont les problèmes posés par la gestion efficace de telles données, puis nous avons proposé des solutions pour chacun d'eux. Ces problèmes sont liés à la taille et la complexité structurelle des données. Cela nous a permis d'implémenter de façon satisfaisante la gestion collaborative de gros corpus et d'introduire l'utilisation de la traduction incrémentale pour réduire le coût de la multilinguisation de corpus, tout en profitant des fonctionnalités développées précédemment (MT, TA, dictionnaires, etc.).

Nous pensons que l'exploitation des Wikis, en particulier XWiki et Tikiwiki, pourrait dans le futur aider à la fois à construire de nouveaux corpus à « coût minimal » et à disposer de ressources précieuses pour le TAL.

Des expérimentations ont été faites pour montrer l'utilité de l'environnement dans les tâches d'évaluation de systèmes de traduction automatique. Les protocoles proposés par les campagnes d'évaluation (par exemple IWSLT) ont été améliorés. Nous avons démontré qu'une présentation « non groupée » des segments à tester permet de diviser en pratique le temps d'une évaluation subjective des résultats de 20 systèmes de TA par 5 ou 6.

Nous avons aussi proposé une méthode qui, nous l'espérons, permettra de montrer que l'évaluation objective par mesure du temps de post-édition pourrait mener à la production de classements par des scores BLEU et NIST bien plus en accord avec les classements subjectifs et avec ceux donnés par la post-édition. Il suffit pour cela d'ajouter aux « références » les (bonnes) traductions obtenues par post-édition des traductions automatiques. Nous n'avons pas pu le faire nous-mêmes car nous ne disposions pas des références utilisées lors de ces campagnes.

Travaux futurs et perspectives

Les méthodes de traduction incrémentale et collaborative basées sur la technologie Wiki et la contribution des bénévoles peuvent être utilisées dans plusieurs applications en traduction automatique ou dans les outils d'aide à la traduction.

Une première application intéressante serait d'intégrer notre environnement dans le processus de localisation des logiciels et systèmes libres (KDE, Gnome, XWiki etc.). Nous pourrions en effet importer les segments à traduire de tous ces logiciels/systèmes et améliorer notre environnement par d'autres fonctionnalités telles que le support de nouveaux formats de documents et de ressources. Cette nouvelle application de notre environnement a été déjà introduite dans le nouveau protocole exposé à la fin de la partie II.

Enfin, nous constatons que les Wiki actuels n'intègrent pas de module favorisant la multilinguisation de leur contenu. Récemment, d'autres projets commencent à apparaître pour répondre à cette question, comme le projet « Cross-Lingual Wiki Engine Project » (Désilets, *et al.*, 2006), mais ils n'en sont qu'à leurs débuts, et, à notre connaissance, les Wiki actuels n'offrent pas cette possibilité. Le développement d'un module générique pour multilingualiser le contenu des Wiki actuels, et ceux du futur, permettrait sans doute de faciliter la production de documents multilingues et leur dissémination.

Bibliographie

(Abekawa, 2007) Abekawa T. and Kageura K. (2007) *A Translation Aid System with a Stratified Lookup Interface*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007) Demos and Posters.

(Agirre, *et al.*, 2000) Agirre E., Arregi X., Artola X., Diaz De Ilarraza A., Sarasola K. and Soroa A. (2000) *A Methodology for Building Translator-oriented Dictionary Systems*. Machine Translation. Vol. 15, pp. 295-310.

(Aït-Mokhtar, *et al.*, 2003) Aït-Mokhtar S., Hagège C. and Sándor Á. (2003) *Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques*. In Proceedings of the TALN-2003 Conference. Batz-sur-Mer, France, pp. 53-66.

(Al-Adhaileh, 1999) Al-Adhaileh M. H. and Tang E. K. (1999) *Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema*. In Proceedings of the Machine Translation Summit VII. Singapore, pp. 244-249.

(Augar, *et al.*, 2004) Augar N., Raitman R. and Zhou W. (2004) *Teaching and Learning Online with Wikis*. In Proceedings of the 21st Australasian Society of Computers In Learning In Tertiary Education Conference, ASCILITE-04. Australia, pp. 95-104.

(Awdé, 2003) Awdé A. (2003) *Comparaison de deux techniques de décodage pour la traduction probabiliste*. Master en informatique. Université de Montréal. Département d'Informatique et de Recherche Opérationnelle. Faculté des arts et des sciences, 99 p.

(Ball, 2003) Ball S. (2003) *Joined-up Terminology - The IATE system enters production*. In Proceedings of the 25th International Conference on Translating and the Computer. London, UK, Vol. 5, 5 p.

(Barbu, 2005) Barbu E. and Mititelu V. B. (2005) *Automatic Building of Wordnets*. In Proceedings of the International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria, pp. 99-106.

(Berment, 2004) Berment V. (2004) *Méthodes pour informatiser les langues et les groupes de langues peu dotées*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble, 270 p.

(Berment, 2005) Berment V. (2005) *Online Translation Services for the Lao Language*. In Proceedings of the First International Conference on Lao Studies. De Kalb, Illinois, USA, pp. 1-11.

(Bey, *et al.*, 2006) Bey Y., Boitet C. and Kageura K. (2006) *The TRANSBey Prototype: An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators*. In Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III), Yuste, E., (Ed.). LREC 2006 - Fifth International Conference on Language Resources and Evaluation. Paris: ELRA / ELDA (European Language Resources Association, European Language Resources Distribution Association). Magazzini del Cotone Conference Centre, Genoa, Italy, pp. 49-54.

(Bey, *et al.*, 2005) Bey Y., Kageura K. and Boitet C. (2005) *A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex*. In Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC19). Taipei, Taiwan, pp. 51-60.

(Bey, *et al.*, 2006) Bey Y., Kageura K. and Boitet C. (2006) *Data Management in QRLex, an Online Aid System for Volunteer Translators*. International Journal of Computational Linguistics and Chinese Language Processing. Vol. 11, pp. 349-376.

(Blanchon, 1991) Blanchon H. (1991) *Problèmes de désambiguïsation interactive en TAO personnelle*. Actes du colloque "l'environnement traductionnel ; la station de travail du traducteur de l'an 2001". Mons, Belgium. Vol. 1, pp. 31-48.

(Blanchon, *et al.*, 1999) Blanchon H., Boitet C. and Caelen J. (1999) *Participation francophone au consortium C-STAR II*. La Tribune des Industries de la Langue et du Multimédia, pp. 15-23.

(Blanchon, 2004) Blanchon H., Boitet C., Brunet-Manquat F., Tomokiyo M., Hamon A., Vo-Trung H. and Bey Y. (2004) *Towards Fairer Evaluations of Commercial MT Systems on BTEC corpora*. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT-04). Kyoto, Japan, pp. 21-26.

(Boitet, 2006) Boitet C., Bey Y., Tomokiyo M., Cao W. and Blanchon H. (2006) *IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations*. In Proceedings of the International Workshop on Spoken Language Translation. Kyoto, Japan, pp. 23-30.

(Boitet, 1982) Boitet C. (1982) *Le point sur ARIANE-78 début 1982*. GETA-CHAMPOLLION, CAP SOGETI-France. Vol. 1:(3).

(Boitet, 2005) Boitet C. (2005) *Gradable Quality Translation Through Mutualization of Human Translation and Revision, UNL-based MT and Coedition*. Research in Computing Science, IPN-CIC. Mexico. (presented at the 2nd Workshop on UNL and Other Interlinguas) (Gelbukh, A. ed.). In Book "Universal Networking Language, advances in theory and applications", Mexico, pp. 393-410.

(Boitet, 2007) Boitet C. (2007) *Corpus pour la TA : types, tailles, et problèmes associés, selon leur usage et le type de système*. Revue française de linguistique appliquée. Vol. XII – 2007, pp. 25-38.

(Boitet, 2004) Boitet C. (2004) *Progress report on building the French BTEC and participating in the MT evaluation campaign (CSTAR project)*. (rapport pour ATR), GETA, CLIPS, & ATR, 10 p.

(Boitet, 1995) Boitet C. and Blanchon H. (1995) *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup*. Machine Translation. Vol. 9, pp. 99-132.

(Boitet, *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. In Book "Computational Linguistics and Intelligent Text Processing"; Springer, Berlin / Heidelberg, pp. 361-373.

(Boitet, *et al.*, 1988) Boitet C. and Zaharin Y. (1988) *Representation Trees and String-Tree Correspondences*. In Proceedings of the Conference on Computational Linguistics (COLING-88). Budapest, pp. 59-64.

(Bouillon, 1992) Bouillon P. and Boesefeldt K. (1992) *Problèmes de traduction automatique dans le sous-langage des bulletins d'avalanches*. Translators' Journal (Meta). Vol. 37:(4), pp. 635-646.

(Brown, *et al.*, 1991) Brown P. F., Lai J. C. and Mercer R. L. (1991) *Aligning Sentences in Parallel Corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, California, pp. 177-184.

(Bryant, *et al.*, 2005) Bryant S. L., Forte A. and Bruckman A. (2005) *Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia*. In Proceedings of the GROUP International Conference on Supporting Group Work. Sanibel Island, Florida, US., pp. 1-10.

(Buffa, 2006) Buffa M. (2006) *Fabien Gandon: SweetWiki: semantic web-enabled technologies in Wiki*. In Proceedings of the International Symposium Wikis 2006, pp. 69-78.

(Buffa, 2006) Buffa M. (2006) *Intranet Wikis*. In Proceedings of the Intranet Web Workshop (WWW Conference). Edinburgh, Scotland, UK, pp. 18-28.

(Buffa, *et al.*, 2006) Buffa M., Crova G., Gandon F., Lecompte C. and Passeron J. (2006) *SweetWiki: Semantic Web Enabled Technologies in Wiki*. In Proceedings of the 1st Semantic Wiki Workshop (SemWiki2006), pp. 69-78.

(Callison-Burch, *et al.*, 2006) Callison-Burch C., Osborne M. and Koehn P. (2006) *Re-evaluating the Role of BLEU in Machine Translation Research*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Trento, Italy, pp. 249-256.

(Carpenter, 2006) Carpenter B. (2006) *Character Language Models for Chinese Word Segmentation and Named Entity Recogntion*. In Proceedings of the 5th ACL Chinese Special Interest Group (SIGHan). Sydney, Australia, pp. 169-172.

(Chen, *et al.*, 2000) Chen J. and Nie J.-Y. (2000) *Automatic Construction of Parallel English-Chinese Corpus for Cross-Language Information Retrieval*. Seattle, Washington, USA, pp. 21-28

(Chen, 2000) Chen J. and Nie J.-Y. (2000) *Parallel Web Text Mining for Cross-Language Information Retrieval*. Actes du Colloque sur la Recherche d'Informations Assistée par Ordinateur (RIAO). Paris, France, pp. 62-77.

(Chenon, 2005) Chenon C. (2005) *Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble, 221 p.

(Choumane, *et al.*, 2005) Choumane A., Blanchon H. and Roisin C. (2005) *Integrating Translation Services Within a Structured Editor*. In Proceedings of the ACM Symposium on Document Engineering (DocEng 2005). Bristol, UK, pp. 165-167.

(Craciunescu, *et al.*, 2004) Craciunescu O., Gerding-Salas C. and Stringer-O'Keeffe S. (2004) *Machine Translation and Computer-Assisted Translation: a New Way of Translating?* http://accurapid.com/journal/29computers.htm. Online Translation Journal. Vol. 8.

(Crestani, *et al.*, 2002) Crestani F., Girolami M. and Van Rijsbergen C. J. (2002) *Building Bilingual Dictionaries from Parallel Web Documents*. In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval. UK, London, pp. 303-323.

(Dagan, 1993) Dagan I. Church K. W. and Gale W. A. (1993) *Robust Bilingual Word Alignment for Machine Aided Translation*. In Proceedings of the Very Large Corpora workshop. Columbus, Ohio, USA, pp. 1-8.

(Daille, *et al.*, 1994) Daille B., Gaussier E. and Langé J.-M. (1994) *Towards Automatic Extraction of Monolingual and Bilingual Terminology*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94) Kyoto, Japan, pp. 712-716.

(Damerau, 1964) Damerau F. (1964) *A Technique for Computer Detection and Correction of Spelling Errors*. Communications of the ACM. 7:(3), pp. 171-176.

(Daoust, 2006) Daoust F. and Yves M. (2006) *Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés*. Actes des 8ème Journées Internationales d'Analyse Statistique des Données Textuelles 2006. Besançon, France : Presses Universitaires de Franche-Comté, pp. 327-340.

(Denoual, 2006) Denoual E. (2006) *Méthodes en caractères pour le traitement automatique des langues*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble; (Thèse préparée à ATR, Kyoto, Japan), 186 p.

(Désilets, *et al.*, 2006) Désilets A., Gonzalez L., Paquet S. and Stojanovic M. (2006) *Translation the Wiki Way*. In Proceedings of the WIKISym 2006. NRC 48736. Odense, Denmark., pp. 19-32.

(Doan-Nguyen, 1998) Doan-Nguyen H. (1998) *Accumulation of Lexical Sets: Acquisition of Dictionary Resources and Production of New Lexical Sets*. In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL'98. Montréal, Canada. Vol. 1, pp. 330-335.

(Doan-Nguyen, 1998) Doan-Nguyen H. (1998) *Techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes*. Thèse INPG, GETA, CLIPS, IMAG, Grenoble, 180 p.

(Doddington, 2002) Doddington G. (2002) *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. In Proceedings of the HLT-NAACL 2002. San Diego, California. Vol. 1, pp. 128-132.

(Edward, *et al.*, 2005) Edward A., Erich F., Neuhold J., Premsmit P. and Wuwongse V. (2005) *Eyes of a Wiki: Automated Navigation Map*. In Proceedings of the 8th International Conference on Asian Digital Libraries (ICADL), Bangkok, Thailand, pp. 186-193.

(Elrufaie, 2004) Elrufaie E. O. (2004) *A Wiki Paradigm to Manage Online Cource Content*. Masters Thesis, The Faculty of California State University, San Bernardino, USA, 68 p.

(Filgueiras, 1994) Filgueiras M. (1994) *A Successful Case of Computer Aided Translation*. In Proceedings of the 4th Conference on Applied Natural Language Processing. Stuttgart, Germany, pp. 91-94.

(Forcada, 2006) Forcada M. L. (2006) *Open-source Machine Translation: an Opportunity for Minor Languages*. In Proceedings of the 5th International Conference on Language Resources and Evaluation LREC. 5th SALTMIL Workshop on Minority Languages: “Strategies for developing machine translation for minority languages”. Genoa, Italy, pp. 1-6.

(Foster, *et al.*, 2002) Foster G., Langlais P. and Lapalme G. (2002) *Text Prediction with Fuzzy Alignments*. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: from Research to Real Users. Tiburon, California, USA, pp. 44-53.

(Foster, 2002) Foster G., Langlais P., Macklovitch E. and Lapalme G. (2002) *TransType: Text Prediction for Translators*. In Proceedings of the ACL-02 Demonstrations Session, Association for Computational Linguistics. Philadelphia, pp. 93-94.

(Och, 2000) Och F.-J. and Ney H. (2000) *Improved Statistical Alignment Models*. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hongkong, China, pp. 440-447.

(Frimannsson, 2005) Frimannsson A. (2005) *Adopting Standards Based XML File Format in Open Source Localisation*. Thesis, Queensland University of Technology, Faculty of Information Technology, School of Software Engineering and Data Communications, Brisbane, Australia, 85 p.

(Gale, 1991) Gale W. A. and Church K. W. (1991) *A Program for Aligning Sentences in Bilingual Corpora*. In Proceedings of the Meeting of the Association for Computational Linguistics (ACL-91), pp. 177-184.

(Gotti, *et al.*, 2006) Gotti F., Langlais P. and Coulombe C. (2006) *Vers l'intégration du contexte dans une mémoire de traduction sous-phrastique : détection du domaine de traduction*. In Proceedings of the Conference sur le Traitement Automatique des Langues Naturelles (TALN). Leuven, Belgium, pp. 483-492.

(Gow, 2003) Gow F. (2003) *Metrics for Evaluating Translation Memory Software*. In Partial Fulfillment of the Requirements for the Degree of MA (Translation). School of Translation and Interpretation University of Ottawa, Faculty of Graduate and Postdoctoral Studies of the University of Ottawa, Ottawa, Canada: University of Ottawa, 135 p.

(Grefenstette, *et al.*, 2004) Grefenstette G., Qu Y. and Evans D. A. (2004) *Mining the Web to Create a Language Model for Mapping between English Names and Phrases and Japanese*. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), pp. 110-116.

(Gupta, 2005) Gupta S. and Kaiser G. (2005) *Extracting Content from Accessible Web Pages*. In Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility (ACM Series). Vol. 88, pp. 26-30.

(Gupta, *et al.*, 2003) Gupta S., Kaiser G., Neistadt D. and Grimm P. (2003) *DOM-based Content Extraction of HTML Documents*. In Proceedings of the 12th International World Wide Web Conference (WWW-03), pp. 207-214.

(Hutchins, 1998) Hutchins J. (1998) *The Origins of the Translator's Workstation*. Machine Translation. Vol. 13, pp. 287-307.

(Hutchins, 2003) Hutchins J. (2003) *Has Machine Translation Improved? Some Historical Comparisons*. In Proceedings of the MT Summit IX conference New Orleans, pp. 181-188.

(Isozaki, *et al.*, 2005) Isozaki H., Sudoh K. and Hajime T. (2005) *NTT's Japanese-English Cross-Language Question Answering System*. In Proceedings of the NTCIR-5 Workshop Meeting. National Institute of Informatics (NII), Tokyo, Japan, pp. 186-193.

(Jeffrey, 1999) Jeffrey A. (1999) *Adapting the concept of "Translation Memory" to "Authoring Memory" for a Controlled Language Writing Environment*. In Proceedings of the International Conference Translating and the Computer. London, pp. 10-11.

(Jones, *et al.*, 2006) Jones M. C., Rathi D. and Twidale M. B. (2006) *Wikifying your Interface: Facilitating Community-Based Interface Translation*. In Proceedings of the 6th ACM Conference on Designing Interactive Systems. University Park, PA, USA, pp. 321-330.

(Jutras, 2000) Jutras J.-M. (2000) *An Automatic Reviser: the TransCheck system*. In Proceedings of the 6th Conference on Applied Natural Language Processing. Seattle, Washington, pp. 127-134.

(Kageura, 2006) Kageura K. (2006) *Improving the Usability of Language Reference Tools for Translators*. In Proceedings of the 10th Annual Meeting of the Japan Association for Natural Language Processing. Japan, pp. 707-710.

(Kay, 1997) Kay M. (1997) *The Proper Place of Men and Machines in Language Translation*. Machine Translation. Vol. 12, pp. 3-23.

(Khadivi, *et al.*, 2006) Khadivi S., Zolnay A. and Ney H. (2006) *Automatic Text Dictation in Computer-Assisted Translation*. In Proceedings of COLING/ACL-2006, Poster Sessions. Sydney, Australia, pp. 467-474.

(Kitamura, *et al.*, 2003) Kitamura M., Murata T., Sukehiro T., Shimohata S., Sasaki M., Matsunaga T. and Nakagawa T. (2003) *Technology and Development on Collaborative Translation Environment "Yakushite.net"*. In Proceedings of IPSJ-03, 65th Annual Conference of Information Processing Society of Japan (IPSJ). Vol. 5, pp. 319-322.

(Kraif, 2001) Kraif O. (2001) *Constitution et exploitation de bi-textes pour l'aide à la traduction*. Thèse UNSA (Université de Nice Sophia Antipolis), 549 p.

(Langlais, 2002) Langlais P., and Lapalme G. (2002) *TRANSTYPE: Development-Evaluation Cycles to Boost Translator's Productivity*. Machine Translation. Vol. 15, pp. 77-98.

(Langlais, 2000) Langlais P., Foster G. and Lapalme G. (2000) *Unit Completion for a Computer-aided Translation Typing System*. Machine Translation. Vol. 15:(4), pp. 267-294.

(Langlais, *et al.*, 2000) Langlais P., Sauvé S., Foster G., Macklovitch E. and Lapalme G. (2000) *Evaluation of TRANSTYPE, a Computer-aided Translation Typing System: A comparison of a theoretical- and a user- oriented evaluation procedure*. In Proceedings of the the Second International Conference On Language Resources and Evaluation (LREC). Athens, Greece. Vol. 2, pp. 641-648.

(Lee, 1999) Lee L. (1999) *Measures of Distributional Similarity*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland College Park, Maryland, USA, pp. 25-32.

(Leuf, 2001) Leuf B. and Cunningham W. (2001) *The Wiki Way: Quick Collaboration on the Web*. Edited by Upper Saddle River, NJ, USA: Addison Wesley.

(Levenshtein, 1966) Levenshtein V. I. (1966) *Binary Codes Capable of Correcting Deletion, Insertions and Reversals*. Soviet Physics Doklady. Vol. 8:(10), pp. 707-710.

(Mangeot-Lerebours, 2001) Mangeot-Lerebours M. (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble, 275 p.

(Mangeot-Lerebours, 2002) Mangeot-Lerebours M. (2002) *An XML Markup Language Framework for Lexical Databases Environments: the Common Dictionary Markup Language*. In Proceedings of the LREC Workshop on International Standards of Terminology and Language Resources Management. Las Palmas, Islas Canarias, Spain, pp. 37-44.

(Mangeot-Lerebours, 2002) Mangeot-Lerebours M. and Sérasset G. (2002) *Frameworks, Implementation and Open Problems for the Collaborative Building of a Multilingual Lexical Database*. In Proceedings of the Building and Using Semantic Networks SEMANET-02 workshop (COLING). Taipei, Taiwan, pp. 9-15.

(Martin, 1990) Martin W. (1990) *User-orientation in Dictionaries: 9 Propositions*. In Proceedings of the BudaLEX´88 Conference. Budapest: Akadémiai Kiadó, pp. 393-399.

(Mauser, *et al.*, 2006) Mauser A. R., Zens E., Matusov H. S. and Ney H. (2006) *The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation*. In Proceedings of the International Workshop on Spoken Language Translation. Kyoto, Japan, pp. 103-110.

(McEwan, 2002) McEwan C. J. A., Ounis I. *Building Bilingual Dictionaries from Parallel Web Documents*. In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval. Vol. 2291, pp. 303-323.

(Melby, 1982) Melby A. K. (1982) *Multi-level Translation Aids in a Distributed System*. In Proceedings of the Coling 82 conference. Prague, Czech, pp. 215-220.

(Melby, 1983) Melby A. K. (1983) *The Translation Profession and the Computer*. The Computer Assisted Language Instruction Consortium Journal. Vol. 1:(1), pp. 55-57.

(Muljadi, *et al.*, 2005) Muljadi H., Takeda H., Araki J., Kawamoto S., Kobayashi S., Mizuta Y., Demiya S. M., Suzuki S., Kitamoto A., Shirai Y., Ichiyoshi N., Ito T., Abe T., Gojobori T., Sugawara H., Miyazaki S. and Fujiyama A. (2005) *Semantic Mediawiki: A User-oriented System for Integrated Content and Metadata Management System*. In Proceedings of the IADIS International Conference on WWW/Internet 2005. Vol. 2, pp. 261-264.

(Muljadi, *et al.*, 2006) Muljadi H., Takeda H., Kawamoto S., Kobayashi S. and Fujiyama A. (2006) *Towards a Semantic Wiki-Based Japanese Biodictionary*. In Proceedings of the 1st Workshop on Semantic Wikis (ESWC2006), pp. 202-206.

(Nerima, *et al.*, 2006) Nerima L., Seretan V. and Wehrli E. (2006) *Le Problème des collocations en TAL*. Nouveaux cahiers de linguistique française. Vol. 27, pp. 95-115.

(Neumüller, 2002) Neumüller M. (2002) *Compact Data Structures for Querying XML*. In Proceedings of the EDBT 2002 PhD Workshop, EDBT, pp. 127-130.

(Nie, *et al.*, 2002) Nie J-Y., Simard M. and Foster G. (2002) Using Parallel Web Pages for Multi-lingual IR. In Book "Evaluation of Cross-Language Information Retrieval Systems". Springer Berlin / Heidelberg, pp. 137-173.

(Nie, 2001) Nie J.-Y. and Cai J. (2001) *Filtering Noisy Parallel Corpora of Web Pages*. In Proceedings of the IEEE Symposium on Natural Language Processing and Knowledge Engineering. Tucson, AZ, pp. 453-458.

(Nie, 2002) Nie J.-Y., Simard M. (2002) Using Statistical Translation Models for Bilingual IR. In Book "Evaluation of Cross-Language Information Retrieval Systems". Springer Berlin / Heidelberg, pp. 137-173.

(Nie, *et al.*, 2001) Nie J.-Y. and Jin F. (2001) *Merging different languages in a single document collection*. In Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum (CLEF). Darmstadt, Germany, pp. 59-62.

(Nie, *et al.*, 1999) Nie J.-Y., Simard M., Isabelle P. and Durand R. (1999) *Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*. In Proceedings of the the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 74-81.

(Pach, 2005) Pach J. (2005) Open Problems Wiki. In Book of Graph Drawing, New York, NY, USA: Springer Berlin / Heidelberg, pp. 508-509.

(Patry, 2005) Patry A. and Langlais P. (2005) *Paradocs: un système d'identification automatique de documents parallèles*. Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Dourdan, France, pp. 223-232.

(Planas, *et al.*, 1999) Planas E. and Furuse O. (1999) *Formalizing Translation Memories.* In Proceedings of the Machine Translation Summit VII. Singapore, pp. 331-339.

(Planas, *et al.*, 2000) Planas E. and Furuse O. (2000) *Multi-level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation*. In Proceedings of the 18th conference on Computational linguistics (COLING-2000). Saarbruecken, Germany. Vol. 2, pp. 621-627.

(Queens, 2005) Queens F. and Recker-Hamm U. (2005) *A Net-Based Toolkit for Collaborative Editing and Publishing of Dictionaries*. Literary and Linguistic Computing Advance Access, Oxford Journal. Vol. 20:(Suppl 1), pp. 165-175.

(Reichling, 1992) Reichling A. (1992) *EURODICAUTOM, ou la terminologie en l'an 2001*. In Book "L'environnement traductionnel - La station de travail du traducteur de l'an 2001", 1992; Presses de l'Université du Québec, pp. 201-206.

(Resnik, 2003) Resnik P. and Smith N. A. (2003) *The Web as a Parallel Corpus*. Computational Linguistics. Vol. 3:(29), pp. 349-380.

(Resnik, *et al.*, 2001) Resnik P., Oard D. and Levow G. (2001) *Improved Cross-Language Retrieval Using Backoff Translation*. In Proceedings of the First International Conference on Human Language Technology Research (HLT-2001). San Diego, CA, pp. 153-155.

(Roukos, 2006) Roukos S. (2006) *Recent Results on MT Evaluation in the GALE Program*. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT-06). Kyoto, Japan, pp. 14-15.

(Sato, *et al.*, 2003) Sato S. and Sasaki Y. (2003) *Automatic Collection of Related Terms from the Web*. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan, pp. 121-124.

(Scott, 2003) Scott B. (2003) *The Logos Model: An Historical Perspective*. Machine Translation. Vol. 18, pp. 1-72.

(Sérasset, 1994) Sérasset G. (1994) *SUBLIM: un système universel de bases lexicales multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble, 194 p.

(Sérasset, 2004) Sérasset G. (2004) *A Generic Collaborative Platform for Multilingual Lexical Database Development.* In Proceedings of the Post-COLING 2004 Workshop on Multilingual Linguistic Resources (MLR2004). Geneva, Switzerland, pp. 73-79.

(Streiter, 2005) Streiter O., Mathiasser M. (2005) *Open Source Framework for Multilingual Computing*. In Proceedings of the Lesser Used Languages & Computer Linguistics. European Academy Bozen, Italy, pp. 189-207.

(Takeuchi, *et al.*, 2007) Takeuchi K., Kanehila T., Hilao K., Abekawa T. and Kageura K. (2007) *Flexible automatic look-up of English idiom entries in dictionaries*. In Proceedings of the MT Summit XI. Copenhagen, Denmark, pp. 451-458.

(Teoh, 2004) Teoh E. H. and Tang E. K. (2004) *User Model for Prototyping Computer-Aided Translation System*. In Proceedings of the 7th International Conference on WWCS (Work With Computing Systems) 2004, Kuala Lumpur, Malaysia.

(Tiedemann, 2004) Tiedemann J. and Nygaard L. (2004) *The OPUS Corpus - Parallel and Free*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. Vol. 5, pp. 1183-1186.

(Torlone, *et al.*, 2003) Torlone R. and Atzeni P. (2003) *Chameleon: an Extensible and Customizable Tool for Web Data Translation*. In Proceedings of the 29th International Conference on Very Large Data Bases. Berlin, Germany. Vol. 29, pp. 1085 -1088.

(Torlone, 2001) Torlone R. and Atzeni P. (2001) *A Unified Framework for Data Translation over the Web*. In Proceedings of the Second International Conference on Web Information System Engineering (WISE 2001), IEEE Computer Society Press. Kyoto, Japan, pp. 350–358.

(Tsai, 2004) Tsai W. (2004) *La coédition langue↔UNL pour partager la révision entre langues d'un document multilingue*. Thèse UJF, GETA, CLIPS, IMAG, Grenoble, 307 p.

(Uchida, *et al.*, 1999) Uchida H., Zhu M. and Della-Senta T. (1999) *A gift for a millenium*, in Report UNL, United Nations University: Institute of Advanced Studies [Geneva: UNDL Foundation] (ed.), *LREC2002*, Tokyo, 62 p.

(Véronis, 2000) Véronis J. (2000) *Evaluation of Parallel Text Alignment Systems - The ARCADE Project*. In Book "Parallel Text Processing - Alignment and Use of Translation Corpora". AA Dordrecht, The Netherlands: Kluwer Academic Publisher, pp. 369-388.

(Véronis, 2000) Véronis J. (2000) *From the Rosetta Stone to the Information Society - A Survey of Parallel Text Processing*. In Book "Parallel Text Processing - Alignment and Use of Translation Corpora". AA Dordrecht, The Netherlands: Kluwer Academic Publisher, pp. 1-24.

(Vo-Trung, 2004) Vo-Trung H. (2004) *Méthodes et outils pour utilisateurs, développeurs et traducteurs de logiciels en contexte multilingue*. Thèse INPG, GETA, CLIPS, IMAG, Grenoble, 224 p.

(Vo-Trung, 2004) Vo-Trung H. (2004) *Réutilisation de traducteurs gratuits pour développer des systèmes multilingues*. In Proceedings of the Conférence Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Fès, Maroc, pp. 111-117.

(Vo-Trung, 2004) Vo-Trung H. (2004) *SANDOH, un outil pour analyser des textes hétérogènes*. In Proceedings of the 5èmes Journées internationales d'analyse statistique des données textuelles (JADT'2004). Paris, pp. 1178-1185.

(Vuillemot, 2006) Vuillemot R. (2006) *Modèles de navigation dans de grands corpus de documents : décomposition, classification et personnalisation*. Actes de IC2006 : 17e Journées Francophones d'Ingénierie des Connaissances. Nantes, France, pp. 28-30.

(Wagner, *et al.*, 1974) Wagner R. A. and Fisher M. J. (1974) *The String-to-String Correction Problem*. Journal of the ACM (JACM). Vol. 21:(1), pp. 168-173.

(Walker, *et al.*, 2001) Walker D. J., Clements D. E., Darwin M. and Amtrup J. W. (2001) *Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality*. In Proceedings of the 8th Machine Translation Summit. Santiago de Compostela, Spain, pp. 369–372.

(Yamron, *et al.*, 1994) Yamron J., Cant J., Demedts A., Taiko D. and Yoshiko I. (1994) *The Automatic Component of the LINGSTAT Machine-Aided Translation System*. In Proceedings of the Workshop on Human Language Technology (HLT'94), Plainsboro, New Jersey, USA, pp. 163-168.

(Yves, 2003) Yves C. (2003) *Convergence in CAT: blending MT, TM, OCR & SR to boost productivity*. In Proceedings of the International Conference Translating and the Computer 25. London.

(Zhang, *et al.*, 2004) Zhang Y. and Vines P. (2004) *Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. The University of Sheffield, UK, pp. 162-169.

(Zhang, *et al.*, 2006) Zhang Y., Wu K., Gao J. and Vines P. (2006) *Automatic Acquisition of a Chinese-English Parallel Corpus from the Web*. In Proceedings of the 28th European

Conference on Information Retrieval (ECIR-06). Imperial College Road, London, pp. 407-419.
(Zimina-Poirot, 2004) Zimina-Poirot M. (2004) *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*, Thèse : Université de la Sorbonne nouvelle, Paris, 328 p.

Signets

Ces signets ont été tous vérifiés le 17/08/2008.

Arabeyes (2007), le projet d'arabisation de Mozilla, *http://www.arabeyes.org/project.php?proj=Mozilla*.
Breen (2007), les dictionnaires en ligne de Jim Breen, *http://www.csse.monash.edu.au/~jwb/wwwjdic.html*.
CNet (2006), Termado : un logiciel de gestion et de publication de lexiques et dictionnaires", *http://xml.coverpages.org/ni2002-06-14-a.html*.
Commercial-MT (2006), une liste de traducteurs automatiques, *http://www.foreignword.com/Technology/mt/mt.htm*.
DéjàVu (2007), un outil commercial d'aide à la traduction, *http://www.atril.com*.
DocBook (2007), un format structuré de livres électroniques, *www.docbook.org*.
DocBook (2006), la bibliothèque DocBook, *http://www-128.ibm.com/developerworks/library/l-docbk.html*.
EU (2008), la terminologie interactive d'Union Européenne (23 languages), *http://europa.eu/rapid/pressReleasesAction.do?reference=IP/07/962&format=HTML&aged=0&language=EN&guiLanguage=en*.
EuroParl (2007), le corpus EuroParl, *http://people.csail.mit.edu/koehn/publications/europarl/*.
FrenchMozilla (2005), le projet de francisation de Mozilla, *http://frenchmozilla.online.fr/*.
Giza++ (2006), un outil libre d'alignement, *http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html*.
GoogleAPI (2006), l'API Google, *http://www.google.com/apis/*.
Googlization (2006), un exemple d'utilisation de l'API Google, *http://www.devx.com/Java/Article/7910*.
Hansard (2008), le corpus Hansard, *http://www.isi.edu/natural-language/download/hansard/*.
HTMLArea (2006), un éditeur WYSIWYG, *http://www.htmlarea.com/*.
Humanitarian-Translation (2006), un site de traductions humanitaires, *http://seattlepi.nwsource.com/local/184671_redcross03.html*.
Hutchins (2000), un article intéressant sur la traduction automatique, *http://nl.ijs.si/eamt00/proc/Hutchins.pdf*
IWSLT (2006), le site du Workshop IWSLT 2006, *http://www.slc.atr.jp/IWSLT2006/*
Japan-News-Corporation (2007), Yakushite.net: un outil de traduction collaborative en ligne, *http://www.japancorp.net/Article.Asp?Art_ID=8404*.
Jimmy (2007), Wikipedia : Social Innovation on the Web, *http://www.wdil.org/conference/pdf/wikiz.pdf*.
LingPipe (2006), un outil de TAL, *http://alias-i.com/lingpipe/demo.html*.
Linguistic-Ressources (2006), une sélection de dictionnaires indispensables pour le français, *http://www.lesannuaires.com/annuaire-dictionnaire.html*.

Lisa (2008), the Localization Industry Standards Association (LISA), *http://www.lisa.com/*.
Logos (2006), le traducteur automatique Logos, *http://www.springerlink.com/media/788y661n1mdqug8a5q06/contributions/r/0/1/3/r01317149734884p.pdf*.
McKay (2007), une liste d'outils libres pour les traducteurs, *http://www.translatewrite.com/foss/atapaper05.pdf*.
MT-Principals (2006), quelques astuces pour commencer la traduction EN-FR, *http://fr.wikipedia.org/wiki/Traduction_de_l%27anglais_vers_le_fran%C3%A7ais*, la traduction selon Wikipedia, *http://fr.wikipedia.org/wiki/Traduction*.
MTPostEditing 2008), la post-édition selon Jeff Allen, *http://www.geocities.com/jeffallenpubs/*.
Multi-User-Text-Editing (2006), un éditeur multiplate-forme et collaboratif, *http://www.moonedit.com/indexen.htm*.
OmegaT (2007), un outil libre d'aide à la traduction, *http://www.omegat.org/omegat/omegat_en/omegat.html*.
Papillon (2006), une base lexicale multilingue, *http://www.papillon-dictionary.org*.
Paxhumana (2006), un site sur la traduction humanitaire, *http://paxhumana.info*.
PENSEE MT (2007), le traducteur automatique intégré à Yakushite.Net, *http://www.oki.com/en/press/1999/z9911e.html#top*.
Pystemmer (2008), Stemmer en Python, *http://sourceforge.net/projects/pystemmer*.
QRselect (2006), un outil d'aide à la traduction JP-EN, *http://apple.cs.nyu.edu/akin/*.
Saha (2005), A Novel 3-Tier XML Schematic Approach for Web Page Translation, *http://www.acm.org/ubiquity/views/v6i43_saha.html*.
Similis (2005), un outil commercial d'aide à la traduction, *http://www.lingua-et-machina.com/*.
TCE (2006), la traduction à la Commission Européenne, *http://europa.eu.int/comm/dgs/translation/*.
Teanotwar (2005), un site humanitaire, *http://teanotwar.blogtribe.org/*.
Termbase (2006), une base terminologique FR-EN des termes Internet, *http://home.uchicago.edu/~kuzmack/dictionary/*.
TM-Principale (2006), les mémoires de traductions selon Wikipedia, *http://fr.wikipedia.org/wiki/M%C3%A9moire_de_traduction*.
TRADOH (2007), un traducteur automatique multilingue avec identification de langues, *http://www-clips.imag.fr/geta/User/hung.vo-trung/traducteur/web_fr/Index.htm*.
Trados (2005), un outil commercial d'aide à la traduction, *http://www.trados.com/*.
Traduct (2006), un site de traductions technique EN-FR, *http://www.traduc.org*.
Translation-Project (2006), site de traduction de l'encyclopédie "Diderot et d'Alembert", *http://www.hti.umich.edu/d/did/call.html*.
Translationwiki (2006), la traduction à la Wiki, *http://www.translationwiki.com*.
Translator's-Corner (2006), une liste classée d'outils d'aide à la traduction, *http://www.geocities.com/fmourisso/CAT.htm*.
Twiki (2006), le site du Wiki TWiki, *http://twiki.org/*.
W3C (2007), le consortium des standards Web, *http://www.w3.org/Consortium/Translation*.
Wikipedia (2007) *http://www.wikipedia.org*.
Wiktionary (2007), dictionnaire libre multilingue, *http://www.wiktionary.org*.
XWiki (2006), un Wiki libre basé sur Java, *http://www.xwiki.com/*.

Yakushite.net, (2007), un environnement de traduction collaborative en ligne, *http://www.yakushite.net*.

Annexe A

Domaines d'application des Wikis

Domaine	Description et quelques exemples
Rédaction de la documentation	- Dotclear : http://dev.dotclear.net. - SlackFR : http://www.slack-fr.org. - EagleFaq : http://faq.eagle-usb.org. - Francophonie : http://blender.doc.fr.free.fr.
Suivi de projets collaboratifs	- Planète couleurs : http://www.planete-couleurs.com - Tela_Insecta : http://www.tela-insecta.org (réseau des entomologistes francophones)
Les encyclopédies et les bases de connaissances en ligne	Cela englobe les encyclopédies généralistes (exemple Wikipedia) et les encyclopédies spécifiques à un domaine telles que : - QuestionsSurLAlimentation : http://www.alainhenry.be. - Web sémantique : Websemantique.org. - Partage de connaissances en gestion : http://www.km-fr.com/wiki/. - Techniques alternatives de vie : http://www.newlimits.org.
Les bases de connaissances des entreprises	Elles permettent de partager des connaissances métier et de communiquer au sein d'une entreprise.
Les Wiki communautaires	Ils rassemblent des personnes autour d'un sujet déterminé à des fins de rencontre, de partage des connaissances, d'organisation, etc. Par exemple : Guide de voyages sur http://wikitravel.org/fr ou échanges sur internet en P2P sur http://www.wikip2p.com.
Les Wiki personnels	Ils sont utilisés comme outils de productivité et de gestion de l'information. Cela va du simple pense-bête au bloc-notes évolué, en passant par des applications aussi variées que la gestion d'agendas, l'historique de documents ou encore les publications Web rapides.
Les traductions de livres publics	Il s’agit de traductions de livres appartenant au domaine public, comme c'est le cas pour Free Culture sur http://blogspace.com/freeculture/Accueil.

Annexe B

Projet DEMGOL : temps de traduction sans utilisation de BEYTrans
« notices de la lettre L »

N	Notice	Minutes	Secondes	Total (secondes)	N	Notice	Minutes	Secondes	Total (secondes)
1	LABDACOS	14 mn	34 s	**874 s**	31	LÉDA	16 mn	0 s	**960 s**
2	LACEDEMONE	26 mn	65 s	**1625 s**	32	LIAGORÉ	6 mn	7 s	**367 s**
3	LAKIOS	11 mn	85 s	**745 s**	33	LIMON	4 mn	35 s	**275 s**
4	LACON	4 mn	96 s	**336 s**	34	LIRIOPÉ	4 mn	8 s	**248 s**
5	LADON	9 mn	41 s	**581 s**	35	LITÒ	4 mn	59 s	**299 s**
6	LAËRTE	2 mn	90 s	**210 s**	36	LÉOS	1 mn	52 s	**112 s**
7	LAÏOS	10 mn	63 s	**663 s**	37	LEONASSA	6 mn	21 s	**381 s**
8	LAMÉDON	20 mn	79 s	**1279 s**	38	LEONTÉE	10 mn	12 s	**612 s**
9	LAMIA	5 mn	74 s	**374 s**	39	LEONTOPHONOS	4 mn	23 s	**263 s**
10	LAMOS	5 mn	54 s	**354 s**	40	LEONTOPHRON	2 mn	15 s	**135 s**
11	LAMPÉTOS	6 mn	66 s	**426 s**	41	LÉPRÉE	8 mn	57 s	**537 s**
12	LAMPÉTIE	6 mn	48 s	**408 s**	42	LESTRYGONS	8 mn	34 s	**514 s**
13	LAMPOS	3 mn	84 s	**264 s**	43	LÉTHÉ	2 mn	3 s	**123 s**
14	LAMPSAKÉ	15 mn	59 s	**959 s**	44	LEUCASPIS	2 mn	19 s	**139 s**
15	LAMPUSA	0 mn	89 s	**89 s**	45	LEUCÉ	2 mn	48 s	**168 s**
16	LAOCOON	5 mn	59 s	**359 s**	46	LEUCIPPE	10 mn	20 s	**620 s**
17	LAODAMANTE	25 mn	54 s	**1554 s**	47	LEUCOS	2 mn	56 s	**176 s**
18	LAODICÉ	10 mn	68 s	**668 s**	48	LEUCOPHANE	4 mn	25 s	**265 s**
19	LAODOCO	3 mn	57 s	**237 s**	49	LEUCON	2 mn	14 s	**134 s**
20	LAOGORAS	2 mn	34 s	**154 s**	50	LEUCONOÉ	5 mn	55 s	**355 s**
21	LAOMÉON	0 mn	46 s	**46 s**	51	LEUCOSIA	2 mn	3 s	**123 s**
22	LAONOME	2 mn	20 s	**140 s**	52	LEUCOTHéa	5 mn	26 s	**326 s**
23	LAOTHOÉ	5 mn	59 s	**359 s**	53	LEUCOTHOé	5 mn	15 s	**315 s**
24	LAPITHES	13 mn	2 s	**782 s**	54	LYCABAS	9 mn	09 s	**549 s**
25	LARINOS	3 mn	85 s	**265 s**	55	LYCAON	9 mn	9 s	**549 s**
26	LAS	5 mn	47 s	**347 s**	56	LYCASTOS	5 mn	20 s	**320 s**
27	LATONE	5 mn	34 s	**334 s**	57	LYCOS	8 mn	25 s	**505 s**
28	LÉAGRE	5 mn	67 s	**367 s**	58	LYCOPHON	8 mn	35 s	**515 s**
29	LÉANDRE	2 mn	77 s	**197 s**	59	LYCOPHRON	1 mn	43 s	**103 s**
30	LEARCO	1 mn	58 s	**118 s**	60	LYOMÈDE	8 mn	27 s	**507 s**

Temps de la traduction (heures) : 7,11 heures

Projet DEMGOL : temps de traduction avec « BEYTrans »

Notices de la lettre M

(Post-édition des traductions automatiques produites par Systran-Web)

N	Notice	NAD	Minutes	Secondes	N	Notice	NAD	Minutes	Secondes
1	MACEDONE	2	6 mn	30 s	16	MEDEIO	0	2 mn	46 s
2	MACHEREO	0	8 mn	10 s	17	MEDO	0	3 mn	24 s
3	MACISTO	3	5 mn	0 s	18	MEDONTE	8	9 mn	34 s
4	MAIA	0	2 mn	45 s	19	MEDUSA	9	15 mn	57 s
5	MAIRA	4	10 mn	40 s	20	MEGAPENTE	9	5 mn	0 s
6	MANIA	2	7 mn	50 s	21	MEGAREO	5	21 mn	22 s
7	MANTICORA	11	28 mn	14 s	22	MEGE	5	10 mn	30 s
8	MANTO	10	8 mn	0 s	23	MELANEO	7	12 mn	27 s
9	MARATO	5	3 mn	56 s	24	MELANIPPO	17	22 mn	12 s
10	MARATONE	0	1 mn	32 s	25	MELANTO	9	21 mn	15 s
11	MARMACE	8	12 mn	12 s	26	MELEAGRIDI	8	12 mn	0 s
12	MARPESSA	6	3 mn	0 s	27	MELEAGRO	6	15 mn	45 s
13	MECISTEO	0	11 mn	25 s	28	MELIA	4	8 mn	0 s
14	MECONE	13	6 mn	30 s	29	MELIBEA	7	10 mn	0 s
15	MEDEA	11	9 mn	30 s	30	MELISSA	5	12 mn	37 s

Temps de la traduction (heures) : 5, 13 heures

* Le temps de la traduction de la notice « MEDUSA » comprend le temps de la recherche et de l'insertion de mots dans le dictionnaire.
* La MT a été consultée deux fois (MARIA, MECONE).
* Aucune consultation n'a été faite durant les traductions des notices.
* NAD : Nombre d'ajouts dans le dictionnaire secondaire.

Projet DEMGOL : liste des entrées ajoutées dans le dictionnaire secondaire

N	Notice	Noms des entrées ajoutées au lexique de traduction	N	Notice	Noms des entrées ajoutées au lexique de traduction
1	MACEDONE	éponyme, Eole	16	MEDEIO	-
2	MACHEREO	-	17	MEDO	-
3	MACISTO	fratello, città, Peloponneso	18	MEDONTE	ucciso, Enea, personaggio, araldo, Itaca, Penelope, Telemaco, Odisseo
4	MAIA	-	19	MEDUSA	Gorgoni, Forci, Ceto, Perseo, sorella, figlie, Stenno, Euriale, Priamo
5	MAIRA	diversi, figli, oppure, sposa,	20	MEGAPENTE	Menelao, schiava, Elena, regno, Sparta, altro, Preto, Tirinto, Argo
6	MANIA	follia, collera	21	MEGAREO	Beota, Poseidone, Enope, Megara, Atene, metropoli
7	MANTICORA	animale, antropofago, fulvo, maschile, femminile, popolazione, fama, timore, cervo, voce, uomo	22	MEGE	Fileo, Ctimene, guerra, Troia, pretendente
8	MANTO	Tiresia, dono, profezia, Apollo, Argivi, Sibilla, cretese, Mopso, tradizione, eponima/eponimo	23	MELANEO	Eurito, Messenia, moglie, altra, Melaneo, Driopi, Ambracia
9	MARATO	Arcadia, Attique, partecipò, Dioscuri, vittoria	24	MELANIPPO	Ares, Triteia, Tritone, sacerdotessa, Atena, Acaia, madre, guerriero, tebano, Astaco, combatté, Tebani, Sette, Tideo, Agrio, Teseo, troiani
10	MARATONE	-	25	MELANTO	Andropompo, cacciato, Eraclidi, Pilo, divenne,re, combattendo, contro, Xanto
11	MARMACE	pretendenti, nozze, Ippodamia, padre, cavalle, fiume, omonimo, popolo	26	MELEAGRIDI	giovani, donne, sorelle, Meleagro, trasformate, Artemide, impietosì, morte
12	MARPESSA	Eveno, fidanzato, Demonice, Ida, facoltà, pretendenti	27	MELEAGRO	Eneo, Etoli,Calidone, Altea, caccia, cinghiale
13	MECISTEO		28	MELIA	eroine, Oceano, Ismene, Inaco
14	MECONE	Ateniese, amato, Demetra, trasformato, racconto, sembra, fonti, greche, dea, scoperto, antico, Sicione, Corinto	29	MELIBEA	Pelasgo, Niobe, sfuggì, fratelli, maschile, Melibeo, bovaro
15	MEDEA	della, ninfa, discendente, Elio, nipote, Circe, maga, filtri, vello, Grecia, Medea	30	MELISSA	Amaltea, Periandro, inviava, enigmi, marito

DTD étendue du format TMX (Translation Memory Exchange)
(DTD TMX version 1.4a)

```
<!ENTITY  lt   "&#38;#60;" >
<!ENTITY  amp  "&#38;#38;" >
<!ENTITY  gt   "&#62;" >
<!ENTITY  apos "&#39;" >
<!ENTITY  quot "&#34;" >
<!ENTITY % segtypes    "block|paragraph|sentence|phrase" >
<!-- Base Document Element -->
<!ELEMENT tmx                (header, body) >
<!ATTLIST tmx
          version            CDATA             #FIXED "1.4" >
<!-- Header -->
<!ELEMENT header             (note|prop|ude)* >
<!ATTLIST header
          creationtool       CDATA                #REQUIRED
          creationtoolversion CDATA               #REQUIRED
          segtype            %segtypes;)          #REQUIRED
          o-tmf              CDATA                #REQUIRED
          adminlang          CDATA                #REQUIRED
          srclang            CDATA                #REQUIRED
          datatype           CDATA                #REQUIRED
          o-encoding         CDATA                 #IMPLIED
          creationdate       CDATA                #IMPLIED
          creationid         CDATA                #IMPLIED
          changedate         CDATA                #IMPLIED
          changeid           CDATA                #IMPLIED >
<!-- Body -->
<!ELEMENT body               (tu*) >
<!-- No attributes -->
<!-- Note -->
<!ELEMENT note               (#PCDATA) >
<!ATTLIST note
          o-encoding         CDATA                #IMPLIED
          xml:lang           CDATA                #IMPLIED
          lang               CDATA                #IMPLIED >
<!-- lang is deprecated: use xml:lang -->
<!-- User-defined Encoding -->
<!ELEMENT ude             (map+) >
<!ATTLIST ude
          name               CDATA                #REQUIRED
          base               CDATA                #IMPLIED >
<!-- Note: the base attribute is required if one or more <map>
     elements in the <ude> contain a code attribute. -->
<!-- Character mapping -->
<!ELEMENT map             EMPTY >
<!ATTLIST map
          unicode            CDATA                #REQUIRED
          code               CDATA                #IMPLIED
          ent                CDATA                #IMPLIED
          subst              CDATA                #IMPLIED >
<!-- Property -->
<!ELEMENT prop            (#PCDATA) >
<!ATTLIST prop
```

```
          type                CDATA                #REQUIRED
          xml:lang            CDATA                #IMPLIED
          o-encoding          CDATA                #IMPLIED
          lang                CDATA                #IMPLIED >
        <!-- lang is deprecated: use xml:lang -->
<!-- Translation Unit -->
<!ELEMENT tu              ((note|prop)*, tuv+) >
<!ATTLIST tu
          tuid                CDATA                #IMPLIED
          o-encoding          CDATA                #IMPLIED
          datatype            CDATA                #IMPLIED
          usagecount          CDATA                #IMPLIED
          lastusagedate       CDATA                #IMPLIED
          creationtool        CDATA                #IMPLIED
          creationtoolversion CDATA                #IMPLIED
          creationdate        CDATA                #IMPLIED
          creationid          CDATA                #IMPLIED
          changedate          CDATA                #IMPLIED
          segtype             %segtypes;)          #IMPLIED
          changeid            CDATA                #IMPLIED
          o-tmf               CDATA                #IMPLIED
          srclang             CDATA                #IMPLIED >
<!—Extended TMX for corpus management-->
<!-- Translation Unit Variant -->
<!—Métadonnées ajoutées pour décrire les UT -->
<!ELEMENT tuv             ((note|prop)*, seg*) >
<!ATTLIST tuv
          xml:lang            CDATA                #REQUIRED
          o-encoding          CDATA                #IMPLIED
          datatype            CDATA                #IMPLIED
          usagecount          CDATA                #IMPLIED
          lastusagedate       CDATA                #IMPLIED
          creationtool        CDATA                #IMPLIED
          creationtoolversion CDATA                #IMPLIED
          creationdate        CDATA                #IMPLIED
          creationid          CDATA                #IMPLIED
          changedate          CDATA                #IMPLIED
          o-tmf               CDATA                #IMPLIED
          changeid            CDATA                #IMPLIED >
<!-- lang is deprecated: use xml:lang -->
<!-- Text -->
<!ELEMENT seg             (#PCDATA|bpt|ept|ph|it|hi|ut)* >
<!—Extended TMX for corpus management-->
<!—id : identificateur unique d’un segment; type : type de segment
(texte, graphe, transcription, mt : nom de la MT; usageuser :
information sur l’utilisateur) -->
<!ATTLIST seg
          id                  CDATA                #REQUIRED
          type                CDATA                #REQUIRED
          mt                  CDATA                #IMPLIED
          creationdate        CDATA                #REQUIRED
          usagecount          CDATA                #REQUIRED >
<!-- Content Markup -->
<!ELEMENT bpt             (#PCDATA|sub)* >
<!ATTLIST bpt
          i                   CDATA                #REQUIRED
          x                   CDATA                #IMPLIED
```

```
          type                  CDATA                #IMPLIED >
<!ELEMENT ept                                   (#PCDATA|sub)* >
<!ATTLIST ept
          i                     CDATA                #REQUIRED >
<!ELEMENT sub
(#PCDATA|bpt|ept|it|ph|hi|ut)* >
<!ATTLIST   sub
          datatype              CDATA                #IMPLIED
          type                  CDATA                #IMPLIED >
<!ELEMENT it                    (#PCDATA|sub)* >
<!ATTLIST it
          pos                   (begin|end)          #REQUIRED
          x                      CDATA               #IMPLIED
          type                   CDATA               #IMPLIED >
<!ELEMENT ph                     (#PCDATA|sub)* >
<!ATTLIST ph
          x                      CDATA               #IMPLIED
          assoc                  CDATA               #IMPLIED
          type                   CDATA               #IMPLIED >
<!ELEMENT hi  (#PCDATA|bpt|ept|it|ph|hi|ut)* >
<!ATTLIST hi
          x                      CDATA               #IMPLIED
          type                   CDATA               #IMPLIED >
<!-- The <ut> element is deprecated -->
<!ELEMENT ut                     (#PCDATA|sub)* >
<!ATTLIST ut
          x                      CDATA               #IMPLIED >
```

Code source : segmentation en utilisant Linpipe

L'interface com.aliasi.sentences.SentenceModel proposé dans LingPipe permet la segmentation à partir d'un ensemble d'items et un ensemble d'espaces. La méthode « boundaryIndices » reçoit les deux ensembles et retourne un ensemble d'indices des items de fin de phrase. Le programme Java suivant montre comment utiliser LingPipe pour la segmentation d'un fichier spécifié en entrée :

```
import com.aliasi.sentences.MedlineSentenceModel;
import com.aliasi.sentences.SentenceModel;
import com.aliasi.tokenizer.IndoEuropeanTokenizerFactory;
import com.aliasi.tokenizer.TokenizerFactory;
import com.aliasi.tokenizer.Tokenizer;
import com.aliasi.util.Files;
import java.io.File;
import java.io.IOException;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.Iterator;
import java.util.Set;
/** Use SentenceModel to find sentence boundaries in text */
public class SentenceBoundaryDemo {
   static final TokenizerFactory TOKENIZER_FACTORY = new
IndoEuropeanTokenizerFactory();
```

```
   static final SentenceModel SENTENCE_MODEL  = new MedlineSentenceModel();
   public static void main(String[] args) throws IOException {
            File file = new File(args[0]);
            String text = Files.readFromFile(file);
            System.out.println("INPUT TEXT: ");
            System.out.println(text);
            ArrayList tokenList = new ArrayList();
            ArrayList whiteList = new ArrayList();
            Tokenizer tokenizer =
TOKENIZER_FACTORY.tokenizer(text.toCharArray(),0,text.length());
            tokenizer.tokenize(tokenList,whiteList);
            System.out.println(tokenList.size() + " TOKENS");
            System.out.println(whiteList.size() + " WHITESPACES");
            String[] tokens = new String[tokenList.size()];
            String[] whites = new String[whiteList.size()];
            tokenList.toArray(tokens);
            whiteList.toArray(whites);
            int[] sentenceBoundaries =
SENTENCE_MODEL.boundaryIndices(tokens,whites);

            System.out.println(sentenceBoundaries.length  + " SENTENCE END TOKEN
OFFSETS");
            if (sentenceBoundaries.length < 1) {
               System.out.println("No sentence boundaries found.");
               return;
            }
            int sentStartTok = 0;
            int sentEndTok = 0;
            for (int i = 0; i < sentenceBoundaries.length; ++i) {
               sentEndTok = sentenceBoundaries[i];
               System.out.println("SENTENCE "+(i+1)+": ");
               for (int j=sentStartTok; j<=sentEndTok; j++) {
                        System.out.print(tokens[j]+whites[j+1]);
               }
               System.out.println();
               sentStartTok = sentEndTok+1;
            }
   }
}
```

Annexe C

Liste non-exhaustive de corpus libres consultables en ligne

Corpus téléchargeables librement
Acquis corpus (JRC) Le socle législatif de l'Union européenne Types : Institutionnel, juridique. Taille : Environ **6.300.000 mots** par langue. Langues : 20 langues officielles de l'UE (cs da de el en es et fi fr hu it lt lv mt nl pl pt ro sk sl sv). Traitements : Segmentation en phrases. Alignement. Format : XCES.
Corpus BAF (Bi-texte anglais français) Types : Institutionnel, technique, scientifique, littéraire. Taille : **400.000 mots** dans chaque langue. Langues : en fr Traitements : Segmentation en phrases. Alignement.
Corpus CARMEL Classiques du récit de voyage (XIXe - début XXe) Type : Littéraire. Taille : 36 ouvrages, **10.000.000 mots**. Langues : en es fr it Traitements : Segmentation en phrases et tokens. Etiquetage morphosyntaxique et lemmatisation. Désambiguïsation sémantique. Identification thématique.
CRATER Multilingual Aligned Annotated Corpus Type : Technique. Taille : **1.000.000 mots**. Domaine : Télécommunications. Langues : en fr es Traitements : Etiquetage des parties du discours. Alignement.
The IJS ELAN - Slovene-English Aligned Corpus Taille : **1.000.000 mots**. Langues : en sl Traitements : Segmentation en phrases, en tokens. Etiquetage morphologique (Multext East tags). Alignement. Format : Standard TMX (Translation Memory Exchange) – XML/TEI P4
OPUS, an open source parallel corpus Types : Technique, institutionnel. Taille : **30.000.000 mots** (en). Langues : 60 langues Traitements : Segmentation en phrases, en tokens. Alignement. Format : XCES Description : - EUconst, le projet de constitution de l'UE (21 langues). - Europarl, comptes rendus du Parlement européen 1996-2003 (11 langues). - Documentation Open Office (6 langues : de en es fr jp sv). Etiquetage des parties du discours. - Manuel de PHP (21 langues). - Messages système de KDE (60 langues). - Manuel de KDE (24 langues).
Swedish political texts (Uppsala Universitet) Textes du gouvernement suédois Langues : de en es fr sv Taille : **11.000 mots**. Format : Textuel.

Traitements : Alignement.
Fournisseur : Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria.
Restrictions : Non disponible pour un usage commercial.

Hansard (Hansard, 2008)
Langues : en fr
Taille : **47.389.000 mots.**
Fournisseur : The Natural Language Group of the USC Information Sciences Institute
Restrictions : Non disponible pour un usage commercial.

Débats de la chambre			**Débats du sénat**		
Paires de phrases	Mots anglais	Mots français	Paires de phrases	Mot anglais	Mots français
1.195 K	18.311 K	19.618 K	233 K	3.879 K	4.153 K

Tanaka
Langues : en jp
Taille : **180.000 paires de phrases.**
Fournisseur : Monash University (Yasuhito Tanaka & Jim Breen, auteur : Yasuhito Tanaka)
Restrictions : Non disponible pour un usage commercial.

XinHua News
Langues **:** en cn
Taille **: 29.000.000 mots**
Restrictions : Non disponible pour un usage commercial.

Corpus interrogeables en ligne

Compara Projet Linguateca
Langues : pt en
Taille : 62 paires de textes (fictions). Plus de 1 million de mots.
Interface : DISPARA System, IMS Corpus Query Processor
Processing : Alignement
Resource provider : Linguateca consortium.

LINEAR B
Langues : de en es fr
Taille : **39.314.085 mots**.
Interface : Moteur de recherche
Traitements : Alignement phrastique et au niveau des mots.

Liste des participants à la campagne d'évaluation IWSLT-06

ATT	AT&T Research
CLIPS	CLIPS-GETA
DCU	Dublin City University
HKUST	Hong Kong University of Science and Technology
IBM	IBM Research
ITC-irst	ITC-irst
JHU	WS06
Kyoto-Univ	Kyoto University
MIT-LL	AFRL MIT Lincoln Laboratory (jointly with Air Force Research Laboratories)
NAIST	Nara Insititute of Science and Technology
NiCT-ATR	National Institute of Information and Communication Technology / Advanced Telecommunication Research Labs
CASIA	NLPR National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
NTT	NTT Communication Science Laboratories
RWTH	RWTH Aachen University
SHARP	Laboratories of Europe Ltd.

TALP_comb	TALP Research Center
TCSTAR	TC-Star Project
UKACMU_SMT	InterAct Research Labs (UKA & CMU)
Washington-U	Washington University
Xiamen-Univ	Xiamen University

Résultat d'évaluation : scores de la campagne IWSLT-06

JE read speech			ASR output									
			official (with case + punctuation)					*additional (without case + punctuation)*				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155226325	RWTH	OPEN	0.2142	5.6502	0.457	0.706735	0.538419	0.2107	5.9364	0.4571	0.73181	0.544503
1155234850	NTT	OPEN	0.1984	5.4843	0.45	0.71083	0.551232	0.1906	5.7956	0.4498	0.739347	0.544439
1155125293	NiCT-ATR	OPEN	0.1899	5.5915	0.4574	0.698404	0.545865	0.1832	5.9428	0.4569	0.721945	0.537013
1155104201	MIT-LL/AFRL	OPEN	0.1891	5.5967	0.4421	0.703773	0.544396	0.1793	5.8474	0.4419	0.734907	0.548873
1155131162	UKACMU_SMT	OPEN	0.1868	5.6343	0.4505	0.729776	0.557492	0.1794	5.9031	0.4509	0.74737	0.5593
1155133582	ITC-irst	OPEN	0.1604	5.4171	0.4397	0.737696	0.569952	0.1617	5.8325	0.4392	0.767176	0.568281
1155230791	SLE	OPEN	0.1599	5.3393	0.4257	0.732526	0.593915	0.1467	5.6063	0.4257	0.777036	0.612982
1155046325	HKUST	OPEN	0.1523	4.9022	0.4283	0.723984	0.581833	0.1647	5.4583	0.4283	0.746029	0.559537
1155168398	Kyoto-U	OPEN	0.1418	4.8804	0.4057	0.744965	0.610351	0.1375	5.2416	0.4047	0.789824	0.628395
1155126275	TALP_comb	OPEN	0.139	4.7672	0.4105	0.742058	0.598199	0.1439	5.1993	0.4101	0.766813	0.587279
1154968590	TALP_tuples	OPEN	0.137	4.9437	0.4133	0.767411	0.60576	0.1397	5.3995	0.4134	0.79615	0.597872
1155135430	NAIST	OPEN	0.1311	4.8372	0.4415	0.851494	0.612372	0.136	5.3846	0.441	0.873737	0.591679
1155125181	TALP_phrases	OPEN	0.128	4.7596	0.4066	0.786361	0.620611	0.1343	5.2119	0.4066	0.816407	0.618639
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155124007	NiCT-ATR	CSTAR	0.2487	6.2778	0.5039	0.656886	0.511796	0.2468	6.7157	0.5032	0.67266	0.49944
1155131950	UKACMU_SMT	CSTAR	0.1841	5.398	0.4316	0.734222	0.568077	0.176	5.5606	0.4319	0.750194	0.571724

JE read speech			Correct Recognition Result									
			official (with case + punctuation)					*additional (without case + punctuation)*				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155226325	RWTH	OPEN	0.2368	5.9183	0.4886	0.685345	0.515105	0.2315	6.2325	0.489	0.710045	0.510728
1155234850	NTT	OPEN	0.2203	5.9077	0.4877	0.690154	0.52166	0.2147	6.3156	0.4873	0.710543	0.506451
1155125293	NiCT-ATR	OPEN	0.2122	5.9494	0.49	0.665719	0.518227	0.2077	6.3325	0.4893	0.682642	0.501853
1155104201	MIT-LL/AFRL	OPEN	0.2099	5.9866	0.4757	0.677691	0.517308	0.1995	6.257	0.4753	0.703922	0.511654
1155131162	UKACMU_SMT	OPEN	0.203	5.9322	0.482	0.701731	0.52641	0.1917	6.1468	0.4824	0.715567	0.519018
1155133582	ITC-irst	OPEN	0.1839	5.8538	0.4744	0.715961	0.542283	0.1882	6.3197	0.4742	0.734216	0.530725
1155230791	SLE	OPEN	0.1726	5.6497	0.457	0.704711	0.56383	0.1601	5.9765	0.4568	0.745407	0.576728
1155168398	Kyoto-U	OPEN	0.1655	5.4325	0.4497	0.709213	0.571467	0.1629	5.8843	0.4491	0.752287	0.583372
1155046325	HKUST	OPEN	0.156	4.875	0.4579	0.724861	0.578684	0.1786	5.7071	0.4578	0.722491	0.528143
1155126275	TALP_comb	OPEN	0.1467	4.9743	0.4382	0.723947	0.578991	0.1566	5.505	0.4383	0.742458	0.55615
1154968590	TALP_tuples	OPEN	0.1461	5.2717	0.4425	0.745624	0.578515	0.1495	5.8068	0.4426	0.767605	0.560742
1155135430	NAIST	OPEN	0.1431	5.2105	0.4664	0.837411	0.58428	0.1508	5.8442	0.4656	0.855201	0.552692
1155125181	TALP_phrases	OPEN	0.137	5.0665	0.4331	0.764481	0.599431	0.1451	5.5877	0.4333	0.792866	0.589125
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155124007	NiCT-ATR	CSTAR	0.2861	6.8327	0.5536	0.610427	0.470026	0.2867	7.3021	0.5529	0.619123	0.453303
1155131950	UKACMU_SMT	CSTAR	0.2007	5.8584	0.483	0.752029	0.568422	0.1956	6.334	0.4831	0.759016	0.551697

Table 17 : Scores d'évaluation JA-EN – dialogues oraux (lecture)

CE spontaneous speech			ASR output									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155228380	RWTH	OPEN	0.1898	5.0523	0.4198	0.711692	0.570519	0.1858	5.2331	0.4197	0.730619	0.568229
1155096930	JHU_WS06	OPEN	0.1807	5.1513	0.4138	0.706072	0.579835	0.1768	5.427	0.4134	0.721653	0.576531
1155103319	MIT-LL/AFRL	OPEN	0.1657	4.2363	0.38	0.69392	0.594289	0.1661	4.3968	0.3798	0.712099	0.590834
1155237182	UKACMU_SMT	OPEN	0.163	4.9732	0.4043	0.715468	0.587299	0.168	5.3175	0.4042	0.728691	0.585076
1155123252	NiCT-ATR	OPEN	0.1591	4.9696	0.4117	0.729099	0.585106	0.1615	5.3592	0.4114	0.748177	0.574682
1155132807	NTT	OPEN	0.1559	4.1801	0.3946	0.702046	0.597209	0.1584	4.5173	0.3945	0.716251	0.588109
1155134618	Xiamen-U	OPEN	0.1505	4.6813	0.3763	0.702426	0.597619	0.1623	4.9573	0.3768	0.714039	0.591011
1155041116	HKUST	OPEN	0.1441	4.6365	0.4238	0.718703	0.593497	0.1653	5.3703	0.4242	0.739327	0.574454
1155165080	ITC-irst	OPEN	0.1422	4.9188	0.4119	0.730687	0.591103	0.1577	5.4799	0.4121	0.754941	0.578239
1155187965	AT&T	OPEN	0.1155	4.1762	0.3584	0.740856	0.630585	0.1229	4.5258	0.3583	0.769892	0.63713
1155115125	NLPR	OPEN	0.107	3.5755	0.3901	0.752778	0.610846	0.1005	3.6311	0.3899	0.775843	0.630421
1155135353	CLIPS-GETA	OPEN	0.0344	2.7374	0.3178	0.87129	0.743063	0.0406	2.8625	0.3184	0.880529	0.720287
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155122729	NiCT-ATR	CSTAR	0.2008	5.4009	0.4502	0.69944	0.562867	0.2039	5.8205	0.4492	0.712915	0.548288
1155083137	UKACMU_SMT	CSTAR	0.1622	5.1865	0.418	0.740256	0.593868	0.1605	5.5303	0.4177	0.761752	0.588642
1155229211	UKACMU_SAMT	CSTAR	0.1566	4.6606	0.3833	0.726147	0.596676	0.1481	4.7742	0.3828	0.75114	0.603999
CE spontaneous speech			Correct Recognition Result									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155228380	RWTH	OPEN	0.2423	6.0961	0.5033	0.666757	0.509164	0.2446	6.4609	0.5031	0.679453	0.492935
1155103319	MIT-LL/AFRL	OPEN	0.2157	6.0537	0.4895	0.650633	0.523646	0.2178	6.4985	0.4892	0.667029	0.51075
1155096930	JHU_WS06	OPEN	0.214	6.0225	0.4802	0.686672	0.542865	0.2184	6.4992	0.4798	0.688366	0.519118
1155132807	NTT	OPEN	0.2135	5.1271	0.4743	0.65469	0.537013	0.2166	5.5115	0.4736	0.661781	0.521411
1155123252	NiCT-ATR	OPEN	0.206	5.8613	0.487	0.683672	0.531392	0.2123	6.3848	0.4862	0.694656	0.510536
1155237182	UKACMU_SMT	OPEN	0.1996	5.7603	0.4729	0.684295	0.545629	0.2045	6.185	0.4726	0.686572	0.534056
1155134618	Xiamen-U	OPEN	0.1976	5.564	0.4783	0.640033	0.527311	0.2162	5.9756	0.4791	0.64128	0.505612
1155165080	ITC-irst	OPEN	0.1837	5.8267	0.4852	0.686188	0.532487	0.1992	6.4263	0.4851	0.70517	0.515908
1155041116	HKUST	OPEN	0.1804	5.3615	0.4915	0.689997	0.548706	0.2038	6.2078	0.4917	0.703732	0.521579
1155187965	AT&T	OPEN	0.1439	4.8954	0.4164	0.705362	0.582178	0.1511	5.2806	0.4165	0.728955	0.583743
1155115125	NLPR	OPEN	0.1284	4.0658	0.4601	0.744201	0.572101	0.1237	4.2242	0.4597	0.767417	0.580604
1155135353	CLIPS-GETA	OPEN	0.0366	2.685	0.3178	0.858339	0.726484	0.0406	2.8625	0.3184	0.880529	0.720287
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155122729	NiCT-ATR	CSTAR	0.2645	6.5274	0.5425	0.637957	0.500277	0.2751	7.086	0.5419	0.638462	0.47659
1155083137	UKACMU_SMT	CSTAR	0.2057	6.0548	0.4987	0.69306	0.537082	0.2103	6.5941	0.4983	0.702533	0.515575
1155229211	UKACMU_SAMT	CSTAR	0.1954	5.7681	0.4642	0.695471	0.549197	0.1918	6.1137	0.4632	0.716292	0.544172

Table 18 : Scores d'évaluation CN-EN – dialogues oraux (spontanés)

AE read speech			ASR output									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155235676	IBM	OPEN	0.2274	5.8466	0.4845	0.604096	0.519818	0.2428	6.4867	0.4842	0.603489	0.495762
1154971705	TALP_tuples	OPEN	0.2136	5.8213	0.4786	0.644797	0.540339	0.2146	6.2598	0.4783	0.652289	0.524044
1155125830	NiCT-ATR	OPEN	0.2117	5.9216	0.4867	0.63541	0.527233	0.2164	6.3959	0.4869	0.640637	0.505552
1155128379	TALP_comb	OPEN	0.2101	5.5583	0.4747	0.63519	0.539492	0.2131	6.0012	0.474	0.636695	0.522438
1155134728	NTT	OPEN	0.2071	4.8403	0.4397	0.646681	0.566524	0.1967	4.7567	0.4384	0.643976	0.553638
1155171401	UKACMU_SMT	OPEN	0.1995	5.3359	0.4513	0.661549	0.563108	0.2086	5.6303	0.4511	0.655308	0.551311
1155127538	TALP_phrases	OPEN	0.1908	5.5448	0.4652	0.652624	0.550444	0.1989	6.0147	0.4646	0.658335	0.529825
1155132472	ITC-irst	OPEN	0.1723	4.7352	0.4186	0.659317	0.566232	0.178	5.1899	0.4182	0.67073	0.561896
1155042025	HKUST	OPEN	0.1477	3.3318	0.392	0.691637	0.594591	0.1584	3.7237	0.3911	0.701861	0.577144
1155056419	DCU	OPEN	0.145	4.5307	0.402	0.70271	0.594865	0.1391	4.7936	0.4	0.716453	0.586984
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155171760	UKACMU_SMT	CSTAR	0.2123	5.8693	0.4875	0.664348	0.550557	0.2234	6.3717	0.4873	0.655578	0.528926
AE read speech			Correct Recognition Result									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155235676	IBM	OPEN	0.2549	6.3769	0.5316	0.566817	0.48249	0.2773	7.1681	0.5314	0.559349	0.448014
1155125830	NiCT-ATR	OPEN	0.2365	6.3521	0.5224	0.611186	0.498561	0.2463	6.8893	0.5229	0.610536	0.473432
1155128379	TALP_comb	OPEN	0.2327	6.0337	0.5091	0.615175	0.5144	0.2395	6.5972	0.5087	0.612122	0.492051
1154971705	TALP_tuples	OPEN	0.2323	6.238	0.5134	0.627078	0.515304	0.2383	6.7958	0.5133	0.628076	0.494075
1155134728	NTT	OPEN	0.2265	5.3316	0.4776	0.627904	0.541468	0.2216	5.3577	0.4758	0.617943	0.519164
1155171401	UKACMU_SMT	OPEN	0.2208	5.9059	0.4932	0.635725	0.531515	0.2349	6.3037	0.4929	0.625825	0.51218
1155127538	TALP_phrases	OPEN	0.2122	6.0177	0.501	0.628217	0.516571	0.222	6.5405	0.5004	0.630598	0.492895
1155132472	ITC-irst	OPEN	0.2005	5.1816	0.4581	0.632154	0.538922	0.2048	5.604	0.4564	0.640152	0.530242
1155042025	HKUST	OPEN	0.1663	3.8863	0.4288	0.675709	0.564738	0.18	4.4473	0.4273	0.681345	0.539727
1155056419	DCU	OPEN	0.1624	4.8902	0.4336	0.685969	0.567762	0.1589	5.29	0.432	0.693548	0.553715
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155171760	UKACMU_SMT	CSTAR	0.242	6.4073	0.5275	0.633797	0.518276	0.2584	6.9741	0.5276	0.616925	0.48619

Table 19 : Scores d'évaluation AR-EN – dialogues oraux (lecture)

CE read speech			ASR output									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155226839	RWTH	OPEN	0.2111	5.4045	0.4432	0.69523	0.553068	0.2032	5.5644	0.443	0.716762	0.554471
1155291791	TC-STAR	OPEN	0.1999	5.5858	0.4596	0.69155	0.547364	0.2076	6.0328	0.4593	0.706488	0.536375
1155098594	JHU_WS06	OPEN	0.1863	5.329	0.4276	0.706984	0.580715	0.1894	5.6685	0.4275	0.714461	0.568927
1155232365	MIT-LL/AFRL	OPEN	0.1861	5.4154	0.4365	0.680442	0.560855	0.1877	5.7698	0.4354	0.702947	0.557487
1155234313	NTT	OPEN	0.1834	4.5306	0.4215	0.684383	0.577053	0.1876	4.9075	0.4212	0.700348	0.566985
1155124539	NiCT-ATR	OPEN	0.1775	5.2286	0.4336	0.714689	0.570451	0.1772	5.6649	0.4323	0.729099	0.55831
1155185909	UKACMU_SMT	OPEN	0.171	5.0768	0.4227	0.707115	0.579163	0.1738	5.3809	0.4222	0.719066	0.575724
1155124036	TALP_comb	OPEN	0.165	4.8933	0.4266	0.697715	0.584579	0.1728	5.3577	0.4264	0.708529	0.5706
1155122619	TALP_tuples	OPEN	0.1624	4.9779	0.4336	0.711726	0.59719	0.1748	5.5128	0.4333	0.720037	0.582349
1155122791	TALP_phrases	OPEN	0.1599	5.1255	0.4307	0.721823	0.588739	0.1676	5.6413	0.4307	0.736284	0.573534
1155135300	Xiamen-U	OPEN	0.1579	5.0115	0.4049	0.695798	0.585706	0.1718	5.3595	0.4052	0.708482	0.576843
1155131203	ITC-irst	OPEN	0.156	5.2207	0.4374	0.720064	0.5786	0.1698	5.7435	0.437	0.740184	0.565402
1155041426	HKUST	OPEN	0.1545	4.7769	0.4456	0.711659	0.582048	0.174	5.4981	0.4457	0.730733	0.563087
1155189312	AT&T	OPEN	0.1226	4.3813	0.3729	0.734056	0.616298	0.1297	4.7122	0.3733	0.762938	0.626844
1155115565	NLPR	OPEN	0.1037	3.6384	0.4073	0.776478	0.609988	0.1022	3.7433	0.4076	0.802199	0.625748
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155121721	NiCT-ATR	CSTAR	0.2155	5.6857	0.4787	0.673348	0.544281	0.2214	6.1453	0.4783	0.681398	0.530454
1155230165	UKACMU_SAMT	CSTAR	0.1685	5.0292	0.4111	0.721987	0.585203	0.163	5.2834	0.4108	0.748615	0.58486
1155083859	UKACMU_SMT	CSTAR	0.1645	5.2372	0.4315	0.736689	0.586606	0.1647	5.6395	0.4308	0.753263	0.584802

CE read speech			Correct Recognition Result									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155226839	RWTH	OPEN	0.2423	6.0961	0.5033	0.666757	0.509164	0.2446	6.4609	0.5031	0.679453	0.492935
1155291791	TC-STAR	OPEN	0.2409	6.4004	0.5182	0.654457	0.498004	0.2421	6.8678	0.5181	0.668064	0.481564
1155232365	MIT-LL/AFRL	OPEN	0.2157	6.0537	0.4895	0.650633	0.523646	0.2178	6.4985	0.4892	0.667029	0.51075
1155098594	JHU_WS06	OPEN	0.214	6.0225	0.4802	0.686672	0.542865	0.2184	6.4992	0.4798	0.688366	0.519118
1155234313	NTT	OPEN	0.2135	5.1271	0.4743	0.65469	0.537013	0.2166	5.5115	0.4736	0.661781	0.521411
1155124539	NiCT-ATR	OPEN	0.206	5.8613	0.487	0.683672	0.531392	0.2123	6.3848	0.4862	0.694656	0.510536
1155185909	UKACMU_SMT	OPEN	0.1996	5.7603	0.4729	0.684295	0.545629	0.2045	6.185	0.4726	0.686572	0.534056
1155135300	Xiamen-U	OPEN	0.1976	5.564	0.4783	0.640033	0.527311	0.2162	5.9756	0.4791	0.64128	0.505612
1155124036	TALP_comb	OPEN	0.1916	5.398	0.4749	0.66703	0.545836	0.2021	5.9698	0.4749	0.671562	0.523817
1155122791	TALP_phrases	OPEN	0.1899	5.803	0.4833	0.687675	0.54533	0.2008	6.4275	0.4832	0.694902	0.520496
1155122619	TALP_tuples	OPEN	0.1883	5.5714	0.4824	0.680412	0.550282	0.2034	6.2119	0.4825	0.682969	0.528145
1155131203	ITC-irst	OPEN	0.1837	5.8267	0.4852	0.686188	0.532487	0.1992	6.4263	0.4851	0.70517	0.515908
1155041426	HKUST	OPEN	0.1804	5.3615	0.4915	0.689997	0.548706	0.2038	6.2078	0.4917	0.703732	0.521579
1155189312	AT&T	OPEN	0.1439	4.8954	0.4164	0.705362	0.582178	0.1511	5.2806	0.4165	0.728955	0.583743
1155115565	NLPR	OPEN	0.1284	4.0858	0.4601	0.744201	0.572101	0.1237	4.2242	0.4597	0.767417	0.580604
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155121721	NiCT-ATR	CSTAR	0.2645	6.5274	0.5425	0.637957	0.500277	0.2751	7.086	0.5419	0.638462	0.47659
1155083859	UKACMU_SMT	CSTAR	0.2057	6.0548	0.4987	0.69306	0.537082	0.2103	6.5941	0.4983	0.702533	0.515575
1155230165	UKACMU_SAMT	CSTAR	0.1954	5.7681	0.4642	0.695471	0.549197	0.1918	6.1137	0.4632	0.716292	0.544172

Table 20 : Scores d'évaluation CN-EN – dialogues oraux (lecture)

IE read speech			ASR output									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155126488	NiCT-ATR	OPEN	0.2989	6.8985	0.5744	0.550034	0.464104	0.3194	7.4724	0.5739	0.534282	0.42654
1155119263	TALP_comb	OPEN	0.2837	6.7065	0.566	0.55352	0.469733	0.3067	7.3139	0.5657	0.537187	0.441328
1154969301	TALP_tuples	OPEN	0.2818	6.8723	0.5764	0.558887	0.470112	0.3067	7.5256	0.5761	0.539096	0.43687
1155131694	MIT-LL/AFRL	OPEN	0.2798	6.8593	0.5679	0.560163	0.465614	0.3007	7.507	0.5678	0.549485	0.431335
1155135070	ITC-irst	OPEN	0.2797	6.6217	0.5592	0.545169	0.466213	0.2969	7.2595	0.5588	0.535623	0.437763
1155027056	Washington-U	OPEN	0.2787	6.9318	0.5853	0.558708	0.467647	0.3168	7.6902	0.5853	0.531756	0.421145
1155233520	NTT	OPEN	0.2769	6.6959	0.5607	0.569973	0.481267	0.2864	7.1949	0.5602	0.542903	0.438815
1155118457	TALP_phrases	OPEN	0.2684	6.6443	0.5634	0.575244	0.483244	0.294	7.2944	0.5631	0.558425	0.449467
1155129266	UKACMU_SMT	OPEN	0.2388	6.1999	0.5376	0.587084	0.496415	0.2577	6.823	0.5371	0.579014	0.470541
1155042542	HKUST	OPEN	0.2374	6.0956	0.5403	0.630773	0.493824	0.2778	7.0994	0.5398	0.627084	0.447516
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155130664	UKACMU_SMT	CSTAR	0.263	6.6617	0.5638	0.576683	0.475417	0.2826	7.3188	0.5633	0.561875	0.446853

IE read speech			Correct Recognition Result									
			official (with case + punctuation)					additional (without case + punctuation)				
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155126488	NiCT-ATR	OPEN	0.3763	8.1318	0.663	0.47382	0.390051	0.412	8.9027	0.6625	0.445099	0.341506
1155131694	MIT-LL/AFRL	OPEN	0.3574	8.0089	0.6669	0.488242	0.392662	0.392	8.8548	0.6669	0.461672	0.338203
1155027056	Washington-U	OPEN	0.3543	8.189	0.7017	0.483471	0.389264	0.4206	9.241	0.7019	0.428617	0.317528
1155135070	ITC-irst	OPEN	0.3497	7.8155	0.6468	0.482177	0.399808	0.3797	8.6186	0.6461	0.456004	0.352706
1155233520	NTT	OPEN	0.3449	7.8259	0.6431	0.507859	0.415732	0.375	8.5266	0.6428	0.461588	0.35743
1155119263	TALP_comb	OPEN	0.3396	7.6405	0.6332	0.496796	0.416311	0.3774	8.4035	0.6328	0.465531	0.37287
1154969301	TALP_tuples	OPEN	0.3331	7.7474	0.6398	0.506168	0.419052	0.3738	8.5922	0.6394	0.469935	0.369025
1155118457	TALP_phrases	OPEN	0.32	7.5248	0.6256	0.5206	0.429816	0.3555	8.3201	0.6254	0.491162	0.381729
1155129266	UKACMU_SMT	OPEN	0.303	7.3011	0.6293	0.522254	0.425916	0.3419	8.1405	0.6286	0.494123	0.37936
1155042542	HKUST	OPEN	0.2964	7.1816	0.6239	0.580833	0.434089	0.3567	8.3486	0.6236	0.558001	0.367741
timestamp	system_name	track	BLEU4	NIST	METEOR	WER	PER	BLEU4	NIST	METEOR	WER	PER
1155130664	UKACMU_SMT	CSTAR	0.3312	7.7622	0.6587	0.509091	0.406532	0.3756	8.6779	0.6583	0.474094	0.356119

Table 21 : Scores d'évaluation IT-EN – dialogues oraux (lecture)

Résumé

Les travaux de recherche présentés dans cette thèse s'inscrivent dans le domaine de la traduction automatique et automatisée. Nous nous intéressons en particulier aux outils d'aide à la traduction sur le Web favorisant la traduction bénévole, non-commerciale et incrémentale. La croissance remarquable de la quantité des documents multilingues disséminés en ligne gratuitement (W3C, Traduct, Mozilla, traducteurs bénévoles de documents en droits de l'homme, etc.) est le résultat d'un travail laborieux de communautés bénévoles qui sont malheureusement technologiquement marginalisées, et privées de toute aide « linguistique » à la traduction. Nous avons effectué une étude approfondie des communautés bénévoles et en avons dégagé les problèmes les plus intéressants et difficiles.

Nous avons construit BEYTrans, un environnement collaboratif et non commercial offrant des services linguistiques et répondant aux besoins spécifiques de ces communautés. Les principales composantes logicielles développées sont : un éditeur multilingue Web intégrant des fonctionnalités linguistiques avancées et variées, un module de recyclage des traductions existantes, et la gestion et le traitement d'une masse de données multilingues. Chacun d'eux résout un problème intéressant, et ils sont entièrement intégrés (avec d'autres fonctionnalités) dans un environnement collaboratif complet, qui a été expérimenté avec succès dans des situations concrètes.

Abstract

The research work presented in this thesis belongs to the machine and machine-aided translation field. We are particularly interested in tools that help and promote online, free, non-commercial and incremental translation. The staggering growth of multilingual document translations on the Web (W3C, Traduct, free translation of human right documents, Mozilla, etc.) is the result of painstaking work of volunteer communities who are unfortunately technologically marginalized and deprived of the possibility to use any linguistic technology and computer-aided translation systems (CAT). We have conducted an in-depth study of volunteer communities and identified the most interesting and challenging problems.

We have built BEYTrans, a collaborative, non-commercial environment that offers linguistic helps and meets the specific needs of these communities. Its 3 main modules are: (i) a multilingual Web editor integrating advanced and varied linguistic functionalities, (ii) a module for recycling existing translations and (iii) a module for managing and processing large multilingual data. Each solves a challenging problem, and they are fully integrated (with other functionalities) in a collaborative environment that has been successfully tested in real-world situations.