Université de la Sorbonne nouvelle – Paris 3
École doctorale : *LANGAGE & LANGUES : Description, Théorisation, Transmission*
Discipline : Sciences du langage

Titre : « APPROCHES QUANTITATIVES DE L'EXTRACTION DE RESSOURCES TRADUCTIONNELLES A PARTIR DE CORPUS PARALLELES »

Auteur : Maria Zimina-Poirot

Jury :

M. Eric Gaussier (Xerox Research Center Europe) – Examinateur
M. Benoît Habert (Université Paris X – Nanterre) – Rapporteur
M. André Salem (Université de la Sorbonne nouvelle – Paris 3) – Directeur de thèse
Mme Monique Slodzian (Institut National des Langues et Civilisations Orientales) – Rapporteur

Date de soutenance : le 26 novembre 2004
Lieu de soutenance : Centre Censier, Université de la Sorbonne nouvelle – Paris 3

Abstract

This research work presents the results of a series of experiments devoted to the development of new tools for intertextual textometric exploration of translation corpora. Various methods of textual statistics have been adapted for use in a multilingual context and put into practice for parallel text processing, such as: *repeated segments extraction*, *characteristic elements computation*, *bi-textual topography*, *multiple co-occurrences*, *factorial analysis*, *automatic classification*, etc. Examples of concrete applications illustrate the use of each of these methods in a multilingual context. These examples are accompanied by sample translation resources obtained on quantitative bases from the parallel French/English corpus of the *Convention for the Protection of Human Rights*. The suggested approach opens up new horizons for automatic exploration of lexical equivalences of translation corpora by a variety of users: translators, foreign language teachers, terminologists, lexicographers, etc.

Keywords: alignment, bi-text, parallel corpora, textometrics, textual statistics, textual topography, translation correspondences.

Résumé

Dans le contexte récent de l'informatisation de la société, chercheurs et praticiens sont confrontés à une croissance spectaculaire des corpus de textes multilingues provenant d'archives de textes traduits, de bases documentaires multilingues numérisées, de sites Web internationaux. Pour des raisons variées, diverses communautés s'intéressent aux données textuelles multilingues. Les *historiens*, les *juristes*, les *philologues* analysent les corpus multilingues avec des outils d'exploration permettant d'observer plus finement les correspondances entre différents volets de corpus comparés. Les *informaticiens* utilisent les ressources obtenues à partir de ces corpus pour améliorer les performances des outils de traduction automatique ou des moteurs de recherche pour le Web. Enfin, les ressources traductionnelles obtenues à partir des corpus multilingues sont utilisées avec profit pour les

études menées dans le cadre de plusieurs disciplines des *sciences du langage* qui s'étendent de la *linguistique contrastive* à la *lexicographie*, de la *traduction assistée par ordinateur* à l'*enseignement des langues étrangères*, de *l'analyse du discours* à la *linguistique computationnelle*.

La croissance spectaculaire des données textuelles multilingues rend toujours plus actuelle la nécessité de disposer d'outils de traitement automatique de corpus dans des langues différentes. Ce travail présente les résultats d'une série de recherches consacrées au développement d'une nouvelle famille d'outils d'exploration textométrique intertextuelle. De nombreuses méthodes de statistique textuelle ont été articulées et adaptées au cadre multilingue : la méthode des *segments répétés*, les *spécificités*, la *topographie bi-textuelle*, les *cooccurrences multiples*, l'*analyse factorielle des correspondances*, la *classification automatique*, etc. L'utilisation de chaque méthode dans le contexte multilingue est illustrée par des exemples d'applications, accompagnés d'échantillons de ressources traductionnelles obtenues à partir du corpus parallèle français/anglais de la *Convention de sauvegarde de Droits de l'Homme*.

Le travail comporte deux grandes parties. La première décrit les enjeux de l'analyse automatique de corpus multilingues ainsi que les acquis obtenus par les principaux courants de recherche du domaine du traitement automatique des langues (chapitre 1-2).

Le chapitre 1 tente de cerner le concept de *parallélisme textuel* dans le contexte multilingue. Le lecteur y trouvera des exemples de corpus parallèles composés de textes sources et de leurs traductions (effectuées par des traducteurs humains) ou de textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source.

Dans la première partie du chapitre 2, sont recensés les problèmes nés dans le contexte de la segmentation de corpus parallèles en équivalences traductionnelles. Des exemples montrent la difficulté de déterminer des mécanismes formels permettant d'automatiser cette segmentation au niveau lexical. La deuxième partie décrit les principales méthodes d'alignement automatique de corpus. On y trouvera la description et la comparaison de quelques grandes familles d'algorithmes d'alignement automatique développés au cours des vingt dernières années.

La deuxième partie (chapitres 3-7) présente les fondements de l'analyse textométrique des corpus multilingues et décrit les applications textométriques mises au point pour l'extraction de ressources traductionnelles à partir de corpus parallèles.

Développées dans le contexte monolingue, les pratiques de l'analyse textométrique de corpus se révèlent particulièrement adaptées à la recherche automatique des équivalences du bi-texte. Dans le cas des corpus parallèles bilingues, la textométrie aide à mettre en relation différents *types* d'unités textuelles entre les deux volets. L'approche quantitative permet d'établir des correspondances aussi bien entre les paragraphes et les phrases, qu'au niveau lexical. Grâce à cette approche, on parvient à mettre en relation des *formes graphiques* isolées, des *lexèmes*, des *structures lexicales récurrentes* sur l'axe syntagmatique, etc.

Les méthodes quantitatives convoquées dans ce travail reposent entièrement sur des ressources construites automatiquement *à base de corpus*. Ces méthodes s'appuient sur des algorithmes qui utilisent les fréquences et les distributions des unités textuelles prises comme points de repère pour l'identification et l'extraction des correspondances.

La comparaison des fréquences des unités textuelles recensées dans les deux volets bilingues du corpus est souvent insuffisante pour détecter les correspondances traductionnelles au niveau lexical. Les différents *sens* dans lesquels un lexème est employé dans un contexte donné induisent la plupart du temps autant de traductions différentes. Les mots dotés d'un large éventail de sens dans le corpus forment des réseaux de correspondances souvent complexes. Ces facteurs entraînent des écarts entre les fréquences des unités équivalentes prises dans des contextes particuliers.

La notion de *résonance textuelle* est alors mobilisée pour mieux cerner les rapports de correspondances entre les lexèmes en fonction des variations contextuelles. Le processus de *résonance textuelle* amorcé par la sélection dans le texte source des sections dans lesquelles les occurrences d'une unité textuelle (*forme, segment répété, patron morpho-syntaxique*) dépassent un seuil fixé, induit une sélection topographique correspondante dans le texte cible et met en évidence des séquences, liées à l'unité de départ, sur le plan de la traduction. Le processus de résonance textuelle peut être enclenché par localisation *topographique* de fragments thématiques du *bi-texte*. Cette exploration topographique s'enrichit des résultats de l'alignement des deux volets bilingues du corpus au niveau de la phrase. Une description automatique des relations d'équivalence multiples entre unités bilingues peut être obtenue par le biais d'appariements statistiques lorsque l'exploration du corpus s'appuie sur un alignement des phrases. Cette approche peut être utilisée pour le repérage des équivalences lexicales y compris dans le cas où leurs fréquences dans le corpus sont peu élevées.

L'exploration topographique de ressources traductionnelles peut être complétée par des approches cooccurrentielles et, notamment, par le calcul des *réseaux de cooccurrences parallèles* (chapitre 6). Les possibilités de navigation intertextuelle ouvertes par cette approche facilitent la mise en évidence de phénomènes traductionnels complexes, relevant de différents niveaux de l'analyse linguistique : la variation des traductions d'un terme en fonction des contextes, le repérage thématique d'équivalences lexicales, la découverte de constellations lexicales parallèles, etc. L'observation de ces phénomènes est susceptible d'enrichir la pratique quotidienne des *traducteurs, lexicographes, terminologues, enseignants en langues étrangères, spécialistes de l'analyse de discours*, etc.

Le dernier chapitre aborde des perspectives de recherche peu explorées jusqu'ici et, en premier lieu, les perspectives d'analyse textométrique de *corpus parallèles catégorisés* (chapitre 7). L'étiquetage de corpus parallèles offre des points d'appui précieux pour l'extraction de ressources traductionnelles du bi-texte. Cependant, une homogénéisation des jeux d'étiquettes morphosyntaxiques utilisés pour la catégorisation de deux volets bilingues d'un corpus parallèle se révèle nécessaire avant l'exploration bi-textuelle.

L'éclairage quantitatif permet de construire des analyses nuancées de ressources textuelles multilingues. Le succès pratique des méthodes d'exploration élaborées au fil de ces recherches, nous a incitée à produire des maquettes de logiciels d'exploration textométrique intertextuelle. Ces maquettes sont fournies sur le Cd-rom qui accompagne ce travail.

Les méthodes quantitatives donnent accès à un réservoir de renseignements précieux qui permettent d'enrichir les pratiques actuelles de l'analyse des données de traduction. Nous sommes convaincue que les outils de la statistique textuelle utilisés à partir de postes de travail équipés de ressources traductionnelles informatisées trouverons très rapidement de nombreuses applications en sciences du langage.

Mots-clés : alignement, bi-texte, corpus parallèles, correspondances traductionnelles, statistique textuelle, textométrie, topographie textuelle.