AIX-MARSEILLE UNIVERSITE

****

Université de Provence

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|

THESE

pour obtenir le grade de

DOCTEUR D'AIX-MARSEILLE UNIVERSITE

Formation doctorale :

Cognition, Langage, Education (mention Traitement Automatique)

Présentée et soutenue publiquement

Par

Stéphanie LEON

le lundi 8 décembre 2008

TITRE :

ACQUISITION AUTOMATIQUE DE TRADUCTIONS D'UNITES LEXICALES COMPLEXES A PARTIR DU WEB

Tome I

Directeur de thèse :

Jean VERONIS

JURY

Mme Béatrice DAILLE (Université de Nantes, examinateur)

Mme Violaine PRINCE (Université de Montpellier 2, rapporteur)

Mme Pascale SEBILLOT (INSA, Rennes, rapporteur)

M. Jean VERONIS (Université de Provence, directeur)

Remerciements

Cette thèse est le fruit d'un travail interdisciplinaire, mêlant à la fois une culture linguistique et informatique, qui m'ont été transmises au contact de personnes dont le panel varié des domaines de recherche ont constitué un éventail d'échanges enrichissants.

Je pense à mon directeur de thèse, Jean Véronis, dont l'implication et le soutien ont été sans limite, depuis mon arrivée à l'Université de Provence. C'est grâce à sa passion et à son enthousiasme pendant mes premiers cours de licence que je me suis orientée vers des études en traitement automatique de la sémantique lexicale. Tout au long de mon parcours universitaire, il a su m'apporter de précieux conseils et m'a transmis son expérience et sa rigueur du travail. Je le remercie également pour son soutien psychologique tout au long de mes années de thèse et ses échanges qui m'ont toujours stimulée et remotivée dans les moments de doute.

En ce qui concerne mon cadre de recherche, je remercie tous les membres de l'ancienne équipe DELIC (Description Linguistique Informatisée sur Corpus), nouvellement TALEP (Traitement Automatique du Langage Ecrit et Parlé), qui m'a accueillie durant mes années de thèse, à l'Université de Provence. Je remercie Estelle Véronis, pour son amitié, son soutien et le partage de son expérience. Elle a toujours su être à l'écoute et m'apporter des conseils avisés lorsque j'en ai eu besoin. Je pense à Laure Brieussel pour son enthousiasme et ses conseils. Je remercie les autres doctorants avec qui j'ai eu le plaisir de travailler au quotidien, et de partager doutes, expérience et bonne humeur, Chrystel Millon pour sa complicité, Alice Carne, Rémi Bove, Emilie Chételat et Loïc Kervajean. Je pense enfin à tous les autres membres de l'équipe.

Je remercie également l'équipe du LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier), qui est mon équipe de recherche d'accueil pour cette année en cours, durant ma dernière année d'ATER. Je remercie tout particulièrement Violaine Prince pour son accueil très chaleureux, pour son enthousiasme stimulant et pour sa participation active en tant que rapporteur au sein de mon jury de thèse. Je pense également

à Mathieu Lafourcade dont l'accueil et la passion pour la recherche m'a apporté un cadre de travail convivial. Je pense également à tous les autres membres de l'équipe, qui m'ont réservé un accueil enthousiaste.

En ce qui concerne mon cadre d'enseignement, je remercie l'équipe du CILSH (Centre Informatique pour les Lettres et Sciences Humaines) de l'Université de Provence, qui a su m'apporter une atmosphère de travail absolument remarquable, par sa bonne humeur et son soutien sans faille, durant mes trois années de monitorat et ma première année d'ATER : Christophe Mathieu, pour son amitié et ses conseils toujours avisés et stimulants, Gérard Della Ragione pour son précieux soutien et son encadrement durant mes premières années d'enseignement, Marie-Thérèse Ponsonnet pour sa bonne humeur communicative, Jean-Luc Péris pour sa présence chaleureuse, sans oublier tous les autres membres de l'équipe. J'ai également une pensée émue à la mémoire d'Henri Tournier.

Je remercie également l'équipe MIAp (Mathématiques Informatique Appliquées) de l'université Montpellier III, où j'ai actuellement le plaisir d'enseigner, pour ma dernière année d'ATER et dans laquelle j'ai été agréablement accueillie. Je remercie Christian Lavergne et Patrice Séébold qui m'ont permis de m'intégrer dans le département MIAp, pour leur accueil amical et chaleureux. Je pense naturellement à mes collègues enseignants d'informatique, Sandra Bringay, Alexandre Pinlou, Sylvain Durand, Joël Quinqueton et Fabrice Philippe, dont la complicité et les nombreux conseils m'ont aidé à m'intégrer rapidement, dans une chaleureuse ambiance, ainsi que tous les autres membres de l'équipe, pour leur sympathique accueil.

Je remercie également Christophe Rey, de l'Université d'Amiens, pour son amitié, ses discussions toujours avisées et le partage de son expérience. Je pense aussi à Louis-Jean Calvet, de l'Université de Provence, avec qui j'ai le plaisir de travailler sur un autre domaine qui me passionne, celui de l'analyse linguistique des textes de chanson. Son expérience, sa passion, ainsi que son extraordinaire culture m'ont beaucoup enrichie et stimulé, au cours de nos entrevues.

Je remercie naturellement Pascale Sébillot et Béatrice Daille, d'avoir accepté, avec Violaine Prince, de faire partie de mon jury de thèse. Leur enthousiasme à toutes les trois, ainsi que

leur expérience et leurs remarques toujours rigoureuses et pertinentes, m'ont offert un cadre particulièrement stimulant pour terminer ma thèse.

Je remercie également Amanda Grey, qui a eu la grande amabilité de s'impliquer dans l'évaluation des traductions obtenues dans mon travail de thèse. Sa rigueur et ses compétences m'ont permis de cibler avec précision les caractéristiques de traduction qui posent des difficultés au traitement automatique et de porter un regard à la fois quantitatif et qualitatif sur mes résultats.

Enfin, je remercie ma famille et mes amis pour leur patience et leur soutien inconditionnel, tout particulièrement mes parents, mon frère Sébastien et Lynda. Je remercie Alizée pour tout ce qu'elle m'a apporté. Un clin d'œil à L-Town et à la planète Mars. Je remercie Bruno pour sa présence et son soutien quotidien et sans limite.

A la mémoire de mon « grand-père » Salvador et de ma tante Jeanine.

Table des matières

Chapitre1. Introduction

1.1 Motivation

Bien qu'elle ait été la première application non-numérique de l'informatique, la traduction automatique a connu des débuts décevants qui ont jeté un discrédit sur cette technologie pendant plusieurs décennies. Toutefois, des progrès considérables ont été accomplis au cours de ces dernières années, en particulier à cause de l'explosion du Web dans un contexte fortement multilingue. A l'heure actuelle, les logiciels de traduction en ligne, accessibles au grand public, permettent de déchiffrer le thème et le contenu global de textes courants dans une autre langue. Des erreurs, parfois grossières, demeurent, et l'on est très loin de traductions de qualité professionnelle, mais les résultats obtenus sont malgré tout intéressants pour un large public souhaitant prendre connaissance d'informations dans des langues totalement inconnues, ou de professionnels cherchant à déchiffrer rapidement des documents dans le cadre de la veille technologique ou stratégique. Ces progrès récents sont essentiellement dus à l'accroissement très important de la couverture des dictionnaires présents dans les systèmes, et à la prise en compte d'un nombre croissant d'expressions composées. Par exemple, le

système *Systran*[1] traduit désormais correctement du français vers l'anglais des expressions figées telles que :

vol à main armée > armed robbery

vol à la roulotte > stealing from parked vehicles

vol à la tire > pick-pocketing

vol à voile > gliding

vol régulier > scheduled flight

Malgré tout, dès que l'on sort de ces listes d'expressions figées, on retombe rapidement dans des erreurs de traductions qui gênent considérablement la compréhension, et lui donnent même parfois un caractère surréaliste. Par exemple, *Systran* utilise la traduction la plus fréquente du mot *vol,* c'est-à-dire *flight* (usage *VOL AERIEN*), dans toutes les autres situations. Si *vol aérien* est correctement traduit (*air flight*), *vol de voitures* est traduit par *flight of cars,* ce qui est totalement incompréhensible pour un anglophone. Pourtant, la co-occurrence *vol-voitures* est un indice désambiguïsateur fort, qui, si elle était correctement enregistrée dans une base de données, pourrait servir à générer des traductions correctes. La combinatoire est toutefois beaucoup plus ouverte qu'avec les expressions figées mentionnées plus haut et la constitution manuelle d'une base de données de co-occurrences lexicales préférentielles, à très grande échelle, est une tâche à peu près impossible. Les dictionnaires bilingues se contentent d'ailleurs de rares indications ponctuelles sur la traduction des unités lexicales complexes, se fiant au jugement du lecteur et à sa connaissance du monde, que l'on ne peut guère espérer d'une machine.

En ce qui concerne l'acquisition automatique de lexiques bilingues, nous distinguons les travaux s'appuyant sur les corpus parallèles ou alignés (Véronis, 2000a) de ceux, plus récents,

[1] http://www.systransoft.com/

s'appuyant sur les corpus comparables (Rapp, 1999, Fung et McKeown, 1997, Fung et Yee, 1998, Morin *et al.*, 2004). Les corpus parallèles souffrent d'un manque de disponibilité. Les corpus comparables, plus accessibles, se limitent généralement à un domaine de spécialité, ce qui restreint la diversité des usages. Le Web, qui génère des besoins considérables en traduction, offre en même temps un réservoir gigantesque de données qui peut être exploité par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que *Google*[1] ou *Yahoo*[2]. Le Web permet de palier les limites des corpus parallèles et comparables : il constitue une base lexicale gigantesque, accessible librement, pour une multitude de domaines et de langues (Kilgarriff et Grefenstette, 2003). L'utilisation du Web permet un changement d'échelle dont les répercussions peuvent être fondamentales pour la compréhension des langues. Toutefois, son utilisation constitue un phénomène récent, non complètement maîtrisé et nécessite des réflexions théoriques et pratiques sur son statut, ses caractéristiques et ses limites pour la recherche linguistique.

1.2 Objectifs et méthodologie

L'objectif de cette thèse est la mise au point de techniques d'extraction automatique d'équivalences bilingues d'unités lexicales complexes[3], à partir du Web, dans un but de construction d'une très grande base de données du français vers l'anglais. De telles bases de données, contenant une quantité massive de traductions, constituent des ressources rares, pourtant fondamentales pour des applications telles que la traduction automatique, la recherche d'information multilingue ou la lexicographie et la terminologie bilingue. Notre méthodologie exploite les différentes facettes du « Web bilingue » afin d'acquérir de façon automatique des équivalences de telles traductions. La base de données constituée n'est pas de

[1] http://www.google.com/

[2] http://www.yahoo.com/

[3] Concernant la combinatoire lexicale, la littérature présente une terminologie disparate et souvent floue. Certains parlent de « préférences lexicales » (Wilks, 1975), de « restrictions de sélection » (Katz et Fodor, 1964), de « termes complexes » (Daille, 1994) ou encore de « collocations » (Benson, 1990, Smadja, 1993, Cruse, 1986). Afin de désigner ce phénomène, nous employons ici le terme d'unité lexicale complexe, plus neutre, défini comme une cooccurrence lexicale entre deux lexèmes liés syntaxiquement.

type dictionnairique, elle ne vise pas à l'exhaustivité et à la description du lexique. Elle constitue un recensement non exhaustif et non fermé, dont l'objectif est une augmentation croissante et quotidienne.

Etant donné l'ampleur du phénomène, nous nous centrons sur un couple de langue (français-anglais), sur une direction de traduction (du français vers l'anglais)[1], et sur deux relations de dépendances syntaxiques en français (limitées à deux « mots-pleins »[2]) :

NOM ADJ

appareil digital

NOM DE(D') NOM

appareil de musculation

Ces champs d'investigation peuvent être élargis par la suite. Nous prenons pour point de départ des unités lexicales complexes en français collectées à partir d'un vaste corpus de pages Web et nous proposons une méthodologie par traitements modulaires, chaque phase étant ciblée sur des caractéristiques particulières de traduction (compositionnalité ou non-compositionnalité, polysémie des unités lexicales, etc.). Nous étudions la langue générale dans son ensemble, celle-ci incluant les domaines de spécialité[3].

Nous espérons apporter modestement quelques techniques afin de construire de façon automatique un vaste lexique bilingue d'unités lexicales complexes attestées, à partir de la « base de textes » du Web. Nous visons à la construction d'un lexique ayant un champ étendu de traitement, tant quantitativement, qu'en matière de diversité des domaines. Ce lexique

[1] Nous parlons de langue source afin de désigner la langue à traduire (pour nous, le français) et de langue cible afin de parler de la langue de traduction (pour nous, l'anglais).

[2] Dans le décompte des mots-pleins, les prépositions telles que *de* ou *d'* ne sont pas prises en compte, c'est-à-dire que *appareil numérique* contient deux mots-pleins, tout comme *appareil de musculation* ou *cachet d'aspirine* (Daille, 1994).

[3] Harris (1991) parle de « sous-langage » , qui est une notion proche.

fonctionne en continu, à partir de données sources à traduire, le lexicographe pouvant à tout moment valider ou modifier les données obtenues, ainsi que rajouter de nouvelles données à traduire, même si celles-ci constituent des néologismes, ou des termes spécialisés, le Web nous permettant un accès aux usages « en temps réel » et de façon quantitative. Notre méthodologie exploite les caractéristiques du Web pour la résolution de difficultés de traduction, ces difficultés étant gérées de façon modulaire : à partir d'une liste à traduire, les traductions obtenues dans la première phase sont éliminées de la liste de départ, et ainsi de suite jusqu'à notre troisième et dernière phase. Dans un premier temps, nous présentons une méthode de validation de traductions candidates basée sur l'étude des fréquences sur le Web, dans la lignée de travaux tels que Grefenstette (1999), Cao et Li (2002) et Léon et Millon (2005). L'hypothèse est que les traductions candidates erronées apparaissent à une faible fréquence sur le Web, contrairement aux traductions correctes. Pour revenir à notre exemple de *vol de voitures*, la traduction *flight of cars* apparaît seulement à une fréquence de 4 sur le moteur de recherche *Google*[1], tandis que *theft of cars* apparaît 36 600 fois. Les résultats de fréquences sélectionnent de façon écrasante la traduction correcte (Léon et Millon, 2005).

Malgré tout, l'absence de prise en compte du contexte lexical constitue une limite, car la méthode des fréquences ne vérifie pas l'équivalence entre l'unité lexicale source et sa traduction, ce qui peut constituer des erreurs pour les cas fortement polysémiques. Par exemple, *group rate*, qui signifie *tarif de groupe* serait une traduction candidate de *cours de formation* par le « jeu » des multiples polysémies de *cours* et de *formation* (Léon et Millon, 2005). Si les fréquences permettent de vérifier l'existence d'une traduction et sont efficaces pour les cas non polysémiques, elles ne sont pas satisfaisantes pour les cas d'ambigüité lexicale. Notre méthode se base sur une détection du degré de polysémie des unités lexicales et propose un module de désambigüisation lexicale pour les cas polysémiques. Notre technique se base principalement sur la notion de « *mondes lexicaux »* à partir du Web. Pour nous, un monde lexical désigne les co-occurrences fréquentes d'une unité lexicale (simple ou complexe) au sein d'une collection de textes[2] (Véronis, 2003, 2004). De tels voisinages, plus larges que le co-occurrent immédiat, peuvent se situer au niveau de la phrase, ou même du

[1] Google, août 2008.

[2] En l'occurrence, les résumés retournés par le moteur de recherche *Yahoo*, en ce qui nous concerne.

paragraphe. Par exemple, le monde lexical de la requête « *caisse centrale* » sur *Yahoo* (*agricole*, *social*, *mutualité*, *crédit*, *banque*, *assurance*, *gestion*, etc.) est proche de celui de sa traduction correcte « *central fund* » (*money*, *pay*, *budget*, *insurance*, *management*, *social*, etc.), contrairement à la traduction erronée « *central case* » (*study*, *law*, *policy*, *enterprise*, *university*, etc.). Notre hypothèse est qu'une comparaison des mondes lexicaux permet de lever un grand nombre d'ambiguïtés lexicales (Léon, 2006). Certains travaux ont montré que l'exploitation des mondes lexicaux[1] permet une désambiguïsation lexicale d'un point de vue monolingue (Sébillot et Pichon, 1997, Pichon et Sébillot, 1999a, Pichon et Sébillot, 1999b, Rossignol et Sébillot, 2003, Véronis, 2003, Véronis, 2004). En traduction, des recherches ont souligné que les co-occurrences immédiates d'un mot cible sont les mêmes d'une langue à l'autre (entres autres (Rapp, 1999)), mais aussi un entourage linguistique plus large (Fung et Yee, 1998, Kikui, 1998, Tanguy, 1999). Ces stratégies ont été appliquées essentiellement sur des corpus terminologiques, ce qui offre une faible diversité des usages d'un mot, à partir de corpus parallèles ou comparables, ce qui limite la quantité des observations. De plus, ils concernent majoritairement des termes simples, tandis que nous nous intéressons aux unités lexicales complexes. Nos travaux présentent des similitudes avec ceux de Lafourcade *et al.* (2004) qui créent des ressources monolingues et bilingues par la construction de vecteurs conceptuels : la démarche adoptée est onomasiologique, c'est-à-dire que les concepts sont donnés a priori via des thesaurus et sont reliés à des items lexicaux. Pour nous, le monde lexical est construit uniquement à partir de données textuelles. Notre démarche est sémasiologique : nous partons des termes pour nous intéresser à leur signification et à leur traduction. A notre connaissance, aucune expérience sur la comparaison des mondes lexicaux n'a été menée en langue générale, sur l'immense base de données que constitue le Web.

Les stratégies présentées, qui constituent les deux premières étapes de notre méthode, l'une pour les unités lexicales non polysémiques, l'autre pour celles qui sont polysémiques se fonde sur une représentation compositionnelle de la traduction : la combinaison des traductions de chaque élément permet d'accéder au sens global. Toutefois, il arrive qu'une traduction ne soit pas transparente. Par exemple, le co-occurrent de *caisse* dans *caisse claire* se traduit par *snare*

[1] La terminologie relative aux mondes lexicaux varie selon les courants théoriques. Nous présentons ces différents courants dans le chapitre 7.

qui signifie littéralement *piège.* Une traduction littérale, à partir d'un dictionnaire ne peut être satisfaisante. De plus, certaines unités lexicales sont très techniques ou récentes et ne sont pas recensées dans les ressources dictionnairiques. Notre troisième phase de traduction propose une méthode afin de résoudre ces difficultés. La méthode se fonde sur une acquisition de pages « partiellement parallèles » sur le Web (Nagata, 2001) et sur un repérage de *cognates* et de bigrammes fréquents. Cette dernière étape permet de combler les lacunes dictionnairiques et de gérer les problèmes de traductions non transparentes. Le schéma ci-dessous récapitule les étapes de traitement de notre méthodologie :

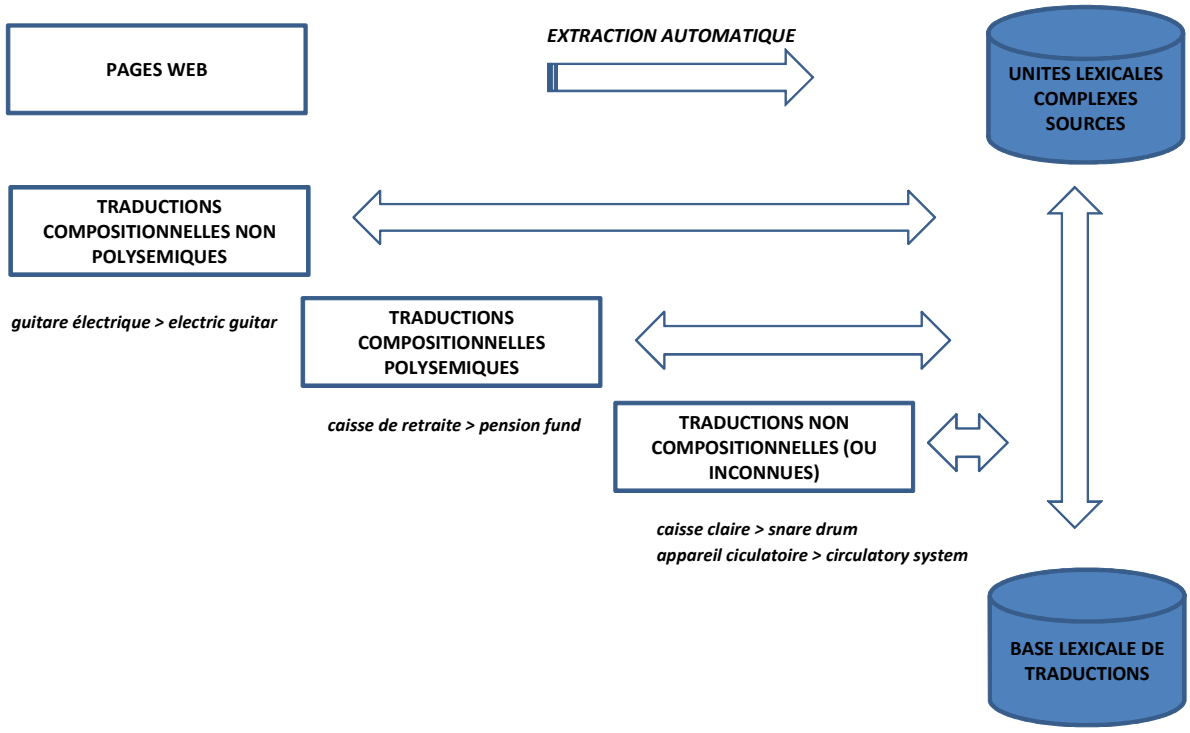


Figure 1. Etapes de traitement

1.3 Domaines d'application

1.3.1 Lexicographie et terminologie

Les dictionnaires bilingues traditionnels contiennent peu d'informations sur les phénomènes d'unités lexicales complexes, se contentant le plus souvent d'indications ponctuelles, bien que la lexicographie moderne vise à modifier cette tendance (Véronis, 2000a). En effet, le recensement de ces unités lexicales est fondamental pour un apprenant qui ne maîtrise pas une langue étrangère, puisque les cooccurrences préférentielles varient entre les langues. Par exemple, la *pluie* est *forte* en français, mais *lourde* (*heavy rain*) en anglais (Mel'cuk, 1997). De plus, de nombreux termes complexes, très techniques et/ou très récents n'ont pas fait l'objet d'un recensement systématique, dans le domaine de la terminologie (*ibid.*). En ce qui concerne la construction des ressources dictionnairiques bilingues, Véronis (2000a) montre que la lexicographie a de plus en plus recours à des corpus électroniques. L'utilité de ces derniers a été mentionnée depuis un certain temps par Hartmann (1980) et Atkins (1990) (Véronis, 2000a). La première utilisation de corpus électroniques en lexicographie remonte à la fin des années 1950, avec le projet du *Trésor de la Langue Française* (Imbs, 1971). Leur utilisation par des maisons d'édition est plus récente, avec le projet COBUILD (Sinclair, 1987a). D'autres projets telle que la compilation du *Oxford-Hachette French Dictionary* s'est appuyée sur des corpus comparables en anglais et en français de plus de 10 millions de mots chacun (Grundy, 1996). Le projet interuniversitaire du *Dictionnaire Canadien Bilingue* fait appel à un corpus de textes comparables complété par un corpus parallèle aligné de près de 50 millions de mots, le *Hansard* (Roberts et Montgomery, 1996). La conception du DEC (*Dictionary of English Collocations*) (Kjellmer, 1994) est basée sur une analyse de fréquence des mots. Les données lexicales extraites du Web pourraient être utiles pour la construction de telles ressources dictionnairiques.

1.3.2 Traduction automatique

La plupart des recherches actuelles en automatisation de traduction se situent dans un continuum entre deux pôles (Véronis, 2000a) : d'une part, la traduction humaine assistée par

des outils informatisés et d'autre part, la traduction automatique assistée par l'homme, tels que des systèmes d'aide à la lecture et à la rédaction par exemple (Li et Cao, 2002, Li *et al.*, 2003a). Tout au long de ce continuum, les bases de traductions d'unités lexicales complexes sont utiles afin de compléter et d'améliorer les ressources traditionnelles. Les systèmes de mémoire de traduction (Kjaersgaard, 1987, Isabelle, 1992, Macklovich, 1992, Picchi *et al.*, 1992) sont basés sur l'idée, proposée par Kay, en 1980, de réaliser une approche progressive de la traduction automatique, dont l'étape de départ serait de s'appuyer sur des exemples de textes (Véronis, 2000a). Le courant de la « traduction automatique basée sur la mémoire » (ou sur les exemples) (Nagao, 1984, Sadler, 1989, Sato et Nagao, 1990, Sumita *et al.*, 1990) avance l'idée d'exploitation de fragments similaires aux portions du texte à traduire et de combinaison de façon adéquate (Véronis, 2000a). La base de données de traductions que nous constituons pourrait être utilisée par des systèmes de mémoire de traductions. Les traductions d'unités lexicales complexes peuvent être réutilisées dans différents contextes. De plus, les mondes lexicaux offrent des informations sur leur contexte qui pourraient être utiles.

1.3.3 Recherche d'information multilingue

Depuis une vingtaine d'années, la recherche d'information multilingue (*cross-language information retrieval*) connaît une explosion grâce au Web (Véronis, 2000a). Il s'agit de formuler une requête dans une langue et obtenir les résultats (ou une partie d'entre eux) dans une autre langue afin d'obtenir des résultats plus précis (*ibid.*). Le présupposé est que l'utilisateur soit capable de déchiffrer les résultats obtenus dans d'autres langues, mais incapable de formuler une requête dans cette langue (*ibid.*). Les techniques de traduction automatique des requêtes posent les mêmes difficultés que pour la Traduction Automatique : problèmes de polysémie et imperfection des dictionnaires. Par exemple, le système de recherche multilingue de *Google*[1] traduit la requête *souris d'agneau* de façon littérale, par *mouse lamb*[2], ce qui offre des résultats faussés pour une requête multilingue en français ciblée vers des résultats en anglais :

[1] http://www.google.fr/language_tools?hl=fr

[2] La traduction attendue est *lamb shank.*

Figure 2. Recherche multilingue Google en anglais de « souris d'agneau »

L'accès à des bases de données lexicales de traductions d'unités lexicales complexes pourrait être utile afin d'améliorer la qualité de traductions de telles requêtes.

1.3.4 Désambiguïsation lexicale

L'accès à la traduction de combinaisons lexicales peut servir à la résolution d'un problème monolingue classique, celui de la désambiguïsation du sens des mots en contexte (Véronis, 2000a). L'idée majeure est que l'ambiguïté lexicale d'une langue est levée par des choix lexicaux différents dans une autre langue (*ibid.*). Par exemple, le nom polysémique *cais*se se traduit en anglais de façon différente en fonction de son sens : *fund* pour l'usage *BANQUE*, *drum* pour l'usage *TAMBOUR*, etc. L'accès à la traduction d'un terme polysémique en contexte permet d'accéder à son sens. Brown *et al.* (1991a) et Gale *et al.* (1993) proposent l'utilisation de corpus parallèles afin de constituer un corpus d'entraînement d'amorçage pour les systèmes de désambiguïsation automatique. L'hypothèse est que des banques d'exemples de traductions en contexte[1] peuvent être précieuses pour la désambigüisation lexicale.

1.3.5 Didactique des langues

Selon Grossmann et Tutin (2003), les études sur l'acquisition d'unités lexicales complexes en langue seconde (Granger, 1998, Howarth, 1998) sont parmi les plus difficiles à maîtriser pour les apprenants. Des lexiques de traductions d'unités lexicales en contexte ont un intérêt en

[1] Au-delà des traductions d'unités lexicales complexes, nous présentons l'accès au monde lexical de ces unités, qui peut également constituer une aide précieuse pour la désambigüisation lexicale.

didactique des langues étrangères, pour l'aide à la maîtrise d'une langue, tant pour l'aide à la production (rédaction, production orale) que pour la compréhension de textes. L'observation de l'utilisation des mots en contexte (par les mondes lexicaux) peut compléter efficacement les outils classiques tels que les dictionnaires et les grammaires (Véronis, 2000a).

1.3.6 Linguistique comparative

Les bases lexicales de traduction peuvent constituer des ressources pour les recherches linguistiques comparatives entre le français et l'anglais (« trous » lexicaux, traduction par d'autres structures, etc.) et pour l'étude théorique de la traduction (Véronis, 2000a). Une application de nos résultats pourrait être de compléter des études de syntaxe comparée telles que celle de (Guillemin-Flescher, 1981), qui portaient sur des données très restreintes : pour l'instant, aucune étude linguistique comparative ne porte sur des données d'une ampleur comparable à celle que nous proposons.

1.3.7 Autres applications

Génération automatique de textes

Les unités lexicales complexes constituent des syntagmes relativement figés qu'il est possible de réutiliser de la même façon dans des contextes divers. Sinclair (1987b) a développé la notion de *idiom principle*, l'idée étant que la langue est faite de blocs préfabriqués :

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments

Ces « blocs » peuvent être considérés comme des « produits semi-finis » (Hausmann, 1979) et réutilisés tels quels à des fins de génération automatique. Par exemple, Smadja et McKweown (1991) ont réalisé un programme de génération automatique de phrases, *Cook*, dans le domaine de la bourse, en anglais, basé sur une utilisation des unités lexicales complexes.

Reconnaissance Optique des Caractères[1]

Les unités lexicales complexes peuvent être utiles afin de désambiguïser deux formes, pour la Reconnaissance Optique des Caractères (OCR). Il s'agit d'une technique, qui à l'aide d'un procédé optique, permet à un système informatique de lire et de stocker automatiquement du texte dactylographié. Lorsque deux mots différents se prononcent de la même manière, la connaissance de son contexte permet de désambiguïser la forme. Par exemple, considérons les formes *farm* et *form* et différents contextes (Church et Hanks, 1990) :

(1) federal credit

(2) some of

Dans le contexte (1), accompagné régulièrement de *federal credit*, nous aurons le terme *form*, tandis que le contexte (2) (*some of*) se rencontre plus fréquemment avec *farm*.

Désambiguïsation syntaxique automatique

L'ambiguïté syntaxique consiste au fait que pour une même phrase, plusieurs analyses syntaxiques sont possibles, en fonction des regroupements établis. Maniez (2001a) montre que l'anglais, par exemple, est propice aux ambiguïtés syntaxiques (appartenance d'un mot à plusieurs catégories grammaticales, pas d'accord pour les adjectifs, etc.). Prenons pour illustration la structure syntaxique (*ibid.*) :

ADJECTIF NOM1 AND NOM2

Weight-reducing diet and chlorthalidone

L'adjectif *weight-reducing* ne s'applique qu'au premier nom *diet*. L'ambiguïté vient du fait qu'une analyse pourrait appliquer la distributivité à la coordination. La désambiguïsation du découpage de ces structures peut être apportée grâce à un recensement des unités lexicales

[1] *Optical Character Recognition*, en anglais.

complexes (dans l'exemple cité, l'unité lexicale complexe serait de la forme syntaxique *ADJECTIF-NOM*).

Résumés automatiques « multilingues »

La tâche de résumé automatique consiste en une reformulation du texte original afin d'en décrire l'essentiel du contenu. Les mondes lexicaux sont utiles pour le repérage automatique des concepts principaux pour la tâche de génération de résumé automatique. Par exemple, SUMMARIST (Hovy et Lin, 1997, Lin et Hovy, 2000) est un système de génération automatique de résumés, qui s'appuie sur une méthode d'acquisition de mondes lexicaux[1].

1.4 Domaines et plan de la thèse

Notre thèse se situe à mi-chemin entre deux domaines, la linguistique, et plus précisément la sémantique lexicale d'une part, et l'informatique d'autre part, plus précisément le Traitement Automatique des Langues et l'extraction automatique d'informations à partir du Web.

En ce qui concerne la **sémantique lexicale,** nous nous intéressons aux aspects lexicologiques des unités lexicales complexes et à leurs critères définitoires. Nous proposons une analyse de différents types d'unités lexicales complexes telles que les *locutions* (et les *termes complexes*) comme par exemple *caisse claire* et les *collocations*, comme par exemple *pluie forte*. Nous nous intéressons à la notion de figement et aux aspects comparatifs des phénomènes de combinatoires lexicales. Nous posons la problématique des rapports entre les unités lexicales complexes repérées dans les textes, les concepts et leurs équivalences dans une langue cible. Nous nous centrons également sur les mondes lexicaux d'un point de vue interlingue et de leur utilité pour la désambigüisation lexicale en traduction.

---

[1] Hovy et Lin (1997) parlent de *signatures thématiques* (*« topic signature »* en anglais), qui est une notion proche. Nous reviendrons sur cette notion dans le chapitre 7.

Du point de vue du **Traitement Automatique des Langues**, nous nous intéressons aux aspects techniques de l'identification automatique des unités lexicales complexes à partir de vastes données textuelles. Nous mettons en place une procédure d'acquisition automatique d'unités lexicales et de leurs traductions à partir du Web. Nous proposons une réflexion sur le statut du Web pour la recherche linguistique. Nous détaillons et utilisons les différentes caractéristiques du « Web bilingue » pour son utilisation en traduction.

Outre l'introduction et la conclusion, notre thèse se divise en huit chapitres, dont les quatre premiers sont un état de l'art des champs qui recoupent notre sujet. Ce dernier se situe à mi-chemin entre plusieurs domaines que nous étudions de façon conjointe : les unités lexicales complexes (aspects théoriques et traitement automatique monolingue et bilingue), l'utilisation du Web pour les recherches linguistiques, l'acquisition de traductions à partir du Web. Les quatre derniers chapitres présentent notre méthodologie d'acquisition de traductions d'unités lexicales complexes à partir du web, analysent les résultats et présentent l'évaluation.

Le **deuxième chapitre** présente les aspects théoriques et définitoires des unités lexicales complexes. Après une mise en perspective des débuts des recherches en Traduction Automatique avec le traitement des unités lexicales complexes, nous montrons qu'il existe plusieurs phénomènes : les *locutions* (ou les *termes complexes*) et les *collocations*. Nous analysons les caractéristiques de ces phénomènes puis les considérons de façon comparative, en montrant que les différences se situent sur un continuum dont les frontières sont floues.

Le **troisième chapitre** présente les méthodes traditionnelles de traitement automatique des unités lexicales complexes, d'abord dans une perspective monolingue, puis dans une perspective bilingue. Nous présentons les limites des méthodes traditionnelles de traduction et montrons que le Web est un outil dont les caractéristiques permettent de palier ces limites.

Le **quatrième chapitre** présente un état de l'art de l'utilisation du Web pour les recherches en Traitement Automatique des Langues (TAL). Nous proposons quelques réflexions théoriques sur les rapports entre l'utilisation de corpus en linguistique et l'utilisation du Web. Nous montrons que malgré une utilisation récente, il existe un panel très varié et prolifique de travaux qui utilisent le Web pour leurs applications.

Le **cinquième chapitre** présente les différentes méthodes d'acquisition automatique de traductions, à partir du web, qu'il s'agisse de construction de corpus parallèles ou d'extraction d'informations à partir du Web.

Le **sixième chapitre** correspond à notre première phase de traduction, celle des combinaisons lexicales compositionnelles non polysémiques, du type de :

guitare électrique > electric guitar

Cette phase est basée sur la prise en compte des fréquences des traductions candidates sur le Web.

Le **septième chapitre** décrit la phase suivante de notre méthodologie, qui traite des traductions compositionnelles polysémiques, comme dans l'exemple :

appareil ménager > household appliance

Cette phase est principalement basée sur la comparaison de mondes lexicaux sur le Web. Nous présentons d'abord les différents aspects théoriques liés à la construction de mondes lexicaux (isotopie sémantique, repérage thématique, etc.), puis nous détaillons notre méthodologie.

Le **huitième chapitre** constitue la dernière phase de notre méthodologie, qui concerne la traduction de combinaisons lexicales non compositionnelles et de mots techniques non recensés dans des ressources dictionnairiques traditionnelles, comme dans les exemples respectifs :

acide folique > folic acid

caisse claire > snare drum

Après quelques réflexions théoriques sur la notion de compositionnalité, nous détaillons cette dernière phase.

Enfin, le **neuvième chapitre** consiste en une évaluation détaillée des résultats, de façon quantitative (en nombre de combinaisons correctes obtenues), et qualitative, en termes de difficultés de traduction et de problèmes résolus. Nous faisons enfin le point sur les apports constatés de notre méthode ainsi que de l'utilisation du Web au sein de notre étude et nous parlons également des limites et des perspectives d'évolution.

Chapitre 2. Vers des unités lexicales complexes pour la traduction

2.1 Introduction

La notion de « mot » est une notion empirique dont les contours sont flous et difficiles à définir, au-delà du critère graphique. Les tentatives de théorisation du « mot » ont suscité (et suscitent encore) de vifs intérêts chez les linguistes qui l'ont rejeté au profit d'autres termes (Léon, 2001). D'un point de vue pratique, les applications en Traitement Automatique des Langues qui nécessitent l'identification d'unités lexicales doivent envisager des unités qui ont une valeur syntaxique et sémantique. Le domaine de l'aide à la Traduction Automatique n'échappe pas à la règle. Un aspect majeur consiste en l'identification des unités lexicales, dont le repérage de « mots » séparés par un espace n'est pas satisfaisant. Une reconnaissance erronée d'une association idiomatique par le système conduit à des résultats qui gênent la compréhension, provoquant parfois des contre-sens. Ainsi, le traducteur en ligne *Systran*[1] propose la traduction anglaise littérale de *feu rouge* :

(1) feu rouge > red light

[1] http://www.systransoft.com/

Cette traduction est incompréhensible pour un anglophone, parce qu'elle ne doit pas être compositionnelle : la traduction correcte est *traffic light* (littéralement *feu de lumière*). Au-delà des phénomènes idiomatiques, la polysémie[1] ou l'homonymie des unités lexicales constitue une difficulté centrale en Traduction Automatique. Par exemple, le nom *appareil* est fortement polysémique[2]. La sélection du nom cible ne peut être effectuée sans connaître son usage. Pour revenir à *Systran*, de nombreuses erreurs en Traduction Automatique sont dues à une absence de désambiguïsation lexicale :

(2) ***appareil*** *ménager > domestic* ***machine***

(3) ***appareil*** *digital > digital* ***apparatus***

Pourtant, les co-occurrents (ici *ménager* et *digital*) sont un indice désambiguisateur fort qui pourrait être exploité pour générer la traduction adéquate. Une solution aux problèmes d'idiomatisme et d'ambigüité lexicale, consiste en la création de vastes lexiques d'équivalences bilingues du type de[3] :

(1) feu rouge > traffic light

(2) appareil ménager > household appliance

(3) appareil digital > digital camera

[1] La polysémie désigne le fait qu'une unité lexicale ait plusieurs sens, entre lesquels il existe un lien étymologique. Dans le cas de l'homonymie, différents sens sont attribués à la même forme graphique, mais il n'existe pas de lien. Dans le cadre de nos travaux, nous ne prenons pas en compte cette distinction, et parlons de polysémie de façon indifférenciée.

[2] Dans notre dictionnaire bilingue *Collins Pocket*, 12 traductions de *appareil* sont recensées, ce qui montre que ce terme est polysémique (Dagan *et al.*, 1991).

[3] Ce problème a été abordé dès les débuts des recherches en Traduction Automatique (Bar-Hillel, 1955, Pottier, 1962c), mais les moyens informatiques sommaires de l'époque ne permettaient pas de traiter le nombre de données que nous pouvons envisager de nos jours.

Les unités lexicales complexes forment une unité syntaxique. Les transformations syntaxiques sont limitées, mais toutes les unités lexicales complexes ne sont pas complètement figées et contigües, ce qui accroit la difficulté de leur traitement automatique :

*(1) feu rouge > *feu très rouge*[1]

feu rouge > le feu est rouge

Parfois, aucune transformation syntaxique n'est possible :

*(2) appareil ménager > *appareil très ménager*

*appareil ménager > *l'appareil est ménager*

Ces unités lexicales complexes forment une unité sémantique, car elles renvoient, le plus souvent, à un référent unique et le sens n'est pas décomposable. Elles doivent être envisagées dans leur globalité et être recensées en tant qu'unités de la langue. Il existe des milliers d'unités lexicales complexes au sein de chaque langue, et une tâche manuelle est impossible, il faut proposer des moyens d'extraction automatique. Malgré une littérature abondante sur la combinatoire lexicale, les contours restent flous et la terminologie disparate. Les unités lexicales complexes posent des problèmes définitoires, de par leurs caractéristiques fluctuantes. Les critères définitoires ne font pas l'unanimité et des exemples prototypiques sont envisagés (Williams, 2001). Nous distinguons deux types d'unités lexicales complexes, les *locutions* (ou *termes complexes*) qui sont des unités lexicales figées telles que *appareil ménager* et les *collocations*, qui sont « semi-figées » telles que *café noir*. Cette distinction n'établit pas des frontières nettes et les critères de définition ne sont pas généralisables à l'ensemble de chaque classe.

Ce chapitre présente un tour d'horizon des aspects théoriques du traitement des unités lexicales complexes, dans une perspective de traduction. Aux Etats-Unis, dans les années 1950-1960, les recherches en Traduction Automatique n'ont pas mis l'accent sur l'analyse des

[1] L'astérisque indique que la séquence est agrammaticale.

combinaisons lexicales (Léon, 2001). Certains travaux ont toutefois énoncé le problème d'un traitement d'unités lexicales complexes. Malgré la place dominante de l'informatique, les préoccupations de Bar-Hillel, premier chercheur en Traduction Automatique, s'intéresse à la traduction d'unités lexicales complexes et montre les prémisses de préoccupations lexicologiques (2.2). Par exemple, une unité lexicale complexe peut se traduire par une unité lexicale simple :

(4) appareil photographique > camera

(5) pomme de terre > potatoe

Les recherches en Traduction Automatique en France, plus tardives, ont conduit des linguistes à définir des unités lexicales. Ces préoccupations tant pour la mécanisation du vocabulaire que pour la Traduction Automatique, ont fait émerger une terminologie nouvelle, disparate, mais révélatrice de questionnements fondamentaux pour l'automatisation de la traduction (Léon, 2001, 2004) (2.3). Après un aperçu des recherches en Traduction Automatique dans les années 1950-1960, pour les traditions américaine et française, sous l'angle des préoccupations du traitement d'unités lexicales, nous définissons notre terminologie. Nous décrivons le phénomène des locutions et des termes complexes (2.4). Au-delà du figement complet, les unités lexicales préférentielles constituent une aide pour la désambiguïsation lexicale en traduction, nous abordons les collocations (2.5). Les frontières entre les deux phénomènes restent floues et nous concluons par une approche comparative et graduelle (2.6). Malgré les différences de critères définitoires, il arrive que nous ayons à parler des deux catégories de façon indifférenciée. Nous parlons d'*unité lexicale complexe*, comme catégorie hyperonyme regroupant les classes des *locutions* et des *collocations*.

2.2 Prémisses de la Traduction Automatique

Aux Etats-Unis, dans les années 1950, l'interaction entre la linguistique et la Traduction Automatique est rare (Léon, 2001). Les difficultés que posent la Traduction Automatique pour le repérage d'unités syntaxiques et sémantiques n'a pas éveillé l'intérêt des linguistes (*ibid.*). La Traduction Automatique est mise à l'épreuve des langages formels par

l'informatique (*ibid.*). L'approche est influencée par le contexte de la seconde guerre mondiale, où des efforts avaient dus être fournis en cryptographie. En 1947, Waever compare le processus de traduction à un processus de chiffrement. Un texte traduit en russe est vu comme un chiffrage de sa version anglaise à l'aide d'un code particulier :

> Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography – methods which I believe succeed even when does not know what language had been coded – one naturally wonders if the problem of translation could conceivably be treated as a problem of cryptography. When I look at an article in Russian, I say : "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode".

Figure 3. Extrait de la lettre de Warren Weaver à Norbert Wiener[1]

Bar-Hillel, philosophe logicien, premier chercheur en Traduction Automatique au MIT (*Massachusetts Institute of Technology*), en 1951, constitue une exception. Dans le premier recueil de travaux en Traduction Automatique de 1955, il s'intéresse aux unités lexicales complexes dont la traduction ne peut être littérale et propose une définition de la notion d' « idiome ». Il cite l'exemple de *red herring* en anglais qui se traduit par une forme graphique en allemand, *Finte*[2]. Selon lui, un idiome est un « bloc de mots » qui fonctionne comme une unité et dont la traduction ne peut être littérale[3] :

> An expression in a given language L is idiomatic within L, with respect to a given monolingual dictionary and a given list of grammatical rules if, and only if, none of the word sequences correlated to the given expression by the dictionary and the list of rules is (sufficiently) synonymous with it.

[1] 4 mars 1947.

[2] *Fausse piste* en français.

[3] Greimas (1960) se réfèrera à Bar-Hillel au sujet de la problématique de traduction des expressions idiomatiques (Léon, 2001, 2004).

Les séquences ne pouvant être traduites de façon littérale, malgré des règles de transformations morpho-syntaxiques, sont des idiomes qui doivent être recensées. Bar-Hillel (1955) envisage l'ajout d'un dictionnaire d'équivalences d'idiomes au sein des systèmes de Traduction Automatique, en amont des dictionnaires traditionnels. Les règles transformationnelles à insérer dans les systèmes selon Bar-Hillel (1955) seraient de deux ordres :

- Les règles de transformations morpho-syntaxiques comme par exemple :

NOM-ADJECTIF (français) > ADJECTIF-NOM (anglais)

appareil digital > digital camera

- Les règles de transformations « idiomatiques » :

roter Hering (allemand)[1] *> Finte (allemand)*

Bar-Hillel (1955) aborde les prémisses du problème des connaissances sémantiques et pragmatiques pour la traduction, qui fera l'objet d'un rapport quelques années plus tard. Pour reprendre l'exemple de *red herring*, il y a un usage en allemand où la traduction n'est pas *finte* mais *roter Hering* (traduction littérale), à savoir l'usage *ŒUVRE d'ART*[2]. Afin de connaître l'usage de *red herring* et la traduction adéquate, il faut disposer de connaissances sémantiques et pragmatiques que l'on ne peut espérer d'une machine. Le problème de l'ambigüité lexicale pour la traduction a été soulevé tôt. En 1949, Waever introduit le besoin de désambiguïsation lexicale pour la tâche automatisée de traduction. Il est impossible pour un lecteur d'accéder au sens d'un mot lorsqu'il est dénué de tout contexte. Lorsqu'on lui donne accès à son voisinage, l'ambiguïté lexicale n'est plus présente. Il préconise un processus qui détermine le sens d'un mot, en prenant en compte son contexte immédiat, dans

[1] Il s'agit de la traduction littérale, en allemand, de l'anglais, *red herring*.

[2] Cette usage fait référence à une peinture de Marc Chagall, *red herring*. Dans cet usage, une traduction littérale en allemand est souhaitable.

une fenêtre de deux mots (contexte gauche et contexte droit). Dans son Memorandum, il envisage un recensement de toutes les séquences possibles de digrammes ou de trigrammes. Hormis quelques expérimentations (Kaplan, 1950), la méthode présentée par Weaver n'a pas obtenu de succès à cause des moyens techniques sommaires de l'époque, problème reconnu par Weaver (1955) (Léon, 2001) :

> It would hardly be practical to do this by means of a generalized dictionary which contains all possible phases 2N +1 words long : for the number of such phases is horrifying, even to a modern electronic computer.

Ces problèmes n'ont pu être surmontés à l'époque. En 1960, Bar-Hillel publie un rapport qui met en lumière les difficultés que pose la traduction, tant sur le plan technologique (la technologie informatique de l'époque était très sommaire) que linguistique. Dans son rapport, il met en avant la sous-estimation des connaissances contextuelles et encyclopédiques mises en jeu dans la traduction. Le célèbre exemple concerne le terme polysémique anglais *pen* dans les deux phrases :

(1) The box is in the pen (la boîte est dans l'enclos)

(2) The pen is in the box (le stylo est dans la boîte)

Afin de traduire le terme *pen*, il faut disposer de connaissances générales sur le monde et cette difficulté avait été sous-estimée. Cette évaluation, venant du premier chercheur recruté dans le domaine, aura un fort impact négatif dans la communauté scientifique. En 1964, l'administration américaine commande un rapport, le rapport ALPAC (*Automatic Language Processing Advisory Commitee*) qui établit un constat d'échec sur les recherches en Traduction Automatique et met fin aux financements et à une majeure partie des recherches dans le domaine.

2.3 Renouveau terminologique

Contrairement aux Etats-Unis ou à l'Angleterre, où les premières recherches en Traitement Automatique des Langues sont celles en Traduction Automatique, c'est la mécanisation du vocabulaire en France qui constitue les premiers travaux en TAL (Léon, 2004a). Le début des recherches en Traduction Automatique commencent dix ans plus tard qu'aux Etats-Unis, à la période de la publication du rapport ALPAC (*ibid.*). Les questionnements sur le statut des unités traitées sont au cœur des travaux en mécanisation du langage (*ibid.*). Ces interrogations donnent naissance à une terminologie nouvelle afin de désigner les unités lexicales complexes[1]. La terminologie propre à chaque linguiste est le reflet de questionnements novateurs mis en relation avec la mécanisation du langage. Ce sont principalement les travaux en Traduction Automatique qui ont donné un nouvel essor aux réflexions lexicologiques. Nous abordons trois auteurs qui ont introduit une nouvelle terminologie pour les unités lexicales complexes : la *lexie* de Pottier, la *synapsie* de Benveniste et le *synthème* de Martinet.

La notion de *lexie* (Pottier)

Pottier (1962a, 1962b, 1962c) pose la définition d'unités lexicales, dans un cadre de Traduction Automatique (Léon, 2001). Il introduit la notion de *lexie*, unité de langue, à la fois unité lexicale et unité de base de la construction syntaxique (*ibid.*). Sa définition est unificatrice (contrairement à la tradition structuraliste) car elle regroupe différents types de lexies (qui ne sont jamais inférieurs au mot graphique) (Léon, 2001) :

- Lexies simples :

pierre

chaise

[1] Nous parlons uniquement des principaux courants théoriques mis en rapport avec les débuts de l'automatisation du langage en France.

- Lexies composées :

bateau-mouche

sous-chef

cheval-vapeur

- Lexies complexes :

chemin de fer

pomme de terre

prendre la mouche

Les critères d'identification sont de divers ordres. D'un point de vue sémantique, le référent est stable et unique. D'un point de vue syntaxique[1], aucune modification n'est possible. Du point de vue interlingue, une lexie simple peut être traduite par une lexie complexe ou inversement[2]. Seule l'identification des lexies complexes pose une difficulté pour la Traduction Automatique, puisqu'elles ne sont pas repérables par un indice graphique (Léon, 2004a). La distinction automatique d'une lexie complexe d'avec un syntagme libre est une tâche compliquée, car les aspects formels sont les mêmes, comme dans les exemples :

(1) cheval de Jean

(2) cheval de course

[1] L'analyse des lexies chez Pottier laisse émerger un début de traitement syntaxique, en proposant une catégorisation qui s'apparente à la notion de *tête* en grammaire syntagmatique (Léon, 2004a). Par exemple, *plaque tournante* est considérée comme un substantif parce que c'est la catégorie hiérarchiquement supérieure (*ibid.*).

[2] Nous retrouvons la problématique initialement soulevée par Bar-Hillel (1955) abordée dans la section 2.1.

D'un point de vue morpho-syntaxique, les exemples (1) et (2) sont équivalents (*NOM-PREP-NOM*). Pourtant, l'exemple (1) constitue une association libre, tandis que l'exemple (2) est une lexie complexe. A mi-chemin entre ces deux types de combinaisons, il existe des zones d'incertitude sur lesquelles nous reviendrons (*ibid.*). Selon Pottier, des critères statistiques permettent de déterminer un degré de lexicalisation (*ibid.*)[1].

La notion de *synapsie* (Benveniste)

Benveniste (1966, 1967) introduit le terme de *synapsie* afin de désigner une unité lexicale complexe, composée de plusieurs lexèmes, dont le signifié global est unique et constant. Son analyse des synapsies s'est opérée dans un contexte d'étude de nomenclature technique. La synapsie correspond à la définition traditionnelle de « mot composé », introduite de la façon suivante dans le *dictionnaire de Linguistique* (Gross, 1996) :

> On appelle mot composé un mot contenant deux, ou plus de deux, morphèmes lexicaux et correspondant à une unité significative : *chou-fleur*, *malheureux*, *pomme de terre* sont des mots composés.

Toutefois, pour Benveniste, les synapsies telles que *machine à coudre* sont à différencier des formes graphiquement soudées, qu'il nomme « conglomérés » (du type de *justaucorps*), ces conglomérés incluant également les mots-composés au sens traditionnel (*timbre-poste*) (*ibid.*). Les synapsies sont proches des lexies de Pottier, mais l'accent est davantage porté sur la structure syntaxique interne des synapsies (Léon, 2004a). Une synapsie est considérée comme la conversion nominale d'un énoncé prédicatif (*ibid.*) :

il garde un asile de nuit > gardien d'asile de nuit [2]

La nature syntaxique des synapsies autorise les expansions (*ibid.*) :

[1] La question de la lexicalisation avait déjà été abordée par Bally (1932) (Léon, 2004a).

[2] Cette proposition avait déjà été abordée dans le *Traité de la formation des mots composés* de Darmesteter (1875).

[[gardien d'asile] de nuit]

Benveniste (1966) décrit la liste des traits caractéristiques des synapsies (Drouin, 2002) :

> Ce qui caractérise la synapsie est un ensemble de traits dont les principaux sont : 1º – la nature syntaxique (non morphologique) de la liaison entre les membres; – 2º l'emploi de joncteurs à cet effet, notamment de et à; – 3º l'ordre déterminé + déterminant des membres; – 4º leur forme lexicale pleine, et le choix de tout substantif ou adjectif; – 5º l'absence d'article devant le déterminant; – 6º la possibilité d'expansion pour l'un ou l'autre membre; – 7º le caractère unique et constant du signifié.

La notion de synthème (Martinet)

Martinet (1960, 1967, 1968) oppose le *syntagme*, dont le choix de combinaison s'effectue librement, au *synthème*. Sur la base de critères syntaxiques, le syntagme est formé de deux ou plusieurs monèmes[1] dont les rapports sont plus étroits entre eux que par rapport aux autres éléments de l'énoncé. Le synthème correspond à un choix unique du locuteur et regroupe les formes construites par composition, par figement et par dérivation (Léon, 2004a) :

lavage

entreprendre

indésirable

pomme de terre

Les monèmes sont dits « libres » dans un syntagme, « conjoints » dans un synthème. Les constituants du synthème ne peuvent pas recevoir de détermination, comme dans l'exemple (Martinet, 1968) :

[1] Selon la terminologie de Martinet, un *monème* est une unité significative minimale, qui n'est pas nécessairement un segment isolable de l'énoncé.

chaise longue (synthème)

*chaise **plus** longue (vs. chaise longue)*

Le critère d'inséparabilité n'est pas obligatoire (Léon, 2001) :

ministre du travail (Synthème)

*ministre **italien** du travail*

Martinet (1985) pose la définition du synthème suivante :

> signe linguistique que la commutation révèle comme résultant de la combinaison de plusieurs signes minima, mais qui se comporte vis-à-vis des autres monèmes de la chaîne comme un monème unique.

Martinet (1968) aborde des cas où la distinction entre syntagme et synthème pose une difficulté. Il cite un exemple qui n'appartient pas à la même catégorie en fonction de son emploi :

*(1) de **jeunes filles** sont arrivées*

*(2) des **jeunes filles** sont arrivées*

Dans la séquence (1), *jeune fille* correspond à un syntagme, dans lequel l'adjectif *jeune* modifie le nom *fille*. Dans la séquence (2), il s'agit d'un synthème, c'est-à-dire un « substantif composé » (Martinet, 1968)[1].

Nous proposons le tableau récapitulatif, inspiré de Léon (2004a) afin de montrer les similitudes et les divergences entre ces courants théoriques. Les lexies de Pottier et les

[1] Martinet (1967) montre que l'article indéfini pluriel est « de » devant un épithète, et « des » devant un substantif.

synapsies de Benveniste sont proches, tandis que le synthème de Martinet est une notion plus large :

	Indésirable	*Pomme de terre*	*Mur du son*	*Ministre du commerce*
Pottier	-	Lexie complexe	Lexie complexe	-
Benveniste	-	Synapsie	-	-
Martinet	Synthème	Synthème	Synthème	Synthème

Figure 4. Tableau comparatif des unités lexicales complexes

2.4 Locutions et termes complexes

2.4.1 Locutions

Si la notion de « mot » est rejetée par les linguistes, c'est parce que ses contours sont difficiles à cerner. D'une part, doit-on considérer que les formes *suis, es*, et *est* constituent trois mots (Polguère, 2003)? D'autre part, il existe des expressions linguistiques complexes qui, bien que constituées de plusieurs unités graphiques, forment une unité lexicale, comme *pomme de terre* qui renvoie à un référent. D'un point de vue diachronique, *pomme de terre* provient de trois unités productives mais ces unités connaissent un figement lexical. Sur l'axe paradigmatique, *pomme de terre* peut commuter avec une unité lexicale simple telle que *patate*. La terminologie qui désigne les unités lexicales complexes figées est disparate et a connu un développement foisonnant :

Expression figée, Expression idiomatique, Figement, Unité polylexicale, Mot composé, Lexie complexe, Locution, Unité phraséologique (Bally, 1909), Synapsie (Benveniste, 1967), Phrasème complet (Mel'cuk, *et al.*, 1998), Unité polylexématique (Corbin, 1997)[1]…

Nous empruntons la terminologie de Polguère (2003) et parlons de *locution* afin de définir les unités lexicales complexes figées, en langue générale. Selon Polguère (2003), une locution désigne une lexie[2] composée d' « expressions linguistiques complexes » :

Une *locution* est une lexie regroupant des expressions linguistiques complexes que seule distingue la flexion.

Une locution forme un « tout lexical » et les éléments qui la constituent perdent leur autonomie de fonctionnement : il est impossible d'insérer de nouveaux éléments au sein d'une locution (i*bid.*). Polguère (2003) recense les types de locutions :

- **les « locutions nominales » :**

fruit de mer

nid de poule

- **les « locutions verbales » :**

rouler sa bosse

passer à tabac

[1] Voir Martins-Baltar (1997) pour une étude de la terminologie des expressions figées (Grossmann et Tutin, 2003).

[2] Pour Polguère (2003), une *lexie* « aussi appelée unité lexicale, est un regroupement 1) de mots-formes ou 2) de constructions linguistiques que seule distingue la flexion. Dans le premier cas, il s'agit de *lexèmes*, dans le second cas, de *locutions* ».

- **les « locutions adjectivales » :**

d'accord

en panne

- **les « locutions adverbiales » :**

au hasard

en vitesse

- **les « locutions prépositionnelles » :**

à propos de

en regard de

Les locutions sont dotées d'une autonomie de fonctionnement et d'un degré de cohésion (toutefois variable en fonction du type de locution) (*ibid.*). Du point de vue de l'interprétation sémantique, le sens global de la locution ne correspond pas à la somme des sens des éléments qui la constituent (non-compositionnalité) comme dans l'exemple de (*ibid.*) :

fruit de mer

Même si le sens peut être interprété de façon métaphorique, il ne s'agit pas d'un « fruit qui pousse dans la mer » (*ibid.*). La non-compositionnalité est d'autant plus perceptible que l'on confronte les locutions au phénomène de traduction. Souvent, la traduction d'une locution n'est pas littérale (*ibid.*) :

fruit de mer > seafood

Une locution est associée à un sens donné, au même titre qu'une unité lexicale simple, et doit bénéficier du statut d'unité (i*bid)* :

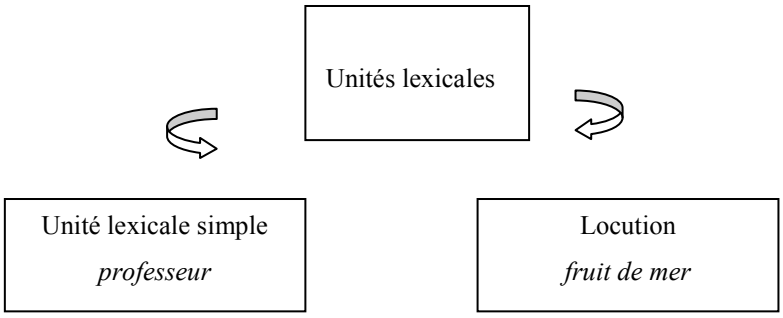


Figure 5. Types d'unités lexicales

2.4.2 Termes complexes

Nous parlons de *terme complexe* afin de désigner une locution relative à un domaine de spécialité (Daille, 1994). Un terme renvoie à un sens spécialisé, mais il n'est pas nécessairement exclusif au domaine (L'Homme, 2005). Selon l'Office de la Langue Française, un terme peut être simple ou complexe et désigne une notion au sein d'un domaine de spécialité (Dubois *et al.,* 1994) (Séguéla, 2001). Il est une «unité lexicale dont le sens peut être appréhendé et décrit en fonction des rapports de cette unité avec un domaine de la connaissance humaine » (L'Homme, 2002). La terminologie ne s'intéresse qu'aux termes qui dénotent des objets ou indiquent des notions. Cet aspect exclut les marques d'énonciation (pronoms personnels, adjectifs possessifs, adverbes de temps et de lieux), ainsi que les mots grammaticaux (*ibid.*). La discipline qui étudie les termes complexes est la terminologie, dont la tâche est de décrire la structuration de la connaissance spécialisée, en étudiant son système de désignation, ainsi que les unités conceptuelles auxquelles il renvoie.

La théorie générale (ou traditionnelle) de la terminologie a été fondée par Wüster, ingénieur autrichien, à la fin des années trente, dans la mouvance du Cercle de Vienne. Elle définit le terme comme le « représentant linguistique d'un concept dans un domaine de connaissances » (Felber, 1987) (Bourigault et Jacquemin, 2000). La vision adoptée est normalisatrice et présente quelques limites (Drouin, 2002). Elle pose une relation biunivoque avec la notion qu'il désigne, c'est-à-dire qu'un terme ne correspondrait qu'à une seule notion et que chaque notion ne pourrait être désignée que par un seul terme. La biunivocité est une utopie : il existe des phénomènes de polysémie et de synonymie qui la remettent en cause. Lorsqu'un terme est polysémique, il renvoie à plusieurs référents, y compris au sein d'un même domaine. Par

exemple[1], dans le domaine de l'agriculture, le terme *agneau* peut désigner « l'animal sur pied» ou « la viande d'agneau ». Un terme peut connaître des synonymes, et d'autres termes peuvent renvoyer à une même notion. Par exemple[1], dans le domaine de l'informatique, les termes *logiciel*, *programme* et *software* peuvent être considérés comme synonymes. Pour la théorie générale de la terminologie, les notions sont considérées comme des entités conceptuelles, et cet aspect prime sur leur représentation linguistique par le biais des termes, considérés comme de simples variables (Wüster, 1981) (Drouin , 2002). La démarche adoptée est onomasiologique : elle consiste à partir de la notion pour trouver le terme qui lui correspond. Nous préférons favoriser l'aspect textuel des termes (démarche sémasiologique). Enfin, l'analyse se situe traditionnellement au niveau du terme seul. Cette vision semble réductrice, car elle ne s'intéresse pas au niveau syntaxique se situant au-delà du mono-terme. Dans nos travaux, nous montrons que les unités lexicales complexes sont fondamentales.

2.4.3 Critères définitoires

Malgré les nombreuses définitions proposées, le terme reste difficile à identifier dans les textes. Nous présentons les principaux critères d'identification du terme relevés dans la littérature.

Critère formel (morpho-syntaxique)

Les termes complexes obéissent à des règles syntaxiques relativement stables, qu'il est possible de décrire afin d'identifier des candidats termes (Daille, 1994). Les structures peuvent être appréhendées en fonction des unités qui peuvent ou non les composer et des joncteurs qui relient ces unités (Drouin, 2002). Le critère syntaxique seul ne permet pas de discriminer une combinaison libre d'un terme pertinent. Pour une même structure, le statut terminologique n'est pas le même (*ibid.*) :

(1) Il utilise un ***langage de programmation***

[1] http://wall.jussieu.fr/~cjuilliard/cours3_deb.htm

*(2) Il parle à cette **fille de programmation***

Les phrases (1) et (2) contiennent la même structure syntaxique (*NOM-PREP-NOM*), mais seule la phrase (1) contient un terme complexe. Les limites de découpage ne sont pas toujours évidentes à identifier. A la gauche du terme complexe, la présence d'un déterminant constitue un indice de limite efficace. Mais la limite reste plus délicate à fixer du côté droit (Guilbert, 1965, Boulanger, 1979, Drouin, 2002). Les structures syntaxiques sont parfois récursives et il n'est pas possible de décrire d'une façon exhaustive de tels phénomènes de récursivité. Une autre difficulté concerne les ambiguïtés de découpage. Considérons le syntagme suivant (Kocourek, 1991) (Drouin, 2002):

Gardien d'asile de nuit

Deux découpages peuvent être envisagés :

[[Gardien d'asile] de nuit]

[Gardien [d'asile de nuit]]

Les critères formels ne permettent pas de pallier ce type d'ambiguïté. Un indice pourrait être des caractéristiques graphiques, telles que la présence de guillemets, ou encore typographiques (gras, italique, etc.). Toutefois, ces indices sont faiblement présents par rapport aux cas d'ambiguïté.

Critère fonctionnel (sémantique)

La propriété sémantique (fonction de l'unité complexe au sein du domaine) est un critère d'identification d'un terme pertinent par rapport à un groupe nominal. Un terme désigne une notion au sein du domaine, d'une façon permanente (Daille, 1995). Ce critère reste difficile à évaluer, notamment pour des non-experts. Une possibilité est la confrontation des termes dans une langue étrangère (*ibid.*). Le critère fonctionnel fait appel à des connaissances générales sur le monde et sur le domaine de spécialité. Le test de l'insertion constitue un indice de

lexicalisation : il n'est pas possible d'insérer un nouvel élément au sein des termes complexes (Guilbert, 1965) (Drouin, 2002).

Critère quantitatif

Un autre critère concerne la fréquence d'apparition de l'occurrence dans les textes. Les calculs de fréquence doivent prendre en considération non seulement la fréquence de l'occurrence, mais l'envisager en fonction de sa répartition dans l'ensemble des textes.

Critère pragmatique

Certains auteurs proposent de décrire le terme en appréhendant sa caractéristique pragmatique : un terme n'apparaît que dans des contextes précis, généralement dans des textes de spécialité (Drouin, 2002). Selon Pearson (1998), les termes sont utilisés dans certaines situations de communication (Drouin, 2002) :

(1) expert/expert

(2) expert/initié

(3) pseudo-expert/non initié

(4) enseignant/élève

De tels contextes sont propices à l'utilisation d'une terminologie relativement stable (*ibid.*).

Nous empruntons à Daille (1995) un récapitulatif des critères d'identification du terme complexe :

- D'un point de vue formel, il doit appartenir à une structure morpho-syntaxique précise.

- Il doit appartenir à un domaine de spécialité, et faire partie intégrante d'un vocabulaire technique.

- D'un point de vue fonctionnel, il doit posséder une traduction unique.

- D'un point de vue statistique, il doit apparaître dans les documents textuels un nombre significatif de fois.

Aucun des critères ne permet d'être complètement systématisé. Ils doivent être envisagés dans leur globalité (Drouin, 2002), mais le statut terminologique de l'unité ne peut jamais être une certitude, sans une étape de validation humaine. Le critère essentiel semble être la relation univoque du terme avec « l'objet » (Bourigault, 1994). Nous mettrons par la suite ces critères en confrontation avec un autre phénomène de combinaison lexicale préférentielle dont les caractéristiques sont proches, les collocations.

2.5 Collocations

Certains mots présentent des affinités et apparaissent fréquemment ensemble, sans constituer des locutions figées. Nous parlons de *café fort* en français, de *strong coffee* en anglais. Ces combinaisons lexicales, bien que préférentielles, ne sont pas totalement figées et peuvent parfois subir des modifications syntaxiques :

Un café très fort

Ce café est fort

Ce type « d'affinité » constitue un phénomène idiomatique, qui n'est pas uniquement déterminé par le sémantisme des constituants et qui varie d'une langue à l'autre. Afin d'exprimer le même sens, nous ne pouvons pas employer une autre combinaison lexicale dont le sémantisme correspondrait :

** powerful coffee*

La co-occurrence n'est pas acceptable d'un point de vue idiomatique. Les collocations[1] constituent un phénomène non nécessairement contigu, avec un degré de figement lexical moins contraint que pour les locutions. Si les locutions doivent avoir le même statut syntagmatique qu'une lexie simple, il n'en va pas de même pour les collocations, qui doivent être recensées à partir de la tête sémantique (on parle de la *base* d'une collocation et de *collocatif* pour son co-occurrent, pour reprendre la terminologie d'Hausmann (1989, 1997)). Il est parfois possible de substituer un élément d'une collocation par un synonyme, mais la combinaison est ressentie comme moins (ou pas du tout) appropriée (Nerima *et al.*, 2006) :

exercer vs. pratiquer une profession

Les collocations constituent un « intermédiaire » entre les expressions figées et les combinaisons libres (Tutin et Grossmann, 2002)[2]. Elles sont souvent considérées comme transparentes en réception (le sens se « devine ») alors que pour un locuteur non-natif, le choix des termes à produire ne va pas de soi (*ibid.*). Avant d'analyser les critères définitoires des collocations, nous introduisons les deux tendances principales qui envisagent le phénomène.

2.5.1 Approche statistique

Bien qu'il ne l'ait pas défini d'une façon précise, Firth est considéré comme l'un des premiers à avoir introduit le concept de collocation afin de désigner deux ou plusieurs mots qui apparaissent fréquemment ensemble, dans un voisinage proche (Firth, 1951) :

> You shall know a word by the company it keeps

[1] Le terme de « collocation » est d'usage courant en anglais, mais d'utilisation récente en français (Grossmann et Tutin, 2003). La littérature présente une terminologie disparate et souvent floue. Certains parlent de *préférences lexicales* (Wilks, 1975), de *restrictions de sélection (*Katz et Fodor, 1964), de *semi-phrasème* (Mel'cuk et al., 1995, Mel'cuk, 1998) ou encore de *collocations* (Benson, 1990, Smadja, 1993, Cruse, 1986).

[2] Mel'cuk (2003) parle également de « locution semi-figée » afin de mettre en valeur cet aspect « intermédiaire ».

Ce phénomène a été étudié dans le cadre d'une première approche, *contextualiste* : le sens d'un mot doit être étudié en fonction des mots avec lesquels ils co-occurrent. Cette approche suit la tradition de Firth, puis de fonctionnalistes anglais tels que Halliday et Sinclair. La notion de co-occurrence habituelle n'est toutefois pas sans poser des difficultés. Que doit-on entendre par habituel, fréquent (Williams, 2001) ? Le critère de la fréquence se mesure dans de nombreux travaux par l'application de formules statistiques à partir de vastes corpus, favorisée par la disponibilité de textes au format électronique (Church et Hanks, 1990, Smadja et McKweown, 1991, Smadja, 1993). L'idée est de collecter les combinaisons lexicales qui « apparaissent ensemble plus fréquemment que par pur hasard » (Smadja, 1993). Il existe de nombreuses mesures d'association permettant d'identifier les co-occurrences les plus fréquentes dans une collection de textes[1]. Les deux courants les plus employés sont le t-score et l'information mutuelle. L'algorithme t-score mesure le degré d'association entre deux éléments, en faisant émerger les combinaisons de fréquence élevée (Clear 1993, Dubreil, 2008) :

> by identifying frequent and very reliable collocations, offers the lexicographer a semantic profile of the node word and a set of particular fixed phrases, grammatical frames and typical stereotyped combinations

L'information mutuelle fait émerger des co-occurrences aux fréquences plus faibles (Church et Hanks, 1990, Dubreil, 2008) :

> compares the probability of observing x and y together (the joint probability) with the probabilities of observing x and y independently (chance)

Même si des outils statistiques peuvent être utiles pour l'extraction automatique, le seul critère statistique est discuté afin de décrire le phénomène collocatif. Les résultats sont dépendants de paramètres tels que la taille du corpus (Nerima *et al.*, 2006) ou le type de mesure adopté (Williams, 2001). Certaines collocations n'apparaissent qu'un nombre réduit de fois dans les corpus (Thoiron et Béjoint, 1989). A l'échelle du Web, la collocation *lire un*

[1] Daille (1994) établit une liste des différents types de mesures adoptées.

livre apparaît 227 000 fois dans les pages françaises indexées par le moteur de recherche *Yahoo*[1]. La collocation *lire une revue* apparaît 2140 fois. Doit-on conclure que *lire un livre* est plus remarquable? Même si le critère de la fréquence est important, il n'est pas suffisant afin d'analyser les contraintes lexicales. Selon Williams (2001), les mesures statistiques permettent de collecter des collocations « candidats » qui doivent ensuite passer par une validation humaine. Haussman (1997) déclare que « tout est idiomatique », parce qu'il est délicat de fixer une limite entre les combinaisons significatives et les combinaisons « banales » et que cette limite se situe certainement sur un continuum (Williams, 2001). Si Clas (1994) affirme qu'une collocation est « une unité de la langue », c'est parce qu'il existe des affinités sémantiques entre les mots que le seul critère de fréquence ne décrit pas.

2.5.2 Approche linguistique

La tradition lexicologique (Cruse 1986) et lexicographique (Hausmann 1989, Mel'cuk 1998) envisagent une conception « restreinte » de la collocation définie comme une association lexicale syntagmatique restreinte entre deux éléments entretenant une relation syntaxique. Dans la lignée de la formalisation élaborée par Mel'cuk (1998), Tutin et Grossmann (2002) posent la définition de la collocation suivante :

> Une collocation est l'association d'une lexie (mot simple ou phrasème[2]) L et d'un constituant C (généralement une lexie, mais parfois un syntagme par exemple *à couper au couteau* dans *un brouillard à couper au couteau*) entretenant une relation syntaxique telle que : C (le collocatif) est sélectionné en production pour exprimer un sens donné en cooccurrence avec L (la base). Le sens de L est habituel.

Cette approche met l'accent sur les paramètres syntaxiques et sémantiques des collocations. Tutin et Grossmann (2002) montrent que cette définition ne permet pas de résoudre le

[1] Juin 2008.

[2] *Locution figée* dans la terminologie de Mel'cuk.

caractère hétérogène des phénomènes collocatifs et proposent une typologie plus fine que nous décrivons dans la section (2.6).

2.5.3 Critères définitoires

La littérature abonde de travaux sur les collocations, mais peu de critères font l'unanimité. Ceux-ci sont fluctuables en fonction des collocations et aucun ne peut être appliqué à l'ensemble de la classe. Bien que le concept soit difficile à formaliser et que les caractéristiques varient selon les auteurs, nous analysons les critères définitoires les plus significatifs.

Critère de l'arbitraire

Benson (1990) parle d' « An arbitrary and reccurrent word combination », afin de souligner le caractère arbitraire des collocations. Selon Mel'cuk *et al.* (1995), il s'agit d'une association de mots conventionnelle qui doit être apprise telle quelle, et qu'il n'est pas possible de prévoir à partir du sens des mots qui la composent. Malgré le caractère transparent de certaines collocations, ainsi qu'une part de motivation sémantique[1], les collocations ne sont pas prédictibles[2]. Par exemple, la *pluie* est *torrentielle*, tandis que les *précipitations* ne le sont pas (Tutin et Grossmann, 2002), la preuve en est la différence entre les langues. Ce n'est qu'en les comparant que nous saisissons l'ampleur du phénomène (Hausmann, 1989) : le sens des constituants est altéré dans des contextes lexicaux précis. Par exemple, « *la pluie est forte en français, mais lourde (=heavy rain) en anglais* » (Mel'cuk, 1997). Cette caractéristique met en évidence l'importance d'un recensement des collocations : un apprenant qui ne maîtrise pas une langue étrangère ne dispose d'aucun moyen pour prédire ces dernières. Il en va de même pour la Traduction Automatique : on ne peut guère espérer de telles connaissances

[1] Certains travaux ont montré qu'il est possible de généraliser des contraintes de sélection à des ensembles de mots clés sémantiquement apparentés (Mel'cuk et Wanner, 1996, L'Homme, 1998).

[2] Il faut exclure du caractère arbitraire l'aspect syntaxique, car les collocations suivent des patrons morpho-syntaxiques précis.

idiomatiques d'une machine. L'aspect arbitraire est un indice quant au degré de signification des collocations, et permet de distinguer celles qui se traduisent librement, de celles qui nécessitent une connaissance idiomatique. Les sélections de restriction collocationnelles peuvent varier d'une langue à l'autre, et une collocation dans une langue peut être un syntagme libre dans une autre (Williams, 2001).

Critère de la transparence

Malgré le caractère arbitraire des collocations, le sens reste interprétable (Cruse, 1986, Hausmann, 1989). Cet aspect ne peut pas être généralisé à tous les cas. Certaines collocations ne sont pas totalement transparentes telles que *peur bleue* ou *colère noire* (Tutin et Grossmann, 2002). Les collocations ne conserveraient pas une totale autonomie de sens et seraient majoritairement semi-compositionnelles : le sens du mot-clé reste le même, mais le co-occurrent acquiert un sens différent (L'Homme, 1998). Il est préférable de dissocier la production et la réception d'une collocation : le sens «se devine », tandis qu'il est impossible pour un locuteur non-natif de produire la collocation adéquate. Pour la Traduction Automatique, ce critère compositionnel est essentiel. Par exemple, *peur bleue* ne peut pas être traduite de façon littérale, mais se traduit par *bad fright* ou *bad scare*.

Critère binaire

Les collocations sont considérées, le plus souvent, comme étant constituées de deux éléments, dont le statut n'est pas le même : il y a collocation lorsqu'un locuteur, voulant produire un sens donné, va sélectionner un co-occurrent de façon non libre, à partir d'une base donnée.

Critère de la dissymétrie (co-occurrence restreinte)

La base est considérée comme autonome car elle conserve son sens habituel, tandis que le collocatif dépend de la base (Haussmann, 1979, 1989, Mel'cuk, 2003). Par exemple, dans la collocation *peur bleue*, la base *peur* conserve son sens habituel, tandis que le collocatif *bleue* acquiert un sens différent de son sens habituel, dans ce contexte lexical précis. La collocation

n'est pas libre, mais apparaît comme une co-occurrence restreinte. A partir du choix de la base, seuls certains co-occurrents peuvent se combiner. Ainsi, pour produire le sens « intense », associée à la lexie *peur*, le co-occurrent va être *bleue*.

2.5.4 Recensement et formalisation

Le recensement des collocations d'un point de vue monolingue a donné lieu à divers ouvrages dont nous citons les plus courants, l'un pour la langue anglaise, le *BBI Dictionary of English Word Combinaison* (Benson *et al.*, 1986), l'autre pour la langue française, le *Dictionnaire Explicatif et Combinatoire du français contemporain* (Mel'cuk *et al*. 1984, 1988, 1992, 1999). D'un point de vue bilingue, nous présentons le projet *PAPILLON*.

BBI Dictionary of English Word Combinaison

Le *BBI Dictionary of English Word Combinaison* (Benson *et al.*, 1986) est un dictionnaire de collocations en anglais, regroupant 18 000 entrées et 90 000 collocations. Deux types de collocations sont recensées :

- la **collocation grammaticale** (ou *colligation*) est définie comme la co-occurrence d'un terme dominant, tel qu'un nom, adjectif, participe ou verbe, et d'une préposition, comme par exemple (Williams, 1999) :

  depend on

  dependence on

- Les **collocations lexicales** combinent plusieurs « mots pleins ». Elles peuvent se présenter sous différentes structures syntaxiques : verbe et nom, adjectif et nom, nom et verbe, nom et nom, adverbe et adjectif, adverbe et verbe.

Dictionnaire explicatif et combinatoire du français contemporain (DEC)

Le *Dictionnaire Explicatif et Combinatoire du français contemporain* (DEC) (Mel'cuk *et al*. 1984, 1988, 1992, 1999) est un dictionnaire de langue qui vise à représenter toutes les connaissances permettant d'employer un mot : description sémantique (définition), syntaxique (régime) et cooccurentielle (Fonctions Lexicales). Il comporte environ un millier d'entrées (noms, verbes, adjectifs, adverbes). Un article du DEC se compose de trois zones majeures :

- **Zone phonologique**

- **Zone sémantique**

- **Zone de combinatoire :**

  - *Stylistique* (contexte textuel, comme par exemple, « littéraire », « familier », etc.) ;

  - *Morphologique* (partie du discours, déclinaison, formes irrégulières, etc.) ;

  - *Syntaxique* ;

  - *Lexicale restreinte* : le DEC indique les substitutions sémantiques possibles sur le plan paradigmatique (synonymes, antonymes, etc.). La modélisation de ces informations sémantiques s'appuie sur la notion de Fonction Lexicale (FL) proposée par (Mel'cuk, 1997) :

> L'expression [d'un] sens peut être décrite par une fonction (au sens mathématique du terme) f qui associe, à tout x pour lequel ce sens peut être exprimé, tous les y possibles : f(x) = y.

Par exemple, pour la fonction *intensité (très),* nous présentons trois lexies x à laquelle sont associées des co-occurrents :

tres(malade) = {très, gravement…}

tres(pleurer) = {amèrement, à chaudes larmes, comme une madeleine…}

tres(pluie) = {grosse, diluvienne, violente…}

La lexie x est nommée l'argument de f, et l'ensemble de ses co-occurrents constitue sa valeur. Dans l'exemple cité, la Fonction Lexicale est modélisée par *très*. Les arguments sont *malade*, *pleurer* et *pluie*. Les valeurs sont *très*, *gravement*, *amèrement*, *à chaudes larmes*, etc.
Dans la théorie Sens-Texte, Mel'cuk (1997) propose un modèle formel de la description sémantique d'une langue. Il distingue deux niveaux de modélisation des phénomènes sémantiques :

- **Choix lexicaux paradigmatiques** : Mel'cuk (1997) décompose le sens des lexies. Prenons la phrase suivante :

Je crois que Pierre est venu, mais je n'en suis pas certain.

La décomposition sémantique de cette phrase s'exprime de la façon suivante (*ibid.*) :

'Je crois que Pierre est venu, mais [tout] en ayant la croyance « Pierre est venu », | je suis disposé à admettre que Pierre n'est pas venu'.

Ces règles formelles constituent des « définitions lexicographiques des lexies citées, ce sont des décompositions sémantiques, ou des formules moléculaires du sens » (*ibid.*).

- **Choix lexicaux syntagmatiques** : il s'agit des collocations. La lexicographie explicative et combinatoire s'appuie sur l'idée que les phénomènes de co-occurrence font appel à un nombre restreint de sens généralisables. Par exemple, le sens *bon* ('tel que le locuteur l'approuve') ne s'exprime pas d'une façon libre, mais dépend de la lexie utilisée (*ibid.*) :

Bon(conseil) = précieux

Bon(temps) = beau

Bon(choix) = heureux

Bon(se porter) = comme un charme

Toutes les expressions qui sélectionnent cette notion sémantique connaissent des contraintes qui en font des collocations. Ces sens généraux constituent des FL. Les FL présentent deux propriétés essentielles (*ibid.*) : elles sont « peu nombreuses » (une soixantaine) et elles sont «universelles », car elles existent dans toutes les langues. Par ailleurs, l'analyse sémantique ne doit pas être trop « poussée », trop nuancée (Mel'cuk, 1988) :

> Le réglage de cet instrument, c'est-à-dire le degré de précision ou de résolution exigé, doit être approprié à la tâche ; cela veut dire, entre autres, que le chercheur ne doit pas être trop précis dans sa recherche des nuances sémantiques.

La théorie de Mel'cuk (1997) présente néanmoins certaines limites. Bien que certaines FL soient généralisées, il n'en va pas de même en ce qui concerne les langues de spécialité, et certains cas semblent plus isolés et de fait, moins « efficaces ». Fontenelle (1996) souligne que :

> Pour formaliser le discours spécialisé utilisé pour parler d'un terme donné, les théories de Mel'cuk ne sont probablement pas les plus appropriées parce qu'elles ne permettent de coder que les relations standard de la langue générale et les langues de spécialité ont le plus souvent recours à des relations très spécifiques.

Ensuite, cette classification peut se présenter utile d'un point de vue scientifique, mais elle ne se révèle pas très accessible pour un utilisateur. C'est pour cette raison que deux versions simplifiées telles que le *DAFLES* (Verlinde *et al.*, 2003) et *DiCo* (Polguère, 2000a, 2003, 2005) ont été proposées. Dans cette lignée, le projet intitulé le *Lexique Actif du Français* (LAF) (Polguère, 2000b) vise à une vulgarisation du DEC pour le grand public.

Projet Papillon

Le projet Papillon[1] est une ressource collaborative qui vise à créer un environnement multilingue de recherche dictionnairique en ligne, comprenant entre autres l'anglais, le français, le japonais, le malais, le lao, le thaï et le vietnamien. Il s'appuie sur des ressources existantes, l'objectif étant de rassembler un maximum de ressources de façon coopérative. La base lexicale distingue trois niveaux différents pour la gestion des dictionnaires existants : les *limbes*, le *purgatoire* et le *paradis* (Mangeot, 2002). Les *limbes* sont constituées de dictionnaires stockés dans leur format original. Le *purgatoire* ne contient que des dictionnaires au format XML mais ayant leur structure d'origine. Le *paradis* contient les volumes consituant le dictionnaire Papillon. La macrostructure du dictionnaire est une structure pivot avec un volume monolingue pour chaque langue et un volume pivot au centre (Mangeot *et al.*, 2003). La microstructure des articles est basée sur la lexicographie explicative et combinatoire issue de la théorie sens-texte (Mel'cuk, 1997) (*ibid.*). Voici un exemple de traductions de collocations pour l'entrée *appareil* du français vers l'anglais :

Figure 6. Traductions de collocations dans le projet Papillon

[1] http://www.papillon-dictionary.org/

2.6 Indices de figement

Entre les associations libres du type de *pomme de Jean* et les locutions figées telles que *pomme d'Adam*, il existe divers degrés de figement, laissant place à des zones d'incertitude[1]. Par exemple, la locution *crise de croissance,* sans être une séquence libre, autorise des critères de séparabilité comme dans le cas de (Léon, 2004a) :

*crise **aiguë** de croissance*

Les deux principaux degrés de figement entre les locutions et les collocations ont été mis en évidence pour la première fois par Bally (1909), sous l'appellation d' « unités phraséologiques » (locutions) et de « groupements usuels » (collocations) (Grossmann et Tutin, 2003). Au-delà de cette distinction, Tutin et Grossman (2002) montrent que les critères définitoires ne s'appliquent pas à toute la classe collocationnelle et proposent une typologie plus fine. Les unités lexicales complexes doivent être envisagées sur un continuum, plutôt que de considérer des frontières nettes pour chaque catégorie. Nous proposons une analyse contrastive de ces phénomènes, en espérant, sinon établir des frontières nettes, du moins éclaircir les différentes réalités de combinatoire lexicale et proposer une typologie plus fine en fonction des critères de figement et envisager des traitements de traduction adaptés.

2.6.1 Opacité sémantique

Le principe de compositionnalité désigne le fait que « le sens d'un tout est une fonction du sens de ses parties et de la façon dont elles se combinent » (Bouillon, 1998). Lorsqu'une unité lexicale complexe est compositionnelle, la signification globale est égale à la somme du sens de ses constituants. D'après Gross (1996), il existe trois types d'opacité sémantique pour les combinaisons de termes :

- L'opacité peut-être totale, c'est-à-dire qu'aucun des constituants ne conserve son sens habituel (Gross, 1996) :

[1] Fontenelle (1997) appelle cette zone floue « fuzzy area ».

clé des champs

- L'opacité peut être partielle, c'est-à-dire que seul l'un des deux constituants perd son sens habituel (*ibid.*) :

clé anglaise

- L'opacité peut être absente et le sens est alors transparent (*ibid.*) :

clé neuve

Les unités lexicales complexes ne sont pas, le plus souvent, interprétables en décomposant le sens habituel de ses éléments, mais les caractéristiques sont variables. En ce qui concerne les locutions, la combinaison n'est jamais totalement transparente, mais il arrive que le sens soit métaphorique. En ce qui concerne les collocations, Tutin et Grossmann (2002) proposent une typologie en fonction de leur degré de figement et du paramètre de la compositionnalité. Nous empruntons à Grossmann et Tutin (2003) une typologie des locutions et des collocations du plus au moins figé.

Locutions figées « opaques »

Les « locutions figées opaques » désignent des locutions dont le sens est totalement opaque, comme par exemple (Grossmann et Tutin, 2003) :

cordon bleu

Dans cette combinaison aucun des éléments n'a conservé son sens habituel.

Locutions figées « imagées »

Les « locutions figées imagées »[1] sont celles dont le sens reste analysable par métaphore ou par métonymie, bien que l'association soit imprédictible (Grossmann et Tutin, 2003) :

œil de bœuf

Collocations « opaques »

Les collocations opaques sont celles dans lesquelles le sens du collocatif prend un sens différent que son sens habituel. L'association est arbitraire, et seule la base conserve son sens habituel. Sur le plan sémantique, la collocation est non transparente (Tutin et Grossmann, 2002) :

peur bleue

Ici, l'adjectif *bleu* ne désigne pas une couleur mais marque l'intensité de la peur (Tutin et Grossmann, 2002). Les collocations opaques sont celles qui sont les plus proches des locutions du point de vue du figement.

Collocations « transparentes »

En ce qui concerne les collocations transparentes, le sens est interprétable, mais le codage de la collocation n'est pas prédictible, comme par exemple (Tutin et Grossmann, 2002) :

faim de loup

[1] Cowie (1981, 1998) parle de « figurative idioms » (Grossmann et Tutin, 2003).

Même si le sens de *faim de loup* est interprétable, une connaissance idiomatique est nécessaire pour produire cette collocation.

Collocations « régulières »

Dans le cas des collocations dites « régulières », l'association est motivée et transparente. En général, le collocatif inclut le sens de la base ou a un sens très générique, comme par exemple (Tutin et Grossmann, 2002) :

grande tristesse

Les collocations régulières sont celles qui sont les plus proches des expressions libres.
Du point de vue de la traduction automatique, cette typologie peut constituer une aide à la classification des phénomènes. Même si le critère d'opacité n'est pas le même entre les langues, il doit être pris en compte, en intégrant des caractéristiques interlingues à cette typologie. C'est ce que nous proposons dans notre méthodologie.

2.6.2 Propriétés transformationnelles

Si les locutions sont nécessairement contigües et ne peuvent pas subir de transformations syntaxiques, le cas des collocations n'est pas généralisable. Certaines collocations connaissent un figement syntaxique, contrairement à d'autres qui s'apparentent presque à des associations libres. En ce qui concerne les propriétés transformationnelles, un syntagme composé d'un nom et d'un adjectif peut traditionnellement faire l'objet de transformations syntaxiques Gross (1996) :

Un livre difficile

La difficulté de ce livre

Ce livre est difficile

Ce type de transformation est complètement impossible avec les locutions (*ibid.*) :

Un cordon bleu

**Le bleu de ce cordon*

**Ce cordon est bleu*

Le cas des collocations est moins généralisable et les critères de transformation ne peuvent pas s'appliquer à l'ensemble de la classe. Certaines collocations autorisent des transformations syntaxiques proches des syntagmes libres :

Une grande tristesse

La grandeur de cette tristesse

Cette tristesse est grande

D'autres collocations autorisent certaines transformations, mais pas d'autres :

Un café noir

**La noirceur de ce café*

Ce café est noir

Enfin, certaines collocations n'autorisent aucune transformation et sont proches du fonctionnement des locutions :

Une peur bleue

**Le bleu de cette peur*

**Cette peur est bleue*

2.6.3 Critère référentiel (Libre actualisation)

Lorsqu'une unité lexicale complexe n'est pas figée, le co-occurrent constitue une modification du nom. Prenons le syntagme libre suivant :

pull-over bleu

L'adjectif épithète *bleu* modifie le nom *pull-over* et lui apporte une caractérisation. En revanche, lorsque la séquence est figée, elle fonctionne comme un tout et le co-occurrent ne constitue pas une modification. Le critère référentiel est un critère définitoire qui s'applique à l'ensemble de la classe des locutions. Les locutions ont une détermination globale, et chaque élément ne peut pas être déterminé séparément (Gross, 1996). Pour reprendre l'exemple de *cordon bleu*, le référent auquel renvoie la locution n'est pas un cordon auquel on apporte une modification, mais l'ensemble de la locution fait référence au signifié « cordon bleu ».

Si la non possibilité de libre actualisation concerne l'ensemble de la classe des locutions, il n'en va pas de même pour les collocations, où seulement une partie de la classe répond au critère, qui est plus délicat à utiliser (Tutin et Grossmann, 2002). Tous les collocatifs n'attribuent pas la même valeur référentielle à la collocation (*ibid.*). Certains collocatifs ont une valeur qualifiante, comme dans l'exemple de (*ibid.*). :

célibataire endurci

L'adjectif *endurci* qualifie le substantif *célibataire* et la valeur référentielle de l'ensemble de la collocation est à opposer à celle d'une locution. Toutefois, il arrive que certains collocatifs aient une valeur typante (*ibid.*). Reprenons l'exemple suivant (*ibid.*) :

café noir

Le collocatif *noir*, bien qu'il indique une propriété du café, fait aussi référence à un type particulier de café. De telles collocations peuvent être perçues comme des unités référentielles tandis qu'elles sont semi-compositionnelles d'un point de vue sémantique (*ibid.*).

2.6.4 Degré de figement

Le figement bloque le plus souvent la possibilité de paradigmes synonymiques. Toutefois, cette caractéristique ne s'applique pas à l'ensemble des unités lexicales complexes. Certaines locutions acceptent une possibilité de paradigme, comme dans l'exemple (Gross, 1996) :

vin rouge

vin blanc

vin gris

Ces locutions connaissent une opacité sémantique, mais une prédication n'est pas possible. D'autres locutions connaissent un sens transparent, comme dans le cas de (Gross, 1996) :

fait historique

Toutefois, certaines modifications syntaxiques sont possibles, et pas d'autres :

ce fait est historique

un fait d'histoire

**un fait très historique*

2.7 Conclusion

Malgré l'importance des unités lexicales complexes et malgré l'abondance de la littérature traitant du phénomène, les critères de définition des locutions et des collocations ne sont pas généralisables. Afin de dissocier une collocation d'une locution, le degré de figement doit être envisagé sur un continuum. Aux extrémités de ce continuum, deux pôles doivent être distingués :

- une *collocation* est constituée d'un terme, accompagnée d'un co-occurrent qui le qualifie comme dans l'exemple :

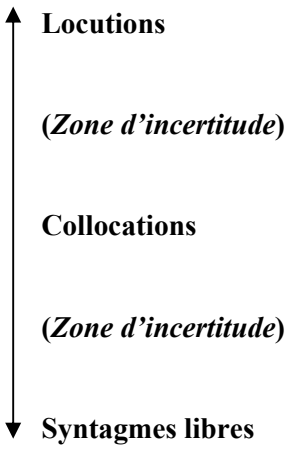
pluie diluvienne

Le nom *pluie* désigne une notion en météorologie, qu'on qualifie comme étant *diluvienne*.

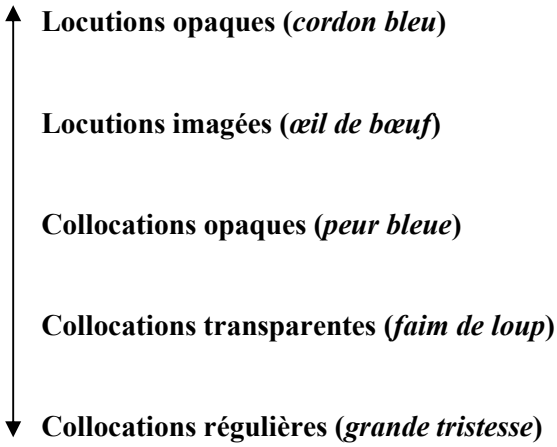
- une *locution* (ou un *terme complexe*) désigne en lui-même une notion :

autan blanc

Ici *autan blanc* renvoie à une seule notion (à savoir le « vent »), et l'adjectif ne caractérise en rien la couleur du vent mais fait partie intégrante du terme. Toutefois, au-delà des cas « extrêmes », les frontières entre les locutions et les collocations restent floues et il est préférable d'envisager les deux phénomènes sur un continuum, du plus au moins figé :



Entre chacune des notions, des zones d'incertitude sont présentes. Une typologie plus fine au sein de chaque classe est nécessaire pour proposer des traitements adaptés. Considérons sur un axe les différents types de locutions et de collocations en fonction du critère de l'opacité sémantique, du plus au moins opaque :



Le figement et l'opacité des unités lexicales complexes va constituer un critère essentiel dans notre méthodologie de traduction. Dans le chapitre suivant, nous présentons les travaux « traditionnels » de traitement automatique des unités lexicales complexes.

Chapitre 3. Traitement automatique des unités lexicales complexes

3.1 Introduction

L'acquisition automatique de traductions d'unités lexicales complexes se heurtent à différentes difficultés (Morin *et al.*, 2004).

Polysémie des unités lexicales

Lorsque plusieurs traductions sont possibles pour une unité lexicale en fonction de son usage, la sélection de la traduction adéquate fait appel à des connaissances générales sur le monde que l'on ne peut guère espérer d'une machine. Par exemple, le nom *appareil* compte douze traductions dans notre dictionnaire bilingue *Collins Pocket*. Le recours aux contextes des unités sources et cibles peut être un élément pour la résolution de l'ambigüité lexicale : *appareil ménager*, *appareil digital*, etc.

Découpages sémantiques différents

Même pour des langues « proches » telles que le français et l'anglais, il n'existe pas de parfaite équivalence de sens entre les mots d'une langue à l'autre. Une conception naïve consisterait à penser que la langue serait comparable à un « répertoire » de mots chacun correspondant à une chose. Il s'agirait de proposer une nomenclature en langue cible équivalente à la langue source. En réalité, les langues possèdent chacune un découpage sémantique qui lui est propre (Saussure, 1916, Martinet, 1960). Si les découpages sémantiques peuvent parfois être les mêmes (comme l'exemple de *souris* qui en français désigne l'usage *ANIMAL* ou l'usage *INFORMATIQUE*, de la même façon que pour l'anglais *mouse*), ce n'est pas systématiquement le cas. Par exemple, en anglais, le mouton (*ANIMAL*), *sheep* est distingué du mouton (*VIANDE*), *mutton* (Saussure, 1916).

Idiosyncrasie

Les collocations, même lorsqu'elles conservent une part de motivation sémantique, se situent principalement du côté des aspects idiosyncrasiques de la langue, plutôt que de celui des régularités (Grossmann et Tutin, 2003).

Non équivalence de longueur de la combinatoire

Les termes complexes ne se traduisent pas nécessairement par une combinaison de même longueur :

- Un terme complexe en français peut être traduit par un terme simple en anglais, ou inversement :

coup de pied > *kick*

appareil photographique > *camera*

- Parmi les termes complexes, la longueur peut varier (Morin *et al.*, 2004) :

essence d'ombre > shade tolerant species

Cette difficulté est décrite sous le terme de « fertilité » dans les travaux de (Brown *et al.*, 1993). Cette caractéristique est rarement prise en compte dans les travaux de traduction de termes complexes, une traduction mot à mot étant la plus souvent adoptée (Morin *et al.*, 2004).

Non compositionnalité

La traduction d'une unité lexicale complexe n'est pas systématiquement traduite par la somme de ses composants (Melamed, 2001). Par exemple, *caisse claire* est traduite par *kick drum*, ou *kick* n'est pas la traduction littérale de *claire*.

Variations linguistiques

Une même combinaison lexicale peut se présenter sous différentes formes suite à des variations morphologiques, syntaxiques ou sémantiques, et doivent être prises en compte dans le processus de traduction (Morin *et al.*, 2004). Par exemple, les termes complexes *aménagement de la forêt* et *aménagement forestier* sont traduits par le même terme anglais *forest management*.

« Trous » lexicaux

Il arrive qu'une unité lexicale au sein d'une langue source n'ait pas une équivalence exacte en langue cible. Par exemple, en anglais, il n'existe pas une traduction littérale de *forcer un barrage*, la traduction dépend du contexte situationnel :

*to **drive** through a roadblock*

*to **run** through a roadblock, etc.*

Méthodes traditionnelles de traduction

Les méthodes d'acquisition automatique de termes à partir de textes, qu'il s'agisse de corpus parallèles ou de corpus comparables, s'appuient généralement sur deux phases plus ou moins dépendantes l'une de l'autre :

- une **extraction monolingue** des unités lexicales complexes en langue source d'une part et en langue cible d'autre part.

- un **alignement ou une mise en correspondance** des unités lexicales complexes.

Dans un premier temps, nous présentons les techniques d'extraction de terminologie monolingue, les principaux courants et les travaux existants (3.2). Si la frontière entre les locutions et les collocations n'est pas nette et doit être envisagée sur un continuum, il en va de même pour la tâche d'extraction automatique. Les méthodes d'identification automatique de locutions et de collocations restent sensiblement les mêmes (L'Homme, 2001) :

> Notons que les cloisons entre extracteurs de collocations et extracteurs de termes ne sont pas étanches : les extracteurs de collocations relèvent des termes complexes ; les extracteurs de termes complexes relèvent forcément des collocations.

Nous décrivons ensuite les techniques d'alignement, qu'il s'agisse de méthodes à partir de corpus parallèles ou à partir de corpus comparables. Malgré des techniques bien rodées, les corpus parallèles restent des ressources rares. Les corpus comparables, plus faciles d'accès, présentent plus de difficultés pour mettre les termes en correspondance. Nous présentons ces méthodes traditionnelles (3.3), avant d'en montrer les limites et de proposer dans le chapitre suivant une gigantesque base de données lexicales exploitable pour l'acquisition automatique de traductions, le Web.

3.2 Méthodes d'extraction automatique

Les recherches en extraction terminologique à partir de corpus sont récentes. L'objectif est de collecter dans des textes des unités simples ou complexes susceptibles d'être des termes pertinents dans un domaine. La phase d'acquisition terminologique peut être considérée comme interactive (Drouin, 2002) :

> Elle est entièrement automatique, mais la chaîne de travail est qualifiée d'interactive dans la mesure où le terminologue valide les résultats obtenus automatiquement par le logiciel

Parmi les groupes nominaux, des filtres linguistiques et/ou statistiques permettent de dégager un ensemble de candidats-termes. Ces premiers résultats contiennent du bruit, seul un certain nombre est pertinent :

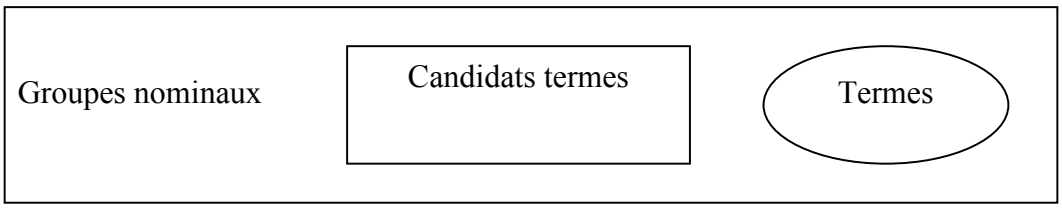


Figure 7. Termes pertinents parmi les groupes nominaux[1]

Les logiciels d'extraction terminologique s'appuient sur diverses méthodes : on distingue ceux qui s'appuient sur des critères statistiques (3.3.1), de ceux basés sur des critères morpho-syntaxiques (3.3.2). Nous parlons enfin des méthodes dites *mixtes*, qui ont recours à la fois à ces deux critères (3.3.3)[2].

[1] Schéma de Sta (1995).

[2] Nous présentons les principaux logiciels d'extraction automatique de terminologie en français, mais nous ne prétendons pas à l'exhaustivité.

3.2.1 Méthodes statistiques

Diverses méthodes statistiques ont été proposées pour la sélection d'unités lexicales complexes au sein d'une langue. Les outils statistiques repèrent les associations préférentielles, sans connaissance linguistique. Parmi les plus connus, citons le logiciel d'Apprentissage Naturel Automatique (**ANA**) (Enguehard, 1993, Enguehard et Panterra, 1995) qui est un logiciel d'acquisition automatique de terminologie pour la construction du thésaurus d'un domaine, à partir d'un vaste corpus de textes bruts. L'architecture du logiciel passe par deux modules, l'un dit de « familiarisation » qui extrait automatiquement des éléments de connaissance sous la forme de listes, l'autre dit de « découverte » qui sélectionne la terminologie du domaine à partir des listes et du corpus de textes.

Le logiciel **MANTEX** (Oueslati, 1999, Rousselot *et al.*, 1996) est un outil d'extraction terminologique qui s'appuie sur le repérage de segments répétés, à partir de textes non étiquetés.

Les méthodes purement statistiques présentent toutefois certaines limites (Daille, 1994, Véronis, 2000a). La rareté d'unités lexicales complexes rend les choix de statistiques délicats. De plus, les unités lexicales complexes « semi-figées » autorisent des transformations linguistiques qui posent les limites de modèles statistiques simples.

3.2.2 Méthodes linguistiques

Afin de palier les limites des modèles statistiques, certaines méthodes proposent une approche linguistique. Les critères morpho-syntaxiques s'appuient sur des connaissances a priori des structures syntaxiques. L'hypothèse est que les termes obéissent à des règles de combinaison stables, et il est possible de définir un nombre limité de schémas morpho-syntaxiques pré-établis (essentiellement des groupes nominaux) repérés d'une façon automatique. Une telle méthode s'appuie sur un certain nombre de présupposés (L'Homme, 2001) :

- les textes en langue de spécialité sont riches de termes représentatifs de la connaissance du domaine.

- Un terme représentatif est utilisé à plusieurs reprises dans le corpus.

- Une majorité de ces termes est composée de noms.

- Nombre de ces termes sont complexes.

- Ces termes complexes font appel à un nombre réduit de structures syntaxiques : il s'agit généralement d'un nom modifié par un autre terme. Les structures syntaxiques principales sont les suivantes (*ibid.*) :

Structures syntaxiques	**Exemples**
Nom + Adjectif	*Intelligence artificielle*
Syntagme Prépositionnel (avec nom)	*Robinet de commande*
Syntagme Prépositionnel (avec verbe)	*Machine à coudre*
Nom + Nom	*Page Web*
Combinaison des séquences ci-dessus	*Temps de conduction auriculaire*

Figure 8. Structures syntaxiques des syntagmes nominaux[1]

L'outil **TERMINO** est une application pionnière de l'acquisition automatique de termes (David et Plante, 1990) (en français ou en anglais). Ce logiciel est basé sur le repérage de syntagmes nominaux qui constituent des candidats termes. La définition des termes se fonde sur les synapsies de Benveniste (1966). Les candidats-termes sont générés à partir des dépendances entre tête et complément au sein de la structure des syntagmes nominaux extraits par l'analyseur.

FASTR (*Filtrage et Acquisition Syntaxique de TeRmes*) (Jacquemin, 1997) est un analyseur syntaxique permettant l'identification de variantes de termes à partir de corpus, à l'aide d'une liste de termes valides fournie en entrée. Les variations sont classées selon trois catégories :

[1] L'Homme (2001).

- **Variantes syntaxiques** :

Mesure de volume et de flux / Mesure de flux

- **Variantes morpho-syntaxiques** :

Flux de sève mesurés / Mesure quotidiennement le flux

- **Variantes sémantico-syntaxiques** :

Evaluation du flux / Mesure de flux

SYMONTOS (Velardi *et al.*, 2001) est un environnement proposant des outils afin de repérer des termes simples et complexes à partir de corpus, et proposer des concepts associés (Bourigault *et al.*, 2004).

Le logiciel **SYNTEX**[1] (initialement Lexter) (Bourigault, 1994, Bourigault et Fabre, 2000) est un outil d'extraction terminologique qui extrait des candidats termes, à partir d'un corpus étiqueté et désambigüisé. Il effectue une analyse syntaxique de surface dédiée au repérage et à l'analyse de syntagmes nominaux. Les candidats termes extraits se présentent sous la forme d'un réseau.

L'introduction de connaissances linguistiques est toutefois relativement coûteuse, et n'est pas indépendante des langues. Divers auteurs ont présenté des approches mixtes, mêlant les stratégies statistiques et linguistiques.

3.2.3 Méthodes mixtes

Afin de pallier les contraintes des méthodes linguistiques ou statistiques, certains travaux mêlent les deux stratégies. On parle de stratégies *hybrides* ou *mixtes* (L'Homme, 2001).

[1] http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html

Le logiciel **ACABIT** (*Automatic Corpus-based Aquisition of BInary Terms*) extrait des candidats termes à partir d'un corpus préalablement étiqueté et désambigüisé (Daille, 1994, 1999). La méthode est basée sur des traitements linguistiques associés à des filtres statistiques :

- **Analyse linguistique** : des séquences nominales sont extraites du corpus étiqueté et sont regroupées sous la forme de candidats termes binaires. Par exemple, *réseau de transit à satellite* constitue deux candidats termes binaires, *réseau de transit* et *réseau à satellite*. Les termes extraits doivent être conformes à un nombre limité de patrons syntaxiques, du type :

  Nom-Adjectif > Emballage biodégradable

  Nom1-Nom2> Diode tunnel

  Nom1 à (Det)Nom2> Assignation à la demande

  Nom1 de (Det) Nom2> Protéine de poissons

  Nom1-Prep(Det)-Nom2 > Multiplexage en fréquence

  Nom1 à Vinf > Viandes à griller

- **Filtre statistique** : les candidats termes sont filtrés au moyen d'un calcul statistique, le *log-likehood ratio* (Dunning, 1993).

XTRACT (Smadja, 1993) est un logiciel d'extraction automatique de collocations basé sur des calculs statistiques, ainsi que sur un filtrage linguistique. L'outil est composé de trois modules :

- **Collocations binaires** : une première phase extrait des couples de mots dont la fréquence est élevée et dont la distance est fixe.

- **Expansion des collocations** : l'étape précédente est répétée de façon itérative afin d'acquérir des collocations de plus grande longueur.

- **Etiquetage** : les collocations sont étiquetées.

Smadja (1993) distingue trois types de collocations : les *collocations prédicatives* (*predicative relations* en anglais) du type de *make/decision*, les *syntagmes figés* (*rigid noun phrases* en anglais) comme *foreign exchange* et les « *phrases à trous* » (*phrasal templates*), comme dans l'exemple :

Temperatures indicate day's high and overnight low to 8 a. m.

FipsCo (Goldman *et al.*, 2001), (Nerima *et al.*, 2003), (Seretan *et al.*, 2004) est un extracteur de collocations basé sur un système d'analyse syntaxique, le système Fips, développé au LATL (Laenzlinger et Wehrli, 1991), (Wehrli, 1997). La méthode s'appuie à la fois sur des critères statistiques (le likelihood ratio) et sur une analyse syntaxique, qui permet le repérage de collocations dont les éléments ne sont pas contigus.

3.3 Méthodes de traductions d'unités lexicales complexes

3.3.1 Corpus parallèles

Un texte parallèle aligné (parfois appelé « bitexte » (Harris, 1988) ou « multitexte ») désigne un ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots. Bien que le recours à des corpus parallèles ne soit pas encore très utilisé pour l'édition de dictionnaires classiques, leur utilisation est largement plus importante dans le domaine de la terminologie et de la conception de lexiques computationnels (Véronis, 2000a). Le schéma suivant donne une illustration d'un corpus parallèle :

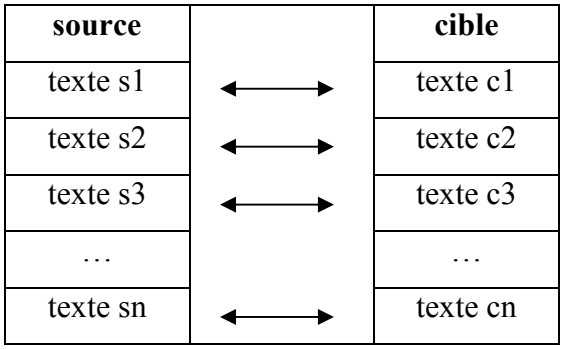


Figure 9. Schématisation d'un corpus parallèle[1]

Les documents parallèles peuvent être des ressources externes, comme par exemple des manuels techniques traduits, des ouvrages traduits (textes religieux, etc.) ou des débats multilingues. Certains travaux ont également recours à des méthodes d'alignement automatique de textes traduits. Nous empruntons à Véronis (2000a) un état de l'art des techniques d'alignement (pour un état exhaustif, se référer à lui).

Alignement de phrases

On distingue généralement deux courants de méthodes d'alignement, l'un dérivant de Kay et Röscheisen (1988) qui s'appuie sur un ancrage lexical, l'autre dérivant de Gale et Church (1993) et de Brown *et al.* (1991b), qui utilisent des méthodes de corrélations de longueurs des phrases. Malgré des méthodes différentes, certaines hypothèses sont proches. Les hypothèses d'alignement s'appuient sur les présupposés suivants (Véronis, 2000a) :

- L'ordre des phrases du texte source et du texte cible sont identiques ou proches.

- Les textes contiennent un nombre réduit de suppressions ou d'adjonctions.

Le courant issu de Kay et Röscheisen (1988) est fondé sur une méthode d'ancrage lexical. L'hypothèse de base de Kay et Röscheisen (1988) est qu'un couple de phrases ne peut être en correspondance que si les mots qui la composent le sont aussi. Les informations utilisées ne

[1] Schéma de Zweigenbaum (http://www.limsi.fr/~pz/p11m2r-2006/corpus-paralleles.pdf).

sont extraites que des textes à aligner, sans autre ressource externe. La méthode prend pour point de départ des phrases candidates avec une probabilité raisonnable de correspondance pour la première et la dernière phrase. Les phrases intermédiaires sont certainement en correspondance dans un couloir diagonal plus ou moins étroit. La méthode compare ensuite la distribution des mots, en partant de l'hypothèse que si un couple de mots a des distributions similaires, la probabilité qu'ils soient une traduction l'un de l'autre est forte. Les mots alignés forment des points d'ancrage permettant d'affiner l'alignement des phrases de départ. Une itération de la procédure permet d'obtenir un alignement maximal. Kay et Röcheisen (1988) montrent que même si un alignement en mots est une tache difficile, un alignement en mots même grossier peut conduire à un alignement en phrases satisfaisant.

Les courants issus de Gale et Church (1993) et de Brown *et al.* (1991b) s'appuient sur une méthode de corrélation des longueurs de phrases. Gale et Church (1993) utilisent une méthode d'alignement qui s'appuie également sur une information extraite des textes. La méthode est fondée sur une comparaison de la longueur des phrases dans le texte source et dans le texte cible. L'hypothèse est que si deux phrases sont la traduction l'une de l'autre, leur longueur doit être proche. La méthode s'appuie sur l'hypothèse d'un rapport constant de longueur de phrases en terme de nombre de caractères. Il est admis que le rapport de longueur de caractères entre deux langues est relativement stable, comme par exemple le fait qu'un texte français a tendance à être plus long que sa traduction anglaise (Véronis, 2000a). Des algorithmes permettent d'effectuer des mesures de dissimilarité entre les phrases du texte source et du texte cible, prenant en compte les phénomènes d'alignement attendus tels que des cas d'omission, d'addition ou de fusion. Brown *et al.* (1991b) applique une méthode qui s'appuie sur le même type d'hypothèses de longueurs de phrases.

De nombreuses méthodes d'alignement de phrases s'appuient sur ces deux hypothèses, une majorité combinant les deux idées. Debili et Sammouda (1992) effectuent un alignement de phrases basé sur un ancrage lexical via un dictionnaire bilingue. Simard *et al.* (1992), Church (1993), Johansson *et al.* (1993) et McEnery et Oakes (1995) ont recours à un ancrage lexical basé sur le repérage de *cognates*, combiné à une méthode dans la lignée de Gale et Church. L'ancrage lexical s'appuie sur le repérage de *cognates*, c'est-à-dire d'unités qui sont identiques en langue source et en langue cible ou qui sont graphiquement proches, comme par

exemple *language* en anglais et *langue* en français (Véronis, 2000a). L'utilisation de *cognates* est surtout préconisé dans le cadre de langues apparentées. Langlais et El-Beze (1997) et Melamed (2000) montrent la nécessité de combiner différents types de critères, tels que par exemple le lexique, les *cognates*, la longueur des phrases.

Alignement de mots et expressions

Dans les méthodes précédentes, l'ancrage lexical constitue un indice d'alignement en phrases. A l'inverse, l'alignement en phrases peut constituer un point de départ pour un alignement plus fin, à savoir un alignement en mots (Dagan et Church, 1994, Resnik et Melamed, 1997, Jones et Somers, 1997, Choueka *et al.*, 2000, Fung, 2000). Toutefois, les phénomènes phraséologiques font que l'alignement en mots est une tache difficile. L'alignement des unités lexicales complexes est d'ailleurs très souvent l'un des buts recherchés, notamment en terminologie. De nombreux auteurs se sont attachés à extraire des unités complexes à partir de textes alignés (Kupiec, 1993, Van Der Eijk, 1993, Dagan, 1994, Gaussier et Lange, 1995). Selon Véronis (2000a), des études plus récentes (Smadja *et al.*, 1996, Melamed, 1997, Hiemstra, 1998) montrent des avancées importantes dans le domaine. Un alignement en mots ou une extraction de lexique bilingue à partir de corpus parallèles peut se diviser en deux grands aspects, premièrement un repérage des unités lexicales complexes en langue source et en langue cible, puis deuxièmement un alignement entre les deux. Les techniques de repérage d'unités lexicales complexes décrites dans la section précédente ont été appliquées avec un certain succès à leur alignement (Daille, 1994, Smadja *et al.*, 1996, McEnery *et al.*, 1997, Blanck, 2000, Piperidis *et al.*, 2000).

Alignement de « segments linguistiques »

Un autre type d'alignement est celui de l'alignement de segments linguistiques supérieurs aux mots ou aux unités lexicales complexes, mais inférieurs à la phrase, à savoir des clauses, des fragments d'arbres syntaxiques ou des « squelettes » de phrases. Ces techniques forment un continuum avec celles de l'alignement en mots ou en expressions. Pour ces techniques, citons les travaux de Kaji *et al.* (1992), de Matsumoto *et al.* (1993), de Grishman (1994) et de

Papageorgiou (1997). Piperidis (2000) et Wu (2000) présentent l'état d'avancement de ce type de techniques.

3.3.2 Outils d'alignement de termes

Les outils d'acquisition de terminologie bilingue exploitent des corpus parallèles pour extraire des termes équivalents en langue source et en langue cible. Le système d'acquisition terminologique de **Van der Eijk** (1993) se compose de deux étapes :

- **Acquisition monolingue** : les textes des langues source et cible sont extraits sur la base de patrons catégoriels.

- **Acquisition bilingue** : les termes extraits sont alignés par une méthode d'analyse des statistiques de cooccurences des termes dans les phrases alignées.

Termight (Dagan et Church, 1994) est un logiciel d'acquisition de terminologie bilingue, pour le français et l'anglais. Il passe également par deux phases d'acquisition :

- **Acquisition monolingue** : lors de l'acquisition monolingue, le repérage se fait à l'aide des patrons morpho-syntaxiques d'unités lexicales simples et complexes, à partir du texte étiqueté. Les unités lexicales sont regroupées à partir de leur tête sémantique. Une interface de validation permet de visualiser le contexte de chaque terme au sein du corpus source et une phase de validation manuelle filtre les candidats-termes.

- **Alignement bilingue** : la mise en correspondance des termes est réalisé à partir d'un algorithme d'alignement au niveau des mots.

TwiC (*Translation of words in context*) (Wehrli, 2004) est un outil d'assistance à la lecture de documents en langues étrangères, par le biais de traduction de mots et d'expression en contexte, basé sur une analyse syntaxique. TwiC traite les unités lexicales complexes telles

que les mots-composés, les locutions et les collocations. Voici un exemple d'interface graphique, pour l'analyse de la phrase[1] :

*A **natural** language interface was developed.*

Suite au mot sélectionné par l'utilisateur, le résultat suivant apparaît :

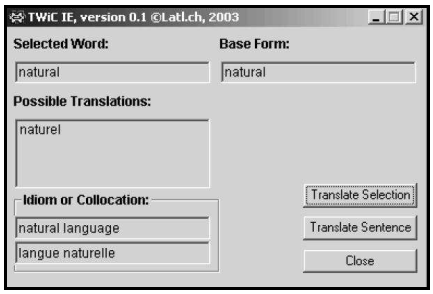


Figure 10. Interface de résultats du logiciel TwiC

L'architecture du logiciel est basée sur différents modules (*ibid.*) :

- **Extracteur de phrases** : le repérage des phrases s'effectue par le biais d'indices typographiques, ainsi que des indices de balises HTML.

- **Identificateur de langue** : un système de tri-grammes est utilisé afin d'identifier la langue du document.

- **Analyseur linguistique** : une analyse morpho-syntaxique avec l'analyseur Fips est effectuée. Elle permet de déterminer l'unité lexicale recherchée.

- **Base de données bilingue** : une base de données bilingue d'équivalences d'unités lexicales simples et complexes est utilisée.

[1] Exemple cité par Wehrli (2004).

- **Interface graphique** : enfin une interface graphique permet d'afficher les résultats de la requête.

Champollion (Smadja *et al.*, 1996) est un outil d'extraction de traductions de collocations à partir d'un corpus parallèle aligné au niveau des phrases. Dans un premier temps, Champollion prend en entrée une collocation en anglais et repère tous les mots qui lui sont fortement associés dans la partie française, à l'aide du coefficient de Dice. L'hypothèse est que la traduction de la collocation source se trouve dans la liste précédemment constituée. Toutes les combinaisons possibles des couples de mots de la listes sont générées et les couples les plus significatifs sont extraits (coefficient de Dice). Les étapes sont répétées de la même façon pour des triplets significatifs, puis pour les séquences de quatre mots, et ainsi de suite. Le logiciel s'arrête lorsque plus aucune séquence ne dépasse le seuil du coefficient de Dice.

3.3.3 Corpus comparables

Les travaux en acquisition automatique de traductions d'unités lexicales complexes se sont principalement basés sur l'exploitation de corpus parallèles (Véronis, 2000a ; Morin *et al.*, 2004). Toutefois, les textes parallèles constituent des ressources rares, surtout pour des couples de langues ne faisant pas intervenir l'anglais (Morin *et al.*, 2004). Des corpus comparables désignent des corpus de langues différentes traitant du même domaine mais non parallèles. Bien que les travaux à partir de corpus comparables constituent un phénomène plus jeune, les avantages, cités dans la littérature, ne sont pas négligeables (Déjean et Gaussier, 2002). D'un point de vue pratique, il est plus facile de collecter des corpus comparables de bonne qualité (Fung, 1998). D'autre part, l'accès à des corpus comparables permet de collecter les usages réels des unités lexicales de la langue cible, et permet d'éviter d'éventuels biais liés à la traduction (Déjean et Gaussier, 2002).

La définition des corpus comparables de (Déjean et Gaussier, 2002) est la suivante :

> Deux corpus de deux langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue L1, respectivement L2, dont la traduction se trouve dans le corpus de langue L2, respectivement L1 .

L'hypothèse sous-jacente des travaux d'acquisition de traductions à partir de corpus comparables est basé sur le principe de la sémantique distributionnelle qui s'attache à décrire le sens des mots, à partir de sa distribution dans un ensemble de contextes (*ibid.*). Si, à partir des corpus parallèles, l'espace de recherche de l'unité lexicale cible se réduit le plus souvent à la phrase, il n'en va pas de même pour les corpus comparables, pour lesquels la traduction recherchée peut se trouver à n'importe quel endroit (*ibid.*). Les méthodes à partir de corpus comparables consistent généralement à collecter l'ensemble des contextes (appelés « vecteurs de contexte ») de chaque unité lexicale, pour les corpus en langue source et en langue cible. Des ressources existantes sont ensuite exploitées afin de traduire les vecteurs de contexte de chaque unité lexicale et de les comparer entre la langue source et la langue cible. Les hypothèses sont les suivantes (Déjean et Gaussier, 2002) :

> (1) Les mots de la langue L1 dont les distributions normalisées sont les plus similaires à la distribution d'un mot donné de la langue L2, sont, avec une forte probabilité, traduction de ce mot.
>
> (2) Deux mots de L1 et L2 sont, avec une forte probabilité, traduction l'un de l'autre si leurs similarités avec les entrées des ressources bilingues disponibles sont proches.

Une majorité des travaux d'acquisition de terminologie bilingue à partir de corpus comparables ont porté sur des termes simples (Morin *et al.*, 2004). Nous pouvons mentionner les travaux de (Fung, 1998) qui extraient des termes simples anglais/chinois, avec une précision de 76% sur les 20 premiers candidats. Les corpus exploités sont le Wall Street Journal et le quotidien japonais Nikkei Financial News. Les travaux de (Rapp, 1999) obtiennent une précision de 89% sur l'extraction des 10 premiers candidats, pour des termes simples anglais/allemand, à partir d'un corpus journalistique. (Déjean et Gaussier, 2002) obtiennent une précision de 84% sur les 10 premiers candidats de couples anglais/allemand, à partir d'un corpus médical.

Toutefois, les travaux d'acquisition de termes complexes, à partir de corpus comparables sont peu courants[1]. En ce qui concerne la traduction de termes complexes en langue de spécialité, (Morin *et al.*, 2004) présentent une méthode, comportant la revue internationale *Unasylva*, consacrée aux forêts et aux industries forestières. Cette approche est une méthode mixte, qui identifie initialement les termes complexes pour chaque langue avec une méthode linguistique (utilisation du logiciel ACABIT (Daille, 1994)), et procède ensuite à un alignement via des méthodes statistiques basées sur le contexte des termes. Le traitement statistique est proche de la méthode proposée par (Déjean et Gaussier, 2002) pour les termes simples. L'idée sous-jacente consiste en la traduction des termes qui sont proches du terme à traduire. L'évaluation de la méthode a été réalisée de façon automatique, via plusieurs lexiques de référence traitant du domaine de spécialité. A partir de ces lexiques, 300 termes français ont été sélectionnés automatiquement, chacun de ces termes devant être présent au moins cinq fois dans le corpus comparable. Les résultats montrent que les termes complexes dont la traduction est compositionnelle sont relativement bien repérés et apparaissent le plus souvent dans les 20 meilleurs candidats. Par contre, les autres termes sont moyennement repérés et n'apparaissent que rarement dans les 20 premiers candidats, bien que les traductions proposées se situent le plus souvent dans le même champ sémantique.

3.4 Conclusion

Nous avons présenté les méthodes traditionnelles de traitement automatique de la terminologie monolingue et bilingue. En ce qui concerne l'acquisition de terminologie bilingue, les techniques d'alignement présentent des résultats qui montrent un certain succès (Daille, 1994, Smadja *et al.*, 1996, McEnery *et al.*, 1997, Blanck, 2000, Piperidis, 2000). Toutefois, les méthodes d'alignement sont coûteuses et l'accès à des textes traduits est rare, surtout pour des langues autres que l'anglais. La taille des corpus parallèles est pour l'instant modeste par rapport aux corpus monolingues (Véronis, 2000a). De plus, les corpus parallèles sont nécessairement biaisés dans leur représentativité , car les textes traduits disponibles

[1] D'autres travaux tels que ceux de Cao et Li (2002) ont recours au Web afin d'acquérir des traductions de termes complexes. Nous parlons des stratégies utilisant le Web pour la traduction dans le chapitre 5.

relèvent de domaines particuliers (textes légaux, textes techniques, textes religieux (Resnik et Melamed, 1997), etc.). Certains genres sont peu représentés, comme par exemple les conversations, les émissions radiophoniques, etc. (Véronis, 2000a).

Les textes parallèles ne constituent pas de véritables actes de discours, puisqu'il s'agit de traductions et sont perçues comme des artefacts (Véronis, 2000a). Des textes originaux sont supposés offrir une phraséologie plus riche que celle d'une langue traduite, que certains nomment « translationese » afin d'en souligner le caractère non idiomatique (Maniez, 2001b).

Bien que le lien entre les traductions soit moins « évident » au sein d'un corpus comparable, puisque la présence d'une traduction n'est pas assurée comme dans les corpus parallèles, l'accès à des corpus comparables reste plus aisé que l'accès à un corpus parallèle de bonne qualité (Fung et Yee, 1998). Les techniques sont toutefois plus récentes et ont moins fait leurs preuves pour l'acquisition d'unités lexicales complexes, se centrant sur l'acquisition de traductions de termes simples. Les travaux se centrent généralement sur des domaines de spécialité (Rapp, 1995, 1999, Fung, 1995, Fung et McKeown, 1997, Fung et Yee, 1998, Diab et Finch, 2000, Morin *et al.,* 2004), ce qui ne favorise pas l'étendue de diversité lexicale que nous recherchons. Dans le chapitre suivant, nous présentons une nouvelle ressource lexicale, le Web, dont les applications en Traitement Automatique des Langues et en acquisition de traduction, bien que récentes, sont de plus en plus nombreuses et présentent un certain succès au vue des avantages qu'il offre, en comparaison avec les ressources traditionnelles.

Chapitre 4. Le Web comme méga base lexicale

4.1 Introduction

Le Web constitue un vaste réservoir de données lexicales, qui peut être exploité par des moyens automatiques, par le biais de moteurs de recherche tels que *Google*[1] ou *Yahoo*[2]. Bien que plus « bruité » que les corpus traditionnels, le Web représente un gigantesque panel d'exemples linguistiques attestés, de genres différents (domaines terminologiques, registres de langues, etc.). Il est le plus vaste et le plus varié des corpus et son multilinguisme est inégalable (Kilgarriff et Grefenstette, 2003). Ses caractéristiques représentent un bouleversement méthodologique pour la linguistique empirique. Malgré la prolifération de travaux qui ont recours au Web depuis la dernière décennie, il est un phénomène nouveau dont les contours restent méconnus, et sort des cadres habituels d'acquisition de terminologie monolingue et bilingue. Il convient de s'interroger sur la place du Web en linguistique, par rapport aux corpus traditionnels (4.2), ainsi que d'analyser ses atouts et ses limites (4.3). Face

[1] http://www.google.fr/

[2] http://www.yahoo.fr/

à la quantité de travaux qui ont recours au Web, nous présentons un tour d'horizon non exhaustif des principaux domaines du Traitement Automatique des Langues (4.4). Les travaux en acquisition automatique de traductions à partir du Web, également prolifiques, feront l'office du chapitre suivant (Chapitre 5).

4.2 Le Web est-il un corpus ?

Avant de s'interroger sur le statut du Web dans la recherche linguistique, il convient de s'interroger sur la définition et le rôle des corpus.

4.2.1 Qu'appelle-t-on « corpus » ?

Il existe des divergences sur la définition d'un « corpus », reflet de variations théoriques sur son statut en linguistique. Malgré des contours flous, la littérature s'accorde sur des caractéristiques générales. McEnery et Wilson (1996) font émerger plusieurs critères :

> In principle, any collection of more than one text can be called a corpus… But the term « corpus » when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, a standard reference.

La première notion est celle de la représentativité. En fonction des textes sélectionnés, un corpus peut-être représentatif d'un état de langue ou de situations linguistiques particulières en vue de leur étude (Duclaye, 2003). Le critère de représentativité est toutefois une notion problématique : de quoi le corpus doit-il être représentatif (Kilgarriff et Grefenstette, 2003) ? Hormis des domaines de spécialité précis, la représentativité de la langue n'est pas concluante, car cette dernière présente des variables qu'il n'est pas possible de prendre en compte dans un corpus (Kilgarriff et Grefenstette, 2003) :

- La représentativité doit-elle se placer du côté de la production ou de la réception de la langue ?

- Doit-elle concerner des textes écrits ou des retranscriptions orales ?

- La réception « passive » du langage fait-elle également partie des événements à considérer ?

- Les citations doivent-elles être considérées comme de nouvelles productions langagières ?

Une notion proche de la représentativité est celle de « corpus de référence ». Selon Sinclair (1996), un corpus de référence a pour objectif de « représenter toutes les variétés pertinentes » d'une langue afin de constituer une base d'analyse linguistique. » Citons le *Brown Corpus*, en anglais, qui regroupe 15 genres différents, ou le *British National Corpus*, qui contient 90% de textes écrits divisés en catégories et 10% de texte parlé. L'idée d'un corpus de référence présente des limites proches de celles de la représentativité.

Un corpus peut-être une sélection de textes organisés selon des critères précis (Sinclair, 1995) :

> a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of language.

La sélection des textes repose sur des critères explicites en fonction des objectifs de la recherche. Selon Habert (2000), des critères extra-linguistiques doivent être ajoutés aux critères linguistiques, permettant d'obtenir les « emplois déterminés » d'une langue (Duclaye, 2003).

L'avènement de textes au format électronique fait émerger une définition, plus vague, celle de « corpus électronique ». Le corpus serait une collection quelconque de textes, au format électronique (Manning et Schütze, 1999) :

> In Statistical NLP, one commonly receives as a corpus a certain amount of data from a certain domain of interest, without having any say in how it is constructed. In such cases, having more training data is normally more useful that any concerns of balance, and onde should simply use all the text that is available.

Un regroupement de textes sans critère précis n'est pas un corpus au sens strict, mais peut être satisfaisant lorsque la nécessité première est la quantité de données. Après avoir fait un tour d'horizon sur la place du corpus en linguistique (4.2.2), nous nous demandons si le Web est un corpus (4.2.3).

4.2.2 Le rôle du corpus dans la recherche linguistique

Une approche du corpus peut être une démarche inductive, qui consiste à expliquer les énoncés du corpus pour en tirer des conclusions générales. Cette démarche est adoptée par des structuralistes américains tels que Harris (1951). Le corpus est un ensemble clos et les procédures de découverte sont strictement internes. Des auteurs tels que Chomsky (1957, 1962) critiquent la vision inductive. Selon lui, la grammaire n'est pas à expliquer à partir d'un corpus, mais à partir de la compétence des locuteurs. La compétence n'est pas un phénomène observable, Chomsky préconise le recours à l'intuition (rationalisme). Pour lui, un corpus ne recense pas tous les faits utiles à une description linguistique. Au jour d'aujourd'hui, il y a une différence d'échelle que la théorie de Chomsky ne pouvait pas prévoir. Les limites d'une telle approche sont que la démarche s'avère normative et non pas descriptive. Elle décrit les faits tels qu'ils devraient être dits, plutôt que tels qu'ils sont dits. Cette méthode fait part de subjectivité : les intuitions ne sont pas les mêmes d'un locuteur à l'autre.

Selon Popper, une collection d'observations ne permet pas d'induire de façon logique une proposition générale. Pour reprendre son célèbre exemple, le fait de ne voir passer que des cygnes blancs ne permet pas d'avoir la certitude qu'il n'existe pas de cygnes noirs. Popper critique une démarche inductive dans le domaine des sciences et préconise un procédé déductif de mise à l'épreuve des théories. Ce processus passe par un mécanisme de prédiction et de réfutation. Dans ce contexte, un corpus en linguistique est un réservoir d'exemples permettant de construire des hypothèses puisqu'on admet que l'intuition n'est pas

satisfaisante. Il constitue un banc de test, qui ne forme pas un ensemble clos et dont de nouveaux exemples peuvent réfuter les théories.

Depuis une vingtaine d'années, la recherche linguistique a pris un tournant empirique avec l'utilisation de plus en plus systématique de corpus (Leech, 1991, McEnery et Wilson, 1996). La linguistique descriptive étudie les faits linguistiques qu'on retrouve fréquemment dans les données réelles, quelque soit le type de données (même si les textes ne correspondent pas à une norme standard). Elle a permis à la linguistique générale d'étendre son champ d'investigation et de concevoir de nouvelles approches de la langue et de la notion de norme. Pour la linguistique empirique, étudier une langue, c'est réunir un ensemble d'énoncés, aussi variés que possible, effectivement émis par des locuteurs de cette langue, à une époque donnée. Il s'agit d'analyser ces énoncés, et d'éventuellement faire apparaître des régularités dans les faits. L'apparition de données massives a permis au Traitement Automatique des Langues de mettre en place des techniques d'apprentissage.

Figure 11. Evolution de la place du corpus en linguistique

4.2.3 Quel statut attribuer au Web ?

Il y a vingt ou trente ans, la constitution d'un corpus électronique était une tâche ardue : saisie et correction des textes, etc. (Habert *et al.*, 1997). Avec l'avènement de la micro-informatique, la situation a radicalement changé (*ibid.*). De plus en plus d'écrits existent directement sous-format électronique et sont exploitables pour la constitution de corpus. Paradoxalement, la définition du corpus s'est obscurcie : la sélection de textes est bouleversée devant la facilité d'accès à des textes électroniques (*ibid.*). L'avènement du Web a constitué un autre bouleversement : les bases de données disponibles ont constitué un nouveau changement

d'échelle qui nécessitent de s'interroger sur ces capacités. Une limite du Web concerne sa « non-représentativité ». Selon Rundell (2000), les types de textes sont hétérogènes : les documents journalistiques et scientifiques sont majoritaires (Duclaye, 2003). Kilgarriff et Grefenstette (2003) montrent que le Web n'est certes pas représentatif, mais les corpus traditionnels ne le sont pas plus :

> We define a corpus simply as a « collection of texts ». If that seems too broad, the one qualification we allow relates to the domains and contexts in which the word is used rather its denotation : a corpus is a collection of texts when considered as an object of language or literary study. The answer to the question "Is the web a corpus?" is yes.

La quantité des données offre une variété de genre plus vaste qu'un corpus traditionnel. Même si le Web ne répond pas aux définitions standard et que les données sont moins contrôlées, elles permettent un changement dont les répercussions peuvent être fondamentales pour la compréhension des langues, à condition de disposer d'instruments d'observation adéquats. Le Web peut être considéré comme un outil d'observation des usages pour le linguiste, en termes à la fois qualitatif (il dispose du contexte réel d'un grand nombre de formes) et quantitatif. Nous parlons à la suite de Habert (2000) de « base de textes » ou de base lexicale, plutôt que de corpus. Divers phénomènes linguistiques sont observables à partir du Web : la quantité des données permet d'observer des phénomènes que des corpus réduits ne permettraient pas d'analyser. Sa dimension considérable vient palier le problème du bruit (Grefenstette, 1999). Pareille à la position du célèbre astronome et physicien Galilée, qui basait ses recherches sur la pratique et l'expérience, le linguiste doit observer le « ciel » linguistique par le biais d'instruments adaptés, c'est-à-dire qui permettent de rapprocher l'observation le plus possible de la réalité. L'utilisation du Web dans un cadre linguistique conduit à ré-appréhender la question du rôle du corpus. Pour nous, le Web est un réservoir d'exemples afin de construire des hypothèses sur la traduction. Les phénomènes de traduction sont des phénomènes complexes et les unités lexicales complexes à observer prolifèrent. Les caractéristiques du Web sont adaptées à nos besoins. Un fait langagier isolé sur le Web ne permet pas de tirer des conclusions. En revanche, nous attribuons à un fait récurrent une valeur linguistique.

4.3 Motivations

4.3.1 Une gigantesque base lexicale

L'argument dominant de l'utilisation du Web est sa taille. Il est difficile de savoir combien de mots sont indexés par les moteurs de recherche dans chaque langue, étant donné le caractère commercialement sensible de cette information, mais des tests indirects (voir Kilgarriff et Grefenstette, 2003) permettent d'estimer à environ 100 milliards le nombre de mots indexés par *Google* pour la seule langue anglaise. Cette quantité est considérable : le *British National Corpus,* qui est de loin le plus grand corpus linguistique au monde, et a servi de base à de nombreuses études (Burnard, 1995), ne comporte que 100 millions de mots, c'est-à-dire une taille environ 1000 fois inférieure. Le changement d'échelle est phénoménal. Voici un exemple, à titre comparatif des fréquences dans le BNC et des fréquences sur le Web d'unités lexicales complexes[1] :

	BNC	**WWW (août 2008)[2]**
medical treatment	414	38 000 000
prostate cancer	39	46 100 000
deep breath	732	20 900 000
acrylic paint	30	3 920 000
perfect balance	38	8 590 000
electromagnetic radiation	39	5 980 000
powerful force	71	5 510 000
concrete pipe	10	1 280 000
upholstery fabric	6	3 020 000
vital organ	46	627 000

Figure 12. Comparaison des frequences entre le BNC et le Web

Même si les fréquences retournées par les moteurs de recherche ne sont que des estimations, elles montrent que les contextes d'étude d'une unité lexicale prolifèrent sur le Web alors

[1] Exemple de Grefenstette (1999), mis à jour pour les fréquences d'Internet.

[2] Fréquences obtenues à partir du moteur de recherche Yahoo. L'utilisation des guillemets est utilisé pour une requête littérale.

qu'ils sont très réduits dans un corpus traditionnel, même « vaste ». Keller et Lapata (2003) montrent que des modèles probabilistes appliqués à diverses applications du TAL présentent de meilleurs résultats lorsqu'ils sont appliqués sur de vastes données, même si les données sont « bruitées ».

4.3.2 Une base lexicale multilingue

En ce qui concerne le multilinguisme du Web, Xu (2000) estime que 71% des pages sont écrites en anglais, 6,8% en japonais, 5,1% en allemand, 1,8% en français, 1,5% en chinois, 1,1% en espagnol, 0,9% en italien et 0,7% en suédois. Le tableau suivant présente une estimation de mots indexés par le moteur de recherche *Altavista*, pour un certain nombre de langues :

Language	Web	Language	Web
Albanian	10,332,000	Catalan	203,592,000
Breton	12,705,000	Slovakian	216,595,000
Welsh	14,993,000	Polish	322,283,000
Lithuanian	35,426,000	Finnish	326,379,000
Latvian	39,679,000	Danish	346,945,000
Icelandic	53,941,000	Hungarian	457,522,000
Basque	55,340,000	Czech	520,181,000
Latin	55,943,000	Norwegian	609,934,000
Esperanto	57,154,000	Swedish	1,003,075,000
Roumanian	86,392,000	Dutch	1,063,012,000
Irish	88,283,000	Portuguese	1,333,664,000
Estonian	98,066,000	Italian	1,845,026,000
Slovenian	119,153,000	Spanish	2,658,631,000
Croatian	136,073,000	French	3,836,874,000
Malay	157,241,000	German	7,035,850,000
Turkish	187,356,000	English	76,598,718,000

Figure 13. Estimation du nombre de mots indexés par Altavista pour différentes langues[1]

[1] Schéma de Kilgarriff et Grefenstette (2003).

A titre comparatif, le *British National Corpus* (BNC) est un corpus exclusivement anglais. Malgré la prédominance de l'anglais sur le Web, le multilinguisme de ce dernier offre des perspectives nouvelles pour la comparaison des langues, certaines langues présentes sur le Web n'étant d'ailleurs pas (ou très peu) disponibles en corpus. Par exemple, De Schryver (2002) montre les perspectives qu'offre le Web pour l'étude de langues africaines.

4.3.3 Une base lexicale évolutive

Le Web présente l'avantage d'être une base de données évolutive, contrairement aux corpus statiques. Il permet d'analyser la langue en « temps réel ». Sajous et Tanguy (2006) présentent, par exemple, une méthode d'acquisition de créations lexicales à partir du Web.

Jacquemin et Bush (2000a, 2000b) utilisent le Web pour la collecte d'Entités Nommées[1] et pour leur classification (selon les pays, les compagnies, les noms d'auteur, etc.). L'intérêt du Web est qu'il est mieux adapté pour le repérage d'Entités Nommées évolutives. Des pages Web sont collectées à partir de requêtes décrivant des patrons susceptibles d'introduire des entités nommées[2] tels que par exemple :

universities such as

Les Entités Nommées candidates sont extraites à partir des pages Web et un filtre permet d'éliminer les résultats bruités.

[1] Les Entités Nommées sont une appellation générique afin de désigner des noms propres référant à des personnes, des lieux ou des organismes (Jacquemin et Bush, 2000b).

[2] Les indices linguistiques choisis sont ceux qui amorcent les collections.

4.3.4 Limites de l'utilisation du Web

Nous relevons deux grandes limites quant à l'exploitation du Web en tant que base de données textuelles, l'une touchant à la qualité des données, l'autre à la performance du protocole d'extraction de co-occurrences lexicales.

Les données textuelles du Web renferment des biais (ou bizarreries) langagiers émis par des locuteurs non natifs de la langue cible ou des locuteurs non-spécialistes de la langue spécialisée. Ces combinatoires lexicales erronées, comme *fumeur lourd* au lieu de *gros fumeur* ou *grand fumeur*, non utilisées par les locuteurs aguerris, "bruitent" le Web en tant que base de données textuelles et doivent être écartées. Grâce à la fréquence de leurs occurrences, des méthodes statistiques permettront d'éliminer (ou réduire) automatiquement celles-ci. Par exemple, la co-occurrence lexicale erronée, en français, de *fumeur lourd*, traduction littérale de l'anglais *heavy smoker* n'apparaît qu'à une fréquence de 73, sur le moteur de recherche *Google*, contrairement à *gros fumeur* (20 700)[1].

D'autre part, les données textuelles du Web sont "brutes", c'est-à-dire qu'aucune information linguistique n'est adjointe. Or, les différents types d'étiquetages (Véronis, 2000b, pour un panorama) appliqués sur les corpus (ou bases textuelles) offrent une aide non négligeable pour le traitement automatique des langues. L'étiquetage morpho-syntaxique détermine la partie du discours (adjectif, nom, verbe, adverbe, etc.) et la morphologie des items textuels (genre et nombre pour les noms et les adjectifs, flexions pour les verbes). Ce processus permet notamment de réduire les ambiguïtés catégorielles. La forme *ferme*, par exemple, peut être un nom, un verbe ou un adjectif. Dans de nombreux cas, l'ambiguïté peut être levée grâce au contexte textuel :

*La **ferme** de mon cousin* **(NOM)**

*Un fromage **ferme*** **(ADJECTIF)**

[1] Google, août 2008.

L'étiquetage morpho-syntaxique autorise une catégorisation des unités lexicales en cadres syntaxiques de type *NOM-ADJECTIF*, *NOM-VERBE*, *ADVERBE-ADJECTIF*, etc. La lemmatisation est un processus qui assigne à chaque occurrence des items textuels sa forme non marquée : la forme masculin singulier pour les adjectifs, le singulier pour les noms, l'infinitif pour les verbes, etc. La lemmatisation permet de rassembler au sein d'une même forme lexicale toutes les variantes morphologiques des lexèmes. Dans notre étude, nous utilisons les deux processus, d'étiquetage morpho-synatxique et de lemmatisation, afin d'extraire les unités lexicales complexes.

4.4 Construction de corpus à partir du Web

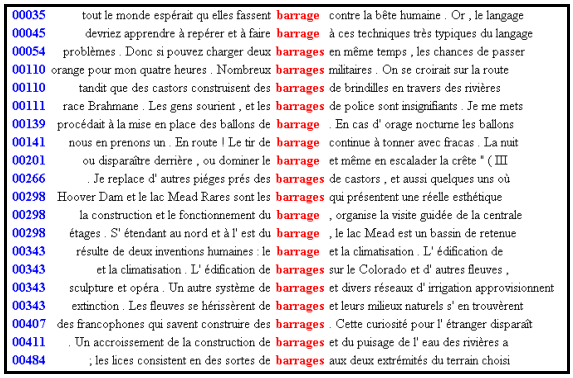
Nous distinguons deux grands courants qui ont recours au Web pour des applications linguistiques, ceux qui collectent des documents pour la constitution de corpus et ceux qui extraient directement des informations sur le Web (fréquences, co-occurrences, etc.). Dans cette section, nous présentons les courants qui ont recours au Web pour la constitution de corpus. Les moteurs de recherche traditionnels n'ont pas été conçus pour des recherches linguistiques. Certains travaux proposent des outils linguistiques afin de construire des corpus à partir du Web ou d'interroger des corpus collectés sur le Web. Le **Sketch Engine** *(SkE ou Word Sketch Engine)*[1] (Kilgarriff, Rychly, Smrz et Tygwell, 2004) est un outil d'analyse linguistique à partir de corpus fournis en entrée, dont certains qui sont proposés en ligne ont été collectés sur le Web, en diverses langues[2]. Le programme offre une fonction de concordancier et une analyse en « Word Sketch » (dépendances syntaxiques et collocations). Les principales fonctionnalités du *Sketch Engine* sont les suivantes :

- **Concordances** : un système de concordances permet d'accéder aux contextes du mot-clé. Les requêtes peuvent être effectuées à partir du lemme, afin d'obtenir les formes

[1] http://www.sketchengine.co.uk/

[2] Les langues disponibles sont entre autres le chinois, l'anglais, le français, l'allemand, l'italien, le japonais, le portuguais, l'espagnol et le slovène.

associées ou à partir d'une forme unique. La catégorie morpho-syntaxique peut être spécifiée. Voici un extrait de concordances pour la requête *barrage* en français :

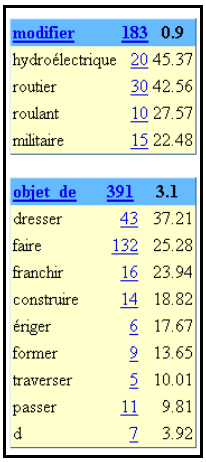


00035	tout le monde espérait qu elles fassent	barrage	contre la bête humaine . Or , le langage
00045	devriez apprendre à repérer et à faire	barrage	à ces techniques très typiques du langage
00054	problèmes . Donc si pouvez charger deux	barrages	en même temps , les chances de passer
00110	orange pour mon quatre heures . Nombreux	barrages	militaires . On se croirait sur la route
00110	tandit que des castors construisent des	barrages	de brindilles en travers des rivières
00111	race Brahmane . Les gens sourient , et les	barrages	de police sont insignifiants . Je me mets
00139	procédait à la mise en place des ballons de	barrage	. En cas d' orage nocturne les ballons
00141	nous en prenons un . En route ! Le tir de	barrage	continue à tonner avec fracas . La nuit
00201	ou disparaître derrière , ou dominer le	barrage	et même en escalader la crête " (III
00266	. Je replace d' autres piéges prés des	barrages	de castors , et aussi quelques uns où
00298	Hoover Dam et le lac Mead Rares sont les	barrages	qui présentent une réelle esthétique
00298	la construction et le fonctionnement du	barrage	, organise la visite guidée de la centrale
00298	étages . S' étendant au nord et à l' est du	barrage	, le lac Mead est un bassin de retenue
00343	résulte de deux inventions humaines : le	barrage	et la climatisation . L' édification de
00343	et la climatisation . L' édification de	barrages	sur le Colorado et d' autres fleuves ,
00343	sculpture et opéra . Un autre système de	barrages	et divers réseaux d' irrigation approvisionnent
00343	extinction . Les fleuves se hérissèrent de	barrages	et leurs milieux naturels s' en trouvèrent
00407	des francophones qui savent construire des	barrages	. Cette curiosité pour l' étranger disparaît
00411	. Un accroissement de la construction de	barrages	et du puisage de l' eau des rivières a
00484	; les lices consistent en des sortes de	barrages	aux deux extrémités du terrain choisi

Figure 14. Extrait des concordances de barrage dans le Sketch Engine

Une requête peut-être affinée en spécifiant son contexte droit et/ou son contexte gauche, sur une fenêtre maximale de dix éléments. Il est possible de limiter la recherche à une sous-partie du corpus, comme par exemple « livres et périodiques » ou « texte oral, gouvernement ».

- **« Word Sketch »** : le *Word Sketch* a été utilisé pour la première fois pour la production du *Macmillan English Dictionary* (Rundell, 2002). Il fournit la liste des dépendances syntaxiques et des collocations dans lesquelles entre le terme, comme l'exemple des relations de modifieurs et d'objet pour *barrage* :

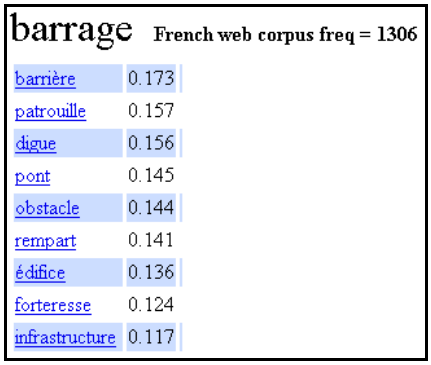


modifier	**183**	**0.9**
hydroélectrique	20	45.37
routier	30	42.56
roulant	10	27.57
militaire	15	22.48

objet_de	**391**	**3.1**
dresser	43	37.21
faire	132	25.28
franchir	16	23.94
construire	14	18.82
ériger	6	17.67
former	9	13.65
traverser	5	10.01
passer	11	9.81
d	7	3.92

Figure 15. Extrait du « word sketch » de barrage

- **Thesaurus :** les mots entrant dans une distribution similaire du mot-clé sont précisés, ce qui offre des classes sémantiques, ici un extrait des mots associés à *barrage* :

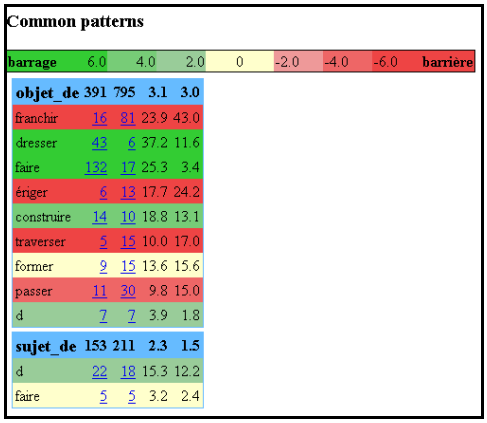


barrage **French web corpus freq = 1306**

barrière	0.173
patrouille	0.157
digue	0.156
pont	0.145
obstacle	0.144
rempart	0.141
édifice	0.136
forteresse	0.124
infrastructure	0.117

Figure 16. Extrait des mots associés à barrage

- **Comparaison des co-occurrences :** à partir de deux mots-clés, il est possible d'obtenir leurs relations syntaxiques communes et celles qui leur sont exclusives, comme dans l'exemple de *barrage* et *barrière* :



Common patterns

barrage	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	barrière
objet_de	**391**	**795**	**3.1**	**3.0**				
franchir	16	81	23.9	43.0				
dresser	43	6	37.2	11.6				
faire	132	17	25.3	3.4				
ériger	6	13	17.7	24.2				
construire	14	10	18.8	13.1				
traverser	5	15	10.0	17.0				
former	9	15	13.6	15.6				
passer	11	30	9.8	15.0				
d	7	7	3.9	1.8				
sujet_de	**153**	**211**	**2.3**	**1.5**				
d	22	18	15.3	12.2				
faire	5	5	3.2	2.4				

Figure 17. Dépendances communes à barrage et à barrière

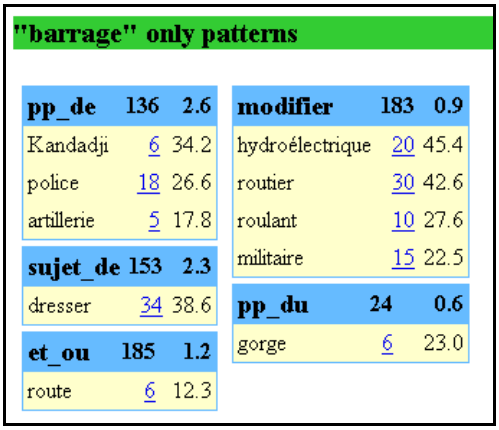


Figure 18. Dépendances spécifiques à barrage

Le système **WebBootCat**[1] (Baroni et Bernardini, 2004) est un outil qui collecte des pages Web via l'API *Google*, pour la construction d'un corpus spécialisé et d'une acquisition terminologique. L'outil prend un ensemble de mots-cibles en entrée (représentatifs du domaine) et collecte les pages Web associées[2]. Une extraction terminologique permet d'élargir les requêtes et le corpus de façon itérative. La collecte de nouveaux mono-termes se fait par une comparaison des fréquences au sein du corpus avec celles d'un corpus de référence. Les termes complexes sont ensuite collectés. Les étapes sont les suivantes[3] :

[1] http://sslmit.unibo.it/~baroni/bootcat.html

[2] Les langues prises en charge sont au nombre d'une trentaine.

[3] Schéma (initialement en anglais) proposé par Baroni et Bernardini (2004).

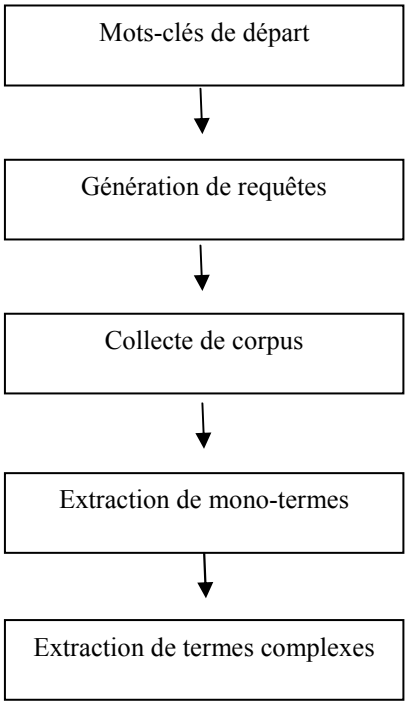


Figure 19. Etapes d'acquisition de corpus via BootCat

L'outil a été appliqué par Baroni et Bernardini (2004) pour la construction de deux corpus en anglais et en italien, dans le domaine de la psychiatrie. Baroni et Ueyama (2004) utilisent *BootCat* pour la collecte de termes spécialisés en japonais, puis pour la construction d'un corpus général en italien et d'un corpus spécialisé (à partir de blogs) en japonais (Baroni et Ueyama, 2006).

Le système **CorpusBuilder**[1] développé par Ghani, Jones et Mladenic est un système qui permet une acquisition automatique de corpus à partir du Web, pour des langues minoritaires telles que le slovène, ou le tagalog, par exemple. La méthode est basée sur l'analyse de deux ensembles de documents fournis en entrée, l'un pertinent pour le corpus à construire, l'autre non pertinent. Les mots-clés pertinents et non pertinents extraits sont respectivement utilisés de façon inclusive ou exclusive pour la génération de requêtes. Les résultats sont utilisés afin de répéter la méthode de façon itérative. Ghani et Jones (2000) et Jones et Ghani (2000)

[1] http://www.cs.cmu.edu/~TextLearning/corpusbuilder/

construisent un corpus en tagalog, à partir de mots-clés pertinents. Ghani *et al.* (2001c) construisent un corpus de slovénien. Ghani *et al.* (2001a, 2001b, 2001d, 2003) appliquent la méthode à différentes langues telles que le slovénien, le croate, le tchèque et le tagalog.

WebCorp[1] (Kehoe et Renouf, 2002, Morley *et al.*, 2003, Renouf, 2003, Renouf *et al.*, 2003, Renouf *et al.*, 2005, Morley, 2006, Renouf *et al.*, 2007, Kehoe et Gee, 2007) est une interface de recherche linguistique vers différents moteurs de recherche (*Google*, *Altavista*, etc.). Les résultats se présentent sous la forme d'un concordancier (contextes et collocations). Il permet de faire des recherches précises telles que la distinction de la casse ou des alternatives de lettres au sein d'un mot (requêtes de « sous-chaînes » telles que « r[u|a]n)[2]. A partir d'une requête, les pages sont collectées, nettoyées et les occurrences sont extraites, offrant un contexte d'une fenêtre maximale de 50 termes à gauche et à droite du terme cible. Les collocations associées à la requête sont également présentées, comme dans l'exemple de *surgery* (Kehoe et Renouf, 2002) :

Collocates of "surgery" (excluding stopwords)

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
laser	57	6	9	9	7	9	3	9	5	31	26
Plastic	47	4	5	1	31	3	1	2		41	6
Society	47	5	21	4		5	1	6	5	30	17
Center	33	8	4		1	5	8	4	3	13	20
Cosmetic	26		4	2	16	1	1		2	22	4
Vascular	23	1	4		13	1	2		2	18	5
Refractive	21				21					21	0
cosmetic	20	1	2	4	11	1	1			18	2
Medicine	20	1		6	4	1	3		5	11	9
Information	20		1		4	12	2		1	5	15
Thoracic	19		5		12			1	1	17	2
General	18	4		1	5	2		6		10	8
University	16		1			2	6	5	2	1	15
Web	14		1	1		1		9	2	2	12

Key Phrases: Plastic surgery Refractive surgery Cosmetic surgery Vascular surgery Thoracic surgery plastic surgery cosmetic surgery Neck surgery brain surgery Brain surgery Pediatric surgery surgery Procedures surgery Handbook

Figure 20. Collocations de surgery extraites par WebCorp

[1] http://www.webcorp.org.uk/

[2] Des requêtes à partir des catégories morpho-syntaxiques ne sont pas possibles.

Dans la même lignée, **KWiCFinder** (*Key Word in Context Web Concordancer*)[1] (Fletcher, 2001, 2002, 2004, 2005, 2007) est un outil qui offre le même type d'options que *WebCorp*. **GoogleLing** (Smarr et Grow, 2002) est également un outil permettant d'intégrer des critères de catégories grammaticales, à partir de différents moteurs de recherche. La méthode de *GoogleLing* est basée sur une conversion de la requête « linguistique » en requête « générale » adaptée au moteur de recherche (*Google*). Par exemple, si la catégorie grammaticale recherchée est un verbe, il peut s'agir d'ajouter des inflexions de verbes. Les pages Web sont ensuite collectées via l'API *Google*, nettoyées et étiquetées. La requête est ensuite identifiée à partir des pages Web collectées. La figure 21 montre le processus général de *GoogleLing*.

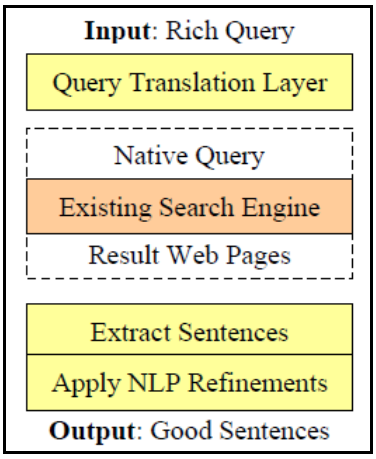


Figure 21. Etapes de traitement de GoogleLing

4.5 Domaines d'application de l'utilisation du Web pour le TAL

Depuis la dernière décennie, de plus en plus de travaux en Traitement Automatique des Langues ont recours au Web, pour des applications très diverses. Nous présentons un bref aperçu des diverses applications du TAL qui exploitent le Web en tant que ressource lexicale. Chaque application étant un domaine très riche en lui même, nous ne visons pas à l'exhaustivité, mais nous proposons un tour d'horizon des possibilités qu'offre le Web dans

[1] http://www.kwicfinder.com/KWiCFinder.html

différents domaines (Volk, 2002). Nos explications sont volontairement simplifiées, car nous présentons une mise en perspective générale des domaines qui ont recours au Web.

4.5.1 Désambiguïsation syntaxique

Certains travaux ont recours au Web afin de résoudre des problèmes d'ambiguïté de rattachement syntaxique. Prenons les exemples suivants (Volk, 2002) :

(1) Peter reads a book ***about computers***

(2) Peter reads a book ***in the subway***

La tâche automatique est confrontée au problème de l'ambiguïté syntaxique de rattachement prépositionnel. Dans la phrase (1), le syntagme prépositionnel (*about computers*) est un attribut du nom *book*, tandis que dans la phrase (2), (*in the subway*) doit être rattaché au verbe (*ibid.*). Une stratégie consiste à utiliser le Web afin de comparer les fréquences de chaque co-occurrence « verbe / préposition / nom2 » (« read, about, computer ») avec « nom 1 / préposition/ nom 2 » (« book, about, computer »)[1] et de voir lesquelles sont les plus fréquentes Volk (2000, 2001). Ces fréquences doivent être mises en rapport avec celles du verbe et du nom lorsqu'ils n'apparaissent pas avec la préposition (*ibid.*). Une telle méthode nécessite un grand nombre de données, et serait difficilement réalisable sans l'apport du Web.

Dans le domaine de la désambiguïsation syntaxique liée au rattachement prépositionnel à partir du Web, les travaux de Volk (2000, 2001) ont été appliqués à l'allemand. Vandeghinste (2002) applique la même stratégie pour le néerlandais. Calvo et Gelbukh (2003) appliquent le même type méthode pour l'espagnol. Lebarbé (2002) utilise également le Web pour l'amélioration d'une méthode de désambigüisation syntaxique. Gala (2003a, 2003b) et Gala et Aït-Mokhtar (2003) présentent une méthode non supervisée d'apprentissage sur le Web, permettant d'améliorer la désambiguïsation du rattachement prépositionnel. Contrairement à

[1] Volk (2001) montre qu'il est plus judicieux d'obtenir les fréquences des trigrammes plutôt que des bigrammes du type verbe+préposition et nom +préposition.

Volk (2000, 2001) qui calcule les fréquences directement à partir d'un moteur de recherche, Gala (2003a, 2003b) collecte un premier corpus à partir du Web qui contient les rattachements ambigus. Ceux-ci sont extraits d'une première analyse et sont générés en tant que requêtes sur le Web, pour la collecte d'un corpus, dont le but est d'extraire automatiquement des informations lexicales (patrons de co-occurrences) et statistiques (poids de cooccurrence statistique) sur ces rattachements. Ces informations sont ensuite utilisées afin de lever les ambiguïtés de rattachement. (Nakov et Hearst, 2005a, 2005b) exploitent le Web pour la désambiguïsation syntaxique de rattachements prépositionnels et de rattachements de syntagmes nominaux, à partir de statistiques dérivées du Web. Rus et Ravi (2006) ont également recours à une approche statistique à partir du Web pour le repérage de syntagmes nominaux dans la tâche de désambiguïsation syntaxique.

4.5.2 Lexicographie

Fujii et Ishikawa (2000) collectent des descriptions encyclopédiques de termes techniques en japonais, à partir du Web. La méthode est basée sur un repérage de patrons linguistiques et de structures HTML susceptibles d'introduire des définitions de termes. Par exemple, la séquence suivante introduit la définition du terme anglais « data mining » :

data mining *is a process that collects data for a certain task, and retrieves relations latent in the data*

Le repérage de la structure « X is Y » permet d'associer la définition Y au terme X. L'acquisition de patrons linguistiques de description est opérée à partir d'une ressource encyclopédique électronique.

4.5.3 Sémantique

Turney et Littman (2003) présentent une méthode de détection de l'orientation sémantique (positive ou négative) d'un ensemble d'unités lexicales de type subjectif, à partir du Web. La méthode est basée sur le calcul des co-occurrences des termes avec un paradigme de termes

positifs et négatifs. L'hypothèse est que les termes ayant les mêmes connotations apparaissent fréquemment ensemble : les termes recensées comme ayant une connotation donnée permettent d'en repérer de nouveaux, par leur contexte lexical (Turney et Littman, 2003). Dans la même lignée, Baroni et Vegnaduzzo (2004) présentent une méthode d'identification d'adjectifs subjectifs à partir du Web, en partant d'une courte liste d'adjectifs subjectifs sélectionnés manuellement. Le Web est exploité afin de collecter les adjectifs qui co-occurrent avec la courte liste créée de façon manuelle (mesure d'information mutuelle).

Turney (2001) présente un algorithme, nommé PMI-IR, pour la détection de synonymes à partir du Web. La méthode est basée sur un calcul d'information mutuelle, à partir des fréquences de couples de termes sur le Web. Les résultats montrent que les termes obtenant les plus hauts scores de co-occurrence ont tendance à être des synonymes. Sato et Sasaki (2003) présentent une méthode d'acquisition de termes thématiquement associés à partir de termes sources, en japonais, à partir du Web. Un corpus contenant les mots cibles est collecté. Les termes associés sont extraits par une méthode d'acquisition terminologique, et filtrés. Dans la même lignée, Baroni et Bisi (2004) ont recours à l'information mutuelle sur le Web pour la tâche de détection de synonymes au sein d'un domaine de spécialité, l'aéronautique. Terra et Clarke (2003) présentent également une mesure de similarité entre termes, par comparaison de leurs co-occurrences, à partir du Web.

(Matsuo *et al.*, 2006) présentent une méthode de classification sémantique de termes, à partir de graphes, nommée *Newman*, qui est basée sur une mesure de similarité à partir de fréquences des vecteurs de co-occurrences sur le Web. Doan *et al.* (2003) proposent un système, le système *GLUE*, basé sur un repérage d'informations disponibles en ligne par d'autres ontologies pour le repérage de similarité entre termes pour la construction d'une ontologie. D'autres travaux ont recours au Web pour la construction ou l'enrichissement d'ontologies (Agirre *et al.*, 2000a, 2000b, 2004a, 2004b, Santamaria *et al.*, 2003, Chung *et al.,* 2006). La méthode étant basée sur l'utilisation de « topic signatures », nous reviendrons sur ces méthodes dans notre chapitre 7.

4.5.4 Désambiguïsation lexicale

L'accès à de très vastes données telles que le Web peut constituer une aide pour la tâche de désambiguïsation lexicale (Gonzalo *et al.*, 2003). Turney (2004) présente un algorithme, à partir du Web, basé sur une collecte de traits sémantiques à partir des probabilités de co-occurrences des mots. Rosso *et al.* (2005) présentent une approche de désambiguïsation lexicale en anglais, à partir du Web, basée sur l'analyse des co-occurrences des termes ambigus.

Une exploitation du Web pour la désambigüisation peut également consister en l'acquisition automatique de corpus annotés sémantiquement. Les travaux de Mihalcea et Moldovan (1999a, 1999b) et Mihalcea (2002) présentent une approche d'acquisition automatique de corpus annoté avec des informations sémantiques. La méthode s'appuie sur les informations de WordNet pour la génération de requêtes (synonymes, définitions) sur des moteurs de recherche traditionnels. Les données collectées sont utilisées pour l'apprentissage de séquences d'exemples désambigüisés en contexte. Chklovski et Mihalcea (2002) présentent le système *Open Mind Word Expert*, qui utilise l'annotation d'utilisateurs en ligne pour la création d'un corpus annoté.

4.5.5 Acquisition de co-occurrences lexicales

Le Web est utile pour l'obervation de co-occurrences monolingues et pour l'acquisition de relations lexicales significatives. Le Web est considéré comme un « miroir » représentatif des phénomènes de co-occurrences lexicales d'une langue. Par exemple[1], la co-occurrence lexicale préférentielle *daunting task* apparaît avec une fréquence de 11 000 000 sur le moteur de recherche *Yahoo*, tandis que des synonymes proches de *task* tels que *job* et *duty* ne constituent pas des co-occurrences lexicales significatives avec *daunting*. *Daunting job*, qui est une co-occurrence lexicale acceptable mais non significative apparaît 110 000 fois. La co-

[1] Exemples cités par Inkpen et Hirst (2002).

occurrence lexicale non acceptable[1] *daunting duty* n'apparaît que 660 fois, ce qui est très peu à l'échelle du Web. Le sémantisme n'est pas une caractéristique pour juger de l'acceptabilité d'unités lexicales complexes idiomatiques : les fréquences sur Web sont un indice révélateur. L'hypothèse est que les co-occurrences non correctes apparaissent peu en comparaison avec les co-occurrences significatives. Certains travaux ont recours au Web pour évaluer de l' « aspect collocationnel » de co-occurrences lexicales collectées ou pour acquérir des collocations à partir du Web. Inkpen et Hirst (2002) évaluent l' « aspect collocationnel » de co-occurrences lexicales, entre synonymes proches. Les collocations sont extraites à partir du corpus BNC. Le Web est utilisé afin d'évaluer leur « aspect collocationnel ». Trois types de collocations sont distinguées : les co-occurrences fréquentes, les faibles co-occurrences (qui restent acceptables) et les co-occurrences impossibles (« anti-collocations »). Keller et Lapata (2003) collectent des bigrammes de type *ADJECTIF-NOM*, *NOM-NOM* et *VERBE-OBJET*, à partir de différents corpus (BNC et NANTC[2]). Le Web est utilisé pour tester leur fréquence. Les résultats montrent que les fréquences sur le Web sont corrélées avec celles des corpus étudiés et avec le jugement d'évaluateurs humains. Seretan *et al*. (2004) collectent des relations syntaxiques de co-occurrences à partir des résumés retournés sur le Web, par des mesures d'association lexicale. Les calculs statistiques sont associés à un filtre syntaxique. Ces travaux collectent un corpus à partir du Web via des noms sources à partir desquels sont extraits les co-occurrents dans des relations de dépendance. Patwardhan et Riloff (2006) collectent des patrons de dépendances syntaxiques relatives à un domaine de spécialité, à partir du Web.

4.5.6 Autres applications

Liu et Curran (2006) ont recours au Web pour la collecte d'un corpus dans le cadre d'un système d'aide à la correction orthographique. Le Web est utilisé afin de détecter les variantes mal orthographiées d'un terme, comme dans l'exemple suivant de *receive* :

[1] Pearce (2001) a introduit le terme d' « anti-collocation » afin de désigner des co-occurrences lexicales non acceptables d'un point de vue idiomatique.

[2] *North American News Text Corpus*.

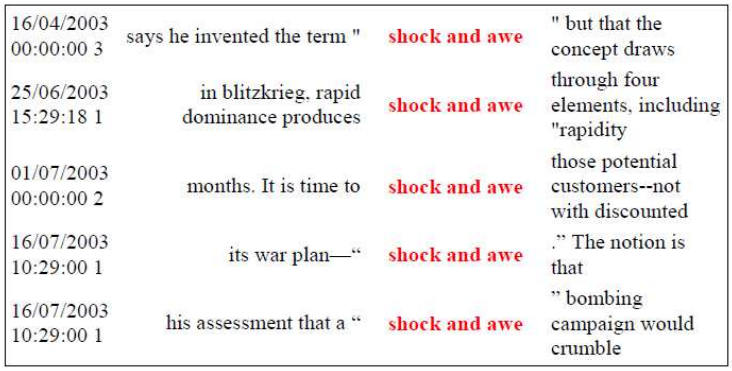
reeceive, recesive, recieive, receivce…

A partir d'un ensemble d'erreurs fréquentes (« confusion set »), la tâche de correction peut être traitée en terme de désambigüisation : il s'agit de sélectionner le terme adéquat en fonction de son contexte. Liu et Curran (2006) montre que le Web est adapté à ce type de travaux, qui nécessite un grand nombre de données.

Le Web peut être utilisé pour observer les évolutions linguistiques d'un point de vue diachronique (Volk, 2002). Par exemple[1], en suisse allemand, la compagnie de téléphone suisse *Swisscom* a lancé un téléphone portable nommé *Natel*. A la même période, les téléphones portables en Allemagne sont nommés *Handy*. En Suisse, ces deux unités lexicales ont été en compétition. Volk (2002) a comparé les fréquences de ces deux unités lexicales avant et après janvier 2000 et a constaté que les fréquences retournées par *Natel* avant janvier 2000 étaient à peu près le double de celles de *Handy*. Après janvier 2000, les fréquences de ces deux unités lexicales ont été à peu près similaires. Ces résultats montrent que l'usage du terme *handy* a nettement augmenté (*ibid.*). Kehoe (2006) utilise WebCorp dans une perspective d'étude diachronique. Les moteurs de recherches traditionnels permettent une recherche avancée, en spécifiant la date du document, mais les options sont fortement limitées. Kehoe (2006) montre que la prise en compte des informations de « dernière modification » constitue une perspective de technique pour la diachronie[2]. A partir de cette information, WebCorp permet de spécifier la date de modification des documents lors d'une requête, soit en indiquant le délai de modification des pages, soit en précisant un intervalle de date. Les résultats sont alors classés en fonction de la date de modification des documents (*ibid.*):

[1] Exemple cité par Volk (2002).

[2] La limite de cette approche est que la date de dernière modification du document ne coïncide pas nécessairement avec sa mise en ligne (Kehoe, 2006). D'autres repères temporels sont proposés tels que la spécification de la dernière révision du document directement dans la page, une précision de la date de copyright, ou la date insérée dans l'URL. Mais ces informations sont faiblement représentées.



16/04/2003 00:00:00 3	says he invented the term "	**shock and awe**	" but that the concept draws
25/06/2003 15:29:18 1	in blitzkrieg, rapid dominance produces	**shock and awe**	through four elements, including "rapidity
01/07/2003 00:00:00 2	months. It is time to	**shock and awe**	those potential customers--not with discounted
16/07/2003 10:29:00 1	its war plan—“	**shock and awe**	.” The notion is that
16/07/2003 10:29:00 1	his assessment that a “	**shock and awe**	” bombing campaign would crumble

Figure 22. Concordances de la requête « shock and awe » classées par date par WebCorp

Une étude de Kehoe (2006) sur le mot anglais *alcopops* montre que les perspectives qu'offre le Web pour les études diachroniques restent intéressantes, en analysant que ce mot est de plus en plus fréquemment employé depuis 1999, alors que son utilisation était peu courante.

Mautner (2005) montre que les caractéristiques du Web offrent des perspectives pour des études en analyse du discours. Zuraw (2006) utilise le Web tel un « corpus phonologique » pour l'étude du Tagalog. Modjeska *et al.* (2003) et Bunescu (2003) ont recours au Web dans le cadre de la résolution d'anaphores.

4.6 Conclusion

Les caractéristiques du Web placent cette gigantesque base lexicale au cœur de domaines très variés en linguistique et en Traitement Automatique des Langues. Même s'il n'est pas un corpus au sens strict, il offre des ressources et des perspectives que le linguiste se doit d'analyser. Il permet de collecter de grandes quantités de textes pour la construction de corpus, ou d'acquérir des informations utiles pour la désambigüisation syntaxique, la lexicographie, la sémantique, la désambigüisation lexicale, la construction de lexiques monolingues et bien d'autres applications dont nous avons citées les plus communes. Le domaine de l'acquisition de traductions n'échappe pas au phénomène. Outre sa taille, son multilinguisme le place au cœur de différentes méthodes pour l'acquisition de données

bilingues. Le chapitre suivant aborde les possibilités qu'offre le Web pour la traduction et présente les différentes techniques employées, telles que l'acquisition de corpus parallèles ou comparables, à partir du Web, ou la collecte d'informations (comme par exemple les fréquences), pour l'aide à la traduction.

Chapitre 5. Méthodes d'acquisition de traductions à partir du Web

5.1 Introduction

Le caractère multilingue du Web le place au cœur d'un courant particulièrement prolifique, celui de l'acquisition de traductions à partir du Web. Les méthodes d'utilisation du Web dans un contexte d'acquisition de traductions sont variées. Nous distinguons cinq grands courants. D'une part, certains travaux présentent des méthodes d'acquisition de « corpus » parallèles à partir du Web (5.2). Certaines méthodes ont recours aux « anchor textes » (5.3). D'autres utilisent le Web tel un « corpus partiellement bilingue » et exploitent des documents linguistiquement mixtes pour le repérage de traductions (5.4). Le Web peut également être considéré comme un « corpus comparable » (5.5). Enfin, certains travaux exploitent les fréquences sur le Web pour l'aide au choix lexical (5.6).

5.2 Acquisition de textes parallèles à partir du Web

5.2.1 Typologie des textes parallèles sur le Web

La diversité des genres, des domaines et des langues présents sur le Web constitue un atout précieux pour les méthodes de construction de corpus parallèles à partir du Web. De nombreux documents du Web sont des textes parallèles[1] (manuels, catalogues, sites administratifs, etc.). Il peut s'agir d'une page Web assortie de sa traduction ou d'un site Web multilingue. Les deux pages suivantes sont par exemple la traduction l'une de l'autre (anglais et espagnol) :

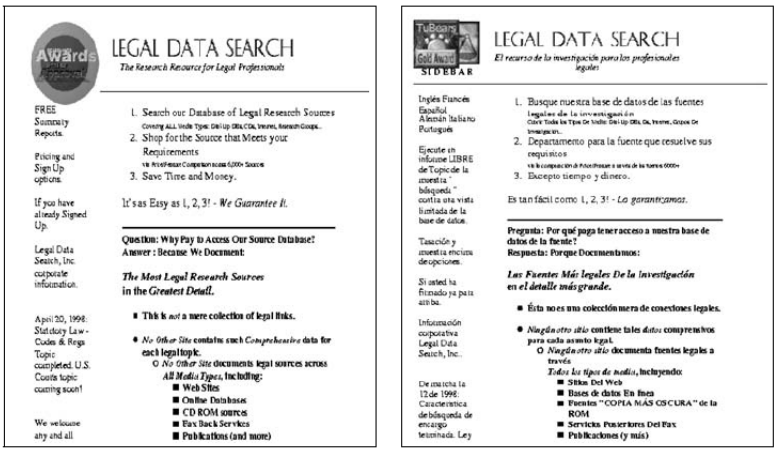


Figure 23. Pages Web parallèles en anglais et en espagnol[2]

Nous empruntons à Resnik (1998) une typologie des principaux documents parallèles présents sur le Web.

[1] Il ne s'agit pas d'un « corpus » au sens propre : les pages Web peuvent être courtes et ne sont pas nécessairement alignées.

[2] Schéma proposé par Resnik (1998). Source des pages : http://www.legaldatasearch.com/.

Pages parentes[1]

Une page parente est un document sur le Web contenant au moins deux liens qui pointent vers des pages qui sont des traductions les unes des autres :

Figure 24. Exemple de page parente[2]

Dans l'exemple ci-dessus, la page parente constitue l'accueil du site « Academy of American and British English ». Il s'agit d'un site multilingue, la version est disponible en six langues. Des moyens automatiques peuvent être utilisés pour repérer des pages parentes, par une analyse des liens hypertextes qui pointent vers des langues différentes. L'accès aux pages traduites permet d'acquérir un corpus parallèle.

Pages «sœurs »[3]

Une page « sœur » est un document monolingue dans une langue donnée qui contient un lien hypertexte vers sa traduction. Dans l'exemple suivant, un lien hypertexte indique explicitement la traduction anglaise (« this page in english ») :

---

[1] « Parent page » en anglais (Resnik, 1998).

[2] Schéma proposé par Resnik (1998). (http://www.academyofenglish.com)

[3] « Sibling page » en anglais (Resnik, 1998).

Figure 25. Un exemple de page «soeur » (du français vers l'anglais[1])

Les liens hypertextes sont analysables afin d'aligner la page « sœur » avec sa traduction. Le plus souvent, le lien de traduction est biunivoque (mais pas de façon systématique) :

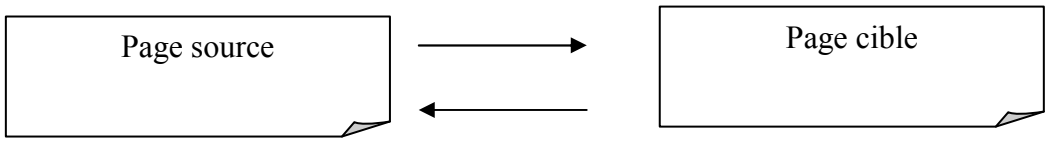


Figure 26. Relation hypertexte biunivoque entre une page fille et sa traduction

Une majorité des textes parallèles sur le Web sont des pages « parentes » ou des pages « sœurs » (Resnik, 1999) :

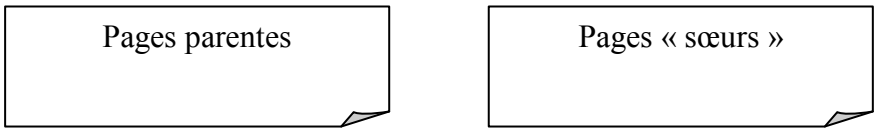


Figure 27. Types de documents parallèles sur le Web

5.2.2 Méthodes d'acquisition

Les méthodes d'acquisition de pages parallèles à partir du Web passent généralement par trois grandes phases, une génération de site candidats, une évaluation de paires candidates, puis un

[1] http://lrs.linbox.org/

filtre des sites candidats. Nous présentons ces phases de façon générale, mais les méthodes existantes n'utilisent pas de façon systématique toutes les caractéristiques présentées.

Localisation de sites candidats

Une première phase consiste en une localisation de « sites candidats », susceptibles de contenir des pages qui sont des traductions. Différentes stratégies de repérage sont utilisées :

- **Analyse des liens hypertextes** : un repérage de liens hypertextes qui pointent vers des pages traduites peut être effectué (Resnik, 1998, 1999, Resnik et Smith, 2003, Nie *et al.*, 1999, Chen et Nie, 2000, Almeida *et al.*, 2002). Les pages parallèles ont pour point commun de contenir des liens hypertextes qui pointent vers la (ou les) traduction(s) de documents. La formulation de différents types de requêtes permet de collecter ce type de documents. Par exemple, la requête suivante permet d'obtenir des pages parentes contenant deux liens hypertextes pointant sur deux pages traduites (en anglais et en français) (Resnik, 1998) :

  anchor : « language1 » AND anchor : « language2 »

  (anchor : « english » OR anchor: "anglais") AND (anchor : « french » OR anchor : "français")

  Ce type de requête permet de collecter des pages qui pointent sur différentes traductions d'une page parente :

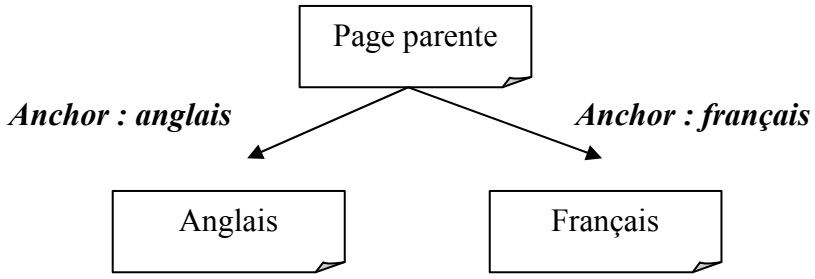


Figure 28. Repérage de pages parentes

Un autre type de pages parallèles repérées concerne les pages « soeurs » (« sibling pages »). Par exemple, les pages françaises retournées par la requête *anchor : « english » OR anchor : anglais* sont des documents contenant un lien vers une page en anglais :

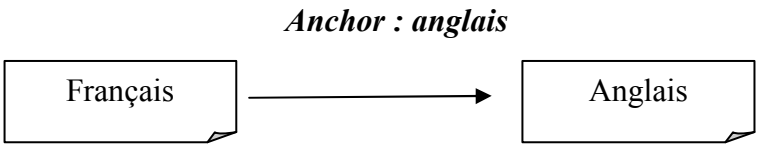


Figure 29. Repérage de pages sœur

Des indications textuelles telles que « this page in english » peuvent apparaitre et constituer des indices supplémentaires de parallélisme.

- **Exploitation du domaine des sites** : le système BITS (Ma et Liberman, 1999) génère une liste de sites Web candidats, fortement susceptibles d'être bilingues en utilisant comme indices les domaines des sites. Par exemple, certains domaines sont susceptibles de contenir des paires de langues données. Ainsi, pour le couple allemand/anglais, les domaines « de » (Allemagne), « au » (Australie) et « lu » (Luxembourg) sont plus susceptibles de contenir des sites bilingues faisant intervenir les langues cibles.

- **Repérage d'URLS similaires** : une phase de normalisation des URLS peut être effectuée afin d'accéder à la traduction d'un document, comme dans l'exemple (Almeida *et al.*, 2002) :

http://www.ex.pt/index_pt.html

http://www.ex.pt/index_en.html

Dans cet exemple, l'extension « pt » indique que le premier lien est en portugais, tandis que l'extension « en » indique que le second lien est en anglais. Lorsque les URLS sont identiques, à l'exception de l'extension de la langue, il est probable que les pages soient des traductions l'une de l'autre (Almeida *et al.*, 2002, Chen *et al.*, 2004) .

- **Comparaison du contenu sémantique des documents** : Chen *et al.* (2004) s'appuie sur une comparaison du contenu sémantique des documents à aligner. La comparaison est fondée sur une liste de termes bilingues, permettant de comparer les termes sources et cibles contenus dans chaque document. Le coefficient de Jacquard, qui mesure le degré de similarité entre deux ensembles, est utilisé afin de comparer les contenus sémantiques.

Génération de sites candidats

Une deuxième phase repère les documents parallèles et filtre les documents non pertinents. Différentes stratégies sont également adoptées :

- **Comparaison des extensions de noms de fichiers** : les noms de fichiers peuvent être des indices de contenus parallèles tels que par exemple « file-fr.html » et file-en.html », ou « fr » et « en » indique que les documents sont respectivement en français et en anglais (Nie *et al*, 1999, Chen et Nie, 2000, Almeida *et al.*, 2002).

- **Comparaison de la structure HTML**: il s'agit de repérer les documents ayant une structure proche (Resnik, 1998, 1999, Resnik et Smith, 2003, Nie et al, 1999). L'idée est que les pages parallèles doivent avoir une structure HTML très proche. Il est également possible de comparer les éléments non textuels (images, liens, etc.) (Almeida *et al.*, 2002). Ce type de stratégie permet un alignement de séquences plus fines, comme dans l'exemple (Resnik, 1998) :

<HTML> <TITLE>Emergency Exit</TITLE> <BODY> <H1>Emergency Exit</H1> If seated at an exit and ⋮	<HTML> <TITLE>Sortie de Secours</TITLE> <BODY> Si vous êtes assis à côté d'une … ⋮

Figure 30. Structures HTML de documents parallèles

- **Comparaison de la longueur des textes** : une comparaison entre la longueur des textes en langue source et en langue cible peut être un indice de traduction. L'hypothèse est que les pages traduites ont une longueur textuelle relativement proche (Resnik, 1998, 1999, 2002, Nie *et al.*, 1999, Chen et Nie, 2000). Resnik (1998, 1999) et Resnik et Smith (2003) procèdent à une comparaison de longueur des textes par alignement de segments.

- **Comparaison du poids des fichiers** : Almeida *et al.* (2002) procède à une comparaison de poids entre le fichier source et le fichier cible.

- **Comparaison de similarité des chaînes de caractères** : Almeida *et al.* (2002) dispose d'un module de comparaison des chaînes de caractères entre le fichier source et le fichier cible.

- **Identification de la langue des documents** : si les modules précédents sont indépendants de la langue, certains travaux ont également recours à un module d'identification de la langue, afin de filtrer les documents qui ne sont pas dans la langue souhaitée (Resnik, 1999, 2003, Ma et Liberman, 1999, Chen et Nie, 2000, Almeida *et al.*, 2002). Resnik (1999) et Resnik et Smith (2003) proposent un repérage automatique de la langue du document basé sur une méthode statistique de comptage des fréquences de caractères, qui permet d'éliminer les pages collectées qui ne sont pas dans la langue attendue. Ma et Liberman (1999) étudient les propriétés linguistiques des sites afin de détecter si le site est monolingue ou multilingue et d'identifier les langues impliquées[1]. Les sites exclusivement monolingues sont supprimés de la liste.

[1] Si plus d'une langue est impliquée dans les 3 ou 4 premiers niveaux d'un site, alors il est admis que le site est au moins bilingue.

Filtre des sites candidats

L'étape d'évaluation de la méthode STRAND (Resnik, 1998, 1999, Resnik et Smith, 2003) permet un alignement des documents au niveau des segments (« chunks »). La méthode STRAND permet donc, outre d'obtenir des paires candidates d'URLs de pages parallèles, de collecter un corpus aligné par segments. Ma et Liberman (1999) ont également recours à un alignement des pages. Un lexique bilingue est utilisé afin d'établir un calcul de similarité entre les unités lexicales sources et cibles pour les pages anglaises et allemandes. La méthode d'Almeida *et al.* (2002) découpe les textes en chunks et les fichiers sont convertis en PML[1]. Un alignement est enfin effectué, via le logiciel EasyAlign[2]. Cet alignement peut ensuite être utile pour des systèmes basés sur les mémoires de traduction (*ibid.*).

Resnik (1998, 1999) et Resnik et Smith (2003) présentent une méthode d'acquisition automatique de documents parallèles à partir du Web[3], le modèle STRAND (*Structural Translation Recognition for Acquiring Natural Data*). Resnik (1998, 1999) évalue la méthode à partir des couples de langues anglais/espagnol et français/anglais. Une version améliorée de STRAND a été appliquée au couple de langues anglais/chinois (Resnik et Smith, 2003). Le schéma résume les étapes du modèle STRAND (Resnik, 1998) :

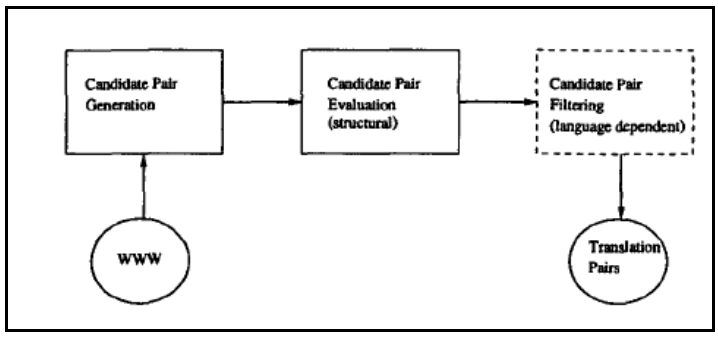


Figure 31. Architecture du modèle STRAND

[1] *Paragraph Markup Language*.

[2] IMS Corpus Workbench, (1994-2002)

[3] Le moteur de recherche utilisé pour cette étude était le moteur *Altavista*.

Dans la même lignée, Nie *et al.* (1999) proposent également une méthode d'extraction automatique de documents parallèles sur le Web en anglais et en français, pour une tâche de recherche d'information multilingue. Les résultats montrent que l'exploitation d'un corpus parallèle issu du Web permet d'améliorer les performances de systèmes de recherche d'informations multilingues. Ma et Liberman (1999) proposent le système BITS (*Bilingual Internet Text Search*) pour l'acquisition de textes parallèle multilingue, à partir du Web. La méthode, basée sur le couple de langue allemand/anglais[1], collecte des pages Web qui contiennent des paires de traduction et les stocke dans une base de données. Les expériences menées avec des paires en allemand et anglais montrent que la méthode offre des résultats satisfaisants, avec un rappel de 97,1% et une précision de 99, 1%. Dans le domaine de la recherche d'information inter-lingue, Chen et Nie (2000) et Kraaij *et al.* (2003) proposent un algorithme, le système PTMiner (*Parallel Text Miner*), dont le but est d'extraire un corpus parallèle à partir du Web. La précision de Chen et Nie (2000) pour le corpus obtenu en anglais/français est de 95% et celle pour le corpus anglais/chinois est de 90%. Dans la même lignée, Nie et Cai (2001) proposent une méthode de nettoyage de corpus parallèle, à partir d'un corpus anglais/chinois, afin d'éliminer les pages Web non-parallèles restantes dans le corpus. Almeida *et al.* (2002) proposent une méthode d'extraction de pages parallèles à partir du Web, par l'utilisation d'une série de modules qui exploitent le « Web bilingue ». Une expérience d'évaluation, basée sur l'alignement de pages en portugais et en anglais, donne une précision de 85%, avec un rappel de 92% (Almeida *et al.*, 2002). Yang et Li (2003) présentent également une méthode de construction de corpus parallèles à partir du Web, pour l'anglais et le chinois, dont la précision de de 99,5% et le rappel de 80,96%. Chen *et al.* (2004) propose le système PTI (*Parallel Text Identification System*) qui détecte des pages parallèles à partir de la comparaison des noms de fichiers et du contenu sémantique des documents. L'évaluation, à partir d'un site gouvernemental multilingue, en anglais et en chinois, offre une précision de 0.93% et un rappel de 0.96%. Le schéma suivant présente l'architecture de PTI (*ibid.*) :

[1] Le système est capable de traiter 13 langues différentes.

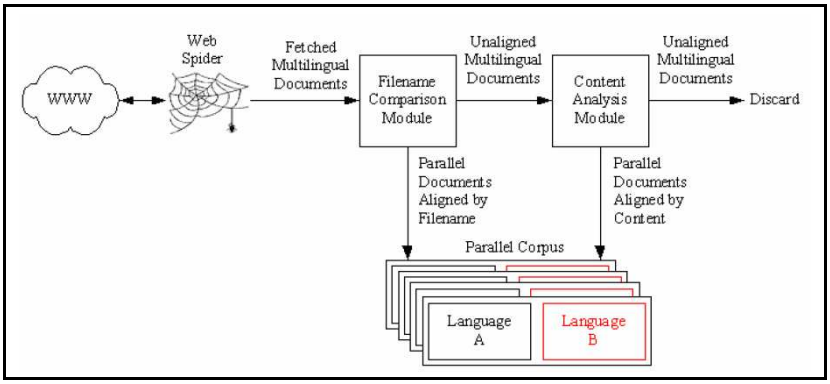


Figure 32. Architecture du système PTI

5.3 Approches basées sur les « anchor textes »

Lu *et al.* (2001, 2002, 2004) présentent une méthode de traduction de requêtes pour la recherche d'informations interlingues, par l'exploitation de « anchor textes »[1]. Un anchor texte est le texte contenu dans le descriptif d'un lien hypertexte, comme dans l'exemple :

Wikipedia

Dans cet exemple, l'anchor texte est "Wikipédia". Les anchor textes sont utilisés par les moteurs de recherche dans la tâche d'indexation. Le contenu des anchor textes peut varier, il peut s'agir de titres, de phrases multilingues, de textes courts, d'acronymes ou même d'URLs (Lu *et al.*, 2001, 2002, 2003). La figure 33 illustre différents anchor textes en de multiples langues qui pointent vers le site du moteur de recherche *Yahoo* :

[1] Le terme anglais est « anchor text ». Il n'existe pas d'équivalent strictement français, bien qu'on puisse parler d'*ancrage*. Nous employons le terme de *anchor texte*, plus proche du terme anglais.

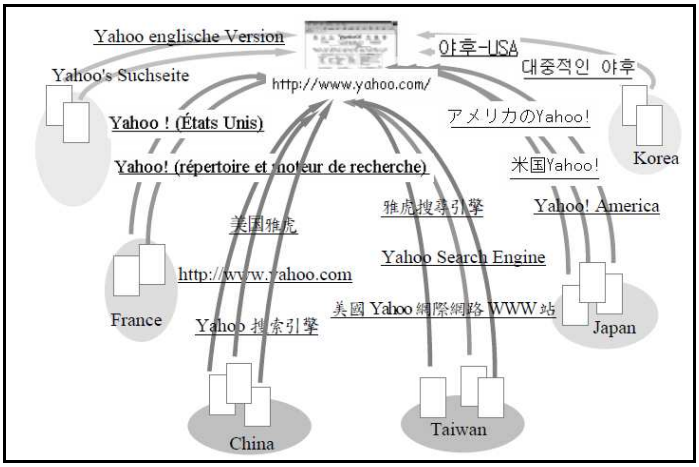


Figure 33. Anchor textes en différentes langues qui pointent sur le site Yahoo[1]

Les « anchor textes » sont propices à la détection d'unités lexicales traduites et peuvent être considérés comme des corpus comparables (*ibid.*). Dans cet exemple, les alias régionaux de l'Entité Nommée *Yahoo* peuvent être détectés par des moyens automatiques. L'objectif de l'approche de Lu *et al.* (2001, 2002, 2003) est de déterminer des stratégies permettant d'identifier automatiquement la traduction d'une requête, à partir des « anchor textes » qui lui sont associés. Le but est d'extraire les traductions candidates qui co-occurrent fréquemment avec la requête source, au sein d'un même anchor texte. Toutefois, les termes qui co-occurrent peuvent être bruitées et le repérage de traductions effectives est une tâche délicate.

L'hypothèse de Lu *et al.* (2001, 2002, 2003) est que les anchor textes qui pointent vers les mêmes pages contiennent certainement des termes proches. Parmi ces termes, certains sont écrits dans des langues différentes et sont susceptibles d'être des traductions l'un de l'autre. Une approche probabiliste est utilisée pour l'identification des traductions. Les expériences de Lu *et al.* (2001) ont montré que 57% des termes testés en requête obtiennent une traduction correcte en chinois dans le top1 des traductions candidates, et 91% dans le top10. Lu *et al.* (2003) ajoute un module qui fait appel à une langue intermédiaire lorsque les traductions ne peuvent pas être extraites de façon directe. Par exemple, afin d'obtenir la traduction du terme

[1] Schéma de Lu *et al.* (2001, 2003).

anglais *Sony*, en chinois simplifié, la traduction est d'abord extraite en chinois traditionnel (Lu *et al.*, 2003) :

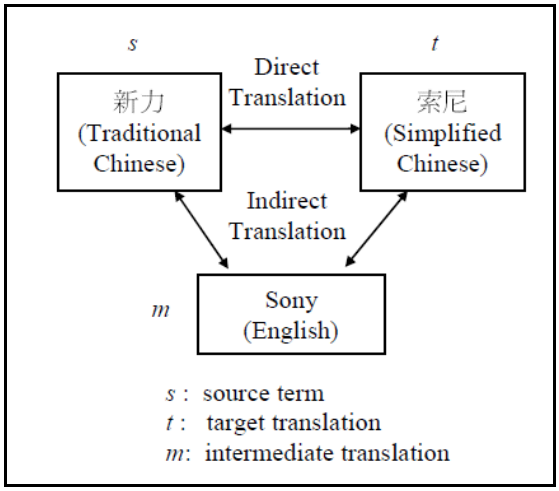


Figure 34. Traductions directes et indirectes

5.4 Acquisition de textes partiellement parallèles à partir du Web

Le terme de « textes bilingues » désigne le plus souvent des textes parallèles, c'est-à-dire un texte en langue source, aligné avec un texte traduit en langue cible, ayant strictement le même contenu (Nagata, 2001). Les méthodes basées sur le repérage de textes parallèles ou de liens hypertextes bilingues sur le Web restent (dans une moindre mesure que pour l'acquisition de corpus parallèles « traditionnels ») victimes d'une limitation des données. Afin de palier cette limite, une solution consiste à collecter des textes « partiellement parallèles » sur le Web, c'est-à-dire des documents mixtes d'un point de vue linguistique, mais qui ne sont pas des textes alignés :

"緑内障が早期発見によって管理可能な病気となったため、黄斑変性(macular degeneration)が先進国の視力障害の主な原因となりつつある。"

Figure 35. Extrait de document partiellement parallèle (japonais/anglais)

Le Web est riche d'un grand nombre de documents « partiellement » bilingues dont les caractéristiques peuvent être variées. Par exemple, la traduction d'un terme peut être présente de façon ponctuelle dans le corps d'un document essentiellement monolingue (*ibid.*) :

*Further support was guaranteed […], the Saudi Fund, France's Central Fund for Economic Cooperation (**Caisse Centrale de Coopération Economique**--CCCE).*

Dans ce type de documents, il est possible d'inférer que le texte entre parenthèses est une traduction du segment précédent (sans occulter d'éventuelles difficultés de segmentation). Ces caractéristiques peuvent être exploitées par des moyens automatiques afin de collecter de nouvelles traductions de termes. De plus, le contexte de l'usage est également disponible (*ibid*.). Ces textes partiellement parallèles sont le plus souvent des documents techniques, dans lesquels la traduction d'un terme technique est précisé, le plus souvent entre parenthèses, à la suite du terme source (ibid.). Cette caractéristique peut être exploitée afin d'extraire des traductions à partir du Web, notamment des traductions absentes de ressources dictionnairiques, parce qu'elles sont trop techniques ou trop récentes (Nagata, 2001).

Il est difficile d'évaluer la quantité de textes partiellement bilingues sur le Web. Cette quantité est dépendante des langues sources et cibles, et varie également en fonction des genres de documents. De plus, il faudrait distinguer lors de l'évaluation entre les termes simples et les termes complexes. Nagata (2001) propose une mesure d'évaluation de la quantité de textes « partiellement parallèles » pour le couple de langues japonais/anglais, en fonction de différents domaines de spécialité. A partir d'un dictionnaire bilingue[1], classé selon 19 catégories (telles que l'aéronautique, l'écologie, etc), 30 paires de termes japonais et anglais (simples et complexes) ont été sélectionnés pour chaque catégorie, et ont été testés en tant que requête sur le moteur de recherche *Google*. Les résultats ont montré que 42% des requêtes ont retournées au moins un document (*ibid.*), ce qui montre que la quantité de textes « partiellement parallèles » n'est pas négligeable.

[1] (NOVA Inc., 2000).

5.4.1 Typologie des textes « partiellement » parallèles sur le Web

En nous appuyant sur la typologie proposée par Nagata (2001) pour le japonais et l'anglais, nous proposons une typologie des textes « partiellement » bilingues sur le Web.

Paragraphes alignés[1]

Les paragraphes alignés sont des documents comportant des paragraphes traduits dans une langue cible. Chaque paragraphe est complètement monolingue, et les paragraphes traduits succèdent les paragraphes sources. Ce type de documents concerne fréquemment des documents officiels destinés à être lus par des locuteurs non natifs, ou des articles scientifiques dans lesquels seuls les titres et les résumés sont traduits :

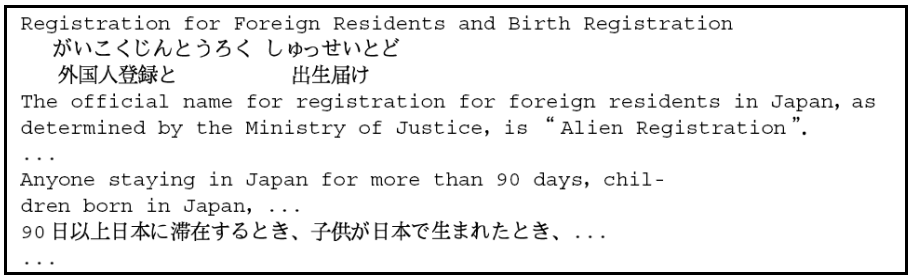

Registration for Foreign Residents and Birth Registration
がいこくじんとうろく しゅっせいとど
外国人登録と 出生届け
The official name for registration for foreign residents in Japan, as determined by the Ministry of Justice, is "Alien Registration".
...
Anyone staying in Japan for more than 90 days, children born in Japan, ...
90日以上日本に滞在するとき、子供が日本で生まれたとき、...
...

Figure 36. Exemple de « paragraphe aligné »[2]

Tables

Le document se présente sous la forme d'une table comprenant des paires d'équivalences de termes. Il s'agit le plus souvent de glossaires bilingues :

[1] « Aligned paragraph format » en anglais (Nagata, 2001).

[2] http://www.pref.akita.jp/life/g090.htm

instrument	bass drum	grosse caisse
instrument	bassoon	basson
instrument	bugle	clairon
instrument	cello	violoncelle
instrument	double bass	contrebasse
instrument	electric guitar	guitare électrique
instrument	English cor	cor anglais
instrument	French horn	cor français
instrument	Gong	gong
instrument	guitar	guitare
instrument	harmonica	harmonica
instrument	harp	harpe
instrument	Jew harp	guimbarde
instrument	kettledrum	timbale
instrument	mandolin	mandoline
instrument	Oboe	hautbois
instrument	organ	orgue
instrument	panpipe	flûte de Pan
instrument	Piano	piano
instrument	piccolo	piccolo
instrument	snare drum	caisse claire

Figure 37. Exemple de format « table »[1]

Texte plein

Les termes en langue cible sont précisés de façon ponctuelle dans le corps d'un document monolingue en langue source :

La caisse claire (snare drum) est constituée d'un fût (shell) , de deux peaux (heads), d'un timbre (snare wires), de deux cercles (hoops), d'un déclencheur (throw-off), de coquilles (lugs) et de vis (screws).Les caisse-claires que l'on rencontre le plus souvent sont fabriquées en érable ou en aluminium, mais il existe une grande variété de matériaux utilisés pour leur fabrication: laiton, bronze, fonte, fer, acier, cuivre, érable, bouleau, bambou, cerisier, tilleul, ou des bois exotiques comme le bubinga, le jarrah, le sheoak, ou encore des matières synthétiques comme l'acrylique, la fibre de verre ou le carbone, des matériaux composite et même du verre ! Les dernières nouveautés en terme de construction de caisse-claires sont les fûts hybrides en bois et métal (DW, edge), ou bois et acrylique (Spaun, hybrid).

Figure 38. Exemple de format « texte plein »[2]

[1] http://www.glossaire.be/english_french/glossaire_multimedia_anglais_francais.htm

[2] http://www.jerrock.com/66/node/154

Une majorité des documents bilingues sur le Web répondent à cette catégorie (*ibid.*). La figure récapitule les différents types de textes partiellement parallèles :

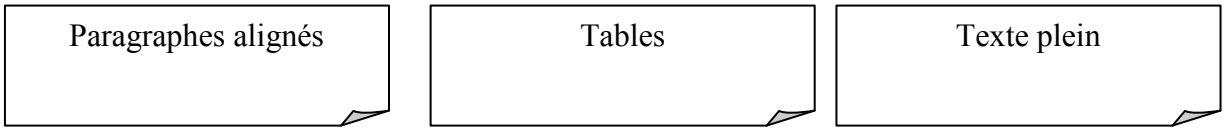


Figure 39. Typologie des documents partiellement parallèles

5.4.2 Méthodes d'acquisition

Une majorité des travaux qui exploite les propriétés du Web « partiellement parallèles » concerne les travaux centrés sur la traduction de langues asiatiques en anglais (Cheng *et al.*, 2004a). Nous présentons ces travaux et nous montrerons par la suite que le Web « partiellement parallèle » peut également être exploité pour des langues telles que le français et l'anglais. Les travaux de Nagata (2001) proposent une méthode d'extraction de traductions de termes spécialisés du japonais vers l'anglais, à partir de documents partiellement parallèles sur le Web. La méthode est basée sur le repérage de documents contenant à la fois le terme source et le terme cible via un moteur de recherche et sur un calcul de distance entre les deux termes au sein du document. Tout d'abord, sont récoltés les 100 premiers documents retournés par un moteur de recherche contenant les termes sources japonais et sont éliminés les documents exclusivement japonais. Pour chaque terme anglais présent dans le document, un calcul de probabilité est estimé, en prenant en compte la distance entre le terme source et la traduction candidate au sein d'un même document, et la traduction candidate obtenant le plus haut score est sélectionnée. Parmi les couples de termes utilisés afin de tester la quantité de documents partiellement parallèles sur le Web, 50 de ceux qui avaient retourné au moins un document sont sélectionnés comme banc de test. Parmi eux, 34 ont retourné des pages partiellement parallèles au sein des 100 premiers résultats. En ce qui concerne l'alignement de termes anglais/japonais, 60% des résultats contiennent un alignement correct dans les 10 premiers candidats termes.

Cheng *et al.* (2004a) présente une méthode de traduction pour l'anglais et le chinois, à partir de requêtes en langue source dont les résultats sont limités à la langue cible. L'hypothèse est que la présence de termes en langue source au sein de pages écrites en langue cible peut être un indice de repérage de documents mixtes. Les étapes de traitement à partir des documents mixtes sont tout d'abord une extraction terminologique, puis un alignement des traductions candidates avec le terme source. L'alignement des traductions est basée sur deux stratégies complémentaires : l'une mesure le taux de co-occurrence sur le Web entre le terme source et la traduction candidate, l'autre compare la similarité des vecteurs de co-occurrences sur le Web entre le terme source et la traduction candidate. Une évaluation, dans le cadre de la recherche multilingue, offre une précision de 46% sur la première traduction candidate pour les requêtes les plus populaires et 58% pour le top 5. En ce qui concerne des requêtes aléatoires, la précison est de 40% pour le top 1 et de 60% pour le top 5. Dans la même lignée, Cheng *et al*. (2004b) proposent une approche basée sur une utilisation « partiellement » multilingue du Web (pages contenant à la fois de l'anglais et des langues asiatiques) afin de construire des lexiques multilingues prenant en compte des variations régionales pour la langue chinoise.

Huang *et al.* (2005) présentent une méthode d'acquisition de traductions chinois/anglais d'Entités Nommées à partir du Web, en exploitant des documents partiellement bilingues. La méthode est basée sur le repérage de traductions au sein de résumés mixtes, à partir de requêtes interlingues « enrichies », c'est-à-dire associant le terme source aux traductions de mots-clés apparentés. Par exemple, le mot-source *Faust* en japonais constitue d'abord une première requête. A partir des résumés retournés, une liste de mots-clés est constituée selon différents critères tels que le taux de co-occurrence du mot source et du mot-clé thématique sur le Web, le recensement de la traduction du mot-clé thématique dans des ressources pré-existantes, la faible quantité de traductions candidates possibles du mot-clé thématique, le fait que le mot-clé thématique soit un nom ou un syntagme nominal. Les mots-clés les plus significatifs sont traduits en anglais (langue cible) et sont générées des requêtes enrichies du type de *Faust(en japonais) Goethe*, comme l'illustre le schéma (Huang *et al.*, 2005) :

Figure 40. Exemple de « requête enrichie »

Des critères phonétiques, sémantiques et statistiques (mesure de la distance entre le mot source et le mot cible au sein des résumés) sont ensuite appliqués à l'extraction des résumés mixtes retournés par ce type de requêtes. Les résultats de traduction offrent une précision de 46% en utilisant les 10 premiers résumés retournés, et de 80% en utilisant 165 résumés. Zhuang et Vines (2004, 2005) utilisent une méthode similaire de traduction chinois/anglais pour la détection de termes inconnus[1]. Wu et Chang (2007) présentent le système *TermMine*, système d'acquisition de « translitérations » de l'anglais vers le chinois. La méthode est basée sur l'expansion de requêtes et la collecte de résumés « mixtes » sur le Web.

5.5 Le Web, un corpus comparable

Fung et Yee (1998) proposent une méthode d'extraction automatique de nouvelles traductions de mono-termes à partir de textes monolingues (journaux) en anglais et en chinois sur le Web.

[1] *OOV terms (Out-Of Vocabulary terms).*

En partant du constat qu'un mot est fortement associé à d'autres mots dans un contexte donné (Rapp, 1995, Fung, McKeown, 1997), la méthode est basée sur une mesure de similarité entre vecteurs de mots : les contextes d'un mot source et d'un mot cible. La mesure calcule le nombre de mots en commun en contextes sources et cibles. Dans un contexte bilingue, les mots communs consistent en une paire bilingue de mots. Les contextes des mots en langue source et en langue cible sont représentés sous la forme de vecteurs de mots. Chaque mot est associé à sa mesure de poids dans le corpus (la mesure utilisée est le TF/IDF).

La méthode d'acquisition automatique de traductions compositionnelles de termes techniques de Tonoike *et al.* (2005) est basée sur la collecte de corpus spécialisés à partir du Web, à partir de termes techniques complexes sources. Le corpus est utilisé afin de valider des traductions candidates générées par la concaténation des traductions (contenues au sein d'une ressource existante) de chaque élément formant un terme complexe. Les termes techniques sources sont catégorisés selon trois groupes, en fonction du nombre de traductions candidates de chaque constituant du terme complexe disponibles au sein d'une ressource bilingue existante (Tonoike *et al.*, 2005). Les trois catégories sont les suivantes :

- Les termes complexes dont les traductions candidates de chaque constituant sont égales à un.

- Les termes complexes dont les traductions candidates de chaque constituant sont supérieures à un : la tâche consiste à sélectionner la traduction appropriée parmi les traductions candidates. La méthode de Tonoike *et al.* (2005) consiste à sélection la combinaison des traductions candidates formant un terme complexe cible la plus fréquente au sein du corpus collecté.

- Ceux dont les traductions candidates ne sont pas recensés au sein du lexique bilingue : la tâche consiste à générer ces traductions.

5.6 Les fréquences sur le Web pour l'aide au choix lexical

5.6.1 Méthodes d'acquisition

L'utilisation la plus répandue des fréquences sur le Web dans la vie courante concerne la vérification orthographique (Kilgarriff et Grefenstette, 2003). Par exemple, la réponse à une hésitation entre les termes anglais *speculater* ou *speculator* nous est donnée par *Google* :

Figure 41. Recherche du terme speculator sur Google (août 2008)

Figure 42. Recherche du terme speculater sur Google (août 2008)

Le principe de vérification lexicale à partir d'un moteur de recherche ne se limite pas à la vérification orthographique. Le même type de stratégie peut être appliqué à la vérification de traduction, lorsque plusieurs choix lexicaux sont candidats.

Grefenstette (1999) est l'un des premiers à avoir mis en rapport l'utilisation des fréquences sur le Web avec le choix lexical pour la traduction. A partir d'unités lexicales complexes (de type *NOM-NOM*) extraites d'un lexique bilingue existant, pour les couples allemand/anglais et espagnol/anglais, dont la traduction est compositionnelle, les traductions de chaque élément sont combinées et les fréquences de chaque traduction candidate sont testées. Prenons pour exemple l'unité lexicale *groupe de travail*. Les traductions candidates de *groupe* sont les suivantes (Grefenstette, 1999) :

groupe > cluster

groupe > group

groupe > grouping

groupe > concern

groupe > collective

Les traductions de *travail* sont les suivantes :

travail > work

travail > labor

travail > labour

Une combinaison de toutes les traductions possibles offre la combinatoire suivante :

work cluster, work group, work grouping, work concern, work collective

labor cluster, labor group, labor grouping, labor concern, labor collective

labour cluster, labour group, labuor grouping, labour concern, labour collective

Les requêtes de chaque traduction candidate donnent les résultats suivants (*ibid.*) :

	WWW count		*WWW count*
labor grouping	4	labor cluster	7
labour concern	8	work grouping	27
labor concern	28	work cluster	112
labor collective	144	labour collective	158
work concern	170	work collective	242
labor group	844	labour group	1131
work group	67238		

Figure 43. Résultats de fréquences des traductions candidates

Les résultats de fréquences sur le Web permettent de sélectionner la traduction adéquate (ici *work group*). Les résultats montrent une précision de 86-87% pour des décisions générées via le plus grand nombre d'occurrences retournées pour chaque traduction candidate sur le moteur de recherche *Altavita.* Les tests de Grefenstette (1999) sont limités à des combinaisons lexicales contraintes déjà traduites au sein d'un dictionnaire, en vue de tester l'utilité du Web pour ce type de tâches d'acquisition de traductions. Wehmeier (2004) propose un outil basé sur une méthode qui se situe dans la même lignée que Grefenstette (1999), pour la catégorie *nom-adjectif* en allemand et en anglais. La méthode de validation est basée sur les résultats de la fréquence la plus élevée pour toutes les traductions candidates, à partir du moteur de recherche *Google*. Les collocations sources ont été extraites à partir d'un échantillon du *British National Corpus*. Une évaluation indépendante des ressources testée à partir d'un échantillon de 100 collocations donne une précision de 67,75 % pour une parfaite intelligibilité et de 82,5% pour une compréhension générale. Contrairement à nous, Grefenstette (1999) et Wehmeier (2004) exploitent des collocations source non issues du Web. Dans la même lignée, Cao et Li (2002) proposent une méthode mixte, dans laquelle les fréquences sur le Web de combinaisons lexicales contraintes candidates sont également exploitées en tant qu'indice de validation. Leur expérimentation est basée sur la traduction de phrases nominales de l'anglais vers le chinois. A partir d'une unité lexicale complexe, les traductions candidates sont recherchées sur le Web et des calculs statistiques sont employées pour la validation. Li *et al.* (2003) développe un système, le « English Reading Wizard » utilisé pour l'aide à la lecture et à la compréhension, basé sur la même méthode que Cao et Li (2002).

5.6.2 Véracité vs. popularité

Nous soulignons une conséquence théorique des méthodes d'acquisition de traductions à partir des fréquences sur le Web et insistons sur deux phénomènes non parfaitement assimilables. La popularité d'un événement retourné par un moteur de recherche ne garantit pas sa véracité. En effet, une expression peu usitée et peu populaire peut avoir une valeur de vérité, tandis qu'une expression populaire peut ne pas avoir cette valeur de véracité. Toutefois, d'un point de vue pratique, il n'est pas évident de palier cette limite théorique. Notre hypothèse est que la fréquence des expressions linguistiques sur le Web est, dans une certaine mesure, le reflet de l'usage. Naturellement, ce « miroir linguistique » ne peut être considéré que comme un « miroir déformant » puisqu'il est dépendant des résultats retournés par les moteurs de recherche. Toutefois, le linguiste se doit d'utiliser les outils qu'il a sa disposition, en gardant conscience de leurs limites. Dans nos travaux, nous utilisons la méthode des fréquences au cours d'une phase de notre méthodologie, tout en gardant conscience que la popularité des événements langagiers retournés par les moteurs de recherche n'est pas un gage « aveugle » de véracité. L'ajout d'autres phases de traitement, ainsi que d'une série de filtres nous permettent d'exploiter l'intérêt pratique de telles méthodes, sans être réellement victimes des effets néfastes de la seule prise en compte des fréquences sur le Web. De plus, le recours à une étape de validation humaine nous permet de contrôler nos données et de tester la validité de notre processus de traitement.

5.7 Conclusion

Notre méthode est un système modulaire, qui s'appuie sur différentes stratégies d'utilisation du « Web bilingue », en fonction des caractéristiques des unités lexicales complexes :

- Utilisation du Web parallèle et « partiellement » parallèle : les caractéristiques du Web parallèle et « partiellement » parallèle sont exploitées dans notre méthode afin de filtrer au préalable les nombreuses traductions candidates générées automatiquement (phase 2, chapitre 7 et phase 3, chapitre 8). Notre hypothèse est qu'une traduction valide doit apparaître au moins une fois en co-occurrence avec l'unité lexicale complexe source au sein d'un même document.

Le Web « partiellement » parallèle intervient également lorsque notre système collecte des résumés « mixtes » (phase 3, chapitre 8). Ces derniers sont utilisés afin de repérer des traductions non compositionnelles ou inconnues des dictionnaires.

- Utilisation du Web en tant que « corpus » comparable : les caractéristiques du Web comparables sont exploitées lors de notre acquisition automatique de mondes lexicaux en langue source et en langue cible (phase 2, chapitre 7). Notre hypothèse est qu'une traduction candidate valide doit avoir un environnement textuel sur le Web (« monde lexical ») proche de celui de l'unité lexicale complexe source.

- Utilisation des fréquences sur le Web : les fréquences sur le Web sont exploitées dans une seule étape de notre phase, celle qui concerne les traductions compositionnelles non polysémiques (phase 1, chapitre 6). Les fréquences sont utilisées afin de prendre une décision de validation ou de non validation lorsqu'un unique choix de traduction candidate ne se présente.

L'originalité de notre approche est, d'une part, de combiner ces différentes stratégies de façon modulaire et d'autre part, d'adapter les traitements en fonction des caractéristiques des unités lexicales complexes sources (compositionnalité, traduction des constituants inconnue de notre dictionnaire, etc.). Notre méthodologie est basée sur des prises de décision, dont les résultats obtenus à chaque étape sont éliminés des unités lexicales restantes à traduire. Nous prenons pour point de départ les seules informations de traduction des constituants (base et co-occurrent) contenues dans notre dictionnaire. La première prise de décision du système est basée sur le nombre de traductions candidates possibles pour chaque constituant. Si chaque constituant n'a qu'une traduction candidate, le traitement consiste à décider si la traduction candidate est valide. En revanche, si chaque constituant connaît plus d'une traduction candidate, il s'agit d'effectuer un choix lexical entre les traductions candidates avant de juger si la combinaison est correcte. Enfin, si aucune des combinaisons traduites candidates n'a été validée, une autre méthode sera employée afin de résoudre l'une des deux difficultés restantes :

- la combinaison traduite est compositionnelle, mais il nous manque l'une (ou les deux) traductions adéquates pour les constituants ;

- la combinaison n'est pas compositionnelle, et il nous faut alors obtenir la traduction adéquate sans passer par une phase de traduction littérale.

Chapitre 6. Architecture et spécification du système d'acquisition des traductions

6.1 Introduction

Notre méthodologie passe par deux grandes phases, l'une d'acquisition d'unités lexicales complexes monolingues, l'autre de traduction. La phase d'acquisition monolingue consiste en la construction d'un très vaste corpus de pages Web, en français dont sont extraites les unités lexicales complexes sources. La phase de traduction est composée d'une architecture modulaire, qui analyse les propriétés des unités à traduire et les regroupe dans le module de traitement adapté. Nous présentons chaque module de traduction dans un chapitre individuel. Ce chapitre traite de notre première phase de traduction, celle qui concerne les unités lexicales non polysémiques. Nous présentons d'abord notre méthode d'acquisition d'unités lexicales complexes sources (en français), à partir d'un vaste corpus de pages Web (6.2). Nous décrivons ensuite le premier module de traduction, qui détecte et traduit les unités lexicales compositionnelles non polysémiques (6.3 à 6.7). Nous analysons enfin les résultats de cette étape (6.8). Notre méthodologie de traduction[1] répond à deux spécificités, le fait d'être constituée d'une architecture modulaire adaptée aux caractéristiques des unités lexicales, et le

[1] Toutes les expériences rapportées dans cette thèse ont été réalisées sous environnement Linux, par la réalisation de scripts écrits en bash et en PERL.

fait de procéder par élimination successive, c'est-à-dire que les unités non traduites dans un module sont reléguées au module suivant.

Architecture modulaire

Notre méthodologie d'acquisition de traductions est adaptée aux caractéristiques linguistiques des unités à traduire. Chaque module est spécifique à une caractéristique donnée. Nous nous centrons sur le degré de polysémie des unités lexicales sources, ainsi que sur leur caractère compositionnel ou non-compositionnel. Notre hypothèse est que la tâche de traduction est dépendante de ces deux critères :

- **Degré de polysémie** : lorsqu'au moins un des constituants de l'unité lexicale complexe est polysémique, la tâche de traduction consiste à sélectionner l'unité lexicale cible adéquate parmi toutes les traductions candidates. Ce choix implique une désambigüisation lexicale de l'unité source. Par exemple, afin de valider la traduction candidate *central fund*, pour *caisse centrale*, il faut connaître l'usage de *caisse (BANQUE)* et le sélectionner parmi de nombreux choix d'usages possibles (*TAMBOUR* → *drum*, VALISE → *case*, etc.). En revanche, lorsque les deux constituants de l'unité lexicale complexe ne sont pas polysémiques, la tâche de traduction ne consiste plus en un choix lexical, mais en une décision de validation ou de non validation. Il s'agit de juger de l' « aspect collocationnel » de la traduction candidate.

- **Compositionnalité** : nous avons montré que certaines unités lexicales complexes sont transparentes du point de vue du sens, et d'autres ne le sont pas. La tâche de traduction doit s'interroger sur le caractère transparent ou non transparent de la traduction. Si la traduction est transparente, une simple combinaison de la traduction de chaque constituant est satisfaisante, à partir de ressources dictionnairiques. En revanche, si la traduction n'est pas transparente, nous utilisons le Web pour collecter la traduction adéquate.

Notre méthode est constituée de modules adaptés à chaque cas.

Eliminations successives

Le traitement modulaire fonctionne également par filtres successifs, c'est-à-dire que les unités lexicales complexes non traduites dans un module sont reléguées au module suivant. Les modules ne fonctionnent donc pas en parallèle, mais de façon successive. Chaque module a accès aux informations de non validation des unités précédentes et les traite de la même façon que les autres unités lexicales qui lui sont attribuées d'office, par leurs caractéristiques. A partir d'une liste d'unités sources, celles qui sont validées dans la première phase sont éliminées de la liste. Les unités restantes sont traitées à l'étape suivante, et ainsi de suite.

La figure 44 synthétise l'ensemble de notre méthodologie :

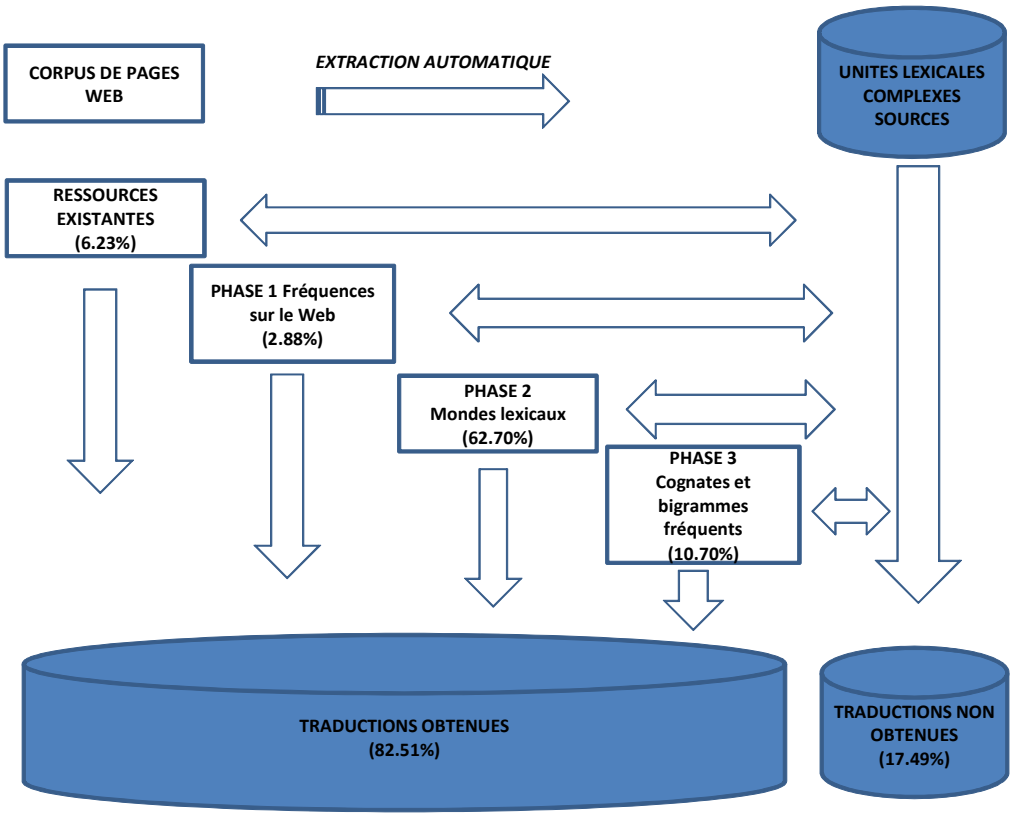


Figure 44. Etapes de traitements

- Dans une première étape, nous collectons des unités lexicales sources, en français, à partir d'un vaste corpus de pages Web.

- Nous générons ensuite toutes les traductions candidates via nos ressources dictionnairiques. Les traductions candidates sont analysées et leurs caractéristiques sont utilisées afin d'affecter chaque unité lexicale source dans le module adapté.

- Si aucun constituant de l'unité lexicale complexe n'est polysémique, nous appliquons une méthode basée sur les fréquences sur le Web. Les unités lexicales complexes non traduites sont rejetées à l'étape suivante.

- Si au moins un des éléments est polysémique, nous procédons à une comparaison des mondes lexicaux en français et en anglais sur le Web, qui vise à une désambigüisation lexicale pour la traduction. Les unités lexicales complexes non traduites sont à nouveau rejetées à l'étape suivante. Nous faisons l'hypothèse qu'un certain nombre des unités non traduites sont non compositionnelles, car aucune des traductions candidates dont nous disposions via le dictionnaire n'a été validée à ce stade.

- Si la traduction est non compositionnelle, ou si l'un des constituants est inconnu de notre dictionnaire, nous appliquons une méthode basée sur une collecte de résumés « mixtes » sur le Web.

6.2 Acquisition automatique d'unités lexicales complexes à partir du Web

6.2.1 Contraintes d'acquisition monolingue

Notre objectif est de proposer une méthodologie d'acquisition automatique de traductions en anglais, à partir d'un grand nombre d'unités lexicales complexes françaises. Afin de proposer un banc de test intéressant, ces unités lexicales complexes doivent être nombreuses et les

usages des termes polysémiques variés. Nos contraintes d'acquisition d'unités lexicales sources sont les suivantes :

- Lorsqu'un nom source est polysémique, les unités lexicales complexes doivent s'inscrire dans différents usages. Par exemple, pour le nom source *appareil*, nous souhaitons obtenir des usages tels que *PHOTOGRAPHIQUE*, *MENAGER*, etc. Nous ne visons pas à l'exhaustivité, mais le banc de test présenté doit être difficile.

- Nous visons à la construction d'une base lexicale de bonne qualité, avec une totale automatisation : les données étant nombreuses, une tâche de validation manuelle serait trop coûteuse.

- Nous visons à la construction d'un lexique quantitativement étendu et évolutif, qui puisse grossir de façon continue.

- Devant la complexité des phénomènes de traduction des unités lexicales, nous nous centrons sur un nombre réduit de relations syntaxiques :

NOM-ADJECTIF

NOM-de (d')-NOM

Une évolution ultérieure sera d'intégrer de nouveaux patrons morpho-syntaxiques au système. Nous laissons volontairement de côté le patron morpho-syntaxique *ADJECTIF-NOM*. Nous faisons l'hypothèse que les unités lexicales de ce type restent souvent ambiguës lexicalement, contrairement aux patrons étudiés. Par exemple, si *grosse caisse* est une unité terminologique dont le sens de *caisse* est désambigüisé, *petite caisse* ne permet pas de désambigüiser *caisse*.

Dans Léon (2006), nous testons un aspect de notre méthodologie dont les données sources sont les « termes associés » retournés par le moteur de recherche *Exalead*, c'est-à-dire les séquences polylexicales qui sont en co-occurrence fréquente avec la requête. L'intérêt de cette

fonctionnalité est de pouvoir affiner les requêtes en incluant et/ou excluant des usages. Voici deux exemples de termes associés, les uns au terme *appareil*, les autres au terme *caisse* :

Figure 45. Termes associés à la requête caisse sur Exalead

Figure 46. Termes associés à la requête appareil sur Exalead

Plusieurs caractéristiques ne répondent pas à nos contraintes actuelles :

- Les usages sont très peu diversifiés. Par exemple, pour la requête *appareil*, seul un usage est représenté (*PHOTOGRAPHIQUE*), ce qui est très faible pour un nom fortement polysémique.

- Les résultats contiennent parfois du bruit.

- La quantité de termes associés n'est pas élevée (une dizaine dans nos exemples).

- Les termes associés ne sont pas nécessairement composés du terme cherché, comme par exemple *grand angle*, pour *appareil*, ce qui limite encore la quantité d'unités lexicales complexes associées à un terme simple.

- Les patrons morpho-syntaxiques ne sont pas tous pertinents pour notre étude, ce qui réduirait davantage notre filtre.

Afin d'obtenir une liste qui réponde à nos critères et qui puisse grossir de façon évolutive, nous optons pour la constitution d'un très vaste corpus de pages Web, collecté à partir de noms simples (*têtes sémantiques*), en français, à partir duquel nous récoltons les unités lexicales complexes associées.

6.2.2 Collecte de pages Web et sous-corpus

Notre point de départ constitue une liste d'unités lexicales simples collectées de façon aléatoire au sein de notre dictionnaire électronique bilingue *Collins Pocket* (français-anglais) à partir desquels sont collectées les pages Web. En l'état actuel de nos travaux, le nombre d'unités simples est au nombre de 1664. Toutefois, la collecte de pages Web continue de grossir. Les unités lexicales simples doivent répondre à certains critères :

- Seule la catégorie grammaticale des noms est conservée. Les noms constituent les *têtes sémantiques* des unités lexicales complexes extraites.

- Les noms composés (typographiquement séparés par un trait d'union) ne sont pas conservés, comme par exemple :

abat-jour

vide-poches

Les noms composés forment une unité lexicale complexe en eux-même, ce qui conduirait à prendre en compte des unités lexicales complexes de longueur plus élevée que l'objectif défini dans le cadre de nos travaux.

- Nous supprimons les unités lexicales complexes qui forment une entrée lexicale en elle-même, pour les mêmes raisons que précédemment, comme dans l'exemple de :

compte rendu

béret basque

- Enfin nous ne tenons pas compte des Entités Nommées (critère typographique d'une majuscule au début du terme), ce qui élimine des termes du type de :

Alsace

Yougoslavie

Pour chaque mono-terme, nous récoltons les pages Web associées par le biais de requêtes, via l'*API Yahoo*[1]. Les requêtes sont formulées au singulier et au pluriel, uniquement pour les pages en français. Elles sont sous la forme d'exclusion du singulier ou du pluriel, afin d'élargir les résultats et se présentent en trois temps pour un même nom, comme dans l'exemple suivant :

appareil -appareils

appareils -appareil

appareil +appareils

Nous collectons les mille premiers résultats de chaque type de requête, ce qui donne environ 2500 pages de résultats par nom simple, quantité variable selon la fréquence d'emploi du nom

[1] http://developer.yahoo.com/

sur le Web. Les pages Web sont ensuite nettoyées automatiquement par le biais de scripts afin d'éliminer le bruit, lié aux caractéristiques des pages Web (rétablissement de caractères dû au codages, lignes vides, adresses Internet, images, PDF, etc.). Les pages Web sont ensuite étiquetées à l'aide du logiciel d'étiquetage morpho-syntaxique *Treetagger*[1] . Nous constituons le sous-corpus étiqueté de chaque mono-terme, en récoltant son contexte de plus ou moins dix termes. Le résultat constitue un sous-corpus des noms sources, qui se présente sous la forme de trois colonnes, avec un terme par ligne et les informations de lemme, de forme et de catégorie grammaticale sur chaque colonne :

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Figure 47. Exemple de résultat obtenu par Treetagger

6.2.3 Extraction d'unités lexicales complexes

Notre méthode d'extraction automatique d'unités lexicales complexes sources est un processus complètement automatique, sans aucun filtre de validation manuelle. Naturellement, les résultats peuvent contenir une part de bruit. Toutefois, nous mettons en place une série de filtres automatiques afin d'éliminer au maximum d'éventuelles unités lexicales complexes erronées. Nous partageons l'idée selon laquelle l'extraction d'unités lexicales complexes doit se baser sur des critères linguistiques et sur des critères de fréquence (Daille, 1994) : une unité lexicale complexe est une co-occurrence préférentielle de termes (donc relativement fréquente), mais surtout elle entre dans une relation de dépendance

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

syntaxique. Pour nous, l'aspect de relation restreinte de dépendance syntaxique est le plus important, car le critère de fréquence est un critère délicat. Notre hypothèse est que le Web peut être utilisé afin d'évaluer l'« aspect collocationnel » de dépendances syntaxiques préalablement collectées. Nous présentons les critères d'identification des unités lexicales complexes, les critères morpho-syntaxiques dans un premier temps, et les critères de fréquence dans un second temps.

Critères morpho-syntaxiques

A partir des sous-corpus étiquetés, nous définissons les patrons morpho-syntaxiques répondant aux relations de dépendances syntaxiques recherchées. Notre méthode se base sur la définition de règles filtrant des éléments catégoriels avant et après l'unité lexicale cible, afin d'éviter des ambigüités de rattachement. Les règles établies n'extraient que des unités lexicales complexes contigües, ce qui présente la limite d'extraire des combinaisons principalement figées et d'obtenir du silence dans le cas de certaines constructions syntaxiques. Dans Léon (2004b), nous avons réalisé un extracteur d'unités lexicales complexes en définissant des règles de patrons catégoriels prenant en compte des éléments non contigus, comme par exemple :

*Le **vent** est **fort** (adjectif attribut)*

*Le **vent**, qui est **fort** (proposition relative)*

A partir de ces constructions syntaxiques, Léon (2004b) extrait l'unité lexicale *vent fort*. Toutefois, nous optons volontairement ici pour un filtre simple, afin d'éviter au maximum des problèmes de bruit et d'obtenir une ressource lexicale de très bonne qualité. Léon (2004b) s'est appuyé sur un corpus terminologique dont la contrainte première est l'exhaustivité des termes complexes collectés, une phase manuelle permettant ensuite de valider les résultats. Dans le cadre de cette thèse, l'extraction porte sur des données beaucoup plus vastes et nous préférons favoriser la qualité des ressources monolingues afin de permettre une totale automatisation et nous centrer sur les méthodes de traduction. Notre hypothèse est que l'extraction de séquences contigües est moins susceptible d'être bruitée qu'une extraction de

dépendances non contigües. Nous présentons les règles morpho-syntaxiques et les contraintes associées[1].

- ***NOM ADJECTIF*** : nous collectons les patrons morpho-syntaxiques du type de *NOM-ADJECTIF*, comme dans les exemples :

NOM ADJECTIF

appareil ménager

caisse claire

parc naturel

Les contraintes associées à ce patron sont les suivantes :

(1) Le syntagme ne doit pas être suivi par un nom afin d'éviter des erreurs de rattachement du type de :

NOM ADJECTIF *NOM*

abonnement haut *débit*

Dans cet exemple, le syntagme collecterait le syntagme *abonnement haut*, ce qui n'est pas pertinent.

(2) Le syntagme *NOM ADJECTIF* ne doit pas être suivi par une préposition suivie d'un nom, comme dans l'exemple :

NOM ADJECTIF ~~*PREPOSITION NOM*~~

[1] Notre méthodologie peut être victime d'éventuelles erreurs d'étiquetage morpho-syntaxique du logiciel *Treetagger*, mais aucun logiciel étiquetage ne présente aucun bruit dans ses résultats.

Abri haut *de gamme*

Dans cet exemple, notre méthode collecterait l'unité *abri haut*.

- ***NOM PREPOSITION NOM*** : nous nous centrons sur la seule préposition *de*, ainsi que sur son extension *d'*. Le nom cible peut apparaître en position 1 ou en position 2, comme dans les deux exemples pour le nom *caisse* :

NOM PREPOSITION NOM

Logiciel de ***caisse***

Caiss*e de retraite*

Les contraintes établies pour ce patron sont les suivantes :

(1) Le syntagme ne doit pas être suivi par une préposition suivie d'un nom :

NOM DE(D') NOM ~~*PREPOSITION NOM*~~

amour de cours *de recréation*

Le syntagme erroné dans cette exemple serait *amour de cours*.

(2) Le syntagme ne doit pas être suivi par un adjectif :

NOM DE(D') NOM ~~*ADJECTIF*~~

Abattement de revenu *imposable*

Dans cet exemple, nous aurions collecté *abattement de revenu.*

(3) Le syntagme ne doit pas être suivi par un nom :

NOM DE(D') NOM -~~NOM~~

Appareil d'Air *France*

Nous aurions ici identifié *appareil d'air*.

Un anti-dictionnaire est appliqué aux patrons obtenus afin d'éliminer des mots généraux ou non pertinents tels que *divers*, *autre*, *différent*, etc. Le schéma suivant présente un graphe de décision appliqué aux patrons morpho-syntaxiques traités :

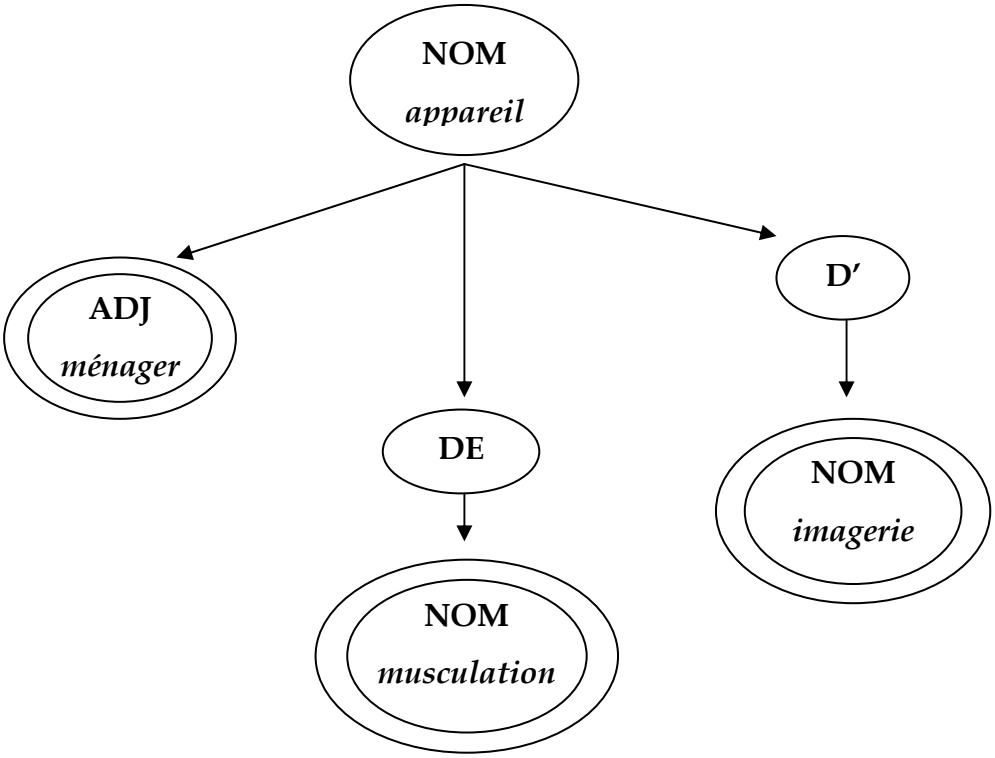


Figure 48. Graphe de décision des patrons syntaxiques pour l'identification des termes complexes

Critères de fréquence

Parmi les patrons morpho-syntaxiques collectés, nous posons plusieurs filtres de fréquence. Ces filtres sont considérés comme une évaluation de l' « aspect collocationnel » des relations

de dépendance syntaxique collectées. Ils permettent également de filtrer d'éventuelles erreurs liées à l'étiquetage morpho-syntaxique.

- La fréquence de l'unité lexicale complexe au sein des pages Web reliées au nom-source correspondant doit être supérieure ou égale à 10. Ce filtre est volontairement peu élevé car notre hypothèse est que les unités lexicales complexes peuvent apparaître un nombre réduit de fois dans les corpus envisagés.

- Les fréquences des unités lexicales complexes sont ensuite testées sur le Web, afin d'évaluer leur aspect « collocationnel ». Nous posons deux critères de fréquences, simples, à partir du Web :

  La fréquence de l'unité lexicale sur le Web doit être supérieure ou égale à 10000.

  La fréquence de l'unité lexicale, précédée par un article (défini ou indéfini), sur le Web doit être supérieure ou égale à 1000. Les requêtes sont du type de :

  « l'appareil ménager » OR « un appareil ménager »

Notre hypothèse est que les unités lexicales pertinentes sont certainement employées avec un article de façon significative. Par exemple, la requête *« l'appareil ménager » OR « un appareil ménager »* compte une fréquence de 43 500, tandis que la requête *« un style d'escalade » OR « le style d'escalade »* apparaît 358 fois. Le choix des valeurs a été déterminé de façon empirique, par observation des données. Ce filtre peut provoquer du silence dans certains cas, mais nous posons volontairement un filtre élevé afin d'obtenir des ressources de très bonne qualité. Cette difficulté est liée à la « zone d'incertitude » des unités lexicales complexes abordée dans le chapitre 2. Un filtre moins élevé aurait généré des unités lexicales complexes dont le statut « collocationnel » est susceptible d'être moins significatif.

6.2.4 Analyse des unités lexicales sources

A l'issue de cette étape, nous obtenons 9664 unités lexicales complexes, en français, associées à 1664 têtes sémantiques (noms simples). Le nombre moyen d'unités lexicales complexes par tête sémantique est de 5,8. Le nombre moyen est beaucoup plus élevé pour les têtes sémantiques polysémiques. Notre corpus de pages Web, quant à lui, est constitué d'environ 4 160 000 pages Web. Le schéma 49 présente le nombre d'unités lexicales sources obtenus par patron morpho-syntaxique :

Patron morpho-syntaxique	**Total**
NOM ADJECTIF	5166
NOM-DE-NOM	2934
NOM-D'-NOM	1564
TOTAL	**9664**

Figure 49. Proportion d'unités lexicales complexes par patron morpho-syntaxique

Le patron morpho-syntaxique *NOM-ADJECTIF* est particulièrement productif, puisqu'il concerne plus de la moitié des cas (53.45%). Le patron *NOM-DE-NOM* concerne, lui, 30.36% des cas, suivi du patron *NOM-d'-NOM* pour 16.18% des cas. Voici un exemple d'unités lexicales complexes associées au nom polysémique *caisse* :

caisse *NOM2*	***d'***	allocation, amortissement, assurance, épargne
***NOM1* caisse**	***de***	bas, logiciel, ticket
caisse *NOM2*	***de***	compensation, dépôt, garantie, mutualité, pension, prévoyance, résonance, retraite, sécurité
caisse *ADJECTIF*	***NOM-ADJ***	autonome, centrale, claire, commune, fédérale, nationale, noire, populaire, primaire, régionale

Figure 50. Unités lexicales complexes associées au nom « caisse »

L'objectif de notre extraction d'unités lexicales complexes n'est pas d'obtenir une liste exhaustive de co-occurrents pour chaque nom, mais d'obtenir ses co-occurrences les plus

significatives, tout en obtenant des usages variés pour les noms polysémiques. Les exemples montrent que les usages sont diversifiés, puisqu'on obtient entre autres, pour *caisse*, les usages *MUSIQUE*, *BANQUE*, *VOITURE*, etc. Voici les unités lexicales associées au mono-terme *appareil* :

appareil *NOM2*	**d'**	état, imagerie,
***NOM1* appareil**	**d'**	catégorie
appareil *NOM2*	**de**	chauffage, contrôle, cuisson, mesure, poche, production, protection
appareil *ADJECTIF*	**NOM-ADJ**	administratif, argentique, auditif,circulatoire, compact, critique, dentaire, digestif, électrique, électroménager, électronique, étatique, génital, gouvernemental, judiciaire, locomoteur, ménager, militaire, mobile, numérique, photo, photographique, politique, portable, productif, réflex, répressif, reproducteur, respiratoire, urinaire

Figure 51. Unités lexicales complexes associées au nom « appareil »

Voici celles associées au nom *parc* :

parc *NOM2*	**d'**	activité, attraction, aventure, exposition
***NOM1* parc**	**de**	gestion, place, projet
parc *NOM2*	**de**	bagatelle, loisirs, production, stationnement, verdure
parc *ADJECTIF*	**NOM-ADJ**	actuel, aquatique, arboré, archéologique, automobile, boisé, botanique, communal, départemental, éolien, fermé, fleuri, floral, forestier, français, historique, hôtelier, immobilier, industriel, informatique, linéaire, locatif, marin, matériel, municipal, national, naturel, nucléaire, olympique, ornithologique, paysagé, paysager, privé, provincial, public, régional, résidentiel, scientifique, social, technologique, thématique, tropical, urbain, verdoyant, zoologique

Figure 52. Unités lexicales complexes associées au nom « parc »

Voici les unités lexicales collectées pour le nom *rapport* :

rapport ***NOM2***	**d'**	activité, analyse, audit, autopsie, avancement, enquête, erreur, étape, étude, évaluation, expert, expertise, information, vraisemblance
NOM1 **rapport**	**de**	formulaire, immeuble, modèle, projet
rapport ***NOM2***	**de**	contraste, contrôle, force, gestion, mission, police, projet, recherche, situation, stage, suivi, synthèse
rapport ***ADJECTIF***	**NOM-ADJ**	annuel, conjoint, définitif, économique, étroit, explicatif, final, financier, global, hebdomadaire, indiqué, intégral, intérimaire, interne, médical, mensuel, mondial, moral, national, officiel, parlementaire, périodique, précédent, public, quotidien, scientifique, semestriel, sexuel, social, sommaire, statistique, technique, trimestriel

Figure 53. Unités lexicales complexes associées au nom « rapport »

Voici les unités lexicales du nom *tour* :

tour ***NOM2***	**d'**	angle, argent, honneur, horizon, ivoire
NOM1 **tour**	**de**	quart
tour ***NOM2***	**de**	babel, chant, chauffe, cochon, contrôle, cou, force, garde, guet, jeu, lit, magie, main, manège, passe, passe-passe, piste, poitrine, scrutin, table, taille, ville, vis
tour ***ADJECTIF***	**NOM-ADJ**	cycliste, final, précédent, rapide

Figure 54. Unités lexicales complexes associées au nom « tour »

Afin de tester notre méthode de traduction, nous réalisons un échantillon aléatoire parmi les unités lexicales complexes obtenues. Le sondage aléatoire est un principe statistique simple qui nous laisse supposer qu'il est représentatif des résultats que nous pourrions obtenir sur un autre échantillon ou sur la totalité de nos unités lexicales complexes. Naturellement, les résultats ne seraient pas complètement les mêmes, mais le sondage aléatoire est une mesure communément admise pour être représentative d'une population donnée. Cet échantillon comporte 1075 unités lexicales complexes, soit 11.12% de la totalité de notre base. La section suivante détaille la première phase de notre méthodologie.

6.3 Présentation de l'approche de traduction

Dans Léon et Millon (2005), nous présentons une méthode d'acquisition automatique de co-occurrences bilingues (français-anglais), du type de *NOM ADJECTIF*, *NOM1 DE NOM2* et *VERBE NOM*(*objet)*, basé sur un processus de validation sur le Web. A partir d'unités lexicales complexes en français, Léon et Millon (2005) génèrent toutes leurs traductions candidates grâce à un dictionnaire électronique. Ces traductions sont automatiquement filtrées à partir des résultats de leurs fréquences, sur le moteur de recherche *Google*. L'évaluation porte sur 10 mots français très polysémiques (*barrage*, *détention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* et *vol)* qui avaient été jugés comme les plus polysémiques parmi 200 noms de fréquence équivalente, lors du projet Senseval (Véronis, 1998) et constituent un banc de test difficile, qui a été utilisé par la suite dans divers travaux. Prenons les co-occurrences lexicales suivantes :

commettre un vol

réserver un vol

En combinant les traductions de chaque élément, les fréquences constituent une aide pour le choix lexical des unités cibles, ici entre *theft* (usage *DELIT*) et *flight* (usage *AERIEN*), par exemple. *Google* permet de valider les traductions correctes, grâce à leur nombre d'occurrences. Par exemple, la requête [*"commit a flight"* OR *"commit the flight"*] retourne seulement 13 résultats. La requête [*"commit a theft"* OR *"commit the theft"*] retourne quant à elle 5110 résultats. Parmi ces deux traductions candidates, les résultats sélectionnent de façon écrasante la traduction satisfaisante (*to commit-theft*), dans la lignée des travaux de Grefenstette (1999) et de Cao et Li (2002) :

	Effectifs absolus		***Effectifs par million***	
	flight	**theft**	**flight**	**theft**
commit	13	5510	0	306
reserve	33 500	3	592	0

Figure 55. Exemples de résultats sur Google (janvier 2005)

L'évaluation de la méthode de Léon et Millon (2005) montre que le Web permet de constituer ou compléter des bases de données lexicales multilingues de bonne qualité, de façon automatique, à partir des fréquences sur le Web. Les résultats sont particulièrement intéressants pour les patrons syntaxiques de type *NOM ADJECTIF* (précision de 97,1 %) et *VERBE NOM*(*objet*) (précision de 88,9 %). La méthode reste imparfaite pour le patron *NOM1 DE NOM2*, mais le taux de précision est honorable (76,9%), surtout étant donné la difficulté volontaire du banc de test choisi (mots très polysémiques). La limite principale de la méthode de Léon et Millon (2005) est d'ordre lexical. Elle concerne l'acquisition de traductions valides, mais non correspondantes à l'unité lexicale source, comme dans l'exemple :

cours de formation > *group rate*

Ici, l'une des traductions candidates du nom polysémique *cours* est *rate* (usage *FINANCE*), tandis que l'une des traductions candidates du nom polysémique *formation* est *group* (usage *COLLECTIVITE*). Bien que ces deux choix lexicaux soient incorrects, la méthode valide cette traduction car *group rate* est une unité lexicale existante, qui signifie *tarif de groupe*. Pourtant, la prise en compte du contexte lexical de l'unité lexicale source, *cours de formation* dans notre exemple et de sa traduction candidate erronée *group rate* est un indice désambiguisateur fort : notre hypothèse est que les mondes lexicaux français et anglais doivent être proches entre une unité lexicale source et sa traduction adéquate, tandis qu'ils doivent être différents lorsque la traduction est erronée. Une observation des résumés retournés par le moteur de recherche *Yahoo*, par exemple, permet d'accéder au contexte lexical et de comparer les mondes lexicaux. Comparons par exemple les résumés retournés pour *cours de formation* et sa traduction correcte*, training course* :

Figure 56. Résumés associés à « cours de formation »

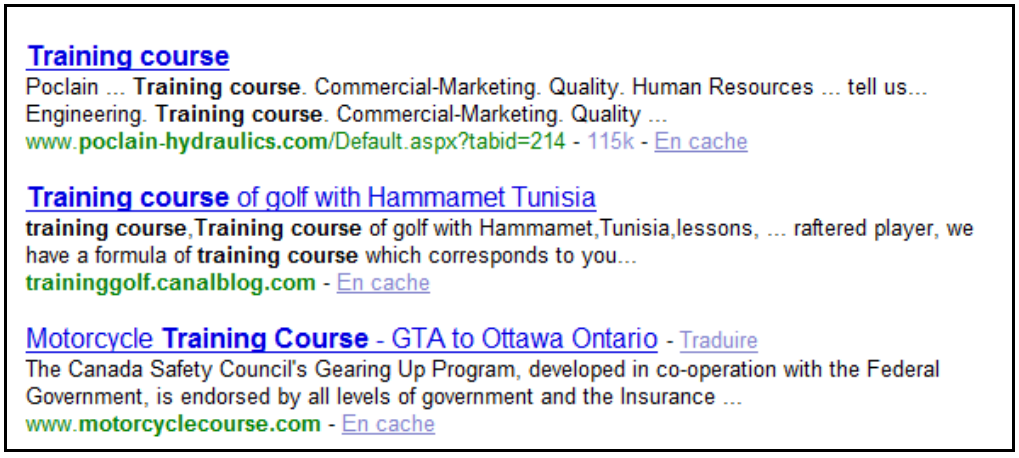


Figure 57. Résumés associés à « training course »

L'accès aux contextes lexicaux permet une comparaison des mondes lexicaux en français et en anglais : nous observons que la thématique est proche. En revanche, le contexte lexical de la traduction erronée, *group rate* est différent :

Figure 58. Résumés associés à « group rate »

Dans le chapitre suivant, nous montrons que les mondes lexicaux retournés par les résumés des requêtes sur un moteur de recherche permettent une désambigüisation lexicale pour la traduction. Toutefois, certaines unités lexicales complexes ne sont pas polysémiques et la tâche de traduction ne consiste pas en un choix lexical, mais en une simple évaluation de l' « aspect collocationnel » de la traduction candidate. Notre hypothèse est qu'une méthode basée sur les fréquences sur le Web est satisfaisante pour des unités lexicales non polysémiques, tandis qu'elle ne permet pas de palier tous les cas d'ambigüité lexicale. Dans notre travail de thèse, notre première phase consiste à traiter les unités lexicales non polysémiques, par une méthode proche de Léon et Millon (2005). Nous évaluons le degré de polysémie des unités lexicales et nous n'appliquons cette phase de méthodologie qu'aux unités non polysémiques. La section suivante détaille premièrement l'utilisation de ressources lexicales existantes pour les traductions déjà répertoriées (6.4), et deuxièmement, elle décrit notre méthode de traduction des unités lexicales non polysémiques (6.5 à 6.8).

6.4 Ressources préalables

Les dictionnaires courants contiennent un nombre restreint d'unités lexicales complexes, généralement les plus figées. Le *Collins Pocket French-English Dictionary*, disponible dans l'équipe sous forme électronique grâce à un accord avec l'éditeur Collins, ne propose de traductions que pour 2.60% des unités lexicales dont nous disposons comme échantillon, telles que :

agence matrimoniale > matrimonial marriage

bain moussant > bubble bath.

cordon ombilical > umbilical cord

Lorsque le dictionnaire propose une traduction, nous ne générons pas de traduction supplémentaire. Un avantage de cette phase est de traiter certains cas d'unités lexicales complexes qui se traduisent en anglais par une unité lexicale simple :

coup de balai > sweep

gant de toilette > flannel

Le plus souvent, ces unités lexicales n'apparaissent que dans un sens de traduction, celui de l'anglais vers le français. Afin d'élargir au maximum notre collecte, nous prenons en compte les deux sens de traduction. La figure 59 donne un exemple de traductions directement recensées dans notre dictionnaire.

PATRON	UNITE LEXICALE	TRADUCTION
NOM ADJECTIF	agence matrimoniale	matrimonial marriage
	bain moussant	bubble bath
	carte postale	postcard
	corde raide	tightrope
	cordon ombilical	umbilical cord
	escalier roulant	escalator
	homicide involontaire	manslaughter
NOM de NOM	plante grimpante	vine
	bouton de rose	rosebud
	coup de balai	sweep
	coup de chaleur	heatstroke
	gardien de nuit	night porter
	joueur de football	football player
	sortie de secours	emergency exit
	tapis de sol	groundsheet
	vin de table	table wine
NOM d' NOM	canalisation d'eau	water main
	char d'assaut	tank
	liste d'attente	waiting list
	papier d'aluminium	aluminium foil
	station d'essence	filling station

Figure 59. Exemples de traductions obtenues avec le dictionnaire Collins Pocket

Il arrive que plusieurs traductions soient proposées pour une même unité lexicale source. Dans ce cas, nous sélectionnons celle qui a la fréquence la plus haute sur le moteur de recherche *Yahoo*, pour la seule langue anglaise. Par exemple, les traductions suivantes sont recensées pour *coup de soleil* :

sunburn (Fréquence : 21 800 000)

sunstroke (Fréquence : 1 980 000)

La traduction *sunburn* est sélectionnée. Notre objectif est de ne proposer qu'une traduction par unité source afin d'évaluer de façon efficace les cas de désambigüisation lexicale lorsqu'un nom est polysémique (étape suivante). Nous appliquons le même principe de choix unique à tout notre lexique.

Le dictionnaire en ligne de *Google*[1] propose, quant à lui, pour 3.62% de traductions des unités lexicales sources à traduire. Nous collectons les traductions existantes. La Figure 60 illustre des traductions obtenues avec le dictionnaire de *Google*.

PATRON	UNITE LEXICALE	TRADUCTION
NOM ADJECTIF	animal domestique	pet
	an prochain	next year
	antenne parabolique	satellite dish
	carte bleue	credit card
	formation continue	continuing education
	fromage râpé	grated cheese
	reprise économique	economic recovery
NOM de NOM	acte de vente	bill of sale
	bague de fiançailles	engagement ring
	farine de maïs	corn starch
	front de mer	sea front
	mal de ventre	tummy ache
	pied de vigne	vine
	rideau de douche	shower curtain
	trafiquant de drogue	dealer

Figure 60. Exemples de traductions obtenues avec le dictionnaire Google

Les traductions obtenues à l'issue de ces deux étapes sont stockées dans notre base de résultat et sont éliminées des traductions restant à traduire. Le dictionnaire *Google* est également préalablement testé afin de collecter les traductions de co-occurrents inconnus de notre dictionnaire : aucun n'ayant été obtenu pour cette phase, nous aborderons ce sujet dans le chapitre suivant.

[1] http://translate.google.com/translate_dict?hl=fr. Le dictionnaire est à distinguer du service de Traduction Automatique et du service de recherche interlingue. Nous ne nous référons qu'à des ressources lexicales dont les résultats ne sont pas générés de façon automatique, afin de ne pas bruiter nos résultats avec des résultats générés par d'autres systèmes automatiques.

6.5 Détection du degré de polysémie

Afin d'évaluer le degré de polysémie des unités lexicales restant à traduire, nous nous appuyons sur le nombre de traductions candidates de chaque constituant recensé au sein du dictionnaire bilingue. Nous faisons l'hypothèse qu'une unité lexicale dont les traductions candidates sont nombreuses est fortement polysémique, en nous appuyant sur le principe de Dagan (1991) :

> les différents sens d'un mot sont déterminés par les différentes traductions de ce mot dans une autre langue

Pour chaque unité lexicale source, nous comptabilisons le nombre de traductions recensées pour ses constituants. Nous ne conservons pour cette phase que celles dont les constituants ne comptent qu'une traduction candidate[1]. Par exemple, *ambiance musicale* est conservée puisque *ambiance* et *musicale* ne comptent respectivement qu'une seule traduction :

ambiance > *atmosphere*

musical > *musical*

ambiance musicale > *musical atmosphere*

Les unités lexicales dont au moins l'un des constituants compte plus d'une traduction ne sont pas conservées. Par exemple, *abandon de poste* n'est pas conservé pour cette phase : *abandon* compte cinq traductions et *poste* en compte huit.

Lorsque la traduction d'au moins l'un des constituants est inconnu, l'unité lexicale n'est pas traitée dans cette étape, comme dans l'exemple :

[1] Nous nous basons sur le dictionnaire que nous utilisons, même si nous avons conscience que certaines traductions peuvent être omises.

*acide **folique***

La traduction du co-occurrent *folique* n'est pas recensée car ce terme est trop technique. La détection de traductions inconnues est traitée dans une phase à cet effet.

Les unités lexicales non polysémiques constituent 5.02% de notre échantillon. Dans cette phase, l'objectif n'est pas de sélectionner une traduction valide parmi des traductions candidates, mais de déterminer si l'unique traduction[1] est valide ou pas.

6.6 Génération de traductions candidates

Nous générons les traductions candidates, à partir de notre dictionnaire électronique. La méthode se fonde sur deux hypothèses, l'une d'ordre morpho-syntaxique, l'autre d'ordre sémantique.

Hypothèse 1 : Critère morpho-syntaxique

Nous fondons l'hypothèse selon laquelle il existe des équivalences de traductions de patrons morpho-syntaxiques régulières entre le français et l'anglais. Ces régularités se formalisent par le biais de règles de transformation syntaxique de la langue source vers la langue cible.

En ce qui concerne le patron source ***NOM1 DE NOM2***, on oppose généralement le type « roman », caractérisé par l'ordre déterminé-déterminant (*bleu foncé*, *point de vue*) du type « germanique », caractérisé par l'ordre déterminant-déterminé (*dark blue*, *viewpoint*) (Chuquet et Paillard, 1987). Ainsi, le patron syntaxique *NOM1 DE NOM2* en français peut être traduit par différentes structures en anglais selon la relation sémantique considérée entre les deux objets. Selon Tournier (1985), le type « roman » *NOM-PREP-NOM* n'est pas

[1] Pour le patron *NOM-de-(d')-NOM*, deux traductions candidates sont proposées car nous prenons en compte deux patrons syntaxiques cibles. Toutefois, il n'existe qu'un choix lexical, qui se manifeste dans deux constructions syntaxiques.

négligeable en anglais, mais le type *NOM-NOM* est dominant. (Chuquet et Paillard, 1987). Nous traitons ces deux types de transformations syntaxiques (*ibid.*) :

- le patron « *NOM2 NOM1 »* en anglais marque une relation étroite entre les deux noms. Nous posons la règle de transformation :

NOM1 DE(D') NOM2 > NOM2 NOM1

caisse de retraite > pension fund

Le processus de composition offre une grande souplesse en anglais et la juxtaposition des composants peut faire intervenir des relations syntactico-sémantiques variées (*ibid.*) :

- Sujet/Verbe ou Verbe/Sujet :

sunshine (the sun shines)

- Objet/Verbe ou Verbe/Objet :

taxi driver (X drives the car)

- Identification :

handlebar (the bar is handle)

- Intrumental :

searchlight (X searches with the light)

- Locatif :

call box (X calls from the box)

Le français exploite moins cette diversité : les composés par juxtaposition, peu nombreux, correspondent presque exclusivement soit à une relation de localisation (*coin cuisine*), soit à la relation verbe/objet (*portefeuille*). Les composés en anglais posent une difficulté d'ambiguïté structurale, quant à la portée de chacun des termes sur les autres, comme dans l'exemple (*ibid.*) :

[[modern history] section]

- la structure « *NOM1 of NOM2* » accorde la priorité à l'élément repéré (*NOM2*)[1], ce qui se traduit par la règle :

NOM1 DE(D') NOM2

fils d'homme > son of man

Ce type de structure explicite la relation entre les deux éléments, par le biais de la préposition (*ibid.*). Cette explicitation est effacée dans la structure *NOM-NOM*.

Le patron syntaxique ***NOM ADJECTIF*** est traduit par le patron *ADJECTIF NOM*, puisque l'adjectif est nécessairement antéposé en anglais :

NOM ADJECTIF > ADJECTIF NOM

appareil digital > digital camera

[1] Le cas du génitif (NOM1's NOM2) n'est pas pris en compte dans le cadre de notre étude. Nous aurions pu également tester le patron morphosyntaxique NOM2NOM1 accolés, mais nous faisons le choix de ne pas traiter ce patron, plus délicat, de par sa structure particulière d'unité lexicale simple. Une évolution ultérieure sera d'ajouter ces deux patrons morphosyntaxiques candidats. Notons que, parmi les erreurs de traduction de type morpho-syntaxique que nous analysons dans nos résultats (chapitre 9), certaines sont dues à une absence du génitif, mais aucune n'a été répertoriée pour l'absence du patron NOM2NOM1.

Hypothèse 2 : Critère sémantique

D'un point de vue sémantique, notre hypothèse est qu'un certain nombre d'unités lexicales complexes sont transparentes du point de vue de la traduction, c'est-à-dire que la somme des traductions permet d'obtenir la traduction adéquate, comme dans l'exemple :

psychologie sociale > social psychology

Si la traduction répond à cette hypothèse, notre filtre automatique la validera. En revanche, si la traduction n'est pas transparente, ou si nos ressources dictionnairiques ne sont pas adéquates (usage non recensé), elle ne sera pas validée et sera soumise à l'étape de traitement suivant.

Nous générons automatiquement toutes les traductions candidates via le *Collins Pocket* selon la méthode de Léon et Millon (2005) et Léon (2006), qui consiste à générer toutes les combinaisons possibles des traductions des éléments simples. Prenons pour exemple :

institut de psychologie

Le *Collins Pocket* donne les traductions suivantes pour les unités lexicales sources *institut* et *psychologie* (unité lexicale source vers unité lexicale cible) :

institut > institute

psychologie > psychology

Notre programme génère la combinatoire en appliquant les règles de transformation syntaxique. Deux traductions candidates sont obtenues pour le patron *NOM-de-NOM* :

institut de psychologie > institute of psychology

institut de psychologie > psychology institute

Afin d'avoir un ensemble de traductions le plus exhaustif possible, nous recensons également les « traductions inversées » des unités lexicales françaises, en recherchant ces dernières lorsqu'elles apparaissent en tant que traduction dans la version *English-French*, ce qui rajoute parfois des traductions supplémentaires, comme pour *vol* :

larceny, robbery, snatch $\rightarrow$ *vol*

6.7 Interrogation automatique du moteur de recherche Yahoo

Le moteur de recherche *Yahoo* est interrogé automatiquement à l'aide de l'interface de programmation d'applications API (*Application Programming Interface*)[1] afin de récupérer le nombre d'occurrences[2] de chaque traduction candidate. Ces fréquences seront utilisées lors de la validation[3]. Pour chaque traduction candidate, nous générons un ensemble de requêtes (voir Figure 57), en considérant les mots de la requête comme une expression exacte, via l'utilisation des guillemets. La recherche est restreinte aux pages Web de langue anglaise.

Patron syntaxique source	**Requête (en langue cible)**
NOM ADJECTIF	"the ADJ NOM" OR "a ADJ NOM"
NOM1 de NOM2	"the NOM_1 of NOM_2" OR "a NOM_1 of NOM_2"
	"the NOM_2 NOM_1" OR "a NOM_2 NOM_1"

Figure 61. Patrons des requêtes des combinaisons lexicales anglaises

[1] http://developer.yahoo.com/search/

[2] Des différences ont été remarquées entre le nombre de résultats renvoyés par l'API et par l'interface Web.

[3] Le choix d'utilisation du moteur de recherche *Yahoo* plutôt que *Google* s'explique par l'observation de résultats de fréquences de *Google* peu fiables dans le cadre de certaines configurations de requêtes (http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html).

Les combinaisons booléennes ramènent un ensemble de résultats qui prend en compte les variations dues aux changements d'article, comme dans l'exemple :

"the American journalist" OR "an American journalist".

L'utilisation d'articles dans les requêtes du patron syntaxique *NOM ADJECTIF* permet également de réduire le problème de l'ambiguïté catégorielle. Par exemple, *complete* peut être un adjectif (*entier, complet, intégral, total*) ou un verbe (*parfaire, compléter*). La collocation *complete restoration* est ambiguë. L'ajout de l'article permet d'éliminer les cas où *complete* est un verbe.

6.8 Validation automatique

Afin de réduire le bruit, un filtre simple est appliqué aux traductions restantes. Nous ne conservons que celles dont la fréquence sur le Web est au moins égale à un dix-millème des occurrences du mot cible[1]. Prenons pour exemple[2] *« messe de minuit »* et deux de ses traductions candidates « *midnight mass »* et *« mass of midnight »* :

Seuil_mass : 764 000 000 / 10 000 = 30400

La collocation « *midnight mass »* (avec une fréquence de 336 000, donc supérieure au seuil limite pour le nom cible *mass*) est retenue, tandis que *« mass of midnight »* (avec une fréquence de 65, donc inférieure au seuil limite pour le nom cible *mass*) est rejetée. Ce filtre provoque évidemment parfois des cas de silence. Notre approche favorise volontairement la précision, car il s'agit de compléter le plus automatiquement possible des ressources existantes. L'augmentation du bruit obligerait à un filtrage manuel des résultats beaucoup plus long et coûteux. Après le filtre automatique sur les fréquences, 34.83% des traductions candidates sont conservées. Cette faible quantité s'explique par le fait que deux patrons

[1] Fréquences toujours limitées aux pages Web en langue anglaise.

[2] Les résultats de fréquence présentées pour cette expérience datent de juillet 2008.

morpho-syntaxiques sont testés pour le patron *NOM-de(d')-NOM*, tandis qu'un seul ne peut être validé par unité lexicale source. Lorsque deux patrons morpho-syntaxiques sont validées pour une même unité source, nous conservons la traduction la plus fréquente :

cycle de vie > life cycle (13 100 000)

cycle de vie > cycle of life (1 070 000)

Dans cet exemple, seule la traduction *life cycle* est conservée, bien que les deux aient été validées par le filtre automatique.

6.9 Analyse des résultats

6.9.1 Proportion de traductions

Les traductions obtenues comptent pour 2.88% de nos unités lexicales à traduire. Parmi les unités lexicales non polysémiques, 57.40% d'entre elles obtiennent une traduction à cette étape. Le tableau suivant présente la proportion de traductions conservées après le filtre automatique.

	Traductions générées	***Filtre automatique*** **Filtre seuil fréquence**	
NOM ADJ	29	20	68,97%
NOM DE NOM	40	8	20,00%
NOM D' NOM	10	3	30,00%
TOTAL	**79**	**31**	**39,24%**

Figure 62. Résultats de la validation des traductions

La figure 63 donne une illustration d'unités lexicales traduites lors de cette phase :

PATRON	UNITE LEXICALE	TRADUCTION
NOM ADJECTIF	drame musical	musical drama
	grange attenante	adjoining barn
	psychologie sociale	social psychology
	transition démocratique	democratic transition
	vent favorable	favourable wind
	vie privée	private life
	village typique	typical village
	vocabulaire médical	medical vocabulary
NOM de NOM	amidon de riz	rice starch
	averse de neige	snow shower
	cycle de vie	cycle of life
	date de fabrication	date of manufacture
	escadron de cavalerie	cavalry squadron
	journaliste de télévision	television journalist
	questionnaire de santé	health questionnaire
	vinaigre de riz	rice vinegar
NOM d' NOM	beurre d'ail	garlic butter
	code d'identification	identification code
	fils d'homme	son of man

Figure 63. Exemples de traductions obtenues avec la phase 1

La figure suivante montre la proportion de traductions obtenues à ce stade de notre méthodologie, par catégorie, et la proportion de traductions restantes à traduire. 2.60% des traductions sont traduites directement par notre dictionnaire, 3.60% sont traduites par le dictionnaire *Google* et 2.88% sont traduites par la phase1 de notre méthode, basée sur les fréquences sur le Web. A ce stade de notre méthode, nous obtenons pour 9.12% de traductions des unités lexicales sources :

Collins	28	2,60%
Google	39	3,63%
Phase1 (Fréquences)	31	2,88%
TOTAL	98	9,12%

Traductions de départ	1075
Traductions restantes	977

Figure 64. Proportion de traductions obtenues

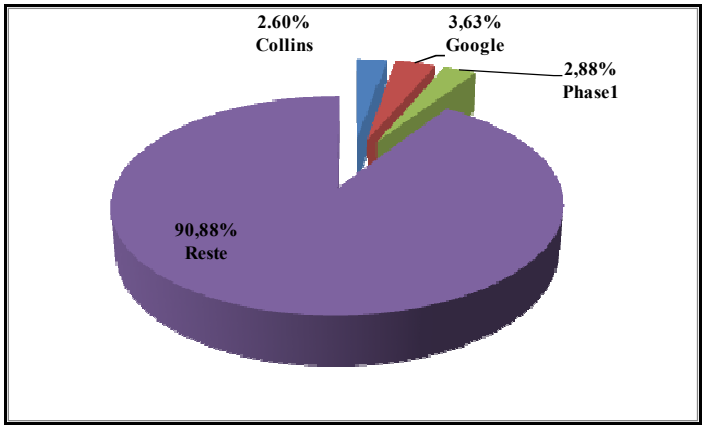


Figure 65. Répartition des étapes

6.9.2 Non validation

Dans cette section, nous analysons les causes de non validation des unités lexicales sources qui correspondent à cette étape.

Restrictions de sélection

Certaines traductions candidates n'ont pas été validées pour des raisons de restriction de sélection, comme dans l'exemple :

*bonheur perdu > * stray happiness*

Dans ce cas, l'adjectif *happiness* signifie *perdu*, mais il ne s'applique pas à des entités générales tel que le bonheur :

animal perdu > stray animal

Un autre exemple concerne l'unité lexicale *retournement de veste* :

*retournement de veste > * upturn jacket*

Dans cet exemple, le nom *upturn* s'applique à la classe des événements, mais pas à la classe des vêtements.

Les fréquences sur le Web permettent de filtrer de façon efficace les difficultés de restriction de sélection, car les co-occurrences erronées pour cette raison connaissent des faibles fréquences. Cruse (1986) fait l'hypothèse d'une directionnalité sélecteur/sélecté, dans le cas de la construction syntaxique tête/modifieur. Ces constructions concernent les syntagmes nominaux endocentriques du type :

X < électronique>

X = <APPAREIL>

Le modifieur *électronique* sélectionnerait le nom (comme dans *appareil < électronique>*, par exemple). Le modifieur est effaçable. On peut parler d'un *appareil* ou d'un *appareil électronique*. Nous nous interrogeons sur une telle directionnalité. Concernant les restrictions entre la tête et le modifieur, il nous semble que c'est plutôt la tête, et non le modifieur qui sélectionne un élément. Il paraît plus productif d'envisager la « collocation » à partir de la tête, comme nous la traitons dans notre base de données :

Appareil > X

X = <ménager, numérique, électrique, électronique, etc.>

Parc > X

X = <naturel, aquatique, botanique, marin, etc.>

Caisse > X

X = <claire, centrale, fédérale, commune, etc.>

Traductions non compositionnelles

Certaines traductions candidates ne sont pas validées car elles sont non compositionnelles, c'est-à-dire qu'il n'est pas possible de traduire une unité lexicale complexe par la somme des traductions de ces constituants. Prenons l'exemple suivant :

souris d'agneau

souris > mouse

agneau > lamb

*souris d'agneau > * lamb mouse*

Dans cet exemple, la traduction de *souris d'agneau* ne doit pas être littérale, mais doit être de la forme :

lamb shank

lamb > agneau

shank > jarret

La co-occurrence lexicale erronée, *mouse shank*, n'apparaît qu'à une fréquence de 34 sur le Web, ce qui nous permet de ne pas la valider. Il en va de même pour l'unité lexicale *chat de gouttière* :

chat > cat

gouttière > gutter

*chat de gouttière > * gutter cat*

Un *chat de gouttière* qui signifie 'un chat qui n'est pas de race' doit être traduit par *ordinary cat* (littéralement *chat ordinaire*) :

ordinary > ordinaire

cat > chat

Structure morpho-syntaxique

Certaines traductions n'ont pas été validées car leur structure morpho-syntaxique est incorrecte ou n'est pas la plus adéquate, comme par exemple :

** Actress of cinema*

** Window of cat*

Choix lexical incorrect

Une autre cause concerne un choix lexical incorrect, qu'il s'agisse de la tête sémantique, comme par exemple :

*Gestion communautaire > * communal management*

Il peut également s'agir d'un mauvais choix lexical au niveau du co-occurrent, comme par exemple :

*Lait de croissance > * growth milk*

Proportion des traductions rejetées (par catégorie)

Les deux figures suivantes présentent la proportion de traductions rejetées au cours de la phase 1, par catégorie de non-validité. Une majorité des cas concerne un choix lexical incorrect, qu'il s'agisse de la tête sémantique (30%), ou du co-occurrent (5%). Un autre type de non validité concerne la non-compositionnalité (9%) et les restrictions de sélection (7%).

Enfin, 19% des traductions non-validées constituent des fausses erreurs (silence). Il est préférable de privilégier l'élimination du bruit, même si les cas de silence sont augmentés. En effet, les cas de silence non validés au cours de la phase 1, par la méthode des fréquences pourront être validées au cours de la phase 2.

Silence	8
Tête incorrecte	13
Co-occurrent incorrect	2
Structure morpho-syntaxique	13
Non-compositionnalité	4
Restriction de sélection	3

Figure 66. Nombre de traductions rejetées (par catégorie)

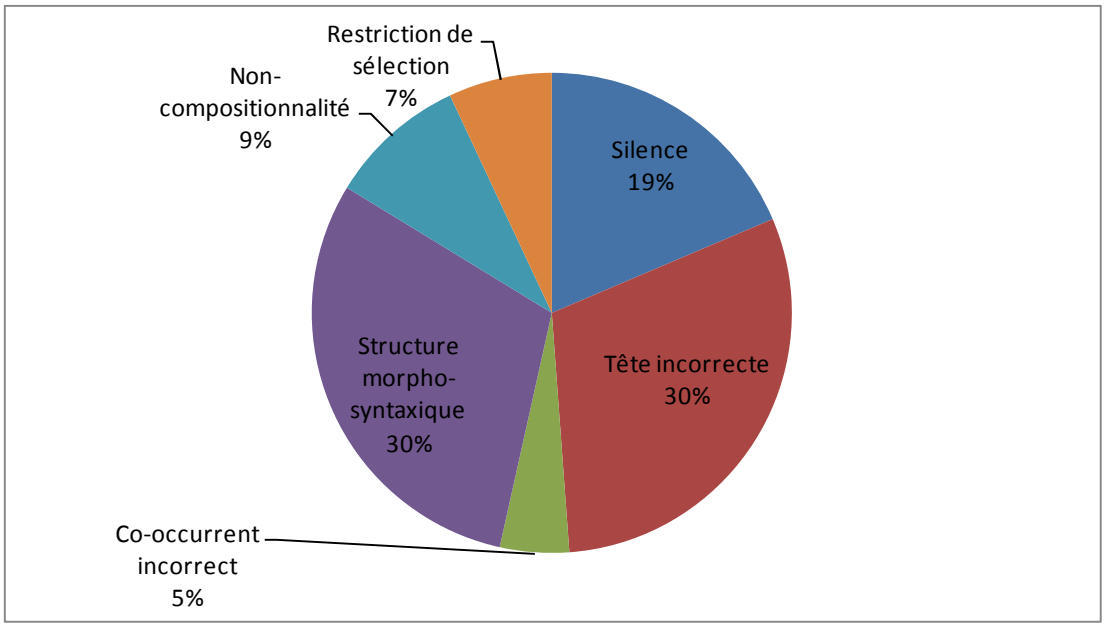


Figure 67. Proportion de traductions rejetées par catégories

Chapitre 7. Traductions compositionnelles polysémiques

7.1 Introduction

Les premières recherches en désambiguïsation lexicale ont eu lieu dans un contexte de Traduction Automatique, ce qui montre l'importance de la désambiguïsation lexicale pour la traduction (Audibert, 2003). Dès 1949, Weaver (1949) aborde dans son Memorandum la nécessité d'une phase de désambiguïsation lexicale pour la traduction par l'ordinateur : il n'est pas possible d'accéder au sens d'un mot ambigu dénué de tout contexte, tandis que l'accès au contexte (aussi bien le contexte gauche que le contexte droit) de ce mot permet d'en désambiguïser le sens. La question reste de déterminer la taille moyenne de la fenêtre de mots utile à la désambiguïsation (Audibert, 2003). Kaplan (1955) montre lors d'une expérience avec sept traducteurs que présenter deux mots à gauche et à droite du mot ambigu n'est pas plus significatif que de présenter la phrase entière (Audibert, 2003). La problématique du contexte d'un mot, dans le cadre des débuts des recherches en traduction automatique est représentative des travaux qui suivront dans le domaine. Dans le chapitre précédent, nous avons présenté une méthode de traduction, basée sur la fréquence sur le Web des traductions candidates, pour les traductions compositionnelles et non polysémiques.

Toutefois, la fréquence ne permet pas de désambiguïser systématiquement les cas de polysémie : l'entourage lexical des collocations n'est pas pris en compte. Les performances des systèmes en recherche d'informations interlingues, tout comme celles des systèmes en traduction automatique sont fortement freinées par le problème de l'ambiguïté lexicale des mots polysémiques ou homographes. Ainsi, la traduction anglaise du terme français *caisse* est différente selon que l'usage concerne, entre autres, l'INSTRUMENT *DE MUSIQUE* (*drum*), la *BANQUE* (*fund*) ou la *VALISE* (*case*). Le manque de désambiguïsation lexicale pour la traduction automatique conduit à des résultats qui gênent souvent considérablement la compréhension. Par exemple, le système de traduction automatique *Systran* traduit le terme complexe *caisse centrale* par *central case*. Pourtant, la polysémie est rendue très faible dès que l'on envisage les mots-clés selon leur co-occurrent (Yarowsky, 1993, Shütze, 1998, Véronis, 2003).

Nous présentons la deuxième phase de notre méthodologie, basée principalement sur une comparaison entre « mondes lexicaux » (ensemble de co-occurrents), à partir du Web. Cette phase est une version améliorée de Léon (2006), qui présente une méthode de comparaison des mondes lexicaux à partir du Web : Léon (2006) construit les mondes lexicaux des unités lexicales complexes sur le Web, puis, génère leurs traductions candidates via un dictionnaire bilingue électronique. Les mondes lexicaux de toutes les traductions candidates sont constitués. Enfin, les mondes lexicaux français et anglais sont comparés, afin de valider la traduction adéquate par filtres statistiques. L'évaluation de Léon (2006), sur 10 mots français très polysémiques montre que l'exploitation des mondes lexicaux des unités lexicales complexes sur le Web permet une acquisition automatique de traductions avec une excellente précision. Nous distinguons deux phénomènes de représentation du contexte d'un mot ou d'une combinaison lexicale :

- Les **dépendances syntaxiques** : il s'agit de mots qui entretiennent une relation de dépendance syntaxique avec le mot cible : sujet/prédicat, prédicat/objet, modification, etc. Ce sont les unités lexicales complexes que nous avons collectées.

- Les **co-occurrences**, à un niveau supérieur, qui sont dans le même entourage lexical mais qui n'entretiennent pas nécessairement une relation de dépendance syntaxique :

nous parlons de *mondes lexicaux*. Ces mondes lexicaux peuvent être utilisés pour lever une ambigüité lexicale. L'exploitation des mondes lexicaux a été exploité dans différents contextes applicatifs, sous des appellations diverses. Nous présentons dans un premier temps les différentes notions théoriques et applicatives relatives aux mondes lexicaux (7.1), avant de présenter notre deuxième phase de méthodologie (7.2).

7.2 Mondes lexicaux : notions théoriques et applicatives

Les mondes lexicaux ont fait l'objet de différentes études, sous des appellations et des applications diverses. Avant de présenter notre méthodologie de désambigüisation lexicale pour la traduction, nous présentons certains courants théoriques dont la problématique est proche de la nôtre : la théorie de l'*isotopie sémantique* (7.1.1), l'utilisation de *mots-clés thématiques* pour la classification thématique (7.1.2), les *vecteurs conceptuels* (7.1.3), les *signatures thématiques* (7.1.4), la *prosodie sémantique* (7.1.5) et la *cartographie lexicale* (7.1.6).

7.2.1 Isotopie sémantique et traduction

L'isotopie, en sémantique structurale, est une notion introduite par Greimas (1986), puis reprise par Rastier (1987) qui en fait un concept central de la sémantique interprétative (Tanguy, 1999). L'isotopie envisage le sens d'une unité lexicale du point de vue des relations qu'elle entretient avec les autres unités lexicales du lexique. Les signifiés des mots du lexique sont nommés des *sémèmes*, dont le sens émerge de leur structure relationnelle. Les relations sémantiques que les sémèmes entretiennent sont appelés des *sèmes* (les *sèmes génériques* sont les éléments de sens partagés par des sémèmes[1], tandis que les *sèmes spécifiques* sont ceux qui les différencient). La présence d'un thème se manifeste par une *isotopie sémantique*, c'est-à-dire la présence répétée d'un même sème entre des mots employés dans une portion de texte (Rossignol et Sébillot, 2003). La définition d'une isotopie passe donc par la définition de

[1] La classe sémantique qui réunit les sémèmes est appelée *taxème*.

classes sémantiques à laquelle on attribue des éléments (Tanguy, 1999). Par exemple[1], les co-occurrences de *soldat*, *char, général* et *offensive* révèlent la thématique de la guerre, car tous les sémèmes de ces noms sont porteurs du sème */guerre/.* Selon Rossignol et Sébillot (2003), d'un point de vue informatique, l'isotopie sémantique naît d'un ensemble limité de mots porteurs du sème indiquant la thématique recherché, qu'ils appellent « mots-clés » (7.1.2).

L'analyse des isotopies sémantiques est formalisée en TAL pour la traduction par Tanguy (1997) et Tanguy (1999). Leurs travaux portent sur un repérage de structures sémantiques générales pour la vérification automatique de traductions. Le logiciel TRACER[2] (Tanguy, 1999) est un module d'aide à la vérification de traduction basé sur la comparaison des isotopies. La méthode se base sur différentes stratégies de vérification traditionnelles (module d'alignement, lexiques bilingues, comparaison des caractéristiques de surface) dont l'un des modules, innovant, se fonde sur une comparaison des structures sémantiques du texte source et du texte cible. La méthode est fondée sur une étude de corrélation entre classes de termes mettant en évidence les principaux thèmes des textes. Voici un exemple de classes sémantiques proches en français et en anglais (Tanguy, 1999) liées au logiciel de traitement de texte. Ces classes sont identifiées par leur sème (en gras) et ont été établies par l'utilisateur sur la base de ses connaissances générales et via un thesaurus :

« Format » *(français) : page, champ, police, marge, en-tête, pieds de page, interligne, gras, italique…*

« Formatting » *(anglais) : edit, format, page, field, case, typing, heading, margin, bold, italic, font…*

« Document » *(français) : symbole, texte, ligne, paragraphe, caractère, phrase, document, lettre, mémo, rapport, graphique…*

[1] Exemple cité par Rossignol et Sébillot (2003).

[2] Le logiciel s'inscrit dans le projet *IDOL (IRS-based Document Localisation)*, plate-forme d'aide à la traduction et à la localisation pour le français, l'anglais et l'arabe.

*« **Document** » (anglais) : symbol, text, typos, document, letter, memo, character, text, page, paragraph…*

Selon Tanguy (1999), une isotopie est une notion plus large que les relations sémantiques définies dans les thésaurus telles que la synonymie. Les classes d'équivalences sémantiques sont des notions plus larges d'un point de vue sémantique, et elles ne sont pas liées aux catégories grammaticales. Contrairement à nous, une intervention humaine est effectuée afin d'assigner des noms aux différentes classes. En ce qui nous concerne, nous ne procédons pas à un repérage thématique : le contexte lexical est utilisé en tant qu'indice pour une comparaison entre le français et l'anglais.

7.2.2 Thème et mots-clés thématiques

Les travaux de Pichon et Sébillot (1999a) et de Rossignol et Sébillot (2003) s'appuient sur les aspects théoriques de la sémantique interprétative (Rastier, 1987). Une méthode de désambiguïsation lexicale communément admise consiste en l'identification du sens d'un mot par l'accès à ses co-occurrents (Yarowsky, 1993, Schütze, 1998). Dans une optique proche, une expérience de Pichon et Sébillot (1999a) montre que la connaissance des thèmes dans lesquels apparaissent les unités lexicales, ainsi que l'accès aux similarités et aux différences de voisinage permet une désambiguïsation lexicale de ces unités lexicales. La méthode est basée sur le calcul d'intersections et de différences ensemblistes entre les mots-clés constituant les contextes. Selon Pichon et Sébillot (1999a), un thème est le « sujet abordé dans les textes ou les segments de textes d'un corpus ». C'est aussi « le niveau de structuration de l'espace sémantique de généralité maximale tel que ne puisse y exister de polysémie » (Rossignol et Sébillot, 2003). L'étude des mots-clés en corpus permet d'accéder aux différents thèmes. Les mots-clés sont des « mots dont l'apparition dans un segment de texte est symptomatique de la présence d'un thème particulier » (Rossignol et Sébillot, 2003). L'expérience de Pichon et Sébillot (1999a), à partir d'un corpus constitué d'articles du journal LE MONDE DIPLOMATIQUE d'environ 7,8 millions de mots, montre qu'il est possible de collecter des séquences de mots ou blocs de contextes significatifs des différences de significations entre mots. Par exemple, l'unité lexicale *guerre* est présente dans deux thèmes différents, *TERRITOIRE* et *NEGOCIATIONS*. Les expériences de Pichon et Sébillot (1999a) permettent d'une part de différencier les divers usages d'une même unité lexicale au sein du

corpus, comme dans l'exemple de *guerre*, présent dans les deux thèmes *TERRITOIRE* et *NEGOCIATIONS*, pour lesquels les co-occurrences sont différentes :

1. pour le thème TERRITOIRE : *américain* 3, *début* 3, *israélo-arabe* 3, *nouveau* 3, *Tchétchénie* 3, *Turc* 3, *Vietnam* 3, *Washington* 3, *interminable* 4, *Irak* 4, *acquisition* 5, *Israël* 5, *jour* 5, *lendemain* 5, *régional* 6, *premier* 9, *froid* 10, *territoire* 10, *civil* 11, *Golfe* 14, *second* 14, *mondial* 17
2. pour le thème NÉGOCIATIONS : *conflit* 3, *début* 3, *Israël* 3, *long* 3, *paix* 3, *premier* 3, *année* 4, *an* 4, *étoile* 4, *Liban* 4, *vainqueur* 4, *nouveau* 5, *commercial* 7, *économique* 7, *second* 7, *lendemain* 8, *mondial* 10, *civil* 16, *froid* 22, *Golfe* 22

Figure 1. Mots-clés différentiels de guerre *pour les thèmes TERRITOIRE et NEGOCIATIONS*[1]

D'autre part, l'extraction de mots-clés thématiques permet de regrouper des unités lexicales dont le thème est proche, comme dans les exemples de *pouvoir*, *autorité* et *gouvernement*, présents sous les thèmes de *TERRITOIRE* d'une part et de *NEGOCIATIONS*, dont les co-occurrences sont fortement similaires pour chaque thème :

1. TERRITOIRE :
   (a) ***pouvoir*** : *état* 7, *local* 7, *soviétique* 7, *année* 8, *exécutif* 9, *parti* 9, *prise* 9, *public* 10, *économique* 11, *président* 11, *nouveau* 12, *place* 12, *arrivée* 17, *politique* 21, *central* 36
   (b) ***autorité*** : *Pékin* 4, *place* 4, *président* 4, *preuve* 4, *région* 4, *transfert* 4, *chinois* 5, *état* 5, *nouveau* 5, *territoire* 5, *gouvernement* 6, *politique* 6, *israélien* 13, *palestinien* 13, *local* 16
   (c) ***gouvernement*** : *fédéral* 7, *occidental* 7, *président* 8, *français* 9, *ministre* 9, *régional* 9, *union* 9, *formation* 10, *politique* 12, *nouveau* 14, *central* 15, *national* 16, *israélien* 32

Figure 2. Mots-clés similaires de pouvoir, autorité *et* gouvernement *pour le thème TERRITOIRE*[2]

[1] Exemple présenté par Pichon et Sébillot (1999a).

[2] Exemple présenté par Pichon et Sébillot (1999a).

2. NÉGOCIATIONS :

(a) ***pouvoir*** : *accession* 8, *an* 8, *armée* 8, *concentration* 8, *pays* 8, *nouveau* 9, *place* 9, *coalition* 10, *contrôle* 10, *gouvernement* 10, *arrivée* 16, *état* 17, *partage* 17, *parti* 17, *achat* 22, *central* 22, *public* 27, *économique* 28, *politique* 50

(b) ***autorité*** : *américain* 3, *frontière* 3, *local* 3, *nouveau* 3, *pays* 3, *pouvoir* 3, *problème* 3, *provisoire* 3, *armée* 4, *autonome* 4, *Cisjordanie* 5, *élu* 5, *état* 5, *gouvernement* 5, *politique* 7, *palestinien* 8

(c) ***gouvernement*** : *actuel* 11, *opposition* 11, *position* 11, *premier* 11, *sandiniste* 12, *accord* 13, *membre* 13, *national* 13, *central* 14, *chef* 14, *occidental* 14, *état* 16, *Bonn* 17, *coalition* 17, *formation* 18, *français* 25, *américain* 26, *nouveau* 26, *pays* 26, *fédéral* 27, *européen* 28, *politique* 37, *israélien* 61

Figure 3. Mots-clés similaires de pouvoir, autorité *et* gouvernement *pour le thème NEGOCIATIONS*[1]

Dans la même lignée, Rossignol et Sébillot (2003) décrivent un système de détection automatique de thèmes à partir d'un corpus non spécialisé, multithématique, reposant sur la notion de mots-clés et de découpage du corpus en sous-corpus thématiques, dans un objectif de désambiguïsation lexicale. Les résultats, obtenus à partir d'un corpus du « Monde diplomatique » montrent une précision de 85% et un rappel de 63%.

7.2.3 Latent Semantic Indexing et Vecteurs conceptuels

Dans le cadre de la représentation du sens, l'équipe Traitement Algorithmique des Langues (TAL) du LIRMM a développé un système d'analyse thématique basée sur la notion de *vecteur conceptuel* (Schwab *et al.*, 2004). Un vecteur conceptuel est la représentation d'idées associées à des segments textuels (documents, paragraphes, syntagmes, etc.). Les vecteurs ont été utilisés en informatique pour la recherche d'information (Salton, 1968) (Schwab *et al.*, 2004). En ce qui concerne la représentation du sens, leur emploi a été utilisé par le modèle LSI (*Latent Semantic Indexing*)[2] (Deerwester *et al.*, 1990). Le modèle LSI est un modèle d'indexation sémantique qui vise à établir des relations entre les documents et les termes qu'ils contiennent, par le biais de concepts. En linguistique, la notion est formalisée par

[1] Exemple présenté par Pichon et Sébillot (1999a).

[2] Analyse Sémantique Latente, en français.

Chauché (1990), dans le cadre des champs linguistiques dans un espace vectoriel. Dans Schwab *et al.* (2004), les vecteurs conceptuels en français sont construits à partir d'un ensemble de notions élémentaires collectées *a priori* dans le Larousse (1992). Lorsqu'un terme est polysémique, il combine différents vecteurs correspondant aux différents sens. Contrairement à nous, les concepts sont donnés a priori et reliés aux items textuels. En ce qui nous concerne, nous ne faisons pas appel à des ressources externes pour la construction de mondes lexicaux : ils sont construits uniquement à partir des mots-clés les plus fréquents collectés dans les données textuelles.

7.2.4 « Signatures thématiques » et « signatures pertinentes »

Une notion très proche des mots-clés thématiques est celle de « signature thématique ». Ce concept a été utilisé dans différents domaines d'application.

Résumés automatiques

SUMMARIST (Hovy et Lin, 1999, Lin et Hovy, 2000) est un système de génération automatique de résumés, qui s'appuie sur une méthode d'acquisition de signatures thématiques (*« topic signature »* en anglais) ou signatures conceptuelles (« concept signatures ») (Hovy et Lin, 1999). La tâche de résumé automatique consiste en une reformulation du texte original afin d'en décrire l'essentiel du contenu, contrairement à un extrait de textes qui consiste en des portions isolées du texte original sans reformulation. La méthode de SUMMARIST est basée sur trois étapes principales : une phase d'identification thématique, une phase d'interprétation sémantique et une phase de génération de résumés. Une signature thématique est définie par Lin et Hovy (2000) comme un vecteur de termes (unités lexicales simples ou complexes) fréquemment associés à un concept, à partir d'un corpus donné, et qui dans la tâche de résumé automatique, regroupe les occurrences des termes avec le concept. Voici un exemple de signatures thématiques en anglais associées au concept *restaurant* (*restaurant-visit* en anglais) Lin et Hovy (2000) :

table, menu, waiter, order, cat, pay, tip

Un concept très proche des « signatures thématiques » est celui de « signatures pertinentes » *(« relevancy signatures »* en anglais), concept introduit par (Riloff, 1996, Riloff et Lorenzen, 1999), développé pour une tâche de résumé automatique. La différence principale entre les « signatures thématiques » et les « signatures pertinentes » est que ces dernières nécessitent un parser (Lin et Hovy, 2000), tandis que les signatures thématiques se basent uniquement sur des calculs statistiques à partir de corpus.

Désambigüisation lexicale

Les signatures thématiques sont utilisés dans le domaine de la désambigüisation lexicale et de l'enrichissement d'ontologies. Agirre *et al.* (2000b, 2001,), Agirre et Lopez (2003) et Agirre et Lopez (2004) utilisent le Web afin d'acquérir les signatures thématiques associées aux concepts de WordNet[1], pour diverses taches de désambigüisation lexicale, d'enrichissement de la description sémantique et des liens thématiques qui relient les concepts. WordNet est un lexique disponible en ligne qui organise les unités lexicales en fonction de leur sens et de leurs relations sémantiques avec les autres unités (synonymie, antonymie, etc.). Par exemple, le nom *waiter* compte deux usages dans WorNet (Agirre *et al.*, 2000b) :

(1) waiter, server – a person whose occupation is to serve at table (as in a restaurant)

(2) waiter – a person who waits or awaits

Pour chaque usage, les signatures thématiques obtenues sont les suivantes (Agirre *et al.*, 2000b) :

waiter(1) : restaurant, menu, waitress, dinner, lunch, counter, etc.

waiter(2) : hospital, station, airport, boyfriend, girlfriend, cigarette, etc.

La méthode d'acquisition de signatures thématiques de Agirre *et al.* (2000b) passe par une acquisition de textes associés à chaque concept de WordNet à partir du Web. Les requêtes

[1] http://wordnet.princeton.edu/

sont construites à partir des informations fournies par WordNet. Voici un exemple de requête, générée pour le premier sens du nom *boy (=male child, boy, child – a younthful male person*) (Agirre *et al.*, 2000b) :

(boy AND (‘altar boy’ OR ‘ball boy OR… OR ‘male person’

AND NOT (‘man’… OR’broth of a boy’ OR #sense 2

‘son OR… OR ‘mama’s boy OR #sense 3

‘nigger’ OR… OR ‘black’) #sense 4

Les textes collectés sont classés en fonction de chaque sens des concepts. Les mots-clés sont extraits pour chaque collection et sont comparés avec ceux des autres collections. Les mots-clés qui ont une fréquence significative dans une collection par rapport aux autres constituent les signatures thématiques. Voici un extrait de signatures thématiques obtenues pour le sens 1 de *boy* (Agirre *et al.*, 2000b) :

child, Child, person, anything.com, Opportunities, Insurance, children, Girl, Person, Careguide, Spend, Wash, enriching, prizes, Scouts, Guides, Helps, Christmas, male, address, paid, age, mother…

Des affinements de construction des signatures thématiques sont apportés dans Agirre *et al.* (2001) tels que le nombre de documents extraits par site, la prise en compte des lemmes, la restriction du contexte aux phrases, et l’utilisation d’un corpus de référence pour l’aide à la validation des termes. Les travaux de Agirre *et al.* (2000b, 2001) et Santemaria *et al.* (2003) montrent que les signatures thématiques sont efficaces pour l’acquisition automatique de sens. Agirre et Lopez (2003) montrent qu’elles peuvent être utilisées pour une classification des sens des mots. Martinez et Agirre (2004) montrent qu’elles sont utiles pour une désambigüisation lexicale. Agirre *et al.* (2004) montrent qu’elles permettent de détecter la similarité entre sens. Klapaftis et Manandhar (2005) développent une méthode de désambigüisation de termes à partir du Web, dans la même lignée que Agirre *et al.* (2000b).

Chung *et al.* (2006) utilisent des signatures thématiques pour la construction d'ontologies à partir du Web. La méthode, nommée, *WebSim*, s'appuie sur deux modèles, l'un de calcul d'information mutuelle : l'hypothèse est que les co-occurrences de termes sont un indice de leur proximité sémantique ; l'autre sur l'étude de similarité entre signatures thématiques[1]. Voici un exemple de signatures thématiques obtenues par *WebSim* :

Term	Features
Oil odbase	ontolog oil web semant confer inform logic descript system odbase knowle languag gobl base proceed databas model
Agent coopis	agent system inform confer cooper univers comput coopi base paper web distribut knowledg data model servic
Agent insurance	insur agent life compani agenc state servic term licens inform financi busi brok onlin quot health nationwid auto
Classification clustering	cluster classif data class analysi method algorithm text inform distanc group list network imag fuzzi vector type similar variabl model hierarch program point document
Clustering architecture	cluster architectur manag server applic network databas servic group avail softwar replic microsoft
Linux automata	linux *automata* program softwar version *cellular* simul org game *life* file comput window java *state model* 3d

Figure 68. Signatures thématiques

7.2.5 Cartographie lexicale

L'étude des co-occurrences de mots peut être utile afin de désambiguïser les différents sens possibles d'un mot, pour la recherche d'information sur le Web (Véronis, 2003). Véronis (2003) propose un algorithme, *HyperLex*, qui permet de déterminer les différents usages d'un mot dans une base textuelle, et de représenter graphiquement les thématiques :

[1] L'auteur n'utilise pas la terminologie de « signature thématique », mais parle en anglais de « features » ou encore de « bag of words ».

> L'algorithme exploite la structure particulière des graphes de cooccurrences entre mots (mots qui apparaissent fréquemment ensemble), qui forment des "petits mondes", un type de graphe qui fait depuis quelques années l'objet de recherches intensives

Nous présentons un exemple, pour le mot *barrage* :

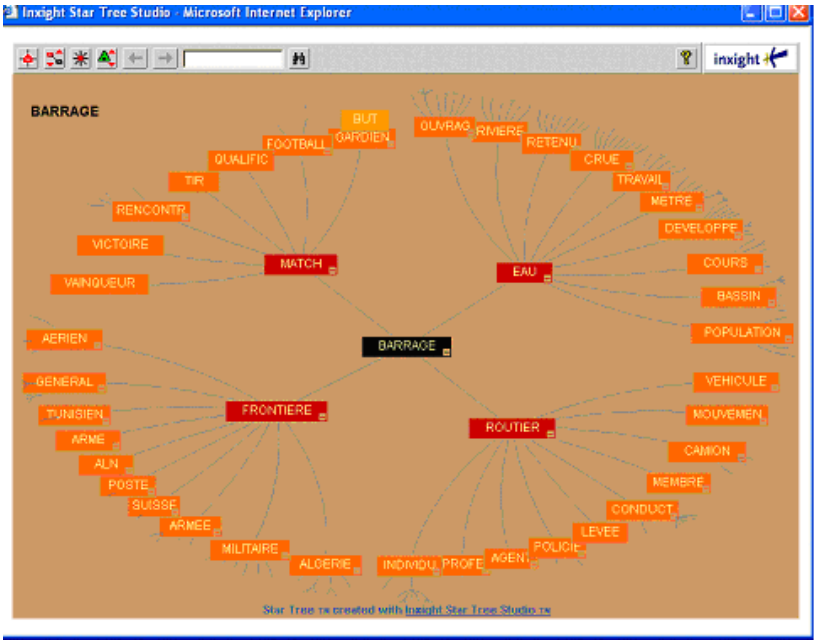


Figure 69. Cartographie « hyperlexicale » pour le mot barrage[1]

Un clic sur chacun des mots permet d'exprimer de nouvelles requêtes. Dans cet exemple, quatre mots apparaissent fortement en co-occurrence avec barrage : *match*, *eau*, *frontière* et *routier*. Ces quatre mots reflètent quatre usages différents du mot *barrage* (*ibid.*). Tous les autres co-occurrents du mot ont toutes les chances d'apparaître en contact avec l'un de ces quatre « mots-racine » (*ibid.*) :

(1) EAU, construction, ouvrage, rivière…

(2) ROUTIER, véhicule, camion, membre…

[1] http://www.up.univ-mrs.fr/~veronis/demos/index.html (Véronis, 2003).

(3) FRONTIERE, Algérie, militaire, efficacité…

(4) MATCH, vainqueur, victoire…

Dans la même lignée, Véronis présente l'outil *Nébuloscope*, qui permet de visualiser sous forme de nuage le « monde lexical » d'une requête sur le Web francophone[1], comme dans l'exemple de *barrage* :

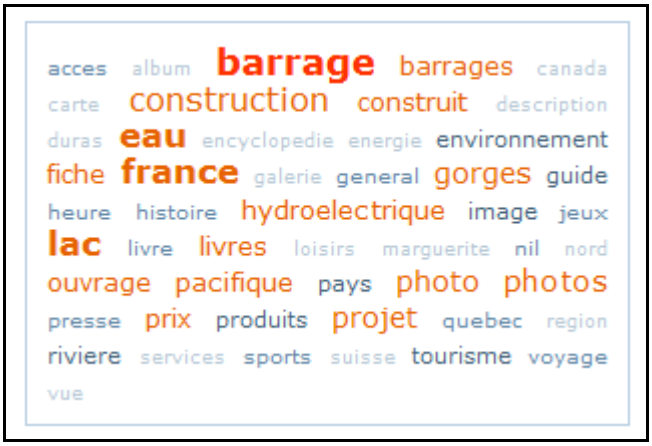


Figure 70. Monde lexical du nom « barrage »

Dans nos travaux, nous parlons de « monde lexical » afin de désigner les co-occurrences fréquentes d'un mot ou d'un terme complexe, à la suite des travaux de (Véronis, 2003). De tels voisinages, plus larges que le co-occurrent immédiat, peuvent se situer au niveau du paragraphe, ou même de la phrase. Toutefois, à la différence de Véronis (2003), ces mondes lexicaux sont exploités dans un contexte de traduction et de comparaison entre le français et l'anglais, pour la sélection de traduction candidate.

7.3 Présentation de l'approche

Dans le chapitre précédent, nous avons montré qu'une méthode de traduction basée sur la fréquence sur le Web des traductions candidates, dans la lignée de travaux tels que Grefenstette (1999) et Cao et Li (2002) est satisfaisante pour la validation d'unités lexicales

[1] http://aixtal.blogspot.com/2006/01/outil-le-nbuloscope.html

non polysémiques. En revanche, la fréquence ne permet pas de désambiguïser les cas de polysémie. Dans Léon (2006), nous montrons qu'une comparaison des mondes lexicaux permet de lever un grand nombre d'ambiguïtés lexicales. Par exemple, voici le monde lexical nominal de la requête « *appareil compact »,* retournée à partir des résumés sur *Yahoo* :

reflex, gamme, zoom, bridge, produit, qualité, canon, photographie, capteur, mode, achat, catalogue, optique, objectif, flash, écran, boîtier, téléphone, affichage, réglage, équipement, traute, amateur, mesure, offre…

Le monde lexical de sa traduction correcte, *compact camera* est très proche :

lens, quality, image, case, film, range, price, market, photography, photo, zoom, size, product, resolution, design, equipment, tripod, line, flash, body, series, shop, technology, sensor, world…

Léon (2006) présente une méthode de traduction automatique d'unités lexicales complexes, fondée sur une comparaison entre mondes lexicaux, à partir du Web. Les mondes lexicaux des unités lexicales complexes sources sont comparés avec ceux des traductions candidates, dans une optique de désambiguïsation lexicale. Une évaluation sur 10 noms français très polysémiques[1] montre que l'exploitation des mondes lexicaux sur le Web permet une acquisition automatique de traductions avec une excellente précision, de 100%. Ces mondes lexicaux peuvent à la fois constituer des ressources en tant qu'aide à la traduction, mais aussi être exploités pour une organisation de la connaissance bilingue de type ontologique. Une limite de Léon (2006) est l'absence d'analyse morpho-syntaxique pour la construction de mondes lexicaux. Notre phase de traitement s'appuie sur une version améliorée de Léon (2006), prenant en compte les aspects morpho-syntaxiques des mondes lexicaux, intégrant des filtres additionnels aux mondes lexicaux et dont les tests s'effectuent à plus grande échelle.

Nous prenons pour point de départ les 977 unités lexicales complexes sources restantes à traduire, après l'application de la phase précédente. Celles-ci comprennent les unités lexicales

[1] Le degré de polysémie a été évalué sur le nombre de traductions candidates par nom au sein du dictionnaire bilingue *Collins Pocket*.

polysémiques, ainsi que celles, non polysémiques qui n'ont pas été validées dans la section précédente.

Les traductions candidates sont générées par la même méthode que celle détaillée en section (6.6). Les co-occurrents dont aucune traduction n'a été trouvée dans notre dictionnaire sont recherchés dans le dictionnaire de *Google*, ce qui nous permet d'acquérir les traductions de co-occurrents, pour 2.66% des traductions de départ. Voici un exemple de traductions absentes de notre dictionnaire, recensées dans le dictionnaire de *Google* :

Français	**Anglais**
alimentaire	food
budgétaire	budget
départemental	departmental
diffusion	broadcast
intéractif	interactive
panoramique	panoramic
pluridisciplinaire	multidisciplinary

Figure 71. Traductions de termes simples (Dictionnaire Google)

Le schéma suivant présente la quantité de traductions candidates générées, par patron morpho-syntaxique. Les patrons *NOM DE NOM* et *NOM D'NOM* ont un nombre moyen plus élevé de traductions candidates, car deux patrons morpho-syntaxiques sont pris en compte en anglais, tandis qu'un seul n'est possible pour le patron *NOM-ADJECTIF*.

	Traductions générées	**Moyenne par unité lexicale complexe française**
NOM ADJ	5514	10
NOM DE NOM	8397	28
NOM D' NOM	4933	32
TOTAL	18844	23

Figure 72. Proportion de traductions candidates par patron morpho-syntaxique

7.4 Filtres préalables

7.4.1 « Web parallèle » ou « partiellement parallèle »

Afin de réduire la quantité de traductions candidates et d'éliminer d'emblée les plus bruitées, nous utilisons un premier filtre, celui du Web « partiellement parallèle ou « parallèle ». Nous avons montré dans le chapitre 5 que les documents multilingues, qu'ils s'agisse de documents intégralement traduits ou de traductions ponctuelles dans le corps d'un document monolingue, sont nombreux sur le Web. Notre hypothèse est qu'une traduction candidate correcte doit apparaître au moins une fois dans le même document que l'unité lexicale source. Afin de tester une éventuelle co-occurrence entre l'unité source et sa traduction candidate, nous testons les couples de traduction par le biais de requêtes, du type de :

« UNITE LEXICALE SOURCE » « TRADUCTION CANDIDATE »

" caisse centrale " " central fund "

Ce type de requête permet de cibler le contenu d'un document parallèle ou d'un document partiellement parallèle. Ce filtre ne garantit pas que le couple entre dans une relation directe de traduction. Par exemple, la requête suivante retourne 932 résultats, ce qui est élevé pour une requête « mixte » :

« caisse claire » « kick drum »

La traduction *kick drum* signifie *grosse caisse*, mais apparaît fréquemment dans des pages où *caisse claire*, sémantiquement proche, est employé, comme dans l'exemple :

D'une façon générale, le couple de traduction correct apparaît de façon plus fréquente, mais cet aspect n'est pas systématique et ne permet pas de sélectionner la traduction adéquate de façon écrasante.

Afin d'éviter au maximum les cas de silence, nous établissons un filtre de fréquence faible : les couples de traduction doivent avoir une fréquence supérieure ou égale à 1. Les fréquences des couples conservées sont classées par ordre décroissant et nous ne conservons que les trois couples les plus fréquents. Après cette étape, il reste 10,2 % des traductions candidates qui avaient été générées. Il est délicat d'évaluer la pertinence exacte d'un filtre basé sur le test du « web parallèle ». En ce qui concerne le bruit, ce filtre est un filtre préalable, et le fait que des traductions erronées soient conservées après ce filtre n'est pas problématique, puisque les filtres suivants permettront une validation plus précise. En ce qui concerne d'éventuels cas de silence, les résultats totaux que nous obtenons à la fin du processus (82,51 % de traductions obtenues) montre que le silence est peu élevé et nous conforte dans l'idée que l'utilisation du filtre basé sur le « web parallèle » nous offre l'avantage d'alléger le processus de notre méthode car il serait trop coûteux de construire un monde lexical pour toutes les traductions candidates générées au départ (nous en obtenons 18 844 avant filtres), sans pour autant que le silence ne soit élevé (seules 17,49% des unités lexicales complexes sources n'ont pas obtenu de traduction).

7.4.2 Rapport des fréquences

Un deuxième filtre est appliqué aux traductions candidates restantes, celui du rapport entre la fréquence sur le Web du terme complexe français et celui des traductions candidates. Par exemple, *« caisse de retraite »* apparaît 157000 fois. *« retirement case »* apparaît 2850 fois, tandis que « *retirement fund* » apparaît 1240000 fois. On exclut les traductions ayant une fréquence inférieure au terme français (le rapport entre le français et l'anglais est d'environ 1/20 sur *Yahoo*). Ce filtre est moins « brutal » que le précédent. A la fin de cette étape, il reste 64,56% des traductions candidates.

7.5 Construction automatique de mondes lexicaux à partir du Web

Nous constituons les mondes lexicaux sur le Web de chacune des combinaisons lexicales à l'aide de requêtes sur *Yahoo*. Puis, nous générons leurs traductions candidates via un dictionnaire bilingue électronique et constituons les mondes lexicaux de toutes les traductions candidates, après différents filtres préalables. Nous comparons enfin les mondes lexicaux français et anglais, par filtres statistiques, afin de valider la traduction adéquate.

7.5.1 Construction automatique de mondes lexicaux en français

Afin de construire les mondes lexicaux, nous utilisons les résumés descriptifs de pages Web retournés par les moteurs de recherche dans le cadre des requêtes (voir figure 71). Ces résumés constituent des paragraphes courts qui permettent de dégager rapidement les mondes lexicaux d'une requête, sans récupérer les pages Web, ce qui constituerait une méthodologie nettement plus coûteuse.

Figure 73. Exemple de résumés retournés par Yahoo pour la requête « caisse centrale »

Le moteur de recherche *Yahoo* est interrogé automatiquement par script via l'interface de programmation d'applications API[1] (*Application Programming Interface*) afin de récupérer les 1000 premiers titres et résumés renvoyés pour chaque requête des termes complexes. Ces dernières sont encadrées par des guillemets afin d'être considérées comme une expression exacte, et sont élargies à leur forme au singulier et au pluriel, en limitant les résultats à la langue française, comme dans l'exemple :

« caisse centrale » -« caisses centrales »

« caisses centrales » -« caisse centrale »

« caisse centrale » +« caisses centrales »

Les résumés sont nettoyés automatiquement par script, afin de rétablir certains problèmes de codage des caractères accentués ou de supprimer des adresses Internet, etc. Ils sont ensuite

[1] http://developer.yahoo.net/

étiquetés morpho-syntaxiquement avec le logiciel Treetagger, afin de filtrer la catégorie morpho-syntaxique des mots à extraire : nous ne conservons dans un premier temps que les noms et les adjectifs, catégories morpho-syntaxiques les plus susceptibles de faire émerger des champs thématiques. Pour chaque unité lexicale complexe, nous sélectionnons de façon automatique les cinquante noms et les cinquante adjectifs les plus fréquents parmi les résumés. Ces mots constituent leur monde lexical. Le choix de longueur du monde lexical s'est opéré par observation, pour déterminer un seuil représentatif. Un « anti-dictionnaire » est utilisé pour supprimer les mots non pertinents tels que des mots liés au Web (comme par exemple *lien*, *blog*, etc.), des verbes supports, etc. Voici pour illustration les mondes lexicaux de « appareil compact » et « appareil militaire », assorties de leurs fréquences absolues :

APPAREIL COMPACT	
NOMS	reflex (137), gamme (112), zoom (103), bridge (102), produit (101), qualité (93), canon (87), photographie (75), capteur (70), mode (69), achat (69), catalogue (50), optique (46), objectif (44), flash (44), écran (42), boîtier (40), téléphone (39), affichage (38), réglage (37)
ADJECTIFS	numérique (532), optique (84), automatique (70), reflex (56), argentique (54), léger (48), pratique (37), matériel (36), puissant (33), portable (33), technique (32), souple (28), classique (28), beau (28), professionnel (27), objectif (27), élégant (27), informatique (26), digital (26), idéal (24)

Figure 74. Mondes lexicaux de « appareil compact »

APPAREIL MILITAIRE	
NOMS	pays (123), guerre (121), avion (120), sécurité (108), membre (103), source (84), existence (77), société (61), réaction (61), vol (60), monde (46), technologie (45), esprit (44), conflit (44), libération (43), transport (42), aviation (42), supériorité (40), droit (39), intégration (38)
ADJECTIFS	civil (112), puissant (110), français (90), étranger (89), américain (78), politique (69), utilisateur (57), aérien (56), médiatique (54), national (40), majeur (40), réel (39), mauvais (39), économique (33), ancien (32), mondial (27), social (26), iranien (26), francophone (26), armé (26)

Figure 75. Mondes lexicaux de « appareil militaire»

7.5.2 Construction automatique de mondes lexicaux anglais

A partir des traductions candidates, nous interrogeons le Web pour la langue anglaise. Voici un exemple de requête pour la traduction candidate de *caisse centrale*, *central fund* :

« central fund » -« central funds »

« central funds » -« central fund »

« central fund » +« central funds »

Le monde lexical des traductions candidates est créé de la même façon que pour les résumés français, dont voici des extraits pour des traductions relatives au nom *appareil* :

COMPACT CAMERA	
NOMS	lens (141), quality (96), image (96), case (96), film (90), range (89), price (89), market (74), photography (64), photo (64), zoom (60), size (55), product (55), resoltion (53), design (51), equipment (47), tripod (46), line (44), flash (44), body (43)
ADJECTIFS	digital (733), optical (80), ultra (76), low (63), wide (53), stylish (50), photographic (50), available (50), top (43), light (39), advanced (35), underwater (33), manual (30), perfect (29), video (28), professional (28), waterproof (27), popular (27), leading (25), simple (24)

Figure 76. Monde lexical de compact camera

MILITARY PLANE	
NOMS	crash (166), aircraft (165), air (141), world (77), fighter (69), time (64), transport (60), area (60), airport (60), security (52), missile (51), fire (49), airplane (49), aviation (46), war (45), pilot (43), jet (43), governement (41), airspace (40), cargo (38)
ADJECTIFS	russian (111), civilian (77), american (73), iranian (66), commercial (48), chinese (48), german (47), strategic (40), iraqi (38), international (37), french (35), least (34), foreign (34), vintage (32), venezuelan (32), free (32), added (31), political (30), turkish (29), regular (29)

Figure 77. Monde lexical de military plane

7.6 Comparaison des mondes de mots français et anglais

Les mondes lexicaux français et anglais sont comparés, par « matching » via le dictionnaire bilingue. Pour chaque mot du nuage lexical français, nous recherchons automatiquement si l'une de ses traductions recensées dans le dictionnaire se trouve dans le nuage lexical anglais. Le nombre de mots communs entre les mondes lexicaux français et anglais est comptabilisé. Si une traduction n'est pas trouvée et si le mot français et le mot anglais sont identiques, l'information est prise en compte, ce qui nous permet de prendre en compte des Entités Nommées comme dans l'exemple, pour le couple *appareil digital/digital camera* :

canon, nikon…

Les Entités Nommées tiennent une place importante parmi les mondes lexicaux. Toutefois, il n'est pas possible de repérer leur équivalence en français et en anglais, sauf lorsque la traduction est la même.

Pour la comparaison des mondes lexicaux, nous utilisons le coefficient de Jacquard, qui mesure le degré de similitude entre deux ensembles. La formule est la suivante :

$$| inter(X,Y) | / | union (X,Y) |$$ [1]

Etant donné les ensembles de termes des mondes lexicaux français (A) et anglais (B), certains termes sont en commun et d'autres n'appartiennent qu'à l'un ou l'autre des mondes lexicaux. Le coefficient de Jacquard établit le rapport entre l'intersection des deux ensembles A et B et l'union de A et B :

- L'intersection de deux ensembles A et B est l'ensemble qui contient tous les éléments qui appartiennent à la fois à A et à B, et seulement ceux-là :

[1] Les scores sont ensuite multipliés par mille afin d'être rendus plus lisibles.

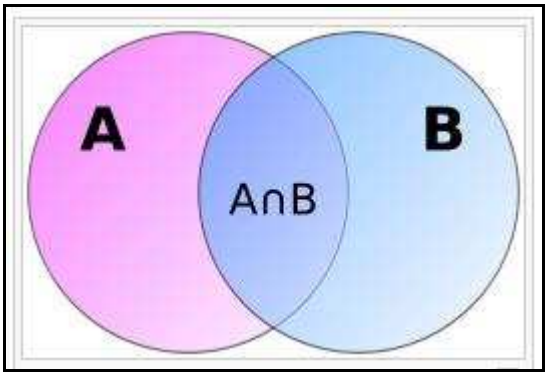


Figure 78. Intersection entre deux ensembles A et B[1]

Pour nous, l'intersection concerne les termes qui apparaissent à la fois dans le nuage lexical français et dans le nuage lexical anglais.

- L'union de deux ensembles A et B est l'ensemble qui contient tous les éléments appartenant soit à A, soit à B. Pour nous, l'union se réfère à tous les termes appartenant au nuage lexical français et à tous les termes appartenant au nuage lexical anglais.

Nos calculs sont appliqués aux mondes lexicaux contenant les noms et aux mondes lexicaux contenant les adjectifs, de façon séparée. Notre hypothèse est qu'un couple de traduction contient un nombre significatif de noms en commun et un nombre significatif d'adjectifs en commun. L'analyse distincte de ces deux ensembles catégoriels permet d'obtenir une analyse comparative plus fine. Nous appliquons plusieurs filtres aux coefficients de Jacquard :

- Le coefficient de Jacquard concernant les mondes lexicaux des noms doit être supérieur ou égal à 40.

- Le coefficient de Jacquard concernant les mondes lexicaux des adjectifs doit être supérieur ou égal à 30.

[1] http://fr.wikipedia.org/wiki/Intersection_%28math%C3%A9matiques%29

- La moyenne des coefficients de Jacquard pour les noms et pour les adjectifs doit être supérieure ou égale à 60.

Ces filtres ont été établis de façon expérimentale, par observation des données. Nous avons constaté que les couples de traduction pertinents ont un nombre significatif de noms et d'adjectifs en commun : les noms sont plus indicateurs de thème que les adjectifs et les couples pertinents contiennent plus de noms en commun que d'adjectifs. Voici un exemple de termes communs au couple *appareil compact/compact camera* :

	APPAREIL COMPACT / COMPACT CAMERA
NOMS	boîtier/case, dimension/size, équipement/equipment, flash/flash, gamme/range, marché/market, mémoire/memory, mesure/time, monde/world, objectif/lens, photographie/photography, produit/product, qualité/quality, sac/bag, série/series, taille/size, technologie/technology, zoom/zoom
ADJECTIFS	automatique/automatic, digital/digital, étanche/waterproof, faible/low, idéal/ideal, léger/light, manuel/manual, optique/optical, parfait/perfect, portable/portable, professionnel/professional, puissant/powerful, rapide/fast, rare/rare

Figure 79. Termes communs pour appareil compact/compact camera

7.7 Analyse des résultats

7.7.1 Proportion de traductions

Le tableau 80 montre la proportion de traductions après chaque filtre. A partir des 977 unités lexicales complexes de départ, 18 844 traductions candidates ont été générées. Après le premier filtre, celui du « web parallèle », il reste 1919 traductions candidates. Après celui du rapport français/anglais, il reste 1239 traductions candidates. Enfin, l'étape de validation qui attribue une unique traduction par terme source, donne 674 traductions, à savoir 69.98% des termes de départ. Cette étape est celle qui offre le plus de traductions, à savoir plus de la moitié de nos données de départ).

Unités lexicales restantes après la phase 1	**Traductions candidates générées**	***Filtre automatique***		
		Filtre Web parallèle, « top 3 »	**Filtre rapport français/ anglais**	**Filtre indice de similarité**
977	18 844	1919	1239	674

Figure 80. Etapes de validation

Les mondes lexicaux obtenus sont en eux-mêmes intéressants, et peuvent probablement être exploités comme ressource bilingue. La figure 81 donne un exemple de traductions obtenues en phase 2, pour les trois patrons morpho-syntaxiques :

PATRON	**UNITE LEXICALE**	**TRADUCTION**
NOM ADJECTIF	accident grave	serious accident
	analyse financière	financial analysis
	crampe musculaire	muscular cramp
	douleur physique	physical pain
	éclat naturel	natural shine
	fumeur invétéré	habitual smoker
	histoire courte	short tale
	immeuble résidentiel	residential building
NOM de NOM	caisse de dépôt	deposit fund
	course de karting	karting race
	cours de morale	ethics class
	disque de platine	platinum record
	football de table	table soccer
	laboratoire de recherche	research laboratory
	licence de psychologie	psychology degree
	mouvement de protestation	protest movement
NOM d' NOM	consommation d'essence	gasoline consumption
	hall d'entrée	entrance hall
	jet d'encre	ink jet
	lettre d'acceptation	acceptance letter
	manque d'amour	lack of love
	méthode d'estimation	assessment system

Figure 81. Exemples de traductions obtenues avec la phase 1

Les schémas suivants montrent la proportion de traductions obtenues, pour chaque phase de la méthodologie, ainsi que la proportion de traductions restantes à traduire :

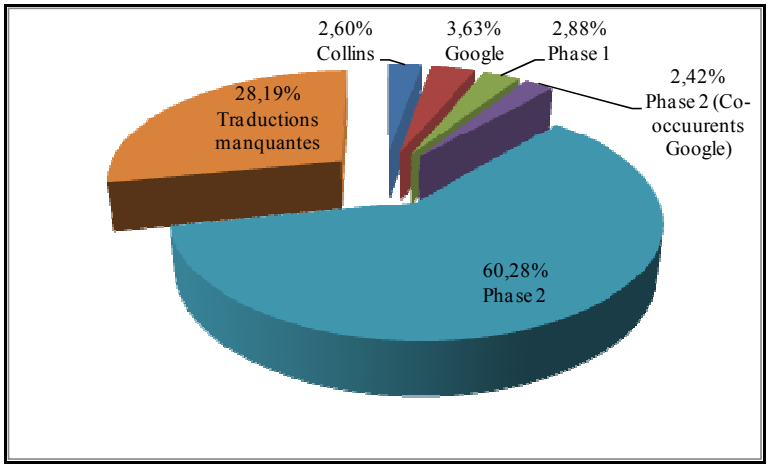


Figure 82. Proportions de traductions obtenues après la phase 2

Collins	28	2,60%
Google	39	3,63%
Phase1 (Fréquences)	31	2,88%
Phase 2 (Mondes lexicaux_Co-occurrents Google)	26	2,42%
Phase 2 (Mondes lexicaux)	648	60,28%
TOTAL	772	71,81%

Traductions de départ	1075
Traductions restantes	303

Figure 83. Quantité de traductions obtenues

7.7.2 Représentativité des mondes lexicaux

La difficulté de notre méthode de construction des mondes lexicaux concerne leur application à la langue générale, ou plutôt à la non-limitation à un domaine de spécialité donné. L'intérêt d'une telle approche est qu'elle nous permet de désambigüiser les différents usages d'un nom polysémique, en fonction de sa co-occurrence au sein d'une unité lexicale complexe, et d'obtenir des mondes lexicaux cohérents pour chaque usage. Le fait de nous limiter à un

unique domaine de spécialité amoindrirait l'apport de cette phase de méthodologie. Toutefois, une difficulté concerne les unités lexicales complexes dont le sens est général et ne s'inscrit pas dans un domaine de spécialité donné. Par exemple, le monde lexical de l'unité lexicale complexe *mois d'absence* est hétérogène, car cette unité n'est pas représentative d'un domaine de spécialité précis :

MOIS D'ABSENCE	
NOMS	membre (28), série (25), match (21), sport (19), musique (18), football (18), championnat (17), santé (16), monde (16), saison (15), foot (14), équipe (14), film (13), voyage (12), succès (12), journée (12), discussion (12), connexion (12), cinéma (12), accueil (12)
ADJECTIFS	français (19), beau (17), francophone (12), ivoirien (10), professionnel (9), jeune (9), bienvenu (9), politique (9), ancien (8), rapide (7), national (7), live (7), informatique (7), social (6), présent (6), longue (6), virtuel (5), sportif (5), public (5), prochain (5)

Figure 84. Monde lexical de « mois d'absence »

Même si ces cas sont peu nombreux, la construction des mondes lexicaux pour les unités lexicales de sens général, ou en tout cas moins spécialisé, posent la difficulté de leur homogénéité. Il est toutefois délicat d'évaluer l'homogénéité d'un monde lexical : cette notion, plutôt intuitive, est difficilement formalisable de façon automatique, à moins d'établir des comparaisons entre mondes lexicaux au sein de la langue source et de regrouper les unités lexicales complexes en français selon des grandes familles thématiques. Nous reviendrons sur cette limite, due au recouvrement de nombreux domaines de spécialité, dans les perspectives (chapitre 10).

Chapitre 8. Traductions non-compositionnelles et inconnues

8.1 Introduction

A ce stade de notre méthode, plusieurs difficultés expliquent l'absence de traduction des unités lexicales manquantes :

1) La traduction n'est pas compositionnelle, et la somme des traductions de chaque élément de l'unité lexicale complexe ne permet pas d'obtenir la traduction adéquate, comme dans l'exemple de :

caisse claire > snare drum (tambour piège)

Les ressources lexicales existantes contiennent peu d'informations sur ces phénomènes idiomatiques, recensant un nombre réduit de ces phénomènes, généralement les plus figés. Le Web (résumés, pages Web) est une ressource utile afin d'acquérir ce genre de traductions.

2) La base et/ou le co-occurrent est recensé dans notre dictionnaire, mais l'usage pertinent n'est pas répertorié, comme pour :

caisse d'épargne > savings bank

Dans cet exemple, l'usage de *caisse* (*BANQUE*) n'est pas répertorié dans notre dictionnaire. Ce phénomène est à dissocier du précédent, car ici, la traduction est compositionnelle, l'une des traductions possibles de *caisse* est *bank*, mais cette traduction n'est pas resencée.

3) La base ou le co-occurrent est un terme technique non recensé dans notre dictionnaire ni dans celui de *Google*, comme dans :

Appareil ***circulatoire***

Dans cet exemple, la traduction de *circulatoire* est inconnue de nos ressources dictionnairiques. Etant donné que la liste des têtes sémantiques a été générée via notre dictionnaire bilingue, seule l'absence de traduction des co-occurrents est concernée. Les co-occurrents dont les traductions sont absentes concernent 7.07% de notre échantillon de départ, comme dans les exemples :

vasculaire

fluorydrique

Les traductions absentes concernent généralement des termes techniques appartenant à un domaine de spécialité.

8.2 Problème de la non-c*ompositionnalité*

8.2.1 Notion de compositionnalité

La notion de compositionnalité désigne le fait que le sens global d'une unité lexicale complexe est décomposable en la somme de sens de ses constituants. Par exemple, l'unité lexicale *pluie forte* est décomposable en accumulant le sens de *pluie* et le sens de *forte*. Toutefois, il arrive que le sens des unités lexicales complexes ne soit plus décomposable,

parce qu'il forme un nouveau « tout lexical ». Cette notion rejoint la définition de *mot composé* (Grévisse, 1975):

> Un mot, quoique formé d'éléments graphiquement indépendants, est composé dès le moment où il évoque dans l'esprit, non les images distinctes répondant à chacun des mots composants, mais une image unique. Ainsi les composés *hôtel de ville*, *pomme de terre*, *arc de triomphe* éveillent chacun dans l'esprit une image unique, et non les images distinctes d'*hôtel* et de *ville*, de *pomme* et de *terre*, d'*arc* et de *triomphe*

Cette notion « d'image unique », qui est l'*unicité du référent* n'est pas systématiquement liée à la notion de compositionnalité. Ainsi, une combinaison lexicale peut désigner un référent unique, et être transparente, comme l'exemple :

barrage hydraulique

Au contraire, elle peut ne pas désigner un référent unique et être opaque, ou « partiellement » opaque, puisque le sens des constituants réunis ne sera pas la somme du sens de chaque constituant pris séparément, comme dans le célèbre exemple de *peur bleue*. Ainsi, il faut encore désigner le figement de la compositionnalité.

Du point de vue de la traduction, la notion de compositionnalité est fondamentale, car c'est elle qui détermine en partie[1] si la traduction peut être littérale. Par exemple, la combinaison littérale *barrage hydraulique* est transparente et se traduit de façon littérale par *hydraulic dam*. En revanche, la combinaison lexicale *peur bleue* ne se traduit pas de façon littérale. Bien sûr, vient se rajouter à cette notion, l'aspect idiomatique qui fait qu'une combinaison lexicale compositionnelle peut s'exprimer d'une autre manière dans une autre langue, tandis qu'une combinaison lexicale non-compositionnelle peut conserver son opacité d'une langue à l'autre. Dans ce chapitre, nous décrivons une méthode permettant d'acquérir les traductions d'unités lexicales complexes qui ne peuvent pas être traduites de façon transparente.

[1] Réserves étant mises sur l'aspect idiomatique des combinaisons lexicales, et donc de la possibilité d'une traduction non littérale malgré une transparence monolingue.

8.2.2 Présentation de la méthode

Notre dernière phase d'acquisition de traductions vise à palier deux difficultés, le problème de la non-compositionnalité d'une part, et celui de co-occurrents inconnus de nos ressources dictionnairiques, parce qu'ils sont trop techniques ou récents, d'autre part. Le point commun de ces deux difficultés est qu'une utilisation de nos ressources dictionnairiques ne peut pas être adaptée, quelle que soit la stratégie adoptée. Le principe de ce module de traduction est de collecter directement les traductions d'unités lexicales complexes, à partir du Web, et plus précisément à partir de résumés « mixtes » sur le Web, dans la lignée de travaux tels que ceux de Nagata (2001). Notre hypothèse de départ est que les requêtes en français (langue source), recherchées dans des pages en anglais (langue cible) ramènent certainement un grand nombre de pages linguistiquement « mixtes », c'est-à-dire contenant des termes dans les deux langues en question. Nous avons présenté dans le chapitre 5 les différents types de documents « partiellement » parallèles sur le Web. Nous pensons qu'une requête en langue source dont les résultats sont limités à la langue cible est propice à la collecte de ce type de document. Par exemple, voici le type de résumés retournés par la requête *caisse claire* limitée aux résultats de pages anglaises :


Jacques Delécluse: Douze Etudes Pour **Caisse Claire** at Musicroom ... - [Traduire cette page]
Sheet Music - £16.95 - For Snare drum (French, English, German and Spanish text).
www.musicroom.com/se/ID_No/0252876/details.html - 42k - En cache - Pages similaires

Translation **caisse claire** in the French-English Collins dictionary - [Traduire cette page]
caisse claire translation French - English : **caisse claire** nf (MUSIQUE) snare drum … .
French - English, Collins dictionary, synonyms, translation.
dictionary.reverso.net/french-english/caisse%20claire - Pages similaires

Figure 85. Résumés « mixtes » associés à « caisse claire »

Dans cet exemple, les résumés « mixtes » contiennent à la fois le terme source, *caisse claire* et sa traduction *snare drum*. La stratégie consiste à mettre en place des méthodes d'identification automatique des traductions candidates au sein des résumés « mixtes ». Cette tâche est délicate car il n'est pas possible d'étiqueter morpho-syntaxiquement ces résumés, d'une façon satisfaisante. Nous nous basons sur deux stratégies de collecte des traductions candidates, à partir des résumés « bruts » : dans un premier temps, nous identifions les

cognates candidats des unités lexicales sources, et dans un second temps, nous repérons les bigrammes les plus fréquents. Ces deux étapes se présentent de la même façon que les étapes précédentes, c'est-à-dire qu'elles sont successives : nous recherchons d'abord tous les cognates des unités lexicales restant à traduire. Nous passons ensuite par plusieurs filtres de validation. Les traductions non obtenues à ce stade passent alors par le module des bigrammes fréquents.

8.3 « Liens morphologiques multilingues » ou cognates

8.3.1 Acquisition de résumés mixtes

Pour chaque unité lexicale complexe restant à traduire, nous récoltons les résumés sur le Web qui leur sont associés, en limitant la recherche à la seule langue anglaise. Le fait de limiter à la langue anglaise des requêtes françaises permet de collecter des textes « mixtes » écrits principalement en anglais et contenant des syntagmes français de façon ponctuelle dans le corps du document (précisés le plus souvent en tant que traductions). A partir des résumés « mixtes » collectés, nous collectons tous les bigrammes. Etant donné que le texte contient plusieurs langues, il est délicat de procéder à un étiquetage morpho-syntaxique. Nous conservons les textes bruts et récoltons les bigrammes les plus fréquents de ces textes. Nous nous centrons volontairement sur les bigrammes candidats, et ne prenons pas en compte les trigrammes, ce qui peut provoquer des cas de silence pour le repérage du patron *NOM-of-NOM*. Nous ne traitons pas non plus le cas où une unité lexicale complexe source se traduirait par une unité lexicale simple en anglais. Toutefois, l'analyse de textes non étiquetés est une tâche délicate et nous faisons le choix de nous centrer sur le patron morpho-syntaxique candidat le plus fréquent. Un anti-dictionnaire est également utilisé. A partir des 303 unités lexicales sources restantes à traduire, 327 815 bigrammes différents sont générés.

Dans un premier temps, nous nous centrons sur le repérage de *cognates*, c'est-à-dire de « d'occurrences qui sont identiques ou se ressemblent graphiquement » (Véronis, 2000a). Il

peut s'agir, par exemple de mots graphiquement apparentés tels que *langue* et *language* (*ibid.*).

Nous nous appuyons sur l'hypothèse de Cartoni (2003) :

> Des langues morphologiquement proches possèdent des régularités morphologiques exploitables

Cartoni (2003) parle également de « lien morphologique multilingue » afin de désigner :

> Dans un cadre multilingue, nous décrivons le lien morphologique comme l'inférence d'une langue par rapport à une autre, existant grâce à un lien intuitif entre deux mots de deux langues proches historiquement

Les cognates peuvent être de deux types :

- Régularités de formes identiques : l'unité lexicale source et l'unité lexicale cible sont identiques :

*salle de **chat** > **chat** room*

*extrait de **code** > **code** snippet*

- Régularités de formes de bases communes : seule la racine des termes est identique :

*astrologie **védique** > **vedic** astrology*

Nous comparons les quatre premières lettres du co-occurrent anglais (premier élément du bigramme) avec celui du co-occurrent français (deuxième élément), comme dans l'exemple de :

*Appareil **circu**latoire > **circu**latory system*

Nous nous appuyons sur les travaux de Simard (1992), qui propose de considérer comme cognates des mots qui ont les mêmes quatre lettres initiales. Cette longueur peut parfois provoquer des cas de silence comme dans l'exemple (Veronis, 2000a) :

gouvernement > *government*

Toutefois, le choix d'une longueur de quatre lettres communes est un compromis afin de limiter des résultats bruités (qui ne sont pas des cognates), susceptibles d'être plus élevés avec un nombre plus réduit de lettres, tout en limitant au maximum le silence. Par exemple, établir une comparaison à partir de cinq lettres aurait provoqué des cas de silence, notamment pour le repérage de termes courts, comme dans l'exemple :

agneau pascal > ***pasc****hal lamb*

Voici un exemple des cinq cognates candidats les plus fréquents pour l'unité lexicale *acide fluorhydrique* :

fluoridrico pharmacy

fluorhydrique theretical

fluo publication

fluoric acid

fluoride acide

A ce stade du traitement, les résultats présentent du bruit, tels que des termes français ou des erreurs de rattachement comme dans :

fluorhydrique theretical

Toutefois, d'autres filtres vont être utilisés. Nous obtenons 8116 traductions avec cognates candidates. Parmi les bigrammes obtenus, nous conservons les 50 bigrammes les plus

fréquents pour chaque collocation source, ce qui nous fait 5178 traductions candidates. La proportion de bigrammes candidats par unité lexicale française conservée à cette étape est de 17.

8.3.2 Filtres des cognates candidats

Les traductions candidates restantes sont testées par le biais de requêtes en couple sur le Web, de la même façon que dans la phase précédente, comme dans l'exemple :

"acide fluorhydrique" " fluoridrico pharmacy"

"acide fluorhydrique" "fluorhydrique theretical"

"acide fluorhydrique" "fluo publication"

"acide fluorhydrique" "fluoric acid"

"acide fluorhydrique"" fluoride acide"

Nous obtenons 2210 traductions candidates restantes après ce filtre. Contrairement à la méthode précédente, nous conservons les dix couples les plus fréquents. Les traductions candidates générées à cette étape sont plus bruitées que celles de la phase précédente qui étaient générées directement via le dictionnaire. Nous conservons plus de traductions afin de palier des cas de silence. Il nous reste 1287 traductions candidates.

Nous utilisons ensuite le filtre du rapport entre les fréquences françaises et anglaises, comme dans la phase précédente. Cette étape nous permet de filtrer un grand nombre de traductions candidates bruitées, comme dans les exemples :

"pression osmotique" (47500) "osmotic pressure" (758000) → Validé

"pression osmotique" (47500) "osmotique figure" (15) → Non validé

"pression osmotique" (47500) "osmotique physique" (7) → Non validé

Il nous reste, après tous les filtres, 292 traductions candidates. Les traductions candidates restantes sont ensuite testées par une comparaison des mondes lexicaux français et anglais, par la même méthode que pour la phase 2, présentée dans le chapitre 7. Les mêmes filtres sont appliqués :

- Le coefficient de Jacquard concernant les mondes lexicaux des noms doit être supérieur ou égal à 40.

- Le coefficient de Jacquard concernant les mondes lexicaux des adjectifs doit être supérieur ou égal à 30.

- La moyenne des coefficients de Jacquard pour les noms et pour les adjectifs doit être supérieure ou égale à 60.

Voici des exemples de mondes lexicaux en commun pour les couples *accident vasculaire / vascular desease* et *parc thématique / theme park* :

ACCIDENT VASCULAIRE / VASCULAR DISEASE	
NOMS	artère/artery, attaque/stroke, cerveau/brain, cœur/heart, décès/death, diabète/diabetes, diagnostic/diagnosis, étude/study, hypertension/hypertension, mort/death, patient/patient, prévention/prevention, risque/risk, santé/health, soin/care, soin/treatment, traitement/traitment
ADJECTIFS	cardiaque/cardiac, chronique/chronic, majeur/major, médical/medical, patient/patient, précoce/early

Figure 86. Mondes lexicaux communs de accident vasculaire/vascular disease

PARC THEMATIQUE / THEME PARK	
NOMS	attraction/attraction, aventure/adventure, billet/ticket, eau/water, entrée/admission, famille/family, film/movie, golf/golf, hôtel/hotel, industrie/industry, monde/world, vacance/vacation, visiste/visit, voyage/trip
ADJECTIFS	animal/animal, célèbre/famous, excitant/exciting, historique/historical, national/national, populaire/popular, professionnel/professional, régional/regional, spécial/special

Figure 87. Mondes lexicaux communs de parc thématique/theme park

A l'issue de cette étape, 89 traductions sont validées, soit 29.37% des unités lexicales de départ pour cette phase, et 8.27% de la totalité de nos données de départ.

Voici un exemple de traductions obtenues par la méthode des cognates :

PATRON	UNITE LEXICALE	TRADUCTION
NOM ADJECTIF	accident vasculaire	vascular disease
	acide aminé	amino acid
	acide nucléique	nucleic acid
	acteur économique	economic actor
	affection neurologique	neurological disease
	alimentation modulaire	modular power
	ambiance thermique	thermal comfort
	anneau gastrique	gastric band
NOM de NOM	cabinet de conseil	consulting group
	cabinet de toilette	toilet water
	casque de protection	protective helmet
	chef de produit	product manager
	étui de protection	protective cover
NOM d' NOM	agent d'exécution	execution platform
	bourse d'excellence	excellence scholarships
	code d'activation	activation code
	musique d'ambiance	ambient music

Figure 88. Traductions obtenues par la méthode des cognates

Le schéma suivant récapitule les étapes de filtres pour la méthode des cognates :

Unités lexicales restantes après la phase 2	Traductions candidates générées	*Filtre automatique*		
		Filtre Web parallèle, « top 10 »	Filtre rapport français/ anglais	Filtre indice de similarité
303	327 815	1287	292	89

8.4 Bigrammes fréquents candidats

Pour chaque unité lexicale complexe restante à traduire, nous collectons les bigrammes les plus fréquents contenus dans les résumés mixtes. Sont exclus de la liste les cognates déjà testés. Notre point de départ est 201 256 bigrammes candidats. Nous conservons les 20 bigrammes les plus fréquents, comme dans l'exemple :

souris d'agneau "lamb shank"
souris d'agneau "geneve pays"
souris d'agneau "detail produit"
souris d'agneau "lamb shanks"
souris d'agneau "weekly letter"
souris d'agneau "anglais discussion"
souris d'agneau "zucchini recipe"
souris d'agneau "weather forecast"
souris d'agneau "username password"
souris d'agneau "train station"
souris d'agneau "touquet restaurant"
souris d'agneau "themes developed"
souris d'agneau "team keep"
souris d'agneau "tapestry founded"
souris d'agneau "station restaurant"
souris d'agneau "soupe fruits"
souris d'agneau "siran chocolate"
souris d'agneau "several themes"
souris d'agneau "scones biscuits"
souris d'agneau "salted nuts"

Figure 89. Bigrammes candidats pour souris d'agneau

Après ce filtre, nous obtenons 4275 bigrammes candidats.

De la même façon que précédemment, les 20 bigrammes pour chaque unité lexicale source sont testés par le biais du Web parallèle et par un filtre du calcul des fréquences françaises et anglaises. Il nous reste 2424 traductions après le filtre du Web parallèle. Nous ne conservons que les 3 couples les plus fréquents, ce qui nous laisse 637 bigrammes.

Les résumés anglais des traductions candidates restantes sont collectés et leurs mondes lexicaux sont comparés avec les mondes lexicaux sources, comme décrits précédemment. Le filtre de Jacquard est toutefois beaucoup plus puissant que dans la méthode des cognates, car les traductions sont susceptibles d'être davantage bruitées (l'accès aux ressources dictionnairiques dans un premier temps, et le repérage des cognates dans un second temps constituaient des indices plus « fiables » que les simples bigrammes) :

- Le coefficient de Jacquard concernant les mondes lexicaux des noms doit être supérieur ou égal à 110.

- Le coefficient de Jacquard concernant les mondes lexicaux des adjectifs doit être supérieur ou égal à 100.

- La moyenne des coefficients de Jacquard pour les noms et pour les adjectifs doit être supérieure ou égale à 130.

Voici deux exemples de mondes lexicaux obtenus pour l'unité lexicale *souris d'agneau* et sa traduction *lamb shank* :

SOURIS D'AGNEAU	
NOMS	restaurant (257), recette (166), cuisine (162), salade (89), plat (66), carte (60), vin (43), légume (43), canard (43), foie (42), tomate (34) chef (34), entrée (32), saumon (31), sauce (31), cœur (30), table (29), huile (28), gigot (28), filet (27)
ADJECTIFS	gras (48), confit (34), vert (30), gastronomique (23), beau (23), frais (21), traditionnel (20), blanc (19), fumé (16), ancien (15), provençal (14), gourmand (12), chaleureux (12), français (11), sec (10), original (10), rôti (9), parisien (9), chaud (9), bienvenu (9)

Figure 90. Monde lexical de « souris d'agneau »

LAMB SHANK	
NOMS	recipe (250), oil (203), salt (202), pepper (192), sauce (187), garlic (157), wine (153), meat (151), flour (117), leg (114), onion (109), tender (106), season (104), dish (104), beef (103), food (93), cup (98), tomato (86), bone (85), restaurant (82)
ADJECTIFS	olive (112), fresh (102), red (93), slow (81), brown (79), delicious (65), whole (51), white (50), grilled (43), dry (43), moroccan (42), black (40), seasoned (39), meaty (36), french (35), greek (34), special (33), rich (33), boneless (33), top (32)

Figure 91. Monde lexical de « lamb shank »

Les mondes lexicaux en commun sont les suivants :

SOURIS D'AGNEAU / LAMB SHANK	
NOMS	ail/galic, carte/menu, chef/chef, cuisine/cooking, cuisine/food, huile/oil, légume/vegetable, oignon/onion, plat/dish, recette/recipe, restaurant/restaurant, sauce/sauce, soupe/soup, tomate/tomato, viande/meat, vin/wine
ADJECTIFS	blanc/white, chaleureux/warm, chaud/hearty, chaud/warm, classique/classic, doux/sweet, frais/fresh, gras/fat, parfait/perfect, particulier/special, riche/rich, sec/cold, sec/dry, spécial/special, tiède/warm, traditionnel/traditional, vert/green

A l'issue de cette étape, 26 traductions sont validées. Nous avons volontairement instauré des filtres plus puissants car cette étape génère davantage de traductions bruitées. Voici un exemple de traductions obtenues par la méthode des bigrammes fréquents:

PATRON	UNITE LEXICALE	TRADUCTION
NOM ADJECTIF	antenne filaire	wire antenna
	appel vocal	voice call
	applique murale	wall lamp
	aurore boréale	northern light
	caisse autonome	social security
	effet indésirable	side effect
	esprit impur	unclean spirit
	plaque signalétique	identification plate
NOM de NOM	brique de lait	milk carton
	facteur de charge	load factor
	industrie de transformation	processing industry
	suprême de volaille	chicken breasts
NOM d' NOM	bloc d'alimentation	power supply
	oxyde d'azote	nitrogen oxide
	souris d'agneau	lamb shank

Figure 92. Traductions obtenues avec la méthode des bigrammes fréquents

Le schéma suivant récapitule les étapes de filtres pour la méthode des bigrammes fréquents :

Unités lexicales restantes après la phase 3 (Cognats)	**Traductions candidates générées**	***Filtre automatique***		
		Filtre Web parallèle, « top 3 »	**Filtre rapport français/ anglais**	**Filtre indice de similarité**
214	201 256	637		26

Figure 93. Etapes de traitement de la méthode des bigrammes fréquents

8.5 Analyse des résultats

8.5.1 Typologie bilingue des unités lexicales complexes

Nous adaptons la typologie monolingue de Tutin et Grossmann (2003) aux cas de la traduction des unités lexicales complexes et proposons une typologie des unités lexicales complexes, d'un point de vue bilingue.

Traductions opaques

Les collocations opaques contiennent des collocatifs imprédictibles sémantiquement, tandis que la base conserve son sens habituel. Dans le cadre de la traduction, plusieurs cas d'altération sémantique sont possibles :

- **Sens altéré de la tête sémantique :** il arrive que le co-occurrent conserve son sens habituel, mais que la tête sémantique soit altérée dans un contexte lexical précis, comme dans l'exemple de *souris d'agneau*, ou *souris* ne peut pas être traduit de façon littérale par *mouse*.

- **Sens altéré du co-occurrent :** la base conserve une des traductions habituelles, mais le co-occurrent n'est pas traduit de façon littérale comme dans l'exemple suivant :

caisse noire > *secret funds*

Le sens de l'adjectif *noir*, combiné à la tête sémantique *caisse* ne désigne pas la couleur, mais a le sens de 'secret'.

- **Sens altéré des deux constituants :** il arrive que le sens des deux constituants soit altéré, comme dans l'exemple :

clé des champs > *free rein*

Traductions transparentes

Les traductions transparentes comportent des collocatifs aisément interprétables, bien qu'étant imprédictibles d'un point de vue lexical, comme dans l'exemple :

pluie forte > heavy rain

Bien que l'unité lexicale *heavy rain* soit interprétable, il n'est pas possible d'accéder à sa traduction de façon littérale, en traduisant *fort*.

Traductions régulières

Les traductions régulières sont des combinaisons dans lesquelles le sens global est déductible et prévisible, et la somme des traductions des constituants est satisfaisante :

allocation familiale > family allowance

Dans le cas des traductions régulières, la difficulté de l'ambigüité lexicale des constituants reste toutefois présente.

8.5.2 Proportions de traductions

Les deux figures suivantes présentent la proportion de traductions obtenues de façon générale, détaillées par étapes. Nous obtenons 82,51% de traductions. La phase qui ramène la plus grande quantité de traductions est la phase 2, basée sur une comparaison des mondes lexicaux. En effet, une majorité des unités lexicales sont polysémiques. Le problème des traductions non transparentes ou inconnues concerne 10,7% des cas traduits (dont 8,28% sont des cognates).

Collins	28	2,60%
Google	39	3,63%
Phase1 (Fréquences)	31	2,88%
Phase 2 (Mondes lexicaux_Co-occurrents Google)	26	2,42%
Phase 2 (Mondes lexicaux)	648	60,28%
Phase 3 (Cognats)	89	8,28%
Phase 3 (Bigrammes fréquents)	26	2,42%
TOTAL	887	82,51%

Traductions de départ	1075
Traductions restantes	188

Figure 94. Proportion de traductions obtenues pour chaque étape

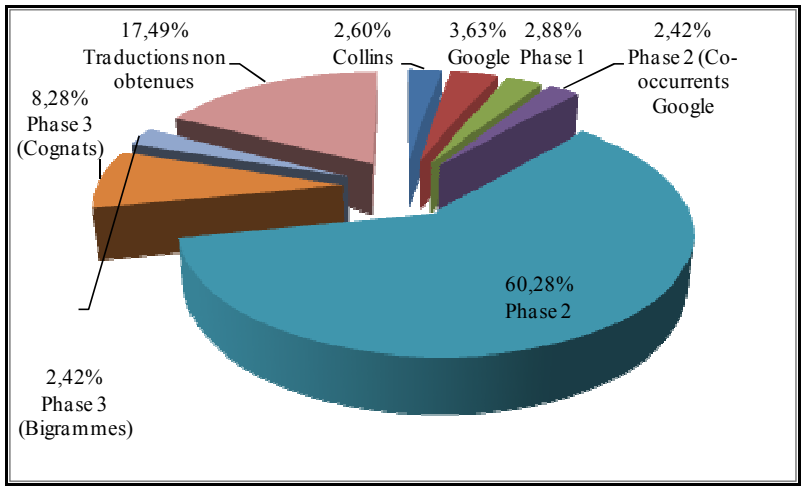


Figure 95. Quantité de traductions obtenues

Chapitre 9. Evaluation

9.1 Evaluation

Au sein de notre échantillon aléatoire, nous évaluons les 887 traductions obtenues. Nous avons opté pour une évaluation manuelle, effectuée par un locuteur bilingue[1]. Nous aurions pu comparer nos résultats avec des systèmes de Traduction Automatique tels que *Systran* ou *Google*, mais nous faisons le choix d'une évaluation humaine, plus fiable, afin de juger efficacement de nos ressources. Nous pourrons envisager lors de futures évolutions une analyse quantitative comparée de nos résultats avec d'autres systèmes, mais dans un premier temps, notre objectif est de faire évaluer à un locuteur bilingue la qualité des ressources collectées automatiquement.

Pour chaque traduction obtenue, stockée dans un document de type *Excel*, nous utilisons le moteur de recherche *Exalead*[2] afin de proposer une illustration contextuelle à l'évaluatrice.
Le travail de l'évaluatrice a consisté à évaluer la qualité de traduction des unités lexicales complexes, dans un sens unique de traduction, du français vers l'anglais. L'évaluatrice a eu le

[1] L'évaluatrice est Amanda Grey, traductrice professionnelle (http://www.amandagrey.com/).

[2] http://www.exalead.fr

choix entre trois appréciations de traduction, à préciser pour chaque unité lexicale complexe traduite :

- A : Bonne traduction ;
- B : Traduction acceptable ;
- C : Mauvaise traduction.

Les liens vers des requêtes en français puis en anglais vers le moteur de recherche *Exalead* sont précisés afin de résoudre d'éventuels cas d'ambigüités et d'offrir un contexte d'emploi lexical des unités sources et cibles. Toutefois, il ne s'agit pas d'évaluer la qualité des résultats retournés par le moteur de recherche, mais uniquement les traductions que nous présentons. Les liens hypertextes constituent une aide pour l'évaluation mais ne sont pas à évaluer.

Voici un échantillon des vingt premières traductions qui ont été évaluées :

	Français	English	Recherche	Evaluation
1	absence temporaire	temporary absence	Recherche	A
2	accès libre	free access	Recherche	A
3	accident grave	serious accident	Recherche	A
4	accident vasculaire	vascular disease	Recherche	C
5	accord de contribution	contribution agreement	Recherche	A
6	accord global	overall understanding	Recherche	C
7	accord mutuel	mutual understanding	Recherche	C
8	achat immédiat	immediate purchase	Recherche	A
9	acide aminé	amino acid	Recherche	A
10	acide nucléique	nucleic acid	Recherche	A
11	acier inox	stainless steel	Recherche	A
12	acte législatif	legislative proceedings	Recherche	B
13	acte de résistance	act of resistance	Recherche	A
14	acteur économique	economic actor	Recherche	A
15	acte de vente	bill of sale	Recherche	A
16	action commune	joint action	Recherche	A
17	action directe	direct action	Recherche	A
18	action internationale	international action	Recherche	A
19	action nouvelle	new share	Recherche	C
20	action stratégique	strategic action	Recherche	A

Figure 96. Extrait des évaluations

Chaque lien hypertexte, intitulé « Recherche », ouvre une fenêtre divisée en deux parties : d'une part la recherche du terme complexe source, limitée aux pages de langue française,

d'autre part, la recherche de la traduction, limitée aux pages de langue anglaise. Voici un exemple de fenêtre pour le couple *absence temporaire/temporary absence* :

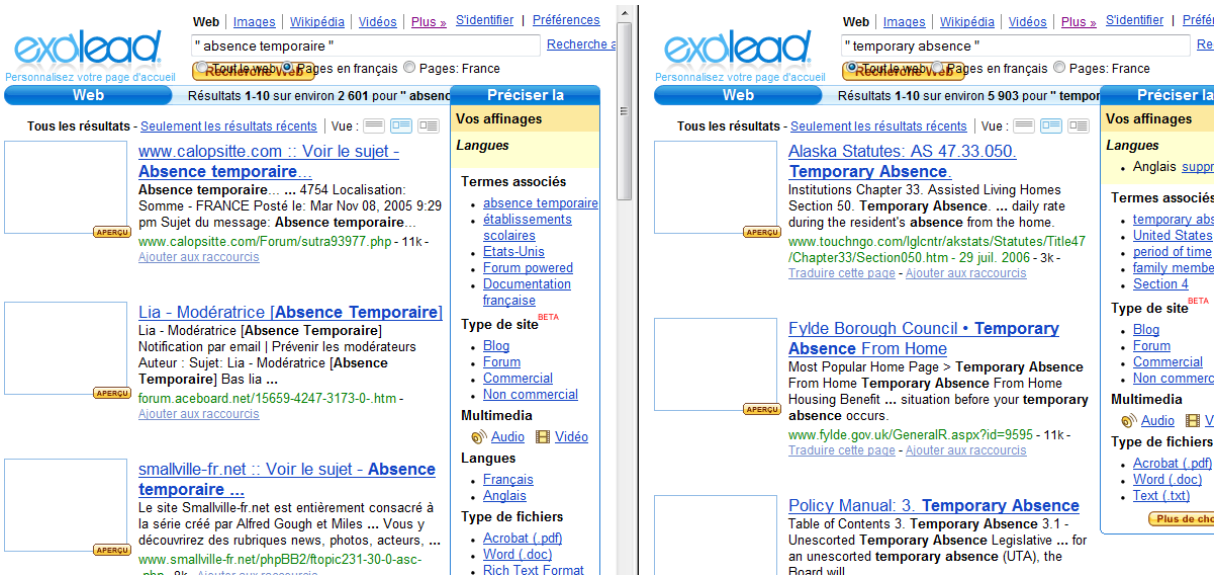


Figure 97. Recherches sur Exalead

La colonne « Evaluation » a été complétée par l'évaluatrice, pour les 887 traductions présentées. Les résultats obtenus montrent que 89,29% des traductions ont été considérées comme correctes par l'évaluatrice (catégorie A) et 5,07% ont été considérées comme acceptables (B), soit en tout 94,36% de traductions jugées comme étant non erronées. Seulement 5,64% de traductions ont été jugées erronées, comme l'illustrent les tableaux suivants :

Catégories d'évaluation	Nombre de traductions	Pourcentage
A	792	89,29%
B	45	5,07%
C	50	5,64%

Figure 98. Proportion de traductions pour les trois catégories

Catégories d'évaluation	Nombre de traductions	Pourcentage
Acceptable	837	94,36%
Non acceptable	50	5,64%

Figure 99. Proportion de traductions acceptables

Ces résultats sont particulièrement satisfaisants puisqu'ils montrent que plus de 94% des résultats sont directement exploitables, sans aucune intervention humaine. Parmi elles, 89% sont totalement satisfaisantes et seulement 5% sont acceptables sans être la traduction la plus satisfaisante.

9.2 Analyse des erreurs

Les erreurs que nous analysons ici concernent exclusivement celles qui ont été relevées via la non acceptation de l'évaluation de l'expert, à savoir le bruit, et non pas celles qui ne sont pas valides (et qui feraient partie des 17,49%).

Nous recensons trois grandes catégories d'erreurs. D'une part, les erreurs lexicales qui consistent en un choix lexical erroné (choix lexical proche mais non équivalent exact ou choix lexical erroné). D'autre part, les erreurs morpho-syntaxiques qui consistent en un choix de structure morpho-syntaxique erroné. Enfin, les erreurs « idiomatiques », c'est-à-dire dont le choix lexical est interprétable, mais non acceptable d'un point de vue « collocationnel ». Nous analysons les types d'erreurs, catégories B (acceptable) et C (erroné) confondues, mais nous signalons entre parenthèse le type de catégorie.

9.2.1 Erreurs lexicales

Les erreurs lexicales désignent un mauvais choix lexical d'au moins un des constituants de l'unité lexicale complexe. Parmi les erreurs lexicales, nous distinguons celles dont le choix lexical est proche d'un point de vue thématique mais non équivalent (il s'agit parfois de la tête

sémantique, parfois du co-occurrent ou de la totalité des éléments), de celles dont le choix lexical est totalement erroné, c'est-à-dire que la désambigüisation lexicale n'a pas été correctement effectuée (il s'agit systématiquement d'un mauvais choix de la tête sémantique).

Choix lexical thématiquement proche (tête sémantique erronée)

Parmi les erreurs lexicales, certaines consistent en un choix lexical de la tête sémantique non équivalent à l'unité lexicale complexe source, mais dont le sens reste proche ou du moins dans la même thématique, comme par exemple :

***Villa** provençale > provencal **style** (B)*[1]

***Expression** orale > oral **communication** (B)*

Dans ces exemples, le sens de la traduction (*style provençal*, *communication orale*) reste proche de l'unité lexicale complexe source (*villa provençale*, *expression orale*), mais ne constitue pas un équivalent exact. Le sens global n'est pas totalement altéré mais l'équivalence n'est pas stricte. Pour nous, le sens d'un choix lexical thématiquement proche est proche de celui de la traduction attendue, contrairement à un choix lexical totalement erroné, dont le sens n'a aucune proximité. Bien que cette répartition entre choix lexical proche et choix lexical erroné ait été établie via une appréciation humaine, de façon manuelle, nous pensons qu'elle est importante, car ces deux types d'erreurs ne doivent pas être mis au même niveau. Lorsque le choix lexical de la traduction est thématiquement proche, un changement paradigmatique entre la traduction attendue et la traduction obtenue n'entrave pas la compréhension globale d'un texte : seules des nuances précises sémantiques sont altérées, comme par exemple, un changement paradigmatique entre *oral expression* et *oral communication*. A l'inverse, lorsque le choix lexical est erroné, comme dans l'exemple :

***Formation** juridique > legal **background** (C)*

La compréhension globale d'un texte comportant la traduction erronée serait entravée.

[1] Ici, on attendrait plutôt une combinaison de trois mots-plein telle que *provencal style villa*.

Voici d'autres exemples de traductions dont le sens de la tête sémantique est quelque peu altéré, sans que cela ne nuise à la compréhension générale de l'unité lexicale complexe :

***Apprentissage** cognitif > cognitive **strategy** (B)*

***Voie** de développement > developing **country** (B)*

***Suprême** de volaille > chicken **breasts** (B)*

***Association** locale > local **society** (B)*

***Extrait** de code > code **snippet** (B)*

***Boulevard** industriel > industry **buisness** (B)*

***Agent** d'exécution > execution **platform** (B)*

***Prestation** supplémentaire > supplementary **benefit** (B)*

***Note** d'application > **industry** note (B)*

Ainsi, la traduction *chicken breasts* (*blancs de poulet*), par exemple, sans être la meilleure traduction possible de *suprême de volaille* reste dans la même thématique et n'altère pas la compréhension globale.

Dans la même lignée, certains choix lexicaux restent dans une thématique plus ou moins proche, avec un co-occurrent correctement traduit, mais le sens de la tête sémantique est totalement altéré, comme dans les exemples :

***Alphabet** arabe > arab **world** (C)*

***Gestion** communautaire > community **wildlife** (C)*

***Profession** de psychologue > psychological **association** (C)*

***Parc** nucléaire > nuclear **energy** (C)*

***Envie** de chocolat > chocolate **cake** (C)*

***Fiche** de vocabulaire > vocabulary **grammar** (C)*

***Accident** vasculaire > vascular **disease** (C)*

***Planification** nationale > national **economic** (C)*

***Lieu** de vote > vote **local** (C)*

***Ambition** présidentielle > presidential **campaign** (C)*

Dans ces exemples, le co-occurrent est correctement traduit, mais les têtes sémantiques ne correspondent pas, telles que respectivement alphabet/monde (world), gestion/vie sauvage (wildlife), profession/association (association), parc/energy, envie/cake (gâteau), fiche/grammar (grammaire), accident/disease (maladie), planification/economic (économique), lieu/local (local) et ambition/campaign (campagne). Ces traductions sont considérées comme incorrectes, mais la thématique reste tout de même relativement proche.
Dans certains cas, l'unité lexicale complexe est ambigüe en français, et sa traduction peut être acceptable en fonction de l'usage retenu, comme dans l'exemple :

Appareil militaire > military plane (B)

Dans cet exemple, la traduction concerne l'usage *AERIEN*, ce qui est acceptable, mais ne concerne pas le seul usage possible de *appareil militaire*.

Il arrive également que la traduction obtenue pour la tête sémantique soit ambigüe et offre un résultat acceptable mais non totalement satisfaisant, comme dans l'exemple :

***Groupe** de jeune > young **party** (B)*

Ici, *party* signifie *groupe*, mais reste ambigu car l'un des usages est *FETE*. La traduction *young party* est donc ambiguë par rapport à l'unité lexicale source car elle peut aussi désigner une « fête organisée par des jeunes », ce qui n'apparaît pas dans l'usage de départ *groupe de jeune*.

Choix lexical thématiquement proche (co-occurrent erroné)

Il arrive que la tête sémantique soit correctement traduite, mais que le co-occurrent soit erroné, même si l'ensemble de la traduction reste thématiquement proche, comme dans les exemples :

*Station de **montagne** > **hill** station (C)*

Ici, le co-occurrent *montagne* est traduit par *hill* qui signifie *colline*. Le sens global est proche mais non totalement équivalent.

Un autre exemple de ce type d'erreur de traduction du co-occurrent concerne :

*Régime **alimentaire** > **atkins** diet (C)*

La tête sémantique est correctement traduite, mais *alimentaire* est traduit par *atkins* qui est une marque de méthode de régime amincissant.

Un autre exemple de mauvaise traduction du co-occurrent est :

*Appartement **meublé** > **room** apartment (C)*

Ici, *room apartment* signifie *pièce d'appartement*.

Une traduction proche mais non complètement satisfaisante concerne l'unité lexicale traduite par :

***Right** government (B)*

Cette traduction reste ambiguë car *right* peut signifier *correct*, *acceptable*. Ici, la traduction équivalente serait :

***Right wing** government*

Une autre erreur concerne l'unité lexicale *général de brigade* :

Team general (B)

Dans cet exemple, le nom *brigade* est traduit par *team* (équipe). Parfois, la traduction du co-occurrent est de sens très proche, mais non parfaitement équivalent :

*Groupe d'**étude** > **practice** group (B)*

Ici, *practice* n'est pas l'équivalent exact d'*étude*, même si le sens général n'est pas altéré.
Un autre exemple particulier concerne l'unité lexicale et sa traduction :

***Ancien** militaire > military **past** (B)*

Dans ce cas, l'adjectif *ancien* n'est pas correctement traduit (nom *past*). Toutefois, notons que l'unité lexicale source est une erreur d'étiquetage entre l'adjectif *ancien* et le nom *ancien*. Nous n'avons pas collecté de structure de type *ADJECTIF-NOM*, mais elle n'a pas été filtrée car l'adjectif *ancien* peut également être un nom. De plus, l'ambigüité morpho-syntaxique concerne aussi le nom *militaire* qui peut également être un nom.

Choix lexical thématiquement proche (ensemble de l'unité lexicale)

Enfin, il arrive que le sens global reste relativement proche, mais que l'ensemble des traductions constituant l'unité lexicale soit erroné et non parfaitement équivalent, comme dans les exemples :

Ampoule économique > energy saving (économie d'énergie) (B)

Village de caractère > holiday rental (location de vacances) (C)

Qualité d'auteur > intellectual property (propriété intellectuelle) (C)

Projection numérique > films festivals (festival de films) (C)

Chambre d'accusation > right violation (violation des droits) (C)

Chiffre d'affaire > press release (communiqué de presse) (C)

Enseignement de base > education system (système éducatif) (C)

Caisse autonome > social security (sécurité sociale) (B)

Bloc d'alimentation > power supply (prise électrique) (B)

Dans ces exemples, les traductions sont erronées, mais appartiennent au même champ lexical que l'unité lexicale source.

Une majorité des erreurs concernant un mauvais choix lexical dans une thématique proche sont relatives à la phase 3 de notre méthodologie, qui collecte les traductions sans accès préalable à une ressource dictionnairique. Les mondes lexicaux sont comparés, ce qui nous garantit une proximité lexicale, mais les traductions de chaque constituant ne sont pas directement comparés avec les constituants sources. Une perspective d'amélioration concerne la vérification d'au moins un des constituants dans notre dictionnaire. Par exemple, pour l'unité lexicale *parc nucléaire*, nous obtenons la traduction :

Nuclear energy

Nous pourrions vérifier si la traduction candidate *nuclear park* (dont nous connaissons la traduction de *park*) n'est pas également présente, car elle aurait plus de probabilités d'être la traduction adéquate.

Choix lexical erroné

D'autres erreurs consistent en un choix lexical totalement erroné, dû à une difficulté d'ambigüité lexicale non résolue, comme dans les exemples :

***Fond** d'aide* **>** *help **back** (C)*

Back s'applique à un fond du type de l'usage *ARMOIRE*.

Une autre erreur d'ambigüité lexicale concerne le nom polysémique *campagne* dans les exemples :

***Campagne** agricole* **>** *agricultural **country** (C)*

***Campagne** publique* **>** *state **country** (C)*

Ici, le nom *campagne* ne s'applique pas à l'usage *NATURE*, mais à l'usage *POLITIQUE/PRESSE*. La traduction attendue pour la tête sémantique est *campaign*.

Un autre exemple concerne la traduction de *accord* par *understanding* :

***Accord** global* **>** *overall **understanding** (C)*

***Accord** mutuel* **>** *mutual **understanding** (C)*

Understanding correspond à l'usage *COMPREHENSION*.

Un autre exemple concerne l'unité lexicale *arbitre de touche* traduite par :

*Touch **judge** (C)*

Ici, *judge* ne s'applique pas à un arbitre dans l'usage du sport.

Une autre erreur concerne la traduction de *action nouvelle* :

*New **share** (C)*

Share s'applique à une action dans le sens de *PART*, ce qui n'est pas approprié dans ce contexte.

En ce qui concerne la traduction de *plateau de fromage*, la traduction obtenue reste acceptable :

*Cheese **set** (B)*

Toutefois, le choix lexical de la tête sémantique n'est pas le plus approprié. La traduction attendue serait :

Cheese board

Une autre erreur de désambigüisation lexicale concerne la traduction :

Prestation** de conseil > consultant **performance

Performance désigne une prestation dans l'usage de *PRESTATION (artiste, sportif)*.

Voici d'autres exemples de mauvaise désambigüisation lexicale de la tête sémantique :

***Voie** commerciale > commercial **road** (C)*

***Histoire** familiale > family **story** (C)*

***Formation** juridique > legal **background** (C)*

Nous pourrons améliorer cette source d'erreurs en affinant nos filtres de comparaison des mondes lexicaux (filtres plus stricts, ajout d'anti-dictionnaires, etc.).

Découpage sémantique différent

La traduction de *terrain* varie en fonction du type de terrain de sport (football, golf, basketball). Dans cet exemple d'erreur, *field* ne s'applique pas au terrain de basketball :

***Terrain** de basket > basketball **field** (C)*

La traduction attendue serait :

Basketball court

9.2.2 Erreurs morpho-syntaxiques

Un autre type d'erreur concerne les cas de traductions ayant un choix lexical correct, mais une structure morpho-syntaxique erronée. Les erreurs morpho-syntaxiques altèrent moins la compréhension globale que les erreurs lexicales, comme dans l'exemple :

Analyse de marché > analysis of market (C)

Ici, la traduction attendue serait du type de :

Market analysis

Parmi les erreurs morpho-syntaxiques, nous distinguons deux cas pour la structure syntaxique source *NOM-DE-NOM* (du type de *analyse de marché*). D'une part, certaines erreurs consistent en un mauvais choix entre les structures de type « roman » (*NOM-NOM*) et de type « germanique » (*NOM OF NOM*). D'autre part, certaines erreurs consistent en une non prise en compte de la structure de type « possessif », faisant intervenir le génitif.

Structures de type « roman » et de type « germanique »

Une difficulté consiste en un choix erroné entre les deux structures morpho-syntaxiques possibles en anglais que nous prenons en compte pour la même structure en français *NOM-DE-NOM*. Nous avons présenté dans le chapitre 6 ces deux types de structures. Par exemple, la traduction de l'unité lexicale *annuaire d'annuaires* est erronée d'un point de vue morpho-syntaxique :

Directory directory (C)

La structure attendue ici serait :

Directory of directories

Cette unité lexicale est naturellement spécifique et délicate à traiter, puisqu'elle est redondante dans les deux éléments lexicaux.

De même, nous obtenons la traduction suivante pour l'unité lexicale *maison de cognac* :

Cognac house (C)

Ici, le terme *maison de cognac* désigne une société qui commercialise du Cognac et la traduction attendue serait :

Cognac

Cette unité lexicale source est délicate, car peu utilisée dans un contexte anglophone.

Les erreurs de choix de structures entre le type « roman » et le type « germanique » concernent majoritairement un choix de type « germanique » alors que le type « roman » (c'est-à-dire l'ordre déterminant-déterminé) serait attendu, comme dans les exemples :

Cycle de vie > cycle of life (B)

Effet de change > effect of exchange (B)

Dans ces exemples de type « germanique », c'est le type « roman » qui est correct :

Life cycle

Exchange effect

Le type « roman » en anglais pose des difficultés de repérage à cause de problèmes d'ambiguïtés de rattachement syntaxique. A l'heure actuelle, nous n'effectuons pas d'analyse syntaxique des traductions candidates testées sur le Web. Une évolution pourra être d'ajouter un module d'analyse morpho-syntaxique, lorsque nous testons les fréquences sur le Web, par exemple (collecte des résumés et analyse morpho-syntaxique des résultats).

Structures de type « possessif »

Nous avons volontairement omis la structure de type « possessif », qui correspond au cas du génitif. Le génitif, qui constitue une relation d'appartenance entre les deux éléments est un cas particulier des constructions de repérage d'un nom par rapport à un autre nom. Cette structure concerne moins de cas et reste délicate à traiter. Par la suite, nous envisageons de la prendre en compte, puisque certaines erreurs morpho-syntaxiques concernent ce type de structure, comme dans les exemples :

Lait de femme > woman milk (C)

Lait de maman > mummy milk (C)

Justice de Dieu > god justice (C)

Dans ces exemples, c'est une structure du type génitif qui est souhaitable :

Woman's milk

Mummy's milk

God's justice

Dans ces exemples, le repère est un être humain (ou un élément plus ou moins assimilé) et le repéré est un objet ou une propriété susceptible « d'appartenir » à ce repère (Chuquet et Paillard, 1987).

Ordre des mots

Un autre type d'erreur concerne un mauvais ordre des mots, comme dans l'exemple :

Lac de garde > garda lake (C)

Cette unité lexicale présente la particularité de désigner un lieu. Contrairement à la majorité des traductions du type « roman » (NOM-NOM), où le second nom correspond au « déterminé », ici c'est la traduction *lake garda* qui est attendue, c'est-à-dire un ordre du type déterminé/déterminant.

Non équivalence de longueur

Une difficulté et source d'erreur concerne une non équivalence de longueur des constituants entre l'unité lexicale source et sa traduction, comme dans l'exemple de :

Métier de vigneron > wine grower (B)

Ici les deux constituants *wine grower* correspondent au seul terme *vigneron*. Si le sens global reste compréhensible, l'unité *métier* n'est pas traduite et devrait être rajoutée.

Absence de déterminant

Il arrive qu'un déterminant soit introduit au sein d'une structure de type « germanique », en anglais. Ainsi, nous obtenons la traduction erronée :

Fruit de saison > season fruit (C)

Non seulement ce n'est pas une structure de type « roman » qui est attendue ici, mais un déterminant doit être introduit au sein de la structure :

Fruit of the season

Une perspective sera de prendre en compte un nombre plus élevé de structures morpho-syntaxiques possibles en langue cible.

9.2.3 Erreurs idiomatiques

Un autre type d'erreur concerne un choix lexical sémantiquement pertinent, mais dont le caractère idiomatique n'est pas pleinement satisfaisant, comme dans l'exemple :

Fête d'anniversaire > anniversary party (B)

Bien que cette traduction soit considérée comme acceptable et reste compréhensible, le choix lexical de *anniversary* ne correspond pas au choix le plus pertinent d'un point de vue idiomatique. Ici, la traduction attendue serait :

Birthday party

Un autre exemple d'erreur idiomatique :

Mariage de princesse > marriage of princess (C)

Ici, le choix lexical attendu est *wedding* au lieu de *marriage*.

Une autre erreur, bien que la traduction reste acceptable, concerne l'unité lexicale :

Balle de ping-pong > ping-pong table (B)

Une traduction préférable serait :

Table tennis table

Une autre traduction acceptable mais non complètement idiomatique est celle de *truc de fou* :

Wild stuff (B)

Ici, une traduction plus idiomatique serait :

Crazy stuff

Une autre erreur de type idiomatique (ainsi que morpho-syntaxique) concerne l'unité lexicale *verre de whisky* :

Whiskey chaser (B)

Ici, *chaser* n'est pas la traduction la plus appropriée pour le nom *verre* (cet usage insiste sur la pluralité des verres qui sont bus). La traduction la plus appropriée serait :

Glass of whisky

9.3 Proportion des erreurs par catégorie

La figure suivante présente la proportion d'erreurs par catégories détaillées. Une majorité des erreurs concerne un choix lexical de la tête sémantique thématiquement proche mais non exactement équivalent (32,63%). Cette difficulté pourra être améliorée en utilisant une vérification au sein d'une ressource dictionnairique pour au moins l'un des constituants. Viennent ensuite les erreurs de structures morpho-syntaxiques (23,16%), qui pourront être améliorées en prenant en compte davantage de structures syntaxiques possibles en langue cible, telles que par exemple le génitif. Enfin, le troisième type d'erreur la plus fréquente concerne une mauvaise désambigüisation lexicale d'un des constituants (14,74%). Cette

difficulté pourra être améliorée en affinant davantage les comparaisons de mondes lexicaux (filtres supplémentaires, différentes mesures[1], etc.).

Type d'erreurs	Nombre de traductions	Pourcentage
Lexical proche(tête)	31	32,63%
Lexical proche(co-occurrent)	7	7,37%
Lexical proche (totalité)	9	9,47%
Désambiguisation lexicale	14	14,74%
Découpage sémantique	1	1,05%
Structure morpho-syntaxique	22	23,16%
Idiomatique	11	11,58%

Figure 100. Typologie détaillée des erreurs

Nous récapitulons les sources d'erreurs par grandes catégories (Figure 98). Une majorité des difficultés est d'ordre lexical (thématique proche mais non équivalente, désambigüisation lexicale), à savoir plus de la moitié des cas d'erreurs (65,26%). Viennent ensuite les erreurs d'ordre morpho-syntaxique (mauvais choix de structure) (23,16%), suivies des erreurs de type idiomatique (11,58%).

Type d'erreurs	Nombre de traductions	Pourcentage
Lexical	62	65,26%
Morpho-syntaxique	22	23,16%
Idiomatique	11	11,58%

Figure 101. Typologie générale des erreurs

Nous catégorisons enfin les erreurs par phase de traitement :

[1] Par exemple, nous avons utilisé la mesure du coefficient de Jacquard, afin de mesurer le degré de similitude entre deux ensembles. Nous pourrions également tester d'autres mesures, telles que par exemple le Cosinus.

Phases	Nombre de traductions	Pourcentage
Phase1	1	1,05%
Phase2	56	58,95%
Phase2_Cooc_Google	1	1,05%
Phase3_Cognats	20	21,05%
Phase3_Bigrammes	13	13,68%
Dico_Google	2	2,11%
Dico_Collins	2	2,11%

Figure 102. Nombre d'erreurs par phase de traitement

Un peu plus de la moitié des erreurs (58,95%) concerne la phase 2 (comparaison des mondes lexicaux). Cet ordre de grandeur s'explique par le fait qu'une majeure partie des traductions est traitée par cette phase. Viennent ensuite les erreurs concernant la phase 3 (21,05% pour la méthode des cognates et 13,68% pour la méthode des bigrammes fréquents). En effet, cette méthode, qui ne s'appuie pas au préalable sur des ressources dictionnairiques, présente la limite d'extraire certaines traductions proches thématiquement mais non parfaitement équivalente à l'unité lexicale source. La figure 100 récapitule la proportions d'erreurs par grandes catégories de nos phases de traitement :

Phases	Nombre de traductions	Pourcentage
Phase 1	1	1,05%
Phase 2	57	60,00%
Phase 3	33	34,74%
Ressources	4	4,21%

Figure 103. Nombre d'erreurs par grandes catégories de phases de traitement

Chapitre 10. Conclusion et perspectives

10.1 Discussion

Notre travail a mis l'accent sur trois types de problématiques que nous avons menées conjointement. D'une part, nous nous sommes interrogée, d'un point de vue linguistique sur le statut des unités lexicales complexes. Nous avons montré que, pour une même structure morpho-syntaxique, le statut linguistique peut être différent, ces différences ne sont pas binaires, mais graduelles. Le statut linguistique des unités lexicales complexes peut être envisagé en rapport avec le signe, c'est-à-dire avec le référent désigné par la globalité de l'unité lexicale. Toutefois, le rapport entre l'unité lexicale complexe et le référent auquel elle renvoie est un rapport complexe dont les frontières sont difficiles à établir. Les co-occurrents, en même temps qu'ils annoncent une caractéristique de l'objet peuvent avoir simultanément une valeur typante, comme dans l'exemple de *café noir*, où *noir* désigne à la fois une propriété du café et le place en même temps dans une catégorie particulière de café. Nous avons mis en relation ces réflexions théoriques avec la tâche d'identification automatique de ces unités lexicales, tâche menée à très vaste échelle, à partir du Web. Nous avons collecté environ 10 000 unités lexicales complexes et notre base continue de s'accroître au quotidien.

Nous nous sommes également intéressée à l'aspect interlingue des unités lexicales complexes, ce qui nous permet d'envisager le statut interlingue du signe : les différences d'une langue à l'autre sont révélatrices des différences de découpage sémantique de la « réalité ». Nous avons montré que les aspects bilingues des unités lexicales complexes étaient variés : traductions compositionnelles ou non compositionnelles. Nous avons proposé une analyse du comportement lexical du phénomène de traduction. Nous avons mis en rapport les aspects linguistiques de la traduction avec son acquisition automatique. Dans nos travaux, nous montrons qu'une analyse linguistique intrinsèque des unités lexicales complexes permet d'apporter un traitement automatique adapté et d'affiner ainsi les méthodes de traduction.

Nous avons également proposé une réflexion à la fois théorique et technique sur l'utilisation du Web dans le cadre d'applications linguistiques. D'un point de vue théorique, nous avons montré que le Web, bien qu'il ne constitue pas une ressource traditionnelle au même titre que les corpus au sens propre, constitue un réservoir gigantesque qui bouleverse les méthodes de travail linguistiques relatives à la compréhension des langues. La fréquence des phénomènes linguistiques n'est pas nécessairement une preuve de validité de la forme linguistique car certains usages valides sont peu usités. Toutefois, la fréquence permet de collecter les phénomènes linguistiques les plus courants, ce qui est utile, non seulement afin de les analyser mais aussi afin de les collecter dans le cadre de la construction de vastes ressources lexicales telles que celle que nous construisons. D'un point de vue technique, nous avons mis en place une méthodologie d'acquisition de ressources lexicales monolingues et bilingues, à vaste échelle, qui présente l'intérêt de pouvoir fonctionner en continu et de grossir de façon quotidienne. Nous avons analysé, puis mis en pratique, les différentes facettes du « Web bilingue », en proposant une méthode « mixte » de stratégies. Les ressources que nous avons collectées jusqu'à présent sont de bonne qualité, avec une précision de traduction très satisfaisante, à savoir 94,4% de traductions acceptables. Le rappel est également particulièrement élevé, avec un taux de 82,5%.

Nous nous sommes également centrée sur l'étude du contexte des unités lexicales complexes et avons défini la notion de « mondes lexicaux ». Nous avons observé ce phénomène à vaste échelle, en collectant les mondes lexicaux directement à partir du Web. Ces mondes lexicaux, en français et en anglais, ont été exploités pour la désambigüisation lexicale pour la

traduction. Toutefois, ces ressources sont intéressantes en elles-mêmes et pourront certainement être exploitées par la suite pour la construction de ressources de type ontologiques.

10.2 Perspectives

Les vastes ressources dont nous disposons grâce au Web nous offre des perspectives d'évolution d'un point de vue quantitatif d'une part (l'augmentation régulière des données va nous permettre d'affiner nos observations en obtenant de nombreux nouveaux cas) et d'un point de vue qualitatif d'autre part, étant donné que nous visons à affiner différents aspects de notre méthode, à savoir la prise en compte des thématiques sur le Web (10.2.1), l'élargissement des structures morpho-syntaxiques étudiées, ainsi que l'ajout de modules syntaxiques (10.2.2) et enfin la prise en compte de relations sémantiques permettant de classer les unités lexicales complexes en familles thématiques et d'organiser les arguments qui constituent les unités lexicales complexes selon des repères sémantiques (10.2.3).

10.2.1 Thématiques de recherche

Nous avons fait le choix de n'extraire qu'une seule traduction par unité lexicale source. Il arrive toutefois que plusieurs traductions soient correctes. Nous pourrions par la suite nous intéresser à un recensement exhaustif de toutes les traductions possibles pour une même unité lexicale source.

D'une façon plus précise, nous pourrions nous intéresser aux domaines de spécialité ou aux thématiques particulières. Par exemple, une traduction satisfaisante en langue générale (l'usage le plus courant) peut être inadéquat dans un domaine de spécialité. Considérons la traduction :

Appareil numérique > digital camera

Bien que l'usage le plus courant soit l'usage *PHOTOGRAPHIE*, la traduction *digital camera* est inappropriée dans certains domaines. Par exemple, dans le domaine médical, la traduction attendue est *digital device*. Une évolution ultérieure sera de nous intéresser aux domaines de spécialité ou aux genres liés à une thématique sur le Web, afin de palier les limites liées à l'ambigüité lexicale. Par exemple, *Wikio*[1] est un portail d'information qui réunit les rubriques de news à partir de différents sites de presses et de blogs. Un alignement d'unités lexicales complexes à partir d'un tel site nous permettrait de cibler des usages précis. Un autre exemple est le site *Wikipédia*[2], portail de recherche encyclopédique et multimédia qui contient de nombreux articles en différentes langues. Les pages traduites disponibles sur *Wikipédia* constitue un réservoir de pages « partiellement » parallèles qu'il serait intéressant d'exploiter à des fins d'alignement. Voici un exemple de pages en français pour l'unité lexicale complexe source *caisse claire*[3] :

Figure 104. Description de « caisse claire » sur Wikipédia

[1] http://www.wikio.fr/

[2] http://fr.wikipedia.org/wiki/Accueil

[3] http://fr.wikipedia.org/wiki/Caisse_claire

A partir de la page en français, un lien pointe vers la description du même terme dans d'autres langues, dont l'anglais, ce qui nous permet non seulement d'avoir accès à la traduction de l'unité lexicale (*snare drum*), mais également de collecter des pages « partiellement » traduites afin d'aligner d'autres termes traitant du même domaine.

Enfin, un autre exemple de ressource thématique pour la traduction concerne les forums liés à une thématique, comme par exemple la médecine. Le fait de cibler sur une thématique en particulier nous permettrait de palier les cas d'ambiguïtés lexicales.

10.2.2 Analyse morpho-syntaxique

Elargissement des patrons morpho-syntaxiques sources

Une autre perspective concerne la prise en compte d'autres patrons morpho-syntaxiques tels que VERBE-OBJET ou SUJET-VERBE. Les relations morpho-syntaxiques du type de VERBE-OBJET (que nous avions commencé à étudié dans Léon et Millon (2005)) constituent des indices désambiguisateurs forts, que nous pourrions exploiter de la même façon que les unités lexicales complexes. Voici des exemples de relations VERBE-OBJET obtenues dans Léon et Millon (2005) pour la co-occurrence source construire-barrage :

V N	construire-barrage	to build-barrage to build-barricade to build-dam to build-roadblock to construct-dam to erect-barricade to erect-roadblock

Figure 105. Exemples de traductions de la relation construire-barrage

Analyse syntaxique pour la traduction

Au-delà de l'élargissement des structures morpho-syntaxiques, la base de données de traductions pourrait être utile à des systèmes de traduction basés sur une analyse morpho-syntaxique des divergences syntaxiques entre langues. Les patrons morpho-syntaxiques bilingues obtenus pourraient être intégrés à des systèmes de traduction tels que *SYGFtoE* (Prince et Chauché, 2006 ; Bonnin et Prince, 2007), qui est un prototype de traduction, basé sur l'analyse des divergences syntaxiques et stylistiques entre la langue source et la langue cible. Le système s'appuie sur une analyse morpho-syntaxique, qui détecte les dépendances de chaque phrase source et construit un arbre de dépendances. Un « transfert » syntaxique est effectué vers la langue cible, via des opérations locales de transformation (connaissance des règles de transformation morpho-syntaxiques de la langue cible). L'accès aux traductions d'unités lexicales complexes contenues dans notre base de données pourrait fournir une aide au système à deux niveaux : d'une part, le possible repérage des séquences lexicalisées pourrait apporter une aide à l'analyse de dépendances morpho-syntaxiques et aux règles de transformations entre la langue source et la langue cible. D'autre part, notre base de données pourrait être utile au « transfert lexical » qui doit être effectué après le transfert syntaxique.

Traductions de taille différente de l'unité lexicale complexe source

A l'heure actuelle, notre système ne dispose que d'une possibilité afin de prendre en compte des traductions de taille différente de l'unité lexicale complexe source. Nous recherchons au préalable si la traduction est recensée dans nos ressources dictionnairiques. Une partie des traductions connues n'est constituée que d'une unité lexicale simple. Au-delà de ces cas, nous ne prenons pas en compte les traductions de taille différente. Pourtant, la taille de la traduction peut être différente, qu'il s'agisse de mono-termes mais aussi de séquences plus longues, comme dans l'exemple (Morin *et al.*, 2004) :

Essence d'ombre > shade tolerant species

Ici, une unité lexicale complexe de deux mots-plein est traduite par une séquence de trois mots-plein. Afin de gérer ce type d'irrégularité de longueur entre la langue source et la langue

cible, nous ne pouvons pas nous appuyer sur l'étape de génération de traduction candidate via un dictionnaire existant. Nous faisons l'hypothèse que ces cas n'obtiendront pas de traductions lors des phase 1 (utilisation des fréquences) et des phase 2 (mondes lexicaux). La prise en compte de ce problème pourrait intervenir lors de la phase 3 (cognates et bigrammes fréquents), lorsque nous collectons les traductions à partir des résumés « mixtes ». Dans notre étude, nous nous sommes volontairement limitée aux bigrammes, mais nous pourrions élargir les traductions candidates collectées et prendre également en compte les mono-termes et les trigrammes au sein des résumés « mixtes ».

10.2.3 Sémantique lexicale

Ajout de ressources externes

Prince et Chauché (2008) présentent une méthode de traduction du français vers l'anglais, basée sur l'exploitation de ressources de type ontologique. Chaque thesaurus utilisé en anglais (*English Roget Thesaurus*) et en français (*Thesaurus Larousse*) est exploité tel un espace vectoriel, dans lequel les entrées monolingues forment un vecteur de concepts associés. Les entrées françaises sont ensuite représentées sous la forme de leur équivalence dans l'espace anglais. La tâche de désambigüisation lexicale consiste à sélectionner le vecteur approprié au sein des vecteurs bilingues par comparaison avec un vecteur contextuel de la phrase source. Par exemple, voici un extrait des concepts anglais attribués à l'entrée française *course* :

« course » : errand, journey, rush, race, racing, travel, stroke, flight path, passage, privateering, shopping

Ainsi, à partir de la phrase source « Les courses de chevaux ont lieu tous les mardis », le vecteur « contextualisé » de *course* dans cette phrase est comparé avec les 11 entrées recensées. Une mesure basée sur le cosinus est utilisée afin de sélectionner la traduction la plus adéquate (*race* ou *racing* dans cet exemple). L'exploitation de ressources externes pour la désambigüisation lexicale pourrait être combinée à notre méthode, afin de mêler des connaissances encyclopédiques (telles que des thésaurus) à des connaissances textuelles (telles que les mondes lexicaux construits à partir de données textuelles).

Amélioration de la comparaison des mondes lexicaux

Il existe de nombreuses mesures afin de mesurer la distance entre deux textes (Brunet, 2003). La distance de Jacquard, qui est la mesure que nous utilisons afin de comparer les mondes lexicaux, permet d'établir le rapport entre les mots communs aux deux textes à comparer et ceux qui n'appartiennent qu'à l'un des deux (*ibid.*). Une limite de cette méthode est de ne pas prendre en compte les éventuelles différences de fréquences au sein de chaque texte. Ainsi, si le partage des fréquences est inégal (par exemple, si l'on trouve dans le texte français une unité qui a une occurrence de 1 et que sa traduction dans le texte anglais a une occurrence de 19), la comparaison est moins efficace que si la répartition des fréquences était équilibrée (par exemple, une occurrence de 10 dans les textes, si l'on imagine naturellement que les textes sont de même longueur). Ainsi, cette mesure présente le risque de privilégier la prise en compte d'unités de faible occurrence au détriment des unités les plus fréquentes. Afin d'améliorer notre comparaison des mondes lexicaux et palier cette limite, nous pourrions prendre en compte le poids (fréquence) des unités lexicales au sein de chaque monde lexical et obtenir ainsi une comparaison pondérée. De nombreuses méthodes peuvent être envisagées afin de tenir compte de la fréquence des unités au sein des textes à comparer[1]. Malgré tout, dans une discussion critique sur le choix d'une méthode de mesure de comparaison entre deux textes, Brunet (2003) montre que quelque soit la méthode utilisée, les résultats ont tendance à être convergeants et que les différences sont peu sensibles.

En ce qui concerne une amélioration ultérieure de la comparaison entre les mondes lexicaux, les ressources déjà obtenues pourraient être exploitées afin d'améliorer notre comparaison des mondes lexicaux, par la prise en compte d'unités lexicales complexes. Par exemple, si une unité lexicale complexe source appartenant à notre base de données apparaît dans le monde lexical français et si sa traduction apparaît dans le monde lexical anglais, nous pourrions « matcher » ces équivalences et la comparaison des unités porterait à un niveau supérieur à l'unité lexicale simple (la comparaison des simples mono-termes constituant une limite).

[1] Pour une description complète et détaillée des mesures permettant de comparer la proximité entre deux textes en prenant en compte les fréquences des unités, se référer à Brunet (2003) et à Labbé et Labbé (2003).

Notre système serait alors basé sur un processus d'apprentissage dont les données collectées seraient exploitées afin d'améliorer le système.

Interprétation automatique des composés nominaux anglais

Fabre et Sébillot (1996) proposent une description sémantique de séquences de composés nominaux anglais de la forme NOM NOM, dans un but d'aide à la structuration et à la lisibilité d'un réseau de termes issus d'une phase d'extraction. La difficulté de l'interprétation des composés nominaux anglais provient de la relation implicite qui relie la simple juxtaposition des deux constituants. Fabre et Sébillot (1996) montrent qu'il est possible d'analyser cette relation à partir de l'information lexicale qui caractérise les constituants et d'obtenir un calcul automatique et une représentation du sens de ces composés. La méthode est basée sur une analyse des propriétés des « noms déverbaux » (dotés d'une structure argumentale) et des « noms rôles » (rôle du constituant modifieur) ainsi que sur des informations relatives à la classe sémantique hiérarchisée des noms. Le calcul automatique passe d'abord par une phase d'identification de la structure argumentale du prédicat et ensuite par une identification du rôle du second constituant. La méthode s'appuie également sur le constat selon lequel la relation entre un nom rôle et son prédicat peut être généralisée à un ensemble de noms appartenant à une classe sémantique commune. L'utilisation de WordNet, combinée à un vaste corpus de noms composés est utilisée afin de mettre en application l'association entre un type sémantique et une relation prédicative. De telles méthodes d'interprétation automatique, utilisées en amont de nos résultats de traduction, pourraient être utiles afin d'affiner notre base de données, grâce à une représentation formelle identique des unités lexicales complexes en français et en anglais. Une telle formalisation serait utile à la vérification de nos traductions et ajouterait des informations sémantiques associées aux unités lexicales complexes.

10.2.4 Autres perspectives

Amélioration du silence

Les cas de silence (17,49% des traductions n'ont pas été obtenues) peuvent correspondre à plusieurs causes. Il peut s'agir de traductions non compositionnelles qui ne sont pas présentes au sein de nos résumés mixtes. En effet, la stratégie de la phase 3 qui consiste à collecter des résumés « mixtes » via une requête française limitée à la langue anglaise ne garantit pas que la traduction soit présente au sein des résumés. Une possibilité d'amélioration de ces difficultés pourra être d'élargir notre collecte, tant quantitativement (en collectant par exemple les pages Web entières et non pas seulement les résumés) que qualitativement. Ainsi, nous pourrions ajouter d'autres stratégies d'acquisition de pages susceptibles de contenir la traduction adéquate, telles que l'exploitation des mondes lexicaux français. Par exemple, nous pourrions générer les traductions des noms et adjectifs les plus fréquents au sein des mondes lexicaux français et utiliser ces traductions en tant que requête. Par exemple, les deux premiers noms[1] du monde lexical français de « futur antérieur » (dont nous n'avons pas obtenu de traduction) sont *verbe* et *anthologie*. Nous pourrions générer des requêtes du type de :

verb +anthology

Ce type de requête nous permettrait de collecter des pages dont le monde lexical est proche du monde lexical source et d'obtenir des pages « comparables » au sein desquelles nous pourrions extraire des traductions candidates, à partir de patrons morpho-syntaxiques définis.

Il arrive également que la traduction adéquate soit présente dans les résumés « mixtes » déjà collectés mais qu'elle ne soit pas validée au cours de l'un des filtres de notre phase 3.

[1] La longueur du nombre de mots-clé à prendre en compte devra être testée.

Parmi la totalité des cas de silence, nous avons évalué à environ 14% la proportion de traductions présentes au sein des bigrammes collectés dans les résumés mixtes, mais non validées au cours de l'un de nos filtres[1].

Il arrive que la traduction correcte soit contenu parmi les couples de traduction testés via le Web parallèle, mais ne soient pas parmi les couples les plus fréquents retenus, ce qui provoque quelques cas de silence, comme dans l'exemple de *cep de vigne*, où la traduction correcte, *wine growing* n'apparaît qu'au huitième rang parmi les couples les plus fréquents. Les traductions correctes non retenues à ce stade comptent pour environ 11% parmi tous les cas de traductions correctes non validées.

Il peut s'agir d'une non-validation de la traduction, par le filtre de comparaison entre la fréquence de l'unité lexicale française et celle de sa traduction. Il peut effectivement arriver que les fréquences d'usage d'une expression ne soient pas proportionnelles entre le français et l'anglais. Par exemple, la traduction *postal bank* (fréquence de 144 000) a une fréquence inférieure à l'unité lexicale source *banque postale* (fréquence de 937 000). Il en va de même pour l'unité lexicale *ballon dirigeable* (fréquence de 132 000), pour laquelle la traduction candidate *dirigible balloon* a une fréquence de 27 500. Parmi les 14% de traductions correctes non validées, nous évaluons à environ 29% le nombre de traductions non validées à ce filtre de fréquence du couple français/anglais.

Un certain nombre de traductions candidates correctes n'ont pas été validées à l'étape de comparaison des mondes lexicaux, pour plusieurs raisons. Il peut s'agir du fait que notre filtre soit trop élevé, mais ce seuil a été fixé afin de palier au maximum le bruit, ce qui provoque naturellement des cas de silence. Par exemple, la traduction candidate *licence plate* pour l'unité lexicale *plaque d'immatriculation* n'a pas été validée lors de la comparaison des mondes lexicaux. Enfin, il peut s'agir d'unités lexicales complexes trop générales pour générer un monde lexical homogène. Dans ce cas, même un seuil de comparaison entre les

[1] Nous évaluons uniquement les traductions valides présentes au sein des bigrammes collectés. Notons qu'une traduction correcte peut toutefois être présente dans les résumés mixtes mais non collectée au sein des bigrammes (à cause de la non-prise en compte des trigrammes par exemple), mais l'évaluation de ces cas serait plus délicate.

mondes lexicaux moins strict n'aurait pas permis une validation. Par exemple, les unités lexicales *monde de douceur* ou encore *mois d'absence* sont des unités pouvant être utilisées dans de nombreux domaines et dont le monde lexical ne peut pas être homogène. Nous avons évalué à environ 59% la proportion de traductions correctes qui n'ont pas été validées au stade de la comparaison des mondes lexicaux, parmi toutes les traductions correctes non validées. Nous avons parlé dans la section 10.2.1 de prendre en compte des thématiques et/ou des domaines de spécialités afin de limiter ce type de problème. Cette évolution peut s'opérer à grande échelle, à partir de thématiques variées. Le fait d'inscrire une unité lexicale complexe au sein d'une thématique donnée nous permettra de palier cette limite. Les deux figures suivantes illustrent la proportion de traductions correctes non validées qui étaient disponibles parmi les bigrammes collectés, classées par catégorie de rejet. Une majorité des cas concerne une non-validation lors de la comparaison des mondes lexicaux (59,26%). Vient ensuite un rejet lors du filtre de la fréquence des couples (29,63%), puis lors du filtre du Web parallèle (11,11%).

Monde lexical	16	59,26%
Comparaison des fréquences	8	29,63%
Web parallèle	3	11,11%

Figure 106. Quantité de traductions correctes non validées

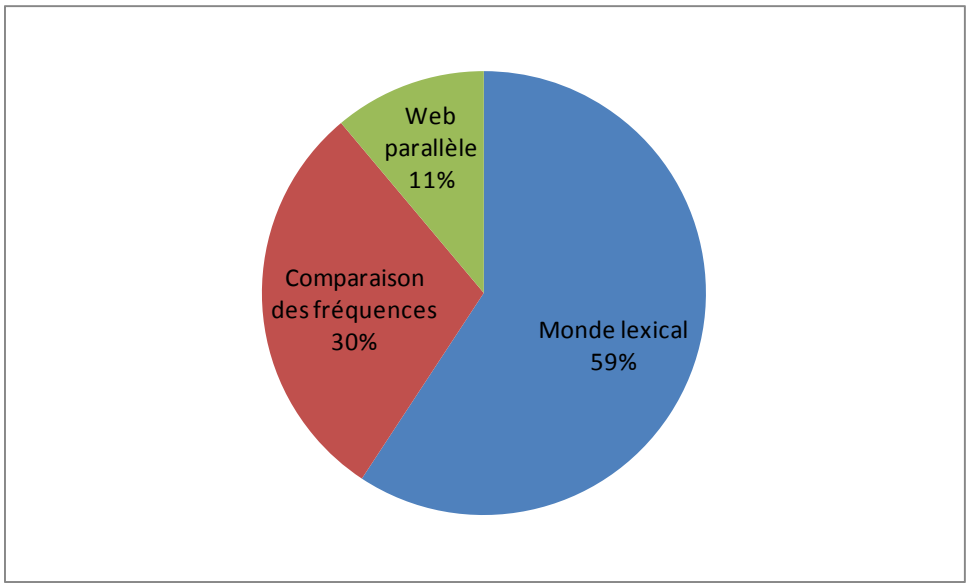


Figure 107. Proportion de traductions correctes non validées, par catégorie

Ressource coopérative

Notre base de données lexicales n'est pas, à l'heure actuelle, une ressource disponible en ligne, car nous voulions dans un premier temps tester de façon locale notre méthodologie. Toutefois, une de nos perspectives concerne la mise en ligne de notre base de données, avec une possibilité d'interaction avec les utilisateurs (corrections, ajout de ressources, etc.). Nous pourrions, par exemple, proposer une base lexicale coopérative, dans la lignée de projets tels que le projet PAPILLON qui permet aux utilisateurs de proposer de nouvelles ressources (ce projet est une base lexicale coopérative, construite à partir de ressources déjà existantes, contrairement à nous), ou encore la recherche interlingue de *Google*, qui permet aux utilisateurs de suggérer d'éventuelles corrections pour les cas où les résultats sont considérées comme étant erronés.

Construction de ressources ontologiques

Les mondes lexicaux pourraient être exploités afin d'organiser les unités lexicales complexes en familles thématiques et construire un lexique bilingue structuré de type ontologique. Par

exemple, les mondes lexicaux de *appareil digital* et de *appareil compact* sont proches. Nous pourrions systématiser les comparaisons de mondes lexicaux entre les unités lexicales complexes monolingues en français d'une part et en anglais d'autre part, et les réunir en grandes familles thématiques. Par exemple, *appareil digital* et *appareil compact* pourraient appartenir à une classe thématique de type *PHOTOGRAPHIE*. Les mondes lexicaux pourraient être exploités afin de construire ces classes thématiques et les nommer, dans la lignée de travaux tels que ceux de Pichon et Sébillot (1999a) et Rossignol et Sébillot (2003), mais à l'échelle du Web, ainsi qu'en ajoutant la dimension bilingue, puisque les familles thématiques seraient constituées pour le français et pour l'anglais.

La classification thématique nous permettrait d'obtenir un réseau de termes hiérarchisés, de type ontologique. Selon (Bourigault et Jacquemin, 2000), une ontologie désigne :

> INGÉNIERIE DES CONNAISSANCES. Ensemble des objets reconnus comme existant dans le domaine. Construire une ontologie, c'est décider de la manière d'être et d'exister des objets.

En terminologie, l'objectif est de regrouper les concepts qui présentent des caractéristiques communes, et d'obtenir ainsi un réseau hiérarchisé de concepts. De plus en plus de travaux visent à obtenir ce type de représentation à partir de données textuelles.

Au-delà des grandes classes thématiques qui regrouperaient les unités lexicales complexes, nous pourrions proposer des repères sémantiques aux sous-classes d'objets qui constituent les unités lexicales complexes. Dans Léon (2003, 2004b), nous montrons que les co-occurrents des unités lexicales complexes (tels que les objets des verbes, par exemple), peuvent être réunis en classes conceptuelles dont les combinaisons offrent des régularités de sélection, dans la lignée de travaux tels que Mel'cuk et Wanner (1996) et L'Homme (1998). Nous avons porté notre étude sur deux domaines de spécialité, la météorologie (Léon, 2003) et le Traitement Automatique des Langues (Léon, 2004b). Par exemple, en météorologie, les noms décrivant un *ELEMENT NATUREL* se combinent de façon régulière avec des adjectifs décrivant un *ETAT* (Léon, 2003) :

ELEMENT NATUREL > ETAT

Ciel bleu

clair

couvert

dégagé

gris

nuageux

orageux

rouge

Air chaud

doux

frais

froid

glacial

humide

nuageux

réchauffé

saturé

sec

De même, pour le domaine du Traitement Automatique des Langues, prenons les exemples des co-occurrences des verbes *phonétiser* et *traduire*. Une majorité des arguments sélectionnés appartiennent à la classe *DONNEES TEXTUELLES* (Léon, 2004b) :

Phonétiser > DONNES TEXTUELLES

Corpus

Forme

Lexique

Mot

Nom

Phrase

Sigle

Terme

Texte

Traduire > DONNES TEXTUELLES

Corpus

Document

Expression

Forme

Lexie

Lexique

Lexème

Message

Mot

Nom

Occurrence

Phrase

Segment

…

Nous pourrions appliquer ces analyses aux données que nous obtenons à partir du Web, et voir s'il est possible de systématiser ces phénomènes sur de plus vastes données, et observer si ces régularités s'appliquent également d'un point de vue bilingue.

Bibliographie

Agirre (2000a). *Exploring automatic word sense disambiguation with decision lists and the Web* Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Content, Luxembourg.

Agirre, E., Olatz, A., Hovy, E., Martinez, D. (2000b). *Enriching very large ontologies using the WWW*. Ontology Construction of the European Conference of AI (ECAI), Berlin, Allemagne.

Agirre, E., Lopez, O. (2004a). *Publicly available topic signatures for all wordnet nominal senses*. Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal.

Agirre, E., Alfonseca, E., Loper, O. (2004b). *Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures* In Second International Wordnet Conference, Czech Republic, Brno.

Almeida, J. J., Simoes, A. M., Castro, J. A. (2002). *Grabbing parallel corpora from the Web*. Sociedade Espanola para el Procesamiento del Lenguaje Natural.

Atkins, B. T. S. (1990). "Corpus Lexicography : The Bilingual Dimension." *Computational Lexicoglogy and Lexicography (Special issue dedicated to Bernard Quemada)* VI.

Audibert, L. (2003). Outils d'exploration de corpus et désambiguïsation lexicale automatique. *Thèse de doctorat (Informatique), Équipe DEscription Linguistique Informatisée sur Corpus (DELIC)*. Aix-en-Provence, Université d'Aix-Marseille I - Université de Provence.

Bally, C. (1909). *Traité de stylistique française*. Paris, Klincksieck.

Bally, C. (1965, 1ère édition 1932). *Linguistique générale et linguistique française*. Berne,

Francke.

Bar-Hillel, Y. (1955). Idioms. *Machine Translation of Languages, Fourteen Essays*. W. N. Locke, Booth, A. Donald. Boston, MIT & John Wiley: 183-193.

Baroni, M., Bernardini, S. (2004). *BootCaT: Bootstrapping corpora and terms from the web*. LREC 2004.

Baroni, M., Vegnaduzzo, S. (2004). *Identifying subjective adjectives through web-based mutual information*. KONVENS 2004, Vienna: ÖGAI.

Baroni, M., Ueyama, M. (2004). *Retrieving japanese specialized terms and corpora from the WWW*. Proceedings of KONVENS 2004.

Baroni, M., Bisi, S. (2004). *Using cooccurrence statistics and the web to discover synonyms in a technical language*. Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).

Baroni, M., Ueyama, M. (2006). *Building general- and special-purpose corpora by Web crawling*. Proceedings of the 13th NœL International Symposium, Language Corpora : Their Compilation and Application.

Baroni, M., Kilgarriff, A. (2006). *Large linguistically-processed web corpora for multiple languages*. EACL, Trento, Italie.

Benson, M., Benson, E., Ilson, R. (1986). *The BBI Combinatory Dictionary of English : A guide to Word Combinations*. Amsterdam, John Benjamins.

Benson, M. (1990). "Collocations and general-purpose dictionaries." *International Journal of Lexicography* 3(1): 23-35.

Benveniste, E. (1966). "Formes nouvelles de la composition nominale." *BSL* 61: 82-95.

Benveniste, E. (1967). "Fondements syntaxiques de la composition nominale." *BSL* 62: 15-31.

Blank, I., Ed. (2000). *Terminology extraction from parallel technical texts*. Parallel Text Processing. Dordrecht, Kluwer.

Bonnin, G., Prince, V. (2007). *Emphasizing Syntax for French to German Machine Translation*. SNLP'07: 7th International Symposium on Natural Language Processing, Chonburi, Thailande, Pattaya.

Bouillon, P. (1998). *Traitement automatique des langues naturelles*. Paris, Bruxelles, Aupelf-Uref – Editions Duculot.

Boulanger, J.-C. (1979). Commentaire de Jean-Claude Boulanger. *Table ronde sur les problèmes du découpage du 260 terme*. Montréal: 169-182.

Bourigault, D. (1994). LEXTER, un Logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de textes, Thèse de doctorat, Ecole des hautes études en sciences sociales.

Bourigault, D., Fabre C. (2000). "Approche linguistique pour l'analyse syntaxique de corpus." *Cahiers de Grammaires, Université Toulouse - Le Mirail* 25: 131-151.

Bourigault, D., Jacquemin, C., Ed. (2000). *Construction de ressources terminologiques*. Industrie des langues. Paris, Hermès.

Bourigault, D., Aussenac-Gilles, N., Charlet, J. (2004). "Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas." *Revue d'Intelligence Artificielle* 18(1).

Brown, P. F., Della Pietra, S., Della Pietra, V. J., Mercer, R. L. (1991a). *Word sense disambiguation using statistical methods*. Actes de 29th Annual Meeting of Association for Computational Linguistics, Berkeley, California.

Brown, P. F., Lai, J. C., Mercer, R. L. (1991b). *Aligning Sentences in Parallel Corpora*. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley.

Brown, P. F., Della Pietra, S., Della Pietra, V. J., Mercer R. L. (1993). "The mathematics of statistical machine translation : parameter estimation." *Computational Linguistics* 19(2): 263-311.

Brunet, E. (2003). "Peut-on mesurer la distance entre deux textes ?" *Corpus, La distance intertextuelle* 2.

Bunescu, R. (2003). *Associative Anaphora Resolution: AWeb-Based Approach*. Proceedings of the EACL-2003 Workshop on the Computational Treatment of Anaphora, Budapest, Hungary.

Burnard, L. (1995). *Users Reference Guide British National Corpus Version 1.0*. Oxford, University Computing Services.

Calvo, H., Gelbukh, A. (2003). *Improving Disambiguation of Prepositional Phrase Attachments Using the Web as Corpus*. CIARP, 2003

Cao, Y., Li, H. (2002). *Base noun phrase translation using web data and the EM algorithm*. International Conference of Computational Linguistics (COLING'02), Tapei, Taïwan.

Chauché, J. (1990). "Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance." *TAL Information*: 17-24.

Chen, J., Nie, J.-Y. (2000). *Parallel Web Text Mining for Cross-Language IR*. In Proceedings of RIAO 2000: Content-Based Multimedia Information Access Paris, France.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., Chien, L.-F. (2004a). *Translating unknown queries with web corpora for cross-language information retrieval*. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval Sheffield, United Kingdom

Cheng, P.-J., Pan, Y.-C., Lu, W.-H., Chien, L.-F. (2004b). *Creating multilingual translation lexicons with regional variations using web corpora*. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Chklovski, T., Mihalcea, R. (2002). *Building a sense tagged corpus with open mind word expert*. Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions.

Chomsky, N. (1957). *Syntactic structures*. La Hague, Mouton.

Chomsky, N. (1962). *The Logical Basis of Linguistic Theory*. in Preprints of Papers from the 9th International Congress of Linguists, Cambridge, Mass.

Choueka, Y., Conley, E. S., Dagan, I., Ed. (2000). *A comprehensive bilingual word alignment system: Accommodating disparate languages: Hebrew and English*. Parallel Text Processing. Dordrecht, Kluwer.

Chung, S., Jun, J., McLeod, D. (2006). *A Web-Based Novel Term Similarity Framework for Ontology Learning* -. OTM Conferences.

Chuquet, H., Paillard, M. (1987). *Approche linguistique des problemes de traduction anglais <-> français*. Gap, Paris, Ophrys.

Church, K., Hanks, P. (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics* 16(1): 22-29.

Church, K. (1993). *Char_align: a program for aligning parallel texts at the character level*. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio.

Clas, A. (1994). "Collocations et langues de spécialité." *Meta* 39(4): 576-580.

Clear, J., Ed. (1993). *From Firth principles: Computational tools for the study of collocation*. Text and technology: In honour of John Sinclair. Amsterdam, John Benjamins.

Corbin, D., Ed. (1997). *Locutions, composés, unités polylexématiques : lexicalisation et mode de construction*.

Cowie, A. (1981). "The treatment of Collocations end Idioms in Learner's Dictionaries." *Applied Linguistics* 2(3): 223-235.

Cowie, A., Ed. (1998). *Phraseology, Theory, Analysis, and Applications*. Clarendon Press. Oxford.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge, Cambridge University Press.

Dagan, I., Alon, Itai, Schwall Ulrike (1991). *Two languages are more informative than one*. Annual Meeting of the Association for Computationnal Linguistics (ACL), Berkeley, Californie.

Dagan, I., Church, K. W. (1994). *Termight : identifying and translating technical terminology*. 4th Conference on Applied Natural Language Processing (ANLP'94), University of Stuttgart, Germany.

Daille, B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques, Université Paris 7. Thèse de Doctorat en Informatique Fondamentale.

Daille, B. (1995). "Repérage et extraction de terminologie par une approche mixte statistique et linguistique." *Revue TAL, Traitements probabilistes et corpus* 36(n°1-2): 101-118.

Darmesteter, A. (1875). *Traité de la formation des mots composés dans la langue française comparée aux autres langues romanes et au latin*. Paris, Honoré Champion.

David, S., Plante, P. (1990). "De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes." *Intelligence Artificielle et Sciences Cognitives au Québec* 3(3): 140-154.

De Schryver, G.-M. (2002). "Web for/as Corpus: A Perspective for the African Languages." *Nordic Journal of African Studies 11(2)* 11(2): 266-282.

Debili, F., Sammouda E. (1992). *Appariement des Phrases de Textes Bilingues*. Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. (1990). "Indexing by latent semantic analysis." *Journal of the American Society of Information Science* 41(6): 391-407.

Déjean, H., Gaussier, E. (2002). "Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables." *Lexicometrica, numéro spécial sur Alignement lexical dans les corpus multilingues*: 1-22.

Diab, M., Finch, S. (2000). *A Statistical Word-Level Translation Model for Comparable Corpora*. Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO).

Doan, A., Madhavan, J., Dhamankar, R.; Domingos, P., Halevy, A. (2003). "Learning to Match Ontologies on the Semantic Web." *VLDB Journal* 12: 303-319.

Drouin, P. (2002). Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés, Thèse de doctorat, Université de Montréal.

Dubois, J. (1973). *Dictionnaire de linguistique*. Paris, Larousse.

Dubois, J., Guespin, L., Giacomo, M., Marcellesi, C. et J.B., Mével, J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Paris, Larousse.

Dubreil, E. (2008). "Collocations : définitions et problématiques." *Texto* XIII(1).

Duclaye, F. (2003). Apprentissage automatique de relations d'équivalence sémantique à partir du Web, Ecole Nationale Supérieure des Télécommunications.

Dunning (1993). "Accurate Methods for the Statistics of Surprise and Coincidenc." *Computational Linguistics* 19(1).

Enguehard, C. (1993). *Acquisition de terminologie à partir de gros corpus*. Informatique & Langue Naturelle, ILN'93, Nantes.

Enguehard, C., Panterra, L. (1995). "Automatic Natural Acquisition of a Terminology." *Journal of quantitative linguistics* 2(1): 27-32.

Fabre, C., Sébillot, P. (1996). *Interprétation automatique des composés nominaux anglais hors domaine : quelles solutions ?* 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA' 96), Rennes.

Felber, H. (1987). *Manuel de terminologie*. Paris, UNESCO.

Firth, J. R., Ed. (1951). *Modes of Meaning*. Papers in Linguistics. Londres, Oxford University Press.

Fletcher, W. (2002). *Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora*. Teaching and Language Corpora 2002.

Fletcher, W. (2004). "Making the Web more useful as a source for linguistic corpora." *LANGUAGE AND COMPUTERS.*

Fletcher, W. (2005). *Towards an Independent Search Engine for Linguists: Issues and Solutions*. Web as Corpus SSMILT Forlì 2005.

Fletcher, W. H. (2001). *Concordancing the Web with KWiCFinder*. Applied Corpus Linguistics 2001.

Fontenelle, F. (1996). *Réseaux sémantiques et dictionnaires bilingues électroniques. Lexicologies dictionnairiques*. Actes du Colloque de Lyon 1995, Beyrouth/Montréal, FMA / AUPELF-UREP.

Fontenelle, T. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen, Niemeyer.

Fujii, A., Ishikawa, T. (2000). "Utilizing the world wide web as an encyclopedia : Extracting

term descriptions from semi-structured text." *Association of Computational Linguistics (ACL)*: 488-495.

Fung, P. (1995). *Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus*. 3rd Annual Workshop on Very Large Corpora, Boston, Massachusetts.

Fung, P., McKeown, K. (1997). *Finding Terminology Translations from Non-parallel Corpora*. Actes de Annual Workshop on Very Large Corpora.

Fung, P., Yee, L. Y. (1998). *An IR approach for translating new words for non-parallel, comparable texts*. Actes de International Conference on Computational Linguistics (COLING).

Fung, P., Ed. (2000). *A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora*. Parallel Text Processing. Dordrecht, Kluwer.

Gala, N., Aït-Mokhtar, S. (2003). *Lexicalising a robust parser grammar using the WWW*. Conference on Corpus Linguistics, Lancaster.

Gala, N. (2003a). Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires, Université de Paris-Sud.

Gala, N. (2003b). *Une méthode non supervisée d'apprentissage sur leWeb pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel*. TALN.

Gale, W., A., Church, K. W. (1993). "A program for aligning sentences in bilingual corpora." *Computational Linguistics* 19(3): 75-102.

Gale, W. A., Church, K. W., Yarowsky, D. (1993). "A method for disambiguating word senses in a large corpus." *Computers and the Humanities* 26: 415-439.

Gaussier, E., Lange, J.-M. (1995). "Modèles statistiques pour l'extraction de lexiques bilingues." *Traitement Automatique des Langues* 36(1-2): 133-155.

Ghani, R., Jones, R. (2000). *Learning a Monolingual Language Model from a Multilingual Text Database*. Ninth International Conference on Information and Knowledge Management (CIKM-2000).

Ghani, R., Jones, R., Mladenic, D. (2001a). *Automatic Web Search Query Generation to Create Minority Language Corpora*. Poster paper in proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001).

Ghani, R. (2001b). Building Minority Language Corpora by Learning to Generate Web Search Queries.

Ghani, R., Jones, R., Mladenic, D. (2001c). *On-line learning for Web query generation: finding documents matching a minority concept on the Web*. Proceedings of the The First

Asia-Pacific Conference on Web Intelligence (WI-2001).

Ghani, R., Jones, R., Mladenic, D. (2001d). *Using the Web to Create Minority Language Corpora*. 10th International Conference on Information and Knowledge Management (CIKM-2001).

Ghani, R., Jones, R., Mladenic, D. (2003). *Building Minority L\anguage Corpora by Learning to Generate Web Search Queries*. KAIS Knowledge and Information Systems.

Gonzalo, J. C., I.; Verdejo, F. (2003). *The web as a resource for WSD*. 1st MEANING Worshop, Espagne.

Granger, S., Ed. (1998). *Prefabriqued Patterns in Andavnced EFL Writing : Collocations and Formulae*. Phraseology : Theory, Analysis and Applications. Oxford, Oxford University Press.

Grefenstette, G. (1999). *The World Wide Web as a Resource for Example-Based Machine Translation Tasks*. ASLIB "Translating and the Computer" conference, Londres, Angleterre.

Greimas, A. (1986). *Sémantique structurale : recherche de méthode*. Paris, PUF.

Greimas, A. J. (1960). "Idiotismes, proverbes, dictions." *Cahiers de lexicologie* 2: 41-61.

Grishman, R. (1994). *Iterative alignment of syntactic structures for a bilingual corpus*. Proceedings of the Second Annual Workshop on Very Large Corpora, Kyoto, Japan.

Gross, G. (1996). *Les expressions figées en français. Noms composés et autres locutions*. Paris, Ophrys.

Grossmann, F., Tutin, A. (2003). "Quelques pistes pour le traitement des collocations." *Travaux et recherches en linguistique appliquée*.

Grundy, V., Ed. (1996). *L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue*. Les dictionnaires bilingues. Louvain-la-Neuve, Duculot.

Guilbert, L. (1965). *La formation du vocabulaire de l'aviation*. Paris, Librairie Larousse.

Guillemin-Flescher, J. (1981). *Syntaxe comparée du français et de l'anglais*. Gap : Ophrys.

Habert, B., Nazarenko, A., Salem, A. (1997). *Les linguistiques de corpus*. Paris, Armand Colin/Masson, U Linguistique.

Habert, B. (2000). *Linguistique sur corpus. Etudes et réflexions*. Perpignan, Presses Universitaires de Perpignan.

Harris, B. (1988). "Bi-text, a new concept in translation theory." *Language Monthly* 54: 8-10.

Harris, Z. (1951). *Methods in Structural Linguistics*. Chicago, University of Chicago Press.

Harris, Z. (1991). *A theory of language and information. A mathematical approach*. Oxford, Oxford University Press.

Hartmann, R. R. K. (1980). Contrastive Textology. Comparative Discourse Analysis in Applied Linguistics (Studies in Descriptive Linguistics 5). J. Gross. Heidelberg.

Hausmann, F. J. (1979). "Un dictionnaire de collocations est-il possible ?" *TraLili* 17(1): 187-195.

Hausmann, F. J. (1989). Le dictionnaire de collocations. *Wörterbücher : ein internationales Hadbuch zur Lexicographie. Dictionaries. Dictionnaires*. R. Hausmann F. J., O.; Wiegand, H. E.; Zgusta, L. Berlin/New-York, De Gruyter: 1010-1019.

Hausmann, F. J., Ed. (1997). *Tout est idiomatique dans les langues*. Langages, La Locution entre langues et usages. Fontenay Saint-Cloud, ENS Editions.

Hiemstra, D. (1998). *Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus*. Proceedings of the eighth CLIN meeting.

Hovy, E., Lin C. Y. (1997). *Automated Text Summarization in SUMMARIST*. Workshop on Intelligent Scalable Text Summarization, Madrid, Espagne.

Howarth, P., Ed. (1998). *The Phraseology of Learners' Academic Writing*. Phraseology : Theory, Analysis and Applications. Oxford, Oxford University Press.

Huang, F., Zhang, Y., Vogel, S. (2005). *Mining key phrase translations from Web Corpora*. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada.

Imbs, P. (1971). *Trésor de la Langue Française. Dictionnaire de la langue du XIXè et du XXè siècles (1789-1960)*. Paris, Editions du CNRS.

Inkpen, D. Z., Hirst, G. (2002). *Acquiring Collocations for Lexical Choice between Near-Synonyms*. Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition - Volume 9, Philadelphia, Pennsylvania.

Isabelle, P. (1992). "La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie." *META* 37(4): 721-737.

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

Jacquemin, C., Bush, C. (2000b). *Fouille du Web pour la collecte d'Entités Nommées*. Actes de la 7eConférence sur le Traitement Automatique des Langues Naturelles.

Jacquemin, C. B., C. (2000a). *Combining Lexical and Formatting Cues for Named Entity Acquisition from the Web*. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics Hong Kong.

Jalabert, F., Lafourcade, M. (2004). *Nommage sens à l'aide de vecteurs conceptuels*. RFIA 2004, Toulouse.

Johansson, S., Ebeling, J., Hofland, K. , Ed. (1993). *Coding and aligning the English-Norwegian parallel corpus*. Languages in Contrast, Papers from a Symposium on Text-based Cross-linguistic Studies. Lund, Lund University Press.

Jones, D. B., Somers, H. L., Ed. (1997). *Bilingual vocabulary estimation from noisy parallel corpora using variable bag estimation*. Recent advances in natural language processing Amsterdam, John Benjamins.

Jones, R., Ghani, R. (2000). *Automatically Building a Corpus of a Minority Language from the Web*.

Kaji, H., Kida, Y., Morimoto, Y. (1992). *Learning translation templates from bilingual text*. Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes, France.

Kaplan, A. (1950). "An experimental study of ambiguity in context." *Mechanical Translation* 1: 1-3.

Katz, J. J., Fodor, J. A., Ed. (1964). *The structure of a semantic theory*. The Structure of Language, chapter 19.

Kay, M., Röscheisen, M. (1988). Text-translation alignment, Technical Report. Xerox Palo Alto Research Center.

Kehoe, A., Renouf, A. (2002). *WebCorp: Applying the Web to linguistics and linguistics to the Web*. WWW2002 Conference, Honolulu, Hawaii.

Kehoe, A., Ed. (2006). *Diachronic Linguistic Analysis on the Web with WebCorp*. The Changing Face of Corpus Linguistics. Amsterdam, Rodopi.

Kehoe, A., Gee, M. (2007) New corpora from the web: making web text more 'text-like'. *Towards Multimedia in Corpus Studies, electronic publication, University of Helsinki*, DOI:

Keller, F., Lapata, M. (2003). "Using the Web to Obtain Frequencies for Unseen Bigrams." *Computational Linguistics* 23(3): 459-484.

Kikui, G. (1998). *Term-list Translation using Mono-lingual Word Co-occurrence Vectors*. Actes de International Conference on Computational Linguistics (COLING).

Kilgarriff, A., Grefenstette, G. (2003). "Introduction to the Special Issue on the Web as

Corpus." *Computational Linguistics* 29(3): 333-348.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). *The Sketch Engine*. EURALEX 2004, Lorient.

Kjaersgaard, P. (1987). *REFTEX. A context-based translation aid*. 3rd conference of the European Chapter of the Association for Computational Linguistics, Copenhagen.

Kjellmer, G. (1994). *A dictionay of English Collocations*. Oxford, Clarendon Press.

Klapaftis, I., Manandhar, S. (2005). *Google & WordNet based Word Sense Disambiguation*. 22 ndICML Workshop on Learning & Extending Ontologies.

Kocourek, R. (1991). *La langue française de la technique et de la science*. Wiesbaden, Brandstetter Verlag.

Kraaij, W., Nie, J-Y., Simard, M. (2003). "Embedding web-based statistical translation models in cross-language information retrieval " *Computational Linguistics, Special issue on web as corpus* 29(3): 381 - 419.

Kupiec, J. (1993). *An algorithm for finding noun phrase correspondences in bilingual corpora*. 31st Annual Meeting of the Association for Computational Linguistics.

L'Homme, M.-C. (2001). *Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe*. L'Impact des nouvelles technologies sur la gestion terminologique.

L'Homme, M. C. (1998). *Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale*. Proceedings EURALEX '98, Université de Liege : Liege.

Labbé, C. L., D. (2003). "La distance intertextuelle." *Corpus, La distance intertextuelle* 2.

Lafourcade, M., Rodrigo, F., Schwab, D. (2004). *Low Cost Automated Conceptual Vector Generation from Mono and Bilingual Resources*. Actes de PAPILLON-2004.

Langlais, P., El-Beze, M. (1997). *Alignement de corpus bilingues : algorithmes et évaluation*. 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF (JST), Avignon.

Lea, D. (2002). *Oxford Collocations Dictionary for Students of English*, Oxford University Press.

Lebarbé, T. (2002). Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative: Université de Caen.

Leech, G. (1991). The state of the art in corpus linguistics. *English corpus linguistics*. A. K. A. B. London: Longman: 8-29.

Léon, J. (2001). "Conception du mot et débuts de la traduction automatique." *Histoire Épistémologie Langage* 23(1): 81-106.

Léon, J., Ed. (2004a). *Lexies, synapsies, synthèmes : le renouveau des études lexicales en France au début des années 1960*. "History of Linguistics in Texts and Concepts" Geschichte der Sprachwissenschaft in Texten und Konzeptionen. Münster, Nodus Publikationen.

Léon, S. (2003). L'extraction automatique des collocations : Une méthode de regroupement par classes conceptuelles. Université de Provence, Aix-en-Provence, Mémoire de maîtrise.

Léon, S. (2004b). Extraction semi-automatique des relations verbe-objet à partir d'un corpus spécialisé : application à la création d'un lexique structuré du TAL. Université de Provence, Aix-en-Provence, Mémoire de DEA.

Léon, S., Millon, C. (2005). *Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web*. Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Dourdan, France.

Léon, S. (2006). *Acquisition automatique de traductions de termes complexes par comparaison de « mondes lexicaux » sur le Web*. Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2006), Louvain, Belgique.

L'Homme, M.-C. (1998). *Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale*. EURALEX '98, Liege.

L'Homme, M.-C. (2001). *Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe*. L'Impact des nouvelles technologies sur la gestion terminologique., Toronto, Université de York.

L'Homme, M.-C. (2002). *What can Verbs and Adjectives can tell us about Terms?* Terminology and Knowledge Proceedings, TKE 2002, Nancy.

L'Homme, M.-C. (2005). "Sur la notion de terme." *Meta* 50(4): 1112-1132.

Li, C., Cao, Y. (2002). Overcoming language barriers in the internet Era. A. F. L. R. A. system. Asia, Report MSR-TR-2002-91 Microsoft Research.

Li, C., Li, H. (2004). *Word translation disambiguation using bilingual bootstrapping*. 40th Annual Meeting of the Association for Computational Linguistics.

Li, H., Cao, Y., Li, C. (2003a). "English Reading Wizard : Mining and Ranking Translations Using Bilingual Data on the Web." *IEEE intelligent systems & their applications* 18(4): 54-59.

Li, H., Cao, Y., Li, C. (2003b). *Using Bilingual Web Data To Mine and Rank Translations*.

IEEE Intelligent Systems.

Lin, C.-Y., Hovy, E. (2000). "The Automated Acquisition of Topic Signatures for Text Summarization." *Actes de COLING Conference*.

Liu, V., Curran, J. R. (2006). *Web Text Corpus for Natural Language Processing*. Proceeding of EACL 2006, 1th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Lu, W.-H., Chien, L.-F., Lee, H.-J. (2001). *Anchor Text Mining for Translation of Web Queries*. Proceedings of the 2001 IEEE International Conference on Data Mining.

Lu, W.-H., Chien, L.-F., Lee, H.-J. (2002). "Translation of Web Queries Using Anchor Text Mining." *ACM Transactions on Asian Language Information Processing (TALIP)* 1(2): 159 - 172.

Lu, W.-H., Chien, L.-F., Lee, H.-J. (2003). "Anchor Text Mining for Translation of Web Queries : A Transitive Translation Approach." *ACM Transactions on Information Systems (TOIS)* 22(2): 242 - 269.

Ma, X., Liberman, M. (1999). *Bits: A method for bilingual text search over the web*. Machine Translation Summit VII, Singapour, Singapour.

Macklovitch, E. (1992). *Corpus-based tools for translators*. 33rd Annual Conference of the American Translators Association, San Diego, California.

Mangeot, M. (2002). *Projet Papillon : intégration de dictionnaires existants et gestion des contributions*. Actes de JST 2002, National Olympic Memorial Youth Center, Tokyo, Japon.

Mangeot, M., Sérasset, G., Lafourcade, M. (2003). "Construction collaborative de données lexicales multilingues, le projet Papillon." *Revue TAL, édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ? (Electronic dictionaries: for humans, machines or both?), Ed. Michael Zock & John Carroll* 44(2): 151-176.

Maniez, F. (2001a). *L'ambiguïté syntaxique due aux structures coordonnées en anglais médical : analyse de la performance d'un logiciel d'aide à la traduction*. TALN 2001, Tours.

Maniez, F. (2001b). "Extraction d'une phraséologie bilingue en langue de spécialité : corpus parallèles et corpus comparables." *Meta* 46(2): 552-563.

Manning, C., Schütze, H. (1999). *Foundations of statistical natural language processing*, MIT Press.

Martinet, A. (1960). *Eléments de linguistique générale*. Paris, Armand Colin.

Martinet, A. (1967). Syntagme et synthème. *La linguistique*. Paris, PUF. 2: 1-14.

Martinet, A. (1968). "Mot et synthème." *Lingua* 21: 294-302.

Martinet, A. (1985). *Syntaxe générale*. Paris, Armand Colin.

Martins-Baltar, M. (1997). *La locution entre langue et usage*. Fontenay, ENS Editions.

Matsumoto, Y., Ishimoto, H., Utsuro, T., Nagao, M. (1993). *Structural matching of parallel text*. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio.

Matsuo, Y., Sakaki, T., Uchiyama, K., Ishizuka, M. (2006). *Graph-based Word Clustering using a Web Search Engine*. 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, Association for Computational Linguistics.

Mautner, G. (2005). "Time to get wired: Using web-corpora in critical discourse analysis." *Discourse & Society* 16(6): 809-828.

McEnery, A., Lange, J.-M., Oakes, M., Véronis, J., Ed. (1997). *The exploitation of multilingual annotated corpora for term extraction*. Corpus Annotation: Linguistic Information from Computer Text Corpora London, Addison Wesley Longman.

McEnery, A. M., Oakes, M.P. (1995). *Sentence and word alignment in the CRATER project : methods and assessment.* . EACL-SIGDAT Workshop, Dublin.

McEnery, T., A. Wilson (1996). *Corpus linguistics*. Edinburgh, Edinburgh University Press.

Mel'cuk, I. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques*. Montréal, Presses de l'Université de Montréal.

Mel'cuk, I., Wanner, L., Ed. (1996). *Lexical Fonctions and Lexical Inheritance for Emotion Lexemes in German*. Lexical Fonctions in Lexicography and Natural Language Processing. Amsterdam / Philadelphia, John Benjamins.

Mel'cuk, I. (1997). Vers une linguistique Sens-Texte, Leçon inaugurale (faite le Vendredi 10 janvier 1997), Collège de France, Chaire internationale.

Melamed, I. D. (1997). *Automatic discovery of non-compositional compounds in parallel data*. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97), Providence.

Melamed, I. D., Ed. (2000). *Bitext maps and alignments via pattern recognition*. Parallel Text Processing. Kluwer, Dordrecht.

Melamed, I. D. (2001). *Empirical Methods for Exploiting Parallel Texts*, MIT Press.

Mel'cuk, I., Wanner, L., Ed. (1996). *Lexical Fonctions and Lexical Inheritance for Emotion Lexemes in German*. Lexical Fonctions in Lexicography and Natural Language Processing. Amsterdam, Benjamins.

Mel'cuk, I., Ed. (1998). *Collocations and Lexical Functions* Phraseology, Theory, Analysis and Applications. Oxford, Clarendon Press.

Mel'cuk, I. (2003). "Collocations : définition, rôle et utilité." *Travaux et recherches en linguistique appliquée*.

Mel'cuk, I. A. C., André; Polguère A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain, Editions Duculot.

Mihalcea, R. (2002). *Bootstrapping large sense tagged corpora*. In Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2002, Canary Islands, Spain.

Mihalcea, R. M., D. I. (1999a). *An automatic method for generating sense tagged corpora*. 16th National Conference on Artificial Intelligence.

Mihalcea, R. M., D. I. (1999b). *A method for word sense disambiguation of unrestricted text*. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland.

Modjeska, N., Markert, K., Nissim, M. (2003). Proceedings of the 2003 conference on Empirical methods in natural language processing - Volume 10.

Morin, E., Dufour-Kowalski, Samuel, Daille, Béatrice (2004). *Extraction de terminologies bilingues à partir de corpus comparables*. Actes de Traitement Automatique des Langues Naturelles (TALN), Fès.

Morley, B., Renouf, A., Kehoe, A. (2003). *Linguistic Research with the XML/RDF aware WebCorp Tool*. WWW2003 Conference, Budapest.

Morley, B., Ed. (2006). *WebCorp: A Tool for Online Linguistic Information Retrieval and Analysis*. The Changing Face of Corpus Linguistics. Amsterdam, Rodopi.

Nagao, M., Ed. (1984). *A framework of mechanical translation between Japanese and English by analogy principle*. Artificial and human intelligence, Elsevier Science Publishers.

Nagata, M. (2001). *Using the Web as a bilingual dictionary*. 39th ACL Worshop on Data-Driven Methods in Machine Translation.

Nakagawa, H. (2001). "Disambiguation of Single Noun Translations Extracted from Bilingual Comparable Corpora." *Terminology* 7(1): 63–83.

Nakov, P., Hearst, M. (2005a). *Search engine statistics beyond the n-gram : Application to noun compound bracket*. CoNLL 2005.

Nakov, P., Hearst, M. (2005b). "Using the Web as an Implicit Training Set: Application to Structural Ambiguity Resolution."

Nerima, L., Seretan, V., Wehrli, E. (2003). "Creating a Multilingual Collocation Dictionary from Large Text Corpora."

Nerima, L., Seretan, V., Wehrli, E. (2006). "Le problème des collocations en TAL." *Nouveaux cahiers de linguistique française* 27: 95-115.

Nie, J.-Y., Simard, M., Isabelle, P., Durand, R. (1999). *Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*. ACM SIGIR'99.

Nie, J.-Y., Cai, J. (2001). *Filtering noisy parallel corpora of web pages*. In IEEE Symposium on Natural Language Processing and Knowledge Engineering, pages, Tucson.

Papageorgiou, H., Ed. (1997). *Clause recognition in the framework of alignment*. Recent advances in natural language processing. Amsterdam, John Benjamins.

Patwardhan, S., Riloff, E. (2006). *Learning Domain-Specific Information Extraction Patterns from the Web*. Proceedings of the Workshop on Information Extraction beyond the Document (ACL-06).

Pearce, D. (2001). *Synonymy in collocation extraction*. In Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistic, Pittsburgh.

Pearson (1998). *Terms in Context*. Amsterdam/Philadelphie, John Benjamins Publishing, .

Picchi, E., Peters, C., Marinai, E. (1992). *A translator's workstation*. 14th International Conference on Computational Linguistics (COLING'92), Nantes.

Pichon, R., Sébillot, P. (1999a). *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience*. Actes de Traitement Automatique des Langues Naturelles (TALN).

Pichon, R., Sébillot, P., Ed. (1999b). *From Corpus to lexicon: from contexts to semantic features*. PALC'99: Practical Applications in Language Corpora, Peter Lang, Lodz studies in Language.

Piperidis, S., Boutsis, S., Papageorgiou, H., Ed. (2000). *From sentences to words and clauses*. Parallel Text Processing. Dordrecht, Kluwer.

Polguère, A. (2000a). "Une base de données lexicale du français et ses applications possibles en français." *Revue de Linguistique et de Didactique des Langues* 21: 75-97.

Polguère, A. (2000b). *Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French*. Actes de EURALEX'2000, Stuttgart.

Polguère, A. (2003). *Lexicologie et sémantique lexicale. Notions fondamentales*, Les Presses de l'Université de Montréal.

Polguère, A. (2005). *Typologie des entités lexicales d'une base de données explicative et combinatoire*. Journée d'étude de l'ATALA « Interface lexique-grammaire et lexiques syntaxiques et sémantiques École nationale supérieure des télécommunications (ENST), Paris.

Pottier, B. (1962a). *Le mot, unité de comportement*. Colloque ATALA Le mot pour la Traduction Automatique et la linguistique appliquée.

Pottier, B. (1962b). "Introduction à l'étude des structures grammaticales fondamentales." *la TA* III(3): 63-91.

Pottier, B. (1962c). "Les travaux lexicologiques préparatoires à la traduction automatique." *Cahiers de lexicologie* 3: 200-206.

Prince, V., Chauché, J. (2006). Translating through divergence : A application to french to english automatic translation. R. L. n. 12758.

Prince, V., Chauché, J. (2008). *Building a Bilingual Representation of the Roget Thesaurus for French to English Machine Translation* Proceedings of the sixth international conference on Language REsources and Evaluation (LREC).

Pu-Jen Cheng, P.-J., Pan, Y.-C.; Lu, W-H., Chien L.-F. (2004b). *Creating multilingual translation lexicons with regional variations using web corpora*. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

Rapp, R. (1995). *Identifying Word Translations in Non-Parallel Texts*. Annual Meeting of the ACL archive, Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, Massachusetts, Association for Computational Linguistics Morristown, NJ, USA.

Rapp, R. (1999). *Automatic Identification of Word Translations from Unrelated English and German Corpora*. Actes de Association for Computational Linguistics (ACL).

Rastier, F. (1987). *Sémantique interprétative*. Paris, Presses Universitaires de France.

Renouf, A., Kehoe, A., Mezquiriz, D., Ed. (2003). *The Accidental Corpus: issues involved in extracting linguistic information from the Web*. Advances in Corpus Linguistics. Amsterdam, Rodopi.

Renouf, A., Ed. (2003). *WebCorp: providing a renewable data source for corpus linguists*. Extending the scope of corpus-based research: new applications, new challenges. Amsterdam, Rodopi.

Renouf, A., Kehoe, A., Banerjee, J. (2005). *The WebCorp Search Engine: a holistic approach to Web text Search*. University of Birmingham.

Renouf, A., Kehoe, A., Banerjee, J., Ed. (2007). *WebCorp: an integrated system for web text*

search. Corpus Linguistics and the Web. Amsterdam, Rodopi.

Resnik, P., Melamed, I.D. (1997). *Semi-Automatic Acquisition of Domain-Specific Translation Lexicons*. Proceedings of the Fifth Conference on Applied Natural Language, Processing (ANLP'97), Washington, DC.

Resnik, P. (1998). "Parallel Strands : A Preliminary Investigation into Mining the Web for Bilingual Text."

Resnik, P. (1999). *Mining the web for bilingual text*. 37th Annual Meeting of the Association for Computational Linguistics.

Resnik, P., Smith, N. A. (2003). "The Web as a parallel corpus." *Computational Linguistics, Special issue on web as corpus* 29(3): 349 - 380.

Resnik, P. S. N. (2002). The Web as a parallel corpus. *Technical Report UMIACS-TR-2002*.

Riloff "Extraction-based text categorization: generating domain-specific role relatonships automatically."

Roberts, R. P., Montgomery, C. (1996). *The Use of Corpora in Bilingual Lexicography*. Actes d'EURALEX '96.

Rossignol, M., Sébillot, P. (2003). "Extraction statistique sur corpus de classes de mots-clés thématiques." *TAL (Traitement automatique des langues)* 44(3): 217-246.

Rosso, P., Montes, M., Buscaldi, D., Pancardo, A., and Villaseñor, A., (2005). *Two Web-based approaches for Noun Sense Disambiguation*. Int. Conf. on Comput. Linguistics and Intelligent Text, Processing, CICLing-2005,, Mexico D.F., Mexico, Springer Verlag, LNCS (3406).

Rundell (2000). "The biggest corpus of all." *Humanising Language Teaching*(3).

Rundell, M., Ed. (2002). *Macmillan English Dictionary for Advanced Learners*, Macmillan.

Rus, V., Ravi, S. (2006). "Towards a base noun phrase parser using web." *Journal of Computing Sciences in Colleges* 21(5): 162-169.

Sadler, V. (1989). Translating with a simulated bilingual knowledge bank, Technical report. BSO/Research. Utrecht.

Sajous, F., Tanguy, L. (2006). *Repérage de créations lexicales sur le Web francophone*. Journée d'étude de l'ATALA Paris.

Salton, G. (1968). *Automatic Information Organisation and Retrieval*. New York, McGrawHill.

Santamaría, C., Gonzalo, J., Verdejo F. (2003). "Automatic Association of Web Directories

with Word Senses." *Computational Linguistics* 23(3): 485-502.

Sato, S., Nagao, M. (1990). *Toward memory-based translation*. 12th International Conference on Computational Linguistics, COLING'90, Helsinki, Finland.

Sato, S., Sasaki, Y. (2003). *Automatic collection of related terms from the Web*. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2, Sapporo, Japan.

Saussure, F. (1916). *Cours de Linguistique générale*, Payot.

Schütze, H. (1998). "Automatic word sense discrimination." *Computational Linguistics* 24(1): 97-124.

Schwab, S., Lafourcade, M., Prince, V. (2004). *Hypothèses pour la construction et l'exploitation conjointer d'une base lexicale sémantique basée sur les vecteurs conceptuels*. JADT 2004, 7es Journées internationales d'Analyse statistique des Données Textuelles, Louvain-le-Neuve, Belgique.

Sébillot, P., Pichon, R. (1997). Acquisition automatique d'informations lexicales à partir de corpus : un bilan. *INRIA*. N. RR-3321.

Séguéla, P. (2001). Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, Université de Toulouse : Thèse de doctorat en informatique.

Seretan, V., Nerima, L., Wehrli, E. (2004). *Using the Web as a Corpus for the Syntactic-Based Collocation Identification*. International Conference on Language Resources and Evaluation (LREC 2004), Lisbonne, Portugal.

Sharoff, S., Ed. (2006). *Creating general-purpose corpora using automated search engine queries*. Wacky! Working papers on the Web as Corpus. Bologna, GEDIT.

Simard, M., Foster, G., Isabelle, P. (1992). *Using cognates to align sentences in bilingual corpora*. Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI), Montréal, Canada.

Sinclair, J., Ed. (1987a). *Looking up: An account of the COBUILD project in lexical computing*. London, Collins.

Sinclair, J., Ed. (1987b). *Collocation : A Progress Report*. Language Topics. Essays in Honour of Michael Halliday, Vol. 2. Amsterdam, John Benjamins.

Sinclair, J. (1996). Preliminary recommandations on corpus typology. E. E. A. G. o. L. E. S. Technical report.

Smadja, F., McKeown, K. (1991). "Using collocations for language generation." *Computational Intelligence* 7(4): 229 - 239.

Smadja, F. (1993). "Retrieving Collocations from Text : Xtract " *Computational Linguistics* 19(1).

Smadja, F., McKeown, K.; Hatzivassiloglou, V. (1996). "Translating Collocations for Bilingual Lexicons: A Statistical Approach." *Computational Linguistics* 22(1): 1-38.

Smarr, J., Grow, T. (2002). "GoogleLing: The Web as a Linguistic Corpus."

Sta, J. D. (1995). "Comportement statistique des termes et acquisition terminologique à partir de corpus." *Revue TAL, Traitements probabilistes et corpus* 36(1-2): 119-132.

Sumita, E., Iida, H, Kohyama, H. (1990). *Translating with examples : a new approach to machine translation*. Actes de International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI'90), Austin, Texas.

Tanguy, L. (1997). Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration informatique d'un modèle de la sémantique interprétative. *Thèse de doctorat*, Ecole Nationale Supérieure des Télécommunication de Bretagne

Tanguy, L. (1999). *Isotopies sémantiques pour la vérification de traduction*. Traitement Automatique des Langues Naturelles (TALN).

Terra, E., Clarke, L. A. (2003). *Frequency Estimates for Statistical Word Similarity Measures*. HLT-NAACL 2003.

Thoiron, P., Béjoint, H. (1989). "Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle." *Meta* 34(4): 661-671.

Tonoike, M. K., Utsuro, T. (2005). "Effect of domain-specific corpus in compositional translation estimation for technical terms."

Turney, P., Littman, M. (2003). "Measuring Praise and Criticism: Inference of Semantic Orientation from Association." *ACM Transactions on Information Systems*.

Turney, P. (2004). *Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities*. Proc. SENSEVAL-3.

Turney, P. D. (2001). *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*. Twelfth European Conference on Machine Learning Berlin: Springer-Verlag.

Tutin, A., Grossmann, Francis (2002). "Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif." *Revue française de linguistique appliquée, Lexique : recherches actuelles* VII: 7-25.

Ueyama, Ed. (2006). *Creation of general-purpose Japanese Web corpora with different search engine query strategies*. WaCky! Working papers on the Web as corpus. Bologna, Gedit.

Van Der Eijk, P. (1993). *Automating the Acquisition of Bilingual Terminology*. Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93), Utrecht.

Vandeghinste, V. (2002). *Resolving PP Attachment Ambiuguities Using the WWW*. CLIN2002 Abstracts, Groningen.

Verlinde, S., Selva, T., Binon, J., Ed. (2003). *Les collocations dans les dictionnaires d'apprentissage : repérage, présentation et accès, dans Les collocations : analyse et traitement*. Travaux et recherches en linguistique appliquée. Amsterdam, de Werelt.

Véronis, J., Ed. (2000a). *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht, Kluwer Academic Publishers.

Véronis, J., Ed. (2000b). *Annotation automatique de corpus : panorama et état de la technique*. Paris, France, Hermès édition.

Véronis, J. (2003). *Cartographie lexicale pour la recherche d'information*. Actes de la Conférence Traitement Automatique des Langues (TALN'2003), Batz-sur-Mer, France, ATALA.

Véronis, J. (2004). "HyperLex: lexical cartography for information retrieval." *Computer Speech & Language* 18(3): 223–252.

Volk, M. (2000). *Scaling up. Using the WWW to resolve PP attachment ambiguities*. Konvens-2000.

Volk, M. (2001). *Exploiting the www as a corpus to resolve pp attachment ambiguities*. Corpus Linguistics 2001.

Volk, M. (2002). "Using the web as a corpus for linguistic research." *Catcher of the Meaning. A festschrift for Professor Haldur Õim*(R. Pajusalu, & T. Hennoste (Eds)).

Weaver, W. ([1949] 1955). Translation. *Machine Translation of Languages, Fourteen Essays*. W. N. Locke, Booth, A. Donald. Boston, MIT & John Wiley: 15-23.

Wehmeir, N. W. (2004). Using web search for machine translation University of Leeds School of Computing

Wehrli, E. (2004). *Traduction, traduction de mots, traduction de phrases*. TALN 2004, Fès.

Wilks, Y. A., Ed. (1975). *Preference Semantics*. The Formal Semantics of Natural Language. Cambridge University Press.

Williams, G. (1999). Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique, Université de Nantes.

Williams, G. (2001). *Sur les caractéristiques de la collocation*. TALN 2001, Tours.

Wu, D., Ed. (2000). *Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammars*. Parallel Text Processing. Dordrecht, Kluwer.

Wu, J.-C., Chang, J. S. (2007). *Learning to find English to Chinese Transliterations on the Web*. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague.

Wüster, E., Ed. (1981). *L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses*. Textes choisis de terminologie. I. Fondements théoriques de la terminologie. Québec, GIRSTERM.

Xu, J. L. (2000). *Multilingual search on the World Wide Web*. In Proceedings of the Hawaii International Conference on System Science (HICSS-33), Maui, Hawaii.

Yang, C., Li, K.W. (2003). "Automatic construction of English/Chinese parallel corpora." *Journal of the American Society for Information Science and Technology* 54(8): 730 - 742.

Yarowsky, D. (1993). *One Sense per Collocation*. Actes de ARPA Human Language Technology Workshop.

Zhang, Y., Vines, P. (2004). *Detection and translation of OOV terms prior to query time*. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval Sheffield, United Kingdom.

Zhang, Y., Vines, P. (2005). *Mining translations of OOV terms from the web through cross-lingual query expansion*. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil.

Zuraw, K. (2006). *Using the Web as a phonological corpus : a case study from Tagalog*. EACL-2006: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics/Proceedings of the 2nd International Workshop on Web As Corpus.