Synergie des approches et des ressources déployées pour le traitement de l'écrit

Emmanuel Morin

▶ **To cite this version:**

Emmanuel Morin. Synergie des approches et des ressources déployées pour le traitement de l'écrit. Sciences de l'ingénieur [physics]. Université de Nantes, 2007. tel-00482893

HAL Id: tel-00482893

https://theses.hal.science/tel-00482893v1

Submitted on 11 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE NANTES
UFR Sciences et Techniques

Synergie des approches et des ressources déployées pour le traitement de l'écrit

Dossier d'Habilitation à Diriger des Recherches

ÉCOLE DOCTORALE STIM 366
SCIENCES ET TECHNOLOGIES DE L'INFORMATION DES MATERIAUX
Spécialité : Informatique

Présentée et soutenue publiquement par

Emmanuel **Morin**

le 30 novembre 2007

au Laboratoire d'Informatique de Nantes Atlantique

devant le jury ci-dessous

Président

M. Viard-Gaudin Christian	*Professeur des universités*	Université de Nantes

Rapporteurs

M. Gaussier Éric	*Professeur des universités*	Université Joseph Fourier
M. Grefenstette Gregory	*Ingénieur de recherche*	Commissariat à l'Énergie Atomique
M. Zweigenbaum Pierre	*Directeur de recherche*	LIMSI CNRS

Examinateurs

Mme Daille Béatrice	*Professeur des universités*	Université de Nantes

Directeur de recherche : Mme Daille Béatrice

Laboratoire LINA - FRE CNRS 2729
Laboratoire d'Informatique de Nantes Atlantique
2, rue de la Houssinière - BP 92208 - 44322 Nantes cedex 3

Remerciements

Je tiens tout d'abord à remercier Éric Gaussier, Gregory Grefenstette et Pierre Zweigenbaum d'avoir rapporté sur ce travail ainsi que Béatrice Daille et Christian Viard-Gaudin d'avoir accepté d'être membres du jury.

Je tiens aussi à remercier les membres de l'équipe Langage Naturel du LINA pour leur soutien et à exprimer ma gratitude à Christine Jacquin, Laura Monceaux et Annie Tartier pour la relecture attentive de ce manuscrit.

Je remercie aussi l'ensemble des étudiants de DEA, Master 2 recherche et thèse côtoyés pour la stimulation procurée : Samuel Dufour-Kowalsky, Nordine Fourour, Lorraine Goeuriot, Freddy Perraud, Emmanuel Prochasson et Sebastián Peña Saldarriaga.

Enfin, un dernier merci aux sans nom et sans-grade, notamment Jim, pour leur soutien...

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Ce mémoire synthétise les activités de recherche menées au sein de l'équipe Traitement Automatique du Langage Naturel (TALN) du Laboratoire d'Informatique de Nantes Atlantique (LINA – FRE CNRS 2729) depuis la fin de ma thèse, en décembre 1999. Les travaux présentés ici se situent au carrefour de l'informatique et de la linguistique. Ils sont tout autant le fruit de collaborations scientifiques stimulantes que de rencontres humaines enrichissantes.

C'est le goût pour la terminologie qui m'a d'abord conduit à collaborer avec Béatrice Daille (équipe TALN, LINA). Notre première coopération était liée à la reconnaissance des entités nommées pour le français (Daille et Morin, 2000). L'originalité de ce travail concernait la définition d'une catégorisation fine et exhaustive des entités nommées ainsi que la prise en compte de leurs variations. Les premiers jalons, que nous avons posés, ont été suivis par Nordine Fourour (2004) dans le cadre de sa thèse sur l'identification et la catégorisation automatiques des entités nommées pour des textes français et du système associé *NEMESIS*. Notre seconde coopération s'inscrit en fouille terminologique multilingue. Au printemps 2002, je cherchais à étendre le mécanisme de variation sémantique monolingue, développé à la fin de ma thèse (Morin, 1999; Morin et Jacquemin, 1999), à un cadre bilingue. De son côté, Béatrice s'intéressait à l'alignement de termes complexes identifiés par son système *ACABIT* (Daille, 2002). En outre, elle avait déjà travaillé en traduction automatique (Langé et al., 1997) et avait étudié des langues variées, notamment le *malgache* pour l'acquisition terminologique (Daille et al., 2000) et le *zarma* pour la constitution d'un corpus spécialisé. De cette curiosité commune pour les aspects multilingues est née une collaboration étroite en acquisition lexicale multilingue où j'ai largement pu apprécier et bénéficier des remarquables intuitions linguistiques de Béatrice. Les objectifs que nous nous étions fixés dénotaient une volonté farouche de s'affranchir des limites de la traduction compositionnelle et du souhait de pouvoir aligner des termes de longueurs différentes. Cela nous a conduit à la spécification, puis au développement d'une chaîne de fouille terminologique multilingue constituée d'un extracteur de termes complexes et d'un module d'alignement fondé sur une méthode statistique exploitant le contexte des termes. Cette chaîne doit beaucoup, notamment du point de vue informatique, au travail préalable réalisé par Samuel Dufour-Kowalski (2003) au cours de son DEA. Depuis lors, nous n'avons eu de cesse de poursuivre ce travail (Morin et Daille, 2004; Morin et al., 2004; Daille et Morin, 2005). D'une part, Béatrice avait l'intuition que la qualité des données textuelles pouvait non seulement suppléer à leur quantité mais qu'elle garantissait aussi celle

des ressources lexicales extraites. Le projet TCAN-DECO[1] a permis de vérifier cette hypothèse tout en montrant l'intérêt de prendre en compte le type du discours lors de la phase de constitution d'un corpus comparable pour obtenir des listes terminologiques de qualité (Morin et Daille, 2006; Morin et al., 2007). D'autre part, le Memorandum of Understanding (MOU) entre le National Institute of Informatics (NII, Japon) et l'Université de Nantes, nous permet de poursuivre ce travail pour des couples de langues à grande distance linguistique et de travailler de manière étroite et enrichissante avec Koichi Takeuchi (Université d'Okayama) et Kyo Kageura (Université de Tokyo).

Une seconde rencontre, au cours de l'hiver 2001, avec Christian Viard-Gaudin de l'équipe Image Vidéo Communication (IVC) de l'Institut de Recherche en Communications et en Cybernétique de Nantes (IRCCyN – UMR 6597) a donné lieu au développement d'un nouvel axe de recherche. Christian, dont les activités de recherche s'inscrivent dans le champ de la reconnaissance de formes, s'intéressait à l'apport d'informations linguistiques pour la reconnaissance de l'écriture manuscrite en ligne. Dans cette optique, une première coopération a été initiée par le co-encadrement du DEA de Freddy Perraud (2002), puis de sa thèse CIFRE (Perraud, 2005) soutenue par la société nantaise Vision Objects[2]. Dans le cadre de ces travaux, nous cherchions à développer des modèles probabilistes de langage associés à un moteur de reconnaissance de l'écriture manuscrite en ligne. En outre, les modèles développés devaient pouvoir s'intégrer dans des engins nomades n'offrant qu'une faible capacité de stockage (Perraud et al., 2003a; Perraud, 2005). Au terme de ce travail, nous avons eu l'idée avec Christian de poursuivre notre collaboration sur un aspect novateur lié à la reconnaissance de mini-messages manuscrits (c'est-à-dire une forme de SMS manuscrits). Ce travail, réalisé dans le cadre du projet AtlanStic-MIMEMA[3], nous a permis d'amorcer une première réflexion sur le sujet et de développer un premier modèle (Prochasson et al., 2007a,b), notamment dans le cadre du Master de Emmanuel Prochasson (2006). La collaboration entre les équipes IVC et TALN et la société Vision Objects est toujours d'actualité, puisque nous débutons deux projets : un premier en technologies logicielles ANR-CIEL[4] et un second dans l'axe multimédia du CPER-MILES[5].

[1] TCAN est un programme interdisciplinaire du CNRS impliquant les départements scientifiques : STIC, SDV et SHS. Le projet DECO (Découverte et Exploitation de Corpus Comparables) dont le LINA est maître d'œuvre a été retenu suite à l'appel à proposition 2004 et couvre les thèmes de *Multilinguisme et diversité culturelle* et *Web sémantique*. Les partenaires associés sont l'INALCO (Paris), XEROX Research Centre Europe (Grenoble) et le NII (Tokyo, Japon).

[2] La société Vision Objects, qui est née en 1998 à la suite d'un essaimage de l'équipe IVC de l'IRRCyN, est aujourd'hui l'un des acteurs majeurs sur le marché de la reconnaissance de l'écriture manuscrite.

[3] AtlanStic est une fédération de recherche (FR CNRS 2819) regroupant les laboratoires LINA (FRE CNRS 2729), IRRCyN (UMR CNRS 6597) et IREENA (EA 1770) dans laquelle le projet MIMEMA (Mini Messages Manuscrits) s'inscrit pour la période 2005/2007.

[4] Le projet ANR-CIEL (Conversion Indexation de l'Ecriture en Ligne) a été retenu suite à l'appel à proposition en technologies logicielles 2006. Il s'agit d'un projet exploratoire ayant pour objectif de concevoir des technologies adaptées au traitement de documents manuscrits complexes (c'est-à-dire des documents contenant à la fois du texte mais aussi des schémas, des tableaux et autres composants de nature bidimensionnelle).

[5] Le projet CPER 2007-2009 MILES (Multimédia – Ingénierie du Logiciel – aidE à la décision – Télécommunication, Détection et LocaliSation) se propose de participer à la création d'un pôle européen de recherche en STIC dans l'Ouest de la France. Il est porté par la fédération de recherche AtlanStic (qui réunit le LINA, l'IRCCyN et IREENA) et associe, outre les laboratoires de la fédération, les laboratoires LERIA et LISA à Angers, et les laboratoires LIUM et LAUM au Mans. Le projet s'articule autour de quatre axes : i) Multimédia, ii) Aide à la décision et logistique, iii) Ingénierie du logiciel, et iv) Télécommunications, Détection et Localisation.

Ce mémoire s'articule donc naturellement autour de ces deux axes de recherche, celui de la fouille terminologique multilingue et celui de la reconnaissance de l'écriture manuscrite en ligne, et se décline en cinq chapitres. Le chapitre 2 est consacré au premier axe. Il repositionne notre travail dans le champ de l'acquisition lexicale multilingue, présente les approches développées et les résultats obtenus. Le chapitre 3 développe le second axe. Il présente les différentes facettes de la reconnaissance de l'écriture manuscrite auxquelles nous nous sommes intéressés et les modèles développés. Le chapitre 4, qui constitue le trait d'union entre ces deux axes, indique la synergie possible entre les approches et ressources déployées. En particulier, nous montrons que les méthodes probabilistes ne sont plus une alternative aux systèmes à base de règles, mais bien complémentaires et que les ressources exploitées doivent être adaptées à la tâche visée. Enfin, nous concluons et proposons différentes perspectives de recherche dans le chapitre 5.

Chapitre 2

Fouille terminologique multilingue

Depuis les travaux fondateurs de Rapp (1995), Tanaka et Iwasaki (1996), puis Fung et McKeown (1997), l'acquisition de lexiques bilingues à partir de corpus comparables a connu un essor important. Cet intérêt pour l'exploitation de corpus comparables est principalement lié aux difficultés rencontrées pour disposer de corpus parallèles lorsqu'il s'agit d'exploiter un matériau textuel qui ne fait pas intervenir l'anglais. En outre, les lexiques bilingues obtenus à partir de corpus parallèles sont quelque peu biaisés. En effet, un corpus parallèle étant constitué d'un texte dans une langue source et de sa traduction dans une langue cible, le vocabulaire rencontré dans la partie traduite est fortement influencé par celui de la langue source en particulier dans les domaines spécialisés.

Dans la suite de ce chapitre, nous commençons par rappeler en section 2.1 les fondements théoriques en acquisition lexicale multilingue. Nous décrivons en section 2.2 la plate-forme de fouille terminologique multilingue développée et les méthodes mises en œuvre. La section 2.3 présente les acquis engrangés à la lumière des différents travaux réalisés. Enfin, la section 2.4 propose une synthèse de ce champ de recherche.

2.1 Fondements en acquisition lexicale multilingue

D'un point de vue théorique, l'acquisition lexicale multilingue se situe dans le prolongement des travaux réalisés dans un cadre monolingue. En cela, elle trouve son ancrage dans l'héritage de la sémantique distributionnelle de Harris (1968) qui met en relation la distribution syntaxique des mots avec leur signification : « *the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities*[1] » (Harris, 1968, p. 12). Cette approche correspond à une vision plus stricte que celle du linguiste anglais Firth (1957) réduite communément à la citation « *You shall know a word by the company it keeps*[2] ». Ainsi le préalable à l'alignement

[1] « *la signification des unités et de leurs relations grammaticales est liée à la restriction imposée sur les combinaisons de ces unités avec d'autres* »

[2] « *On peut juger un mot à ses fréquentations* »

lexical bilingue est l'appréhension d'un mot au regard de ses fréquentations[3] dans un contexte monolingue. Un retour au cadre monolingue s'impose donc comme préalable à l'acquisition lexicale multilingue.

2.1.1 Acquisition lexicale monolingue

Dans un cadre monolingue, la mise en œuvre de l'hypothèse distributionnelle (Harris, 1968) repose sur l'identification d'« affinités du premier ordre » : « *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word*[4] » (Grefenstette, 1994b, p. 279). Pour ce faire, il est nécessaire de disposer d'un « contexte d'observation » des mots afin d'identifier des traits pertinents. Ce contexte est étudié à travers le prisme d'une *fenêtre* de granularité plus ou moins importante. Il peut s'agir d'un document, d'un paragraphe, d'une phrase ou simplement de quelques mots. En ce qui concerne l'identification des traits pertinents, il est d'usage de se limiter aux « mots pleins » par opposition aux « mots vides » (articles, conjonctions, prépositions...) qui ne sont pas pris en compte. En outre, une opération préalable de lemmatisation du texte est souvent réalisée pour se ramener à la forme canonique des mots pleins. La figure 2.1.a est une illustration des traits observés pour le mot *forêt* dans une fenêtre contextuelle de 7 mots (c'est-à-dire trois mots avant et après celui-ci). L'ensemble des traits observés et le nombre associé d'occurrences pour le texte étudié forment alors un « vecteur de contexte » pour le mot visé comme cela est illustré en figure 2.1.b. Ce vecteur sera d'autant plus fourni, et par conséquent représentatif de la variété des usages par rapport aux fréquentations observées pour un mot donné, que le texte étudié sera de taille importante.

À ce niveau, une attention particulière doit être accordée à la taille du contexte d'observation : « *La taille de la fenêtre dépend des relations sémantiques que l'on recherche, les cooccurrences à petite, moyenne et grande distance tendant respectivement à faire ressortir des expressions figées ou semi-figées [...], des contraintes de sélection [...] et des mots appartenant au même champ sémantique* » (Habert et al., 1997, p. 179). Ce paramètre doit être manié à bon escient. En amont de toute approche distributionnelle, il va conditionner fortement les traits observés et les relations identifiées. Il n'existe malheureusement pas, à notre connaissance, une étude précise sur ce phénomène. Dans le meilleur des cas, la taille de la fenêtre est fixée de manière plus ou moins empirique.

Afin d'identifier les mots caractéristiques du contexte lexical et de supprimer l'effet induit par leur fréquence, l'association entre le mot observé et ses cooccurrences est normalisée sur la base d'une mesure de récurrence contextuelle comme l'Information Mutuelle (Fano, 1961) ou le Taux de vraisemblance (Dunning, 1993). Les figures 2.1.c et 2.1.d donnent un aperçu des vecteurs obtenus après normalisation du vecteur de contexte du mot *forêt* pour les deux précédentes mesures[5]. Ici encore, la mesure de récurrence contextuelle utilisée influence

[3] L'idée de *fréquentation* semble ici sémantiquement plus appropriée que celle de *voisin* dans la mesure où elle prend en considération pour un mot donné ceux avec lesquels il ne souhaite pas être vu, et non uniquement ceux avec lesquels il est vu.

[4] « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné.* »

[5] Dans ces exemples, les hapax ne sont pas pris en compte car trop peu représentatif de la variétés des usages.

Etant donné l'évolution démographique et les tendances de la production et de la consommation, la demande de bois de la forêt boréale ne peut qu'augmenter.

consommation	demande	bois	forêt	boréal	pouvoir	augmenter

(a) Fenêtre contextuelle

tropical	1708
forestier	1168
aménagement	846
naturel	768
bois	678
humide	615
pays	504
département	472
exploitation	459
développement	434
zone	428
production	428
service	407
terre	370
produit	348
million	347
utilisation	344
valeur	325
arbre	323
région	318
...	...

(b) Vecteur observé

oligarchique	5,4
luquillo	5,4
frc	5,4
wytham	5,2
mayombe	5,2
subméditerranéen	5,2
semi−décidue	5,2
submontagnard	5,2
holdsworth	5,2
melaleuca	5,2
semi−décidu	5,1
dw	5,0
onf	5,0
wallaba	5,0
sempervirent	4,9
semidécidu	4,9
fout−adjalon	4,9
panamensis	4,9
écorégion	4,9
gaulois	4,9
...	...

(c) Vecteur normalisé (Information Mutuelle)

tropical	1460,3
humide	770,9
aménagement	538,8
département	496,2
naturel	434,1
dense	415,0
superficie	249,7
domanial	225,4
sempervirent	205,5
durable	186,4
ombrophile	179,7
destruction	168,9
million	161,4
artificiel	152,9
hectare	152,5
productif	142,7
exploitation	129,9
rôle	122,9
protection	118,0
commission	113,5
...	...

(d) Vecteur normalisé (Taux de vraisemblance)

FIG. 2.1 – Observation, représentation et normalisation du contexte du mot *forêt* pour le corpus [SYLV]

considérablement les observables. Ainsi, l'Information Mutuelle a tendance à faire remonter les associations avec les mots de faible fréquence (Daille, 1994).

En raison de sa facilité de mise en œuvre, mais aussi de sa robustesse face aux données bruitées, cette technique de cooccurrence a souvent été utilisée en lexicographie (Church et Hanks, 1990; Smadja, 1993; Grefenstette, 1994a) ou encore en recherche documentaire (Rajman et al., 2000).

Cette technique sert aussi de point d'appui à l'identification d'« affinités du second ordre » : « *Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar*[6] » (Grefenstette, 1994b, p. 280). La mise en exergue de ces « mots similaires » repose sur

[6] « *Les affinités du second ordre dévoilent quels mots partagent les mêmes environnements. Les mots partageant des affinités du second ordre n'ont pas besoin d'apparaître ensemble, mais leurs environnements sont*

l'exploitation d'une mesure de distance vectorielle comme le Cosinus (Salton et Lesk, 1968) ou le Jaccard (Tanimoto, 1958) pour évaluer le nombre de traits partagés ou non par deux mots. De cette manière, il est possible de fournir des associations de mots (Hindle, 1990; Ruge et Schwarz, 1991) ou de construire des classes de mots (Agarwal, 1995; Habert et al., 1996). Comme la similitude conceptuelle exploitée ici est un « lien neutre », les classes sémantiques obtenues n'ont pas de signification *a priori* ni de représentant privilégié. Elles regroupent des entités linguistiques hétérogènes dont les liens sémantiques sont parfois lointains (Morin, 1999).

2.1.2 Acquisition lexicale bilingue

Les recherches en acquisition de lexiques bilingues à partir de corpus se sont initialement concentrées sur l'exploitation de corpus parallèles, c'est-à-dire un ensemble de textes accompagnés de leur traduction (Véronis, 2000). À partir de textes alignés phrases à phrases, des techniques symboliques (Carl et Langlais, 2002), statistiques (Gaussier et Langé, 1995) ou mixtes (Daille et al., 1994) sont mises en œuvre pour aligner les mots et les constituants de ces phrases. Cette mise en correspondance au niveau des structures phrastiques ou lexicales d'un corpus parallèle n'est plus opérationnelle sur un corpus comparable. En outre, les textes parallèles demeurent des ressources rares, principalement pour des couples de langues ne faisant pas intervenir l'anglais. Leur utilisation soulève aussi le problème de la qualité des traductions obtenues : un corpus parallèle étant composé de textes sources et traduits, le vocabulaire rencontré dans les traductions est fortement influencé par le vocabulaire de la langue source, en particulier pour les domaines scientifiques ou techniques.

Pour ces différentes raisons, des travaux récents se tournent vers l'exploitation de corpus comparables : « *sets of texts in different languages, that are not translations of each other*[7] » (Bowker et Pearson, 2002, p. 93). Le terme *comparable* signifie que ces textes partagent des caractéristiques communes comme la période, le domaine, le thème, le support médiatique, le type de discours, etc. (Baayen, 1994). Dans un domaine scientifique, cette proximité thématique peut s'interpréter par un vocabulaire scientifique commun, les termes, reflétant les concepts du domaine étudié. Cette propriété du partage lexical est centrale en exploitation de corpus comparables. Une définition très applicative de celle-ci a été donnée par Déjean et Gaussier (2002, p. 7) : « *Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1* ».

Les principaux travaux visant à l'automatisation de l'acquisition de lexiques bilingues à partir de corpus comparables se focalisent sur l'alignement de termes simples bilingues. Un terme simple bilingue consiste dans la langue source comme dans la langue cible en un seul mot plein. Ainsi, le terme simple français *manteau* se traduit dans le domaine de la foresterie par le terme simple anglais *mantle*, dans le domaine de la marine par *shield* ou dans l'habillement par *coat*. Les méthodes employées pour extraire ces couples bilingues héritent globalement des acquis engrangés dans un cadre monolingue. Elles s'appuient sur l'analyse du contexte de ces mots simples et sur l'hypothèse qu'un mot et sa traduction partagent un même contexte

semblables. »

[7] « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres* »

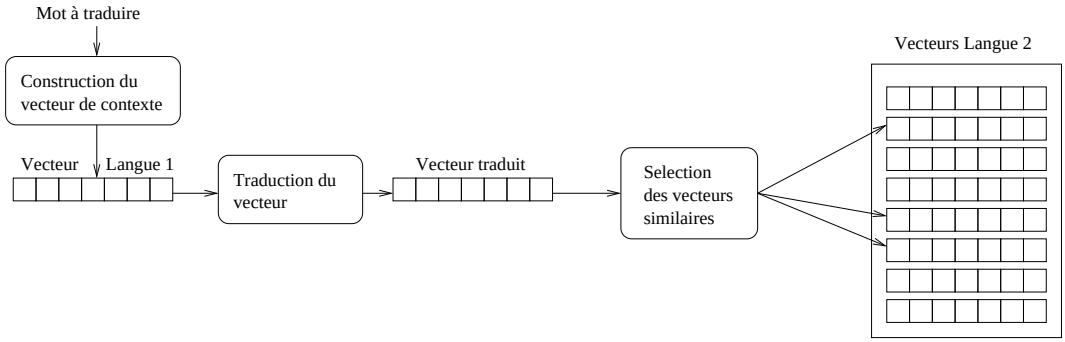


FIG. 2.2 – Approche par traduction directe

lexical. Ainsi pour nos trois couples, trois contextes différents sont attendus. Nous pouvons imaginer que ces contextes pourraient être représentés par les éléments lexicaux suivants donnés en français :

- *manteau/mantle* : végétation, forêt, bois...
- *manteau/shield* : bateau, mer, construction...
- *manteau/coat* : vêtement, froid, mettre...

À chaque mot dont la traduction doit être identifiée dans le texte cible est associé un ensemble de mots qui forment son vecteur de contexte. Les traductions potentielles du mot sont celles dont les vecteurs de contexte en langue cible sont les plus proches du vecteur de contexte en langue source traduit au sens d'une mesure vectorielle. La traduction du vecteur de contexte source est obtenue en traduisant chaque élément de ce vecteur à l'aide d'un dictionnaire bilingue. Cette méthode, qualifiée d'« approche par traduction directe » par Déjean et Gaussier (2002), est illustrée en figure 2.2.

Avec cette méthode, Fung (1998) extrait des couples de termes simples anglais/chinois avec une précision de 76 % sur les 20 premiers candidats proposés en exploitant deux ans du Wall Street Journal et du quotidien chinois Nikkei Financial News. Rapp (1999) porte cette précision à 89 % sur les 10 premiers candidats en exploitant des couples de termes simples anglais/allemand à partir d'un corpus journalistique de 85 millions de mots. Ici l'évaluation porte sur des mots très fréquents de la langue comme *bébé*, *femme*, *musique*, *religion*... là où Fung (1998) utilise des mots moins fréquents et aussi des noms propres. En ce qui concerne l'alignement de groupes nominaux relevant du domaine général, une première approche a été proposée pour des termes anglais/japonais par Shahzad et al. (1999) où l'évaluation se limite à une liste de dix termes. Une deuxième approche plus aboutie, qui s'appuie sur l'algorithme EM (*Expectation-Maximisation*), a été développée par Cao et Li (2002) pour l'extraction de groupes nominaux bilingues anglais/chinois. Ces derniers obtiennent une précision de 91 % sur les 3 premiers candidats en exploitant le web. Ces résultats, qui restent inférieurs à ceux obtenus avec un corpus parallèle, permettent néanmoins d'envisager des applications concrètes notamment en recherche d'information interlangue.

En ce qui concerne les domaines spécialisés, les résultats obtenus pour des couples de termes simples sont moins significatifs (cf. tableau 2.1). Le premier obstacle est bien sûr lié à la difficulté de disposer de corpus aussi volumineux que pour les domaines de langue

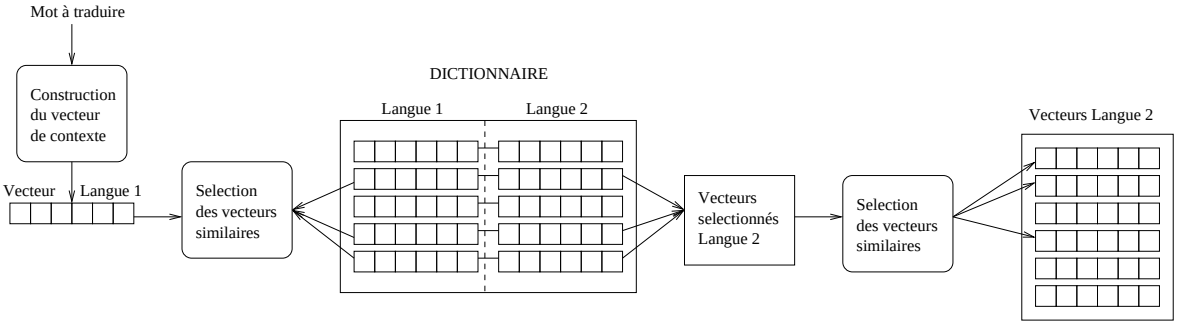


FIG. 2.3 – Approche par similarité interlangue

générale qui ne permet pas de construire des vecteurs de contexte fortement nourris. Le second s'explique par l'appauvrissement du vecteur de contexte d'une unité à traduire lors de son transfert en langue cible si un trop grand nombre d'éléments ne peuvent être traduits.

Référence	*Domaine/Langues/Taille (Mmots)*	TOP_{10}	TOP_{20}
Chiao et Zweigenbaum (2002)	médical/français-anglais/1,2	61	94
Déjean et Gaussier (2002)	médical/français-allemand/0,1	44	57
Morin et Daille (2006)	diabète/français-japonais/1,6	51	60
Morin (2007)	sylviculture/français-anglais/4,9	43	47

TAB. 2.1 – Résultats d'alignements de termes simples pour des corpus comparables spécialisés

Une amélioration de l'approche par traduction directe a été proposée par Déjean et Gaussier (2002) pour pallier l'insuffisance des ressources bilingues utilisées lors de la traduction des vecteurs de contexte et donc d'éviter d'avoir trop de contextes lexicaux non traduits. Cette méthode appelée « approche par similarité interlangue » associe comme contexte des mots à traduire les contextes des mots présents dans le dictionnaire qui pourront être traduits (cf. figure 2.3). Avec cette méthode, Déjean et Gaussier (2002) obtiennent pour des termes simples français/allemand une précision pour les 10 et 20 meilleurs candidats de 43 % et 51 % pour un corpus médical de 100 000 mots (respectivement 44 % et 57 % avec la méthode directe) et de 79 % et 84 % pour un corpus des sciences sociales de 8 millions de mots (respectivement 35 % et 42 % avec la méthode directe).

2.1.3 Plaidoyer pour les termes complexes

Comme nous venons de le rappeler, la fouille terminologique multilingue à partir de corpus comparables s'est d'abord focalisée sur l'exploitation de corpus de langue générale avant de se tourner vers des corpus de langue de spécialité. Dans les deux cas, les ressources bilingues extraites sont le plus souvent limitées à des listes terminologiques constituées de termes simples. Or en domaine spécialisé, les termes complexes sont plus précis et reflètent tout autant de la spécificité du domaine que les termes simples (Wagner, 1991). Il est donc surprenant que si peu de travaux s'intéressent à ce problème tant il semble pertinent par nature.

Nous nous intéressons donc à l'extraction de termes complexes et à leur traduction pour des corpus spécialisés et cherchons à prendre en compte les problèmes suivants :

Fertilité Les termes simples et complexes ne se traduisent pas systématiquement par un terme de même longueur. Par exemple, le terme complexe *peuplement forestier* est traduit en anglais par le terme simple *crop* et le terme *essence d'ombre* par le terme *shade tolerant species*. Ce problème bien connu, décrit par Brown et al. (1993) sous le terme de *fertility* (*fertilité*), est rarement pris en compte dans le processus d'acquisition de lexiques bilingues. Une hypothèse de traduction « mot à mot » étant le plus souvent adoptée.

Non-compositionnalité Quand un terme complexe est traduit par un terme de même longueur, la traduction n'est pas toujours obtenue par la simple traduction de ses composants (Melamed, 2001). Par exemple, le terme *plantation énergétique* est traduit en anglais par *fuel plantation*, où *fuel* n'est pas la traduction de *énergétique*. Cette propriété est aussi connue sous le terme de « non-compositionnalité ».

Variation Un même terme peut se présenter sous différentes formes suite à des variations morphologique, syntaxique ou encore sémantique (notamment dans le cas de la synonymie) (Jacquemin, 1999, 2001; Daille, 2003b). Les variations des termes doivent donc être prises en compte dans le processus de traduction. Par exemple, les termes français *aménagement de la forêt* et *aménagement forestier* sont traduits par le même terme en anglais : *forest management*.

Ces trois problèmes de fertilité, non-compositionnalité et variation, sont au cœur même de nos travaux en fouille terminologique bilingue focalisés sur l'alignement de termes complexes. Dans le prochaine section, nous proposons et détaillons la méthode mixte mise en œuvre pour cette tâche et la plate-forme associée.

2.2 Plate-forme de fouille terminologique multilingue

La plate-forme de fouille terminologique multilingue que nous avons développée permet à partir d'un corpus comparable en deux langues de proposer une liste de termes simples et complexes et leurs traductions candidates. L'architecture présentée en figure 2.4 est modulaire et est constituée d'un extracteur de termes dans chaque langue et d'un programme d'alignement lexical.

2.2.1 Extraction terminologique

Les termes complexes sont identifiés au sein de chaque partie monolingue du corpus comparable à l'aide de l'outil d'acquisition terminologique fonctionnant sur le français, l'anglais et le japonais : *ACABIT*[8]. Outre sa disponibilité dans le domaine public, *ACABIT* possède la spécificité de prendre en compte les variantes de termes complexes (flexionnelles, morphologiques, syntaxiques, morpho-syntaxiques) (Daille, 2003b). Il ne nécessite aucune ressource linguistique propre et est indépendant du domaine traité. Il est robuste car il permet de trai-

[8] `http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/`

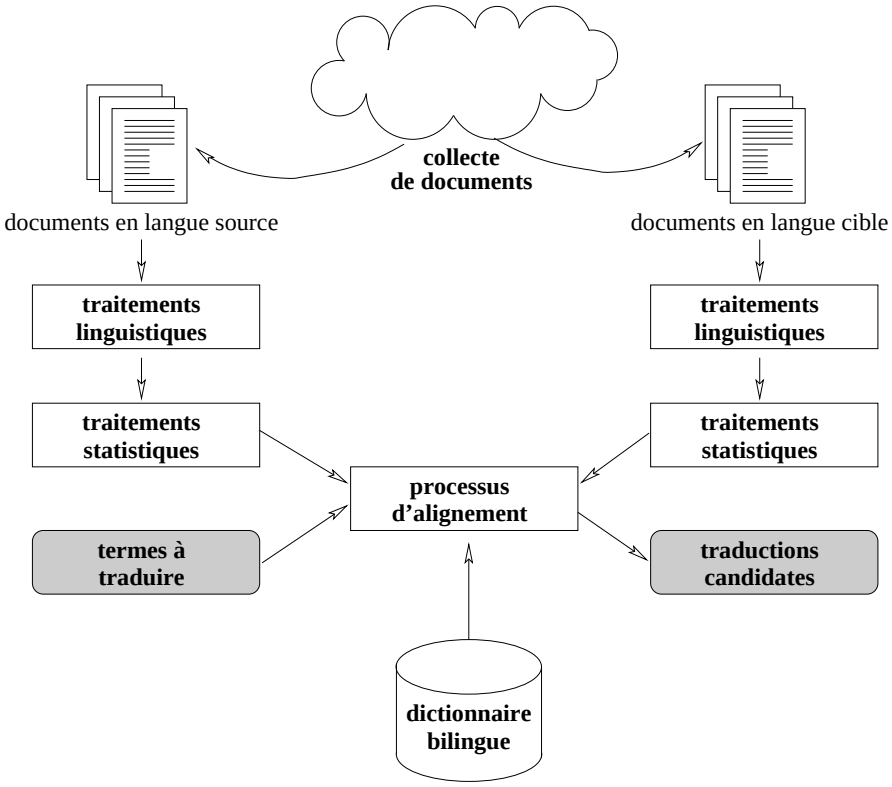


FIG. 2.4 – Plate-forme de fouille terminologique multilingue

ter des corpus de plusieurs millions de mots. *ACABIT* s'applique sur un corpus ayant subit les pré-traitements suivants :

- nettoyage des données indésirables (tableaux, caractères de contrôle, etc.) ;
- segmentation en occurrences de formes et de phrases ;
- étiquetage linguistique : assignation aux formes de leur étiquette grammaticale et de leur lemme.

ACABIT effectue en premier une analyse syntaxique locale : il parcourt le corpus étiqueté et relève les occurrences de candidats termes ou de leurs variations qui épousent des structures morpho-syntaxiques prédéfinies. Les différentes occurrences qui réfèrent à un terme binaire ou à l'une de ses variantes sont listées sous un même candidat terme. Ainsi le candidat terme *produit forestier* a été identifié sous les différentes occurrences suivantes :

- **forme de base** : *produit forestier* ;
- **variante graphique** : *produit fo-restier*, *pro-duit forestier* ;
- **variante flexionnelle** : *produits forestiers* ;
- **variante syntaxique (modification)** : *produit non forestier*, *produit alimentaire forestier*, *produit fini d'origine forestière*, *produit ligneux non forestier* ;
- **variante syntaxique (coordination)** : *produit halieutique et forestier*, *produit agricole ou forestier*, *le produit et le service forestier*.

De même, les candidats termes *produit de la forêt*, *produit agroforestier*, *non-produit agroforestier* et *sous-produit forestier*, *sous-produit de la forêt* ont aussi été identifiés.

Dans une deuxième phase, *ACABIT* réalise un regroupement sémantique des candidats termes à l'aide des opérations suivantes :

Fusion de deux candidats termes Deux candidats termes sont fusionnés s'ils correspondent à des variantes morphologiques dérivationnelles synonymiques. Ainsi les candidats termes *produit de la forêt* et *produit forestier* n'en forment plus qu'un.

Dissociation de certaines variantes Les variantes syntaxiques qui induisent une distance sémantique sont extraites de la liste des variantes du candidat terme et sont considérées comme de nouveaux candidats termes. Il s'agit des modifications par insertion d'un adverbe de négation qui induisent une relation d'antonymie comme *produit non forestier* avec *produit forestier* et des insertions d'adjectifs relationnels qui induisent une relation d'hyponymie comme *produit alimentaire forestier* avec *produit forestier* (Daille, 2003a).

Regroupement de candidats termes Tous les candidats termes liés par morphologie dérivationnelle ou par variation induisant une distance sémantique forment un groupement de candidats termes. Ainsi, seront regroupés les candidats termes : *produit forestier/produit de la forêt*, *produit non forestier*, *non-produit agroforestier*, *produit agroforestier*, *sous-produit forestier/sous-produit de la forêt*, *produit alimentaire forestier* et *produit forestier*.

Dans la suite du traitement, nous considérons comme variantes de termes uniquement celles qui sont listées sous le même candidat terme. Cette opération, qui correspond à une normalisation terminologique, améliore le découpage du texte en unités de sens au même titre que la lemmatisation au niveau morphologique. De cette manière, nous ne traduisons pas un terme mais une classe d'équivalences de termes. À l'issue de cette première phase, les termes complexes identifiés par *ACABIT* dont la fréquence est supérieure à deux occurrences sont considérés comme une seule unité textuelle, dans le cas contraire ils sont décomposés en mots simples.

2.2.2 Alignement terminologique

Le processus d'alignement lexical peut être réalisé soit en adoptant la « méthode par traduction directe » (Fung, 1998; Peters et Picchi, 1998; Rapp, 1999) soit la « méthode par similarité interlangue » (Déjean et Gaussier, 2002; Morin et Daille, 2004). Nous détaillons précisément ici le principe de chacune de ces méthodes, en mettant l'accent sur le processus de transfert des termes de la langue source vers la langue cible qui en est la clef de voûte.

2.2.2.1 Méthode par traduction directe

Notre implémentation de la méthode par traduction directe, illustrée en figure 2.5.a, se décompose en quatre étapes :

étape 1 : Identification des contextes lexicaux

Pour chaque langue du corpus comparable, le contexte de chaque unité lexicale[9] i est

[9]Nous employons ici le terme *unité lexicale* pour désigner la forme lemmatisée d'un mot, d'un terme simple

extrait en repérant les unités qui apparaissent autour de i dans une fenêtre contextuelle de n mots[10]. Afin d'identifier les unités lexicales caractéristiques des contextes lexicaux et de supprimer l'effet induit par la fréquence des unités lexicales, nous normalisons l'association entre les unités lexicales sur la base d'une mesure de récurrence contextuelle comme *Information Mutuelle* (Fano, 1961) ou *Taux de vraisemblance* (Dunning, 1993) (cf. équations 1 et 2 et tableau 2.2). Après normalisation, à chaque élément j du vecteur de contexte de l'unité i nous attachons le taux d'association $assoc^i_j$.

étape 2 : Transfert d'une unité à traduire

Le transfert d'une unité k à traduire de la langue source à la langue cible repose sur la traduction de chacun des éléments de son vecteur de contexte au moyen d'un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de l'unité k l'ensemble des traductions proposées (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Dans le cas où l'élément ne figure pas dans le dictionnaire, il ne sera pas exploité dans le processus de traduction.

En fonction de l'adéquation du dictionnaire bilingue avec le corpus d'étude, plus ou moins d'éléments du vecteur de contexte seront traduits. L'unité à traduire sera d'autant plus discriminante en langue cible que le nombre d'éléments traduits de son vecteur de contexte sera important.

étape 3 : Identification des vecteurs proches de l'unité à traduire

Le vecteur de contexte v_k ainsi traduit est ensuite comparé à l'ensemble des vecteurs de contexte de la langue cible en s'appuyant sur une mesure de distance vectorielle comme *Cosinus* (Salton et Lesk, 1968) ou *Jaccard* (Tanimoto, 1958) (cf. équations 3 et 4).

étape 4 : Obtention des traductions candidates

En fonction des précédentes valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour l'unité k.

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

TAB. 2.2 – Table de contingence

$$IM(i,j) = log\frac{a}{(a+b)(a+c)} \tag{1}$$

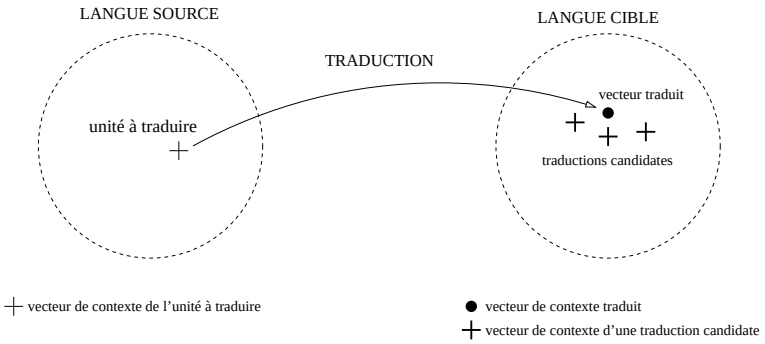
$$\begin{aligned} TV(i,j) = \; & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \end{aligned} \tag{2}$$

ou d'un terme complexe. Par abus de langage, le terme *unité simple* fait référence à un mot ou un terme simple et *unité complexe* à un terme complexe.

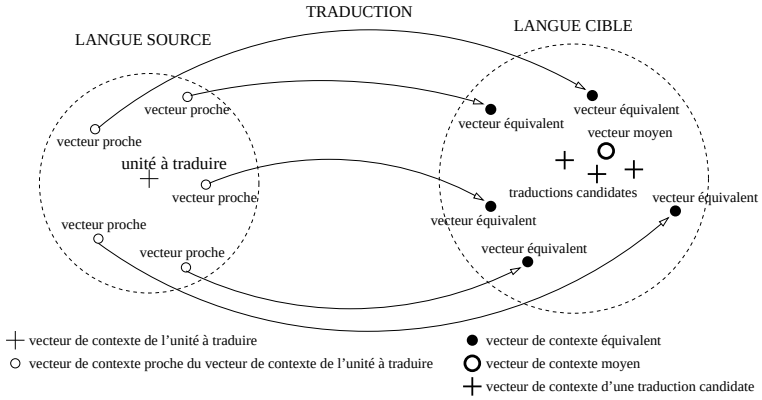
[10]En ce qui concerne les mots vides, ils ne sont pas exploités dans le processus d'alignement.

$$cosinus_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t assoc_t^{l\,2}} \sqrt{\sum_t assoc_t^{k\,2}}} \tag{3}$$

$$jaccard_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t assoc_t^{l\,2} + \sum_t assoc_t^{k\,2} - \sum_t assoc_t^l \, assoc_t^k} \tag{4}$$



(a) Approche par traduction directe



(b) Approche par similarité interlangue

FIG. 2.5 – Processus de transfert d'une unité à traduire de la langue source à la langue cible

2.2.2.2 Méthode par similarité interlangue

Dans le cadre de la méthode directe, le transfert du vecteur de contexte d'une unité k à traduire de la langue source à la langue cible repose sur la traduction de chaque élément de v_k au moyen d'un dictionnaire bilingue. En fonction de l'adéquation du dictionnaire bilingue avec le corpus d'étude plus ou moins d'éléments de v_k vont être traduits. Ainsi, si aucun élément de v_k ne peut être traduit, le vecteur transféré sera vide et il ne sera pas trouvé de traduction pour l'unité k. Dans ce type d'approche – où il y a toujours des éléments de v_k qui ne peuvent être traduits – l'unité k perd naturellement de son potentiel de discrimination dans la langue cible. Pour limiter cet effet, l'usage d'un dictionnaire spécialisé éventuellement associé à un dictionnaire de langue générale est souvent incontournable (Chiao et Zweigenbaum, 2003).

Afin d'éviter les inconvénients de la traduction directe des vecteurs de contexte avec un dictionnaire bilingue, à savoir l'inadéquation des ressources bilingues au corpus et l'impossibilité de traduire certains éléments des vecteurs de contexte, nous avons proposé de réaliser le transfert d'une unité k à traduire en nous appuyant sur les autres unités de la langue source qui lui sont « proches » (Morin et Daille, 2004; Morin et al., 2004). Les unités proches que nous cherchons à identifier, qui correspondent à des « affinités du second ordre » (Grefenstette, 1994b), n'apparaissent pas ensemble mais partagent le même environnement lexical. Par exemple, les unités *pieu*, *piquet* et *poteau* ont certainement des vecteurs de contexte très similaires sans pour autant apparaître dans les vecteurs de contexte des uns et des autres. Le principe proposé par Déjean et Gaussier (2002) consiste alors à s'appuyer sur les unités qui partagent le même environnement lexical que l'unité k à traduire pour assurer le transfert direct de v_k de la langue source à la langue cible. Le dictionnaire va ainsi permettre de « traduire » les vecteurs de contexte dans leur globalité et non élément par élément. De cette manière, les vecteurs de contexte transférés ne perdent pas de leur potentiel de discrimination en langue cible. Dans cette méthode, appelée « approche par similarité interlangue », le dictionnaire bilingue est mieux exploité puisque des traductions candidates peuvent être proposées pour une unité k à traduire même si aucun élément de v_k ne peut être traduit. Ainsi, les inconvénients de la méthode par traduction directe sont contournés.

Dans Déjean et Gaussier (2002), les vecteurs de contexte sont calculés pour chaque unité simple du corpus et pour chaque entrée de la ressource bilingue. Dans le cas où l'entrée du dictionnaire bilingue est une unité simple, le vecteur de contexte associé est celui issu du corpus. S'il s'agit d'une unité complexe, le vecteur de contexte est calculé comme étant la conjonction des vecteurs de contexte des mots simples le composant puisque les vecteurs de contexte sont constitués d'unités simples. Par exemple, si la ressource bilingue comporte comme entrée le terme complexe *forêt boréale*, son vecteur de contexte sera uniquement composé des unités qui se trouvent à la fois dans les vecteurs de contexte de *forêt* et de *boréale* extraits du corpus. Dans notre approche, nous pouvons calculer directement à partir du corpus les vecteurs de contexte des unités simples et complexes.

Notre implémentation de cette méthode, illustrée en figure 2.5.b, reprend les étapes 1 et 4 de la méthode par traduction directe (cf. section précédente) et adapte les étapes 2 et 3 de la manière suivante :

étape 2 : Transfert d'une unité à traduire

Le transfert d'une unité k à traduire repose, dans un premier temps, sur l'identification en langue source des unités lexicales dont les vecteurs de contexte lui sont similaires en exploitant une mesure de distance vectorielle comme Cosinus ou Jaccard. Le dictionnaire bilingue est ensuite utilisé pour assurer la traduction directe des unités similaires à k. Nous obtenons ainsi des vecteurs de contexte équivalents attestés en langue cible. Dans le cas où le dictionnaire bilingue propose plusieurs traductions pour une unité, comme il est essentiel de prendre en compte l'ensemble des vecteurs de contexte issus des différentes traductions, nous en réalisons l'union[11]. Le vecteur de contexte résultant est alors composé de l'ensemble des éléments des différents vecteurs de contexte originaux. Si plusieurs vecteurs ont un élément commun, alors le taux d'association de

[11] Cette technique correspond à l'approche couramment adoptée lorsque les traductions ne sont pas ordonnées dans le dictionnaire bilingue.

cet élément sera le plus grand des taux d'association des différents éléments. De cette manière, nous prenons en compte l'ensemble des traductions possibles et pas seulement la plus courante.
Ici encore, si l'unité lexicale n'est pas présente dans le dictionnaire, elle ne sera pas exploitée dans le processus de traduction. Néanmoins, dans le cadre de cette méthode la traduction n'altère pas le vecteur de contexte puisque ce dernier est transféré directement de la langue source à la langue cible. Si une unité ne peut être traduite, elle ne sera pas prise en compte dans la recherche des vecteurs proches de l'unité à traduire. Cela ne pénalise pas, *a priori*, la recherche des unités similaires en langue cible.

étape 3 : Identification des vecteurs proches de l'unité à traduire
En s'appuyant sur les vecteurs de contexte équivalents précédemment obtenus, nous calculons en langue cible un vecteur de contexte moyen. Celui-ci est obtenu par le barycentre des vecteurs de contexte équivalents pondéré par le coefficient de similarité issu de la langue source (c'est-à-dire le score donné par la mesure de distance vectorielle). Ce vecteur de contexte moyen est ensuite comparé à l'ensemble des vecteurs de contexte de la langue cible en exploitant une mesure de similarité comme Cosinus ou Jaccard.

Notre approche diffère de celle proposée par (Déjean et Gaussier, 2002) dans la mesure où pour identifier les vecteurs proches de l'unité à traduire nous ne comparons pas l'ensemble des vecteurs équivalents aux vecteurs de contexte originaux de la langue cible. Notre approche, qui permet de faire passer les calculs d'une complexité polynomiale à une complexité linéaire, ne pénalise pas à mon sens la recherche des unité similaires puisque le vecteur de contexte moyen de la langue cible peut être interprété comme le vecteur de contexte de la meilleure traduction candidate.

2.3 Résultats engrangés en acquisition lexicale multilingue

Les différents résultats engrangés en acquisition lexicale multilingue sont le fruit d'expériences réalisées pour des domaines spécialisés et des langues variées, celui de la sylviculture — noté [SYLV] — pour des textes français/anglais, d'une part, et celui du diabète — noté [DIAB] — pour des documents français/japonais, d'autre part. À défaut de disposer d'un corpus comparable de référence, qui puisse servir à la communauté pour comparer les approches développées et résultats obtenus, nous avons toujours eu le souci de décrire avec précision les données utilisées (corpus, dictionnaires et listes de référence) et de comparer systématiquement nos résultats avec ceux que nous obtenions pour l'alignement de termes simples. Pour une description détaillée des données exploitées, nous renvoyons les lecteurs aux articles publiés dans la revue TAL (Morin et Daille, 2004, 2006) ainsi qu'à l'annexe A.

À la lumière des différentes expériences réalisées et des travaux étudiés, cette section présente et commente les différentes acquis que nous avons engrangés dans un contexte multilingue.

2.3.1 Se méfier des apparences

Si d'un point de vue méthodologique, l'alignement de termes complexes se situe dans le prolongement des méthodes développées pour l'acquisition de termes simples, les outils et les développements informatiques nécessaires sont plus lourds à mettre en œuvre. En outre, la prise en compte des termes complexes dans le processus d'alignement augmente les dimensions de l'espace vectoriel associé. D'un autre côté, l'apport attendu par cette prise en compte mérite une attention particulière.

2.3.1.1 Prendre en compte les aspects pratiques et techniques

En ce qui concerne les aspects informatiques, il ne faut pas négliger le temps nécessaire à la mise en place de la plate-forme de fouille terminologique multilingue. D'une part, cette chaîne nécessite dans sa version minimaliste, des outils d'étiquetage et de lemmatisation de corpus comme préalable à l'extraction terminologique. En outre, ces outils doivent être disponibles pour des langues variées et pouvoir fonctionner de concert. En ce qui concerne les données informatiques manipulées, il faut travailler avec des documents de formats distincts (principalement HTML et PDF) mais plus encore avec des encodages de fichiers variés. Dans la mesure où nous travaillons avec le japonais, le format le plus opportun est bien sûr l'UNICODE. Néanmoins, ce nouveau standard de codage est loin d'être la norme sur le web. Par exemple lorsque nous avons constitué la partie japonaise du corpus [DIAB], nous avons constaté que 70 % des textes collectés étaient encodés sous Shift-JIS (encodage par défaut sous Windows), 10 % sous EUC-JP (encodage UNIX/LINUX), 5 % sous ISO-2022-JP (encodage standard des industries japonaises), l'UTF-8 ne représentant que 1 %. À noter que nous n'avons pu normaliser 14 % des documents qui présentaient un encodage inconnu.

Le fonctionnement de ces outils en cascade est aussi une source de bruit, puisque les erreurs de segmentation, d'étiquetage et de lemmatisation se propagent et s'amplifient d'outils en outils pour finalement se révéler au niveau de l'outil d'extraction de terminologie. Par exemple, une analyse réalisée sur 100 termes complexes choisis aléatoirement, qui ne sont pas des hapax, dans la partie française du corpus [DIAB], a mis en évidence que 17 % des termes proposés par *ACABIT* sont inexploitables : 9 % sont des termes mal orthographiés ou segmentés (p. ex. *diabtique de type 2*), 4 % des termes anglais (p. ex. *young adult*), 4 % des termes incomplets (p. ex. *ministère chargé*) ou encore des termes incohérents (p. ex. *laitier lait*). La même analyse réalisée pour des mots simples donne des résultats proches : 10 % sont des mots mal orthographiés ou segmentés ou encore des mots anglais.

Un autre aspect plus technique, lié à l'espace vectoriel construit, mérite une attention particulière. En effet, lorsque l'on s'intéresse à l'alignement de termes complexes, les unités lexicales manipulées sont tout autant des mots ou termes simples que des termes complexes. Ainsi, l'intégration des termes complexes dans les vecteurs de contexte augmente le nombre de dimensions de l'espace vectoriel tout en diminuant la représentativité numérique des unités simples. Par exemple, le terme simple *débardage*, issu de la partie française du corpus [SYLV], passe d'une fréquence de 544 dans un espace vectoriel limité aux unités simples à 144 dans une espace intégrant des unités simples et complexes puisqu'il apparaît dans différentes termes complexes comme *débardage mécanique*, *piste de débardage*, *technique de débardage*. Pour

« contourner cet obstacle » et limiter la construction des vecteurs de contexte à des unités simples, deux approches sont possibles. Prenons par exemple la phrase suivante : « *La forêt boréale perdrait ainsi 37 pour cent de sa superficie [...]* » et le terme complexe *forêt boréale* à traduire. Une première « *approche amont* » consiste à s'appuyer sur l'identification en corpus du terme complexe pour construire son vecteur de contexte, qui serait alors limité aux unités *perdre* et *superficie*, puis à associer aux autres unités des contextes composés d'unités simples (p. ex. pour *superficie* : *forêt*, *boréal* et *perdre*). Une deuxième « *approche aval* » consiste pour chaque unité simple à lui associer uniquement les unités simples de son contexte (p. ex. *forêt* : *boréal*, *perdre*, *superficie* ; *boréal* : *forêt*, *perdre*, *superficie* ; etc.), puis à construire le vecteur de contexte d'un terme complexe à traduire comme étant la conjonction des vecteurs de contexte des mots simples le composant (ce qui correspond à la technique utilisée par Déjean et Gaussier (2002)). Dans les deux cas, le vecteur de contexte associé à *forêt boréale* est globalement le même (à quelques occurrences liées au décalage de la fenêtre contextuelle). Mais pour l'approche amont l'espace vectoriel comporte toujours les contextes initialement calculés pour les unités *forêt* et *boréale* et pour l'approche aval nous conservons les contextes des composants dans le contexte des unités *perdre* et *superficie*. Pour la méthode par similarité interlangue, ce type de subterfuge risque de faire ressortir les composants d'une unité complexe à traduire comme étant les unités les plus similaires. À ce jour, nous n'avons pas réalisé d'expériences précises sur ce point afin de déterminer la situation la plus favorable.

Dans la mesure où les termes complexes sont moins ambigus, plus précis que les mots ou termes simples et très présents dans les textes spécialisés, leur prise en compte dans le calcul des contextes lexicaux ne devrait même pas porter à discussion. Pourtant la réalité est bien loin de confirmer cette affirmation...

2.3.1.2 Vérifier l'apport des termes complexes à l'alignement lexical

Afin de vérifier l'apport des termes complexes à l'alignement, nous avons réalisé une expérience simple en nous appuyant sur les termes simples de la liste de référence (*i.e.* [lexique 1] – cf. annexe A) associée au corpus [SYLV] pour la méthode directe. D'un côté, les vecteurs de contexte sont limités à des unités simples, et de l'autre, les vecteurs de contexte sont composés d'unités simples et complexes.

Le tableau 2.3 présente pour l'approche par traduction directe les résultats de cette expérience où nous indiquons, pour les deux configurations visées, le nombre de traductions trouvées (NB_{trad}), la position moyenne et son écart type pour les traductions trouvées dans la liste ordonnée de traductions proposées (MOY_{pos}) et (ECT_{pos}), ainsi que le nombre de traductions trouvées dans les dix et vingt meilleures positions (TOP_{10} et TOP_{20}).

	NB_{trad}	MOY_{pos}	ECT_{pos}	TOP_{10}	TOP_{20}
Unités simples	62	20,3	34,9	43	47
Unités simples et complexes	50	12,5	19,9	33	39

TAB. 2.3 – Apport des unités complexes à l'alignement

Si nous nous appuyons sur le nombre de traductions trouvées ainsi que sur les valeurs de

TOP_{10} et TOP_{20}, les résultats obtenus en intégrant les termes complexes dans la construction des vecteurs de contexte sont globalement inférieurs à ceux décelés sans eux. Il est néanmoins délicat de comparer directement entre eux ces résultats dans la mesure où nous ne travaillons plus dans le même espace vectoriel. Afin de mieux comprendre la nature des résultats obtenus en intégrant les termes complexes au contexte lexical et aussi d'affiner notre intuition initiale, nous avons cherché à vérifier un ensemble d'hypothèses pour le corpus [SYLV] que nous restituons ici.

Hypothèse 2.3.1 *« Les termes complexes ne sont pas suffisamment présents dans les vecteurs de contexte. »*

Pour vérifier cette première hypothèse, nous avons mesuré la proportion d'unités simples par rapport aux unités complexes dans les vecteurs de contexte en nous limitant aux 100 premières entrées. Cette proportion est d'environ 50 % en langue source comme en langue cible. En outre, le nombre d'unités pivots de notre corpus est de 7 352 unités simples pivots et de 6 769 unités complexes pivots sur un nombre total de 55 013 unités simples et complexes. Il apparaît donc que cette hypothèse est non fondée puisque les termes complexes sont aussi nombreux que les unités simples dans les vecteurs de contexte.

Hypothèse 2.3.2 *« Les termes complexes ne sont pas des éléments "porteurs de sens" dans les vecteurs de contexte. »*

Le fait que les termes complexes soient suffisamment présents dans les vecteurs de contexte ne permet pas de juger de leur potentiel de discrimination par rapport aux unités simples. Afin de vérifier cette hypothèse, nous avons inspecté les vecteurs de contexte des termes du lexique de référence dans le cas d'une construction limitée à des unités simples, puis d'une construction intégrant des unités simples et complexes. D'une manière générale, les vecteurs de contexte construits à partir d'unités simples et complexes sont plus précis que ceux construits uniquement à partir d'unités simples. Ceci est essentiellement dû à une meilleure représentation du contexte lexical des mots. Par exemple, le vecteur de contexte du terme *débardage* comporte le terme complexe *tracteur à chenille* qui est plus discriminant que les mots pleins le composant (cf. colonnes 1 et 2 du tableau 2.4). Les termes complexes sont bien des éléments « porteurs de sens ». Ils semblent donc pertinents dans la construction des vecteurs de contexte.

Afin de conforter cette affirmation, nous avons réalisé une nouvelle expérience où nous limitons la description des vecteurs de contexte aux seules unités complexes. D'une manière générale, les termes complexes des vecteurs de contexte ainsi décrits ont un « rapport sémantique » plus ou moins proche et ne semblent pas correspondre à une distribution aléatoire (cf. colonne 3 du tableau 2.4). D'un point de vue numérique, 12 traductions correctes ont été identifiées à partir du lexique de référence parmi les 100 premières traductions candidates (cf. tableau 2.5). Les résultats de traduction obtenus ici — qui doivent être interprétés avec prudence dans la mesure où les termes complexes sont en moyenne moins fréquents que les unités simples — sont bien en deçà de ceux décelés précédemment (cf. tableau 2.3). Dans cette expérience, nous avons constaté que peu de termes complexes des vecteurs de contexte

Unités simples	*Unités simples et complexes*	*Unités complexes*
tracteur	groupement	tracteur à chenille
câble	abattage	coût de le abattage
distance	chargement	traction animal
abattage	transport	utilisation de le traction
groupement	tracteur à chenille	équipement mécanique
arche	coût de le abattage	progrès ultérieur
montée	distance	progrès réalisé
chenille	manuel sur le méthode	homme au brélage
coût	tronçonnage	homme dirigeant
piste	tracteur	énergie économique

TAB. 2.4 – Premières entrées du vecteur de contexte de *débardage*

	NB_{trad}	MOY_{pos}	ECT_{pos}	TOP_{10}	TOP_{20}
Unités complexes	12	17,5	28,2	6	8

TAB. 2.5 – Apport des unités complexes à l'alignement en se limitant aux unités complexes

de la langue source sont transférés en langue cible. Nous devons donc déterminer l'importance de ce phénomène dans le processus de traduction.

Hypothèse 2.3.3 *« Les termes complexes des vecteurs de contexte de la langue source ne sont pas correctement transférés en langue cible ».*

Pour évaluer la quantité d'unités complexes des vecteurs de contexte de la langue source transférée en langue cible, nous avons considéré les 100 premières entrées des vecteurs de contexte associées au [lexique 1] (dans cette expérience les vecteurs de contextes sont décrits par des unités simples et complexes). Pour chaque vecteur de contexte associé à un terme du lexique de référence, nous avons ainsi compté le nombre d'unités simples et complexes traduites et identifiées en langue cible. Le tableau 2.6 illustre cette opération de transfert pour les dix premières entrées du vecteur de contexte associé au terme *débardage*. Nous constatons que la proportion de termes complexes dans les vecteurs de contexte traduits passe de 50 % à environ 20 %. Plus de la moitié des termes complexes ne sont pas conservés pendant la phase de transfert. Comparativement, 92 % des unités simples des vecteurs de contexte de la langue source sont conservées en langue cible.

En ce qui concerne la qualité des termes complexes traduits, nous ne l'avons pas jugée particulièrement mauvaise, quoique limitée le plus souvent aux traductions compositionnelles ajoutées au dictionnaire. Dans l'ensemble, il ne semble pas y avoir de décalage sémantique fortement marqué. Les difficultés de traduction des termes complexes semblent essentiellement liées à l'insuffisance de dictionnaire. Par exemple, le terme *tracteur à chenille* du vecteur de contexte du terme *débardage* ne peut être traduit directement à partir de notre dictionnaire bilingue, son transfert en langue cible repose sur la traduction compositionnelle de ses

Vecteur de contexte français	*Vecteur de contexte traduit*
groupement	group, grouping
abattage	logging
chargement	load, loading, consignment
transport	carriage, haulage, ferrying, transport, transportation
tracteur à chenille	crawler tractor
coût de le abattage	
distance	distance, detachment, length, gap, spacing
manuel sur le méthode	
tronçonnage	sawing
tracteur	tractor

TAB. 2.6 – Exemple de transfert des premières entrées du vecteur de contexte de *débardage*

composants. Dans ce cas, puisque notre dictionnaire bilingue propose *tractor* pour *tracteur* et *caterpillar* pour *chenille*, la seule traduction obtenue est *caterpillar tractor*. Or cette traduction ne correspond pas à celles usitées le plus souvent en langue cible, à savoir *crawler* et *crawler tractor*. Nous sommes confrontés ici à deux problèmes importants, d'une part, un terme ne se traduit pas systématiquement par un terme de même longueur (le terme complexe *tracteur à chenille* est traduit par le mot simple *crawler*) (Brown et al., 1993), et d'autre part, la traduction d'un terme complexe ne s'obtient pas simplement par la traduction de ses composants (*tracteur à chenille* est aussi traduit par *crawler tractor*, où *crawler* n'est pas la traduction de *chenille*) (Melamed, 2001).

Il semble donc que les difficultés de transfert des termes complexes soient une explication à leur manque d'apport dans le processus d'alignement. Néanmoins, nous devons préciser si les termes complexes transférés de la langue source à la langue cible participent ou non à l'identification des traductions candidates.

Hypothèse 2.3.4 *« Le nombre de termes communs entre le vecteur original et le vecteur traduit est trop faible. »*

Pour vérifier cette dernière hypothèse, nous avons compté le nombre de termes communs obtenus lors de la comparaison du vecteur traduit avec les vecteurs originaux de la langue cible. À partir des 100 premières entrées des vecteurs de contexte des termes simples du lexique de référence, nous avons trouvé en moyenne 14,7 termes communs en ne considérant que la meilleure traduction candidate (celle qui a le plus de termes communs). Si nous considérons toutes les traductions candidates, nous obtenons une moyenne de 8,1 termes communs sur 100. Sachant que nous avons environ 20 % de termes complexes dans les vecteurs traduits, nous ne pouvons espérer trouver plus de deux ou trois termes complexes communs entre les vecteurs source et cible. Ceci semble être une explication cohérente à l'inefficacité des termes complexes dans les vecteurs de contextes. Pour compléter cette analyse, nous devrions aussi vérifier que la traduction effective des termes complexes engendre bien un gain dans la précision des résultats. Si cette condition, plus délicate à vérifier, n'était pas avérée, cela

engendrerait plutôt une remise en cause du processus de sélection des termes complexes que du processus d'alignement lequel est intrinsèquement lié à la capacité du dictionnaire à assurer un pont entre les langues cible et source.

2.3.2 Maîtriser et métisser les indices

La chaîne de fouille terminologique multilingue développée fait intervenir de nombreux paramètres. Les plus importants sont la taille de la fenêtre contextuelle et le choix des mesures de récurrence contextuelle et de distance vectorielle. Il est malheureusement assez difficile de statuer précisément sur le rôle de chaque paramètre en raison des temps de calculs nécessaires et du manque de traçabilité de la chaîne de traitement. Néanmoins, l'utilisation de paramètres croisés permet de faire émerger des traductions préférentielles en supprimant du bruit dans les solutions proposées.

Le tableau 2.7 présente les résultats bruts d'une expérience réalisée avec le [lexique 1] (composé de termes simples), le [lexique 2] (composé de termes complexes) et le corpus [SYLV] pour la méthode par similarité interlangue. L'analyse de ces résultats indique que les termes des [lexiques 1 et 2] sont moyennement repérés, mais plus encore, n'apparaissent que rarement dans les 20 premiers candidats (quoique les traductions proposées pour un terme donné se situent le plus souvent dans le même champ sémantique, cf. tableau 2.8).

	NB_{trad}	AVG_{pos}	$STDDEV_{pos}$
[lexique 1]	56	32,9	23,7
[lexique 2]	63	30,7	26,7

TAB. 2.7 – Évaluation du processus d'extraction de terminologies bilingues

Une analyse manuelle des traductions candidates pour différentes configurations de paramètres a permis de mettre en évidence différents comportements intéressants :

- Des termes correctement traduits pour une configuration donnée de paramètres ne le sont pas forcément dans une autre, et inversement, des termes non traduits dans la première configuration peuvent être traduits correctement dans la deuxième. De ce fait, il est difficile de déterminer quelle est la configuration à adopter.
- Plus précisément, pour un terme donné, les premières traductions candidates pour des configurations variées de paramètres sont souvent différentes. Par exemple, pour le terme *pâte à papier* (*paper pulp*), les 50 meilleures traductions candidates d'une vingtaine de configurations différentes n'ont qu'un trentaine d'éléments en commun.
- Par extension, la bonne traduction d'un terme donné apparaît souvent à des positions très différentes. Cette position variant en fonction des configurations.

À partir de ce constat et afin de capturer davantage de bonnes traductions, nous avons choisi de prendre en compte les résultats donnés par différentes configurations de paramètres et non par une seule. Pour cela, le processus de traduction est exécuté sur un ensemble de configurations différentes et nous ne retenons que les x meilleures traductions candidates pour chacune d'entre elles. Ces différents résultats proposés sont alors fusionnés en sommant les

degré de humidité (# occ. 41)	*gaz à effet de serre* (# occ. 33)	*papeterie* (# occ. 178)
humidity	carbon	newsprint
saturation	carbon cycle	paper production
aridity	atmosphere	raw material
evaporation	**greenhouse gas**	mill
saturation deficit	greenhouse	pulp mill
rate of evaporation	global carbon	raw
atmospheric humidity	atmospheric carbon	manufacture
water vapor	emission	**paper mill**
joint	sink	manufacturing
dry	carbon dioxide	capacity
hot	fossil fuel	printing
rainy	fossil	paper manufacture
temperature	carbon pool	factory
moisture control	mitigate	paperboard
meyer	global warming	fiberboard
party	climate change	bagasse
atmospheric	atmospheric	paper-making
dryness	dioxide	board
monsoon	sequestration	material supply
joint meeting	quantity of carbon	paper pulp

TAB. 2.8 – Traductions candidates ordonnées obtenues pour trois termes du [lexique 2]

scores de traductions associés à chaque candidat pour les différentes configurations. Ainsi, nous ne retenons que les candidats les mieux « classés » sur l'ensemble des configurations. Cela permet d'améliorer nettement la position des traductions correctes parmi l'ensemble des traductions candidates. Il convient cependant de choisir un x qui n'est pas trop grand, par exemple 20, afin d'améliorer le comportement et de supprimer du bruit inutile.

Nous avons ainsi fusionné différentes configurations, en particulier en faisant varier les tailles des vecteurs de contexte et de similarité et les mesures d'association et de similarité[12]. Les résultats obtenus par cette méthode, présentés dans le tableau 2.9 et en figure 2.6, améliorent globalement le processus d'extraction de termes bilingues. Les résultats du [lexique 1] sont sensiblement identiques à ceux obtenus pour la méthode directe (pour les 10 et 20 meilleurs candidats 43 % et 47 %). En ce qui concerne les résultats du [lexique 2], ils sont d'une qualité légèrement supérieure à ceux du [lexique 1].

Une approche similaire est proposée par Déjean et al. (2002) qui croisent les méthodes directe et par similarité interlangue. Le croisement des deux méthodes repose sur la combinaison linéaire des probabilités de traduction d'un mot individuellement (moyenne harmonique).

[12]Pour des raisons de temps de calcul, nous ne faisons pas varier la taille de la fenêtre contextuelle. Nous limitons aussi les unités contenues dans les vecteurs de contexte à des termes simples, puisque nos expériences semblent indiquer que cette configuration est la plus intéressante.

	NB_{trad}	AVG_{pos}	$STDDEV_{pos}$	TOP_{10}	TOP_{20}
[lexique 1]	59	16,2	15,9	41	51
[lexique 2]	63	14,8	22,3	45	55

TAB. 2.9 – Évaluation du processus d'extraction par combinaison de paramètres

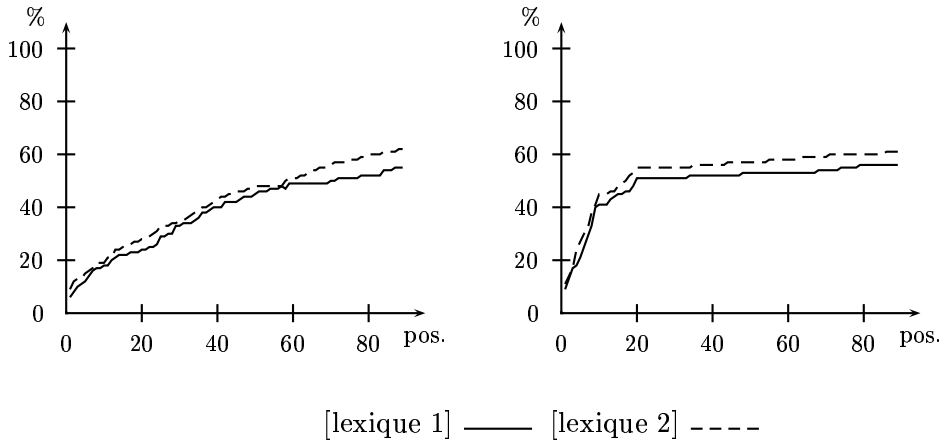


FIG. 2.6 – Évolution de la position des traductions candidates sans (à gauche) puis par combinaison de paramètres (à droite)

Avec cette approche, les résultats obtenus sont meilleurs que ceux décelés pour les méthodes seules.

La simple méthode que nous avons proposée pour améliorer la qualité des traductions candidates par fusion de différentes configurations ne permet qu'un gain de précision des traductions candidates et non une amélioration de la couverture. Dans le cadre de sa thèse Chiao (2004) introduit une hypothèse de « symétrie distributionnelle » pour l'approche par traduction directe qui est formulée ainsi :

« *[...] si deux mots sont proches dans une direction de traduction ainsi que dans l'autre alors ils ont de plus fortes chances d'être traductions l'un de l'autre que s'ils ne sont proches que pour une seule direction de traduction.* » (Chiao, 2004, p. 53)

Cette méthode, qui permet aussi une amélioration de la précision des traductions candidates, n'influence pas le rappel. Il serait intéressant de déterminer s'il existe une corrélation entre un terme et la capacité des différentes méthodes à proposer une bonne traduction pour celui-ci. Des critères comme le degré de polysémie d'un terme ou sa fréquence sont intéressants mais ne seraient être suffisants.

Sadat et al. (2003) utilisent une technique similaire pour filtrer leurs résultats et y ajoutent une condition sur la catégorie syntaxique attendue de la traduction. Ce double filtrage permet un gain d'environ 12 %, toujours en précision, pour la recherche d'information interlangue.

2.3.3 Décrire et croiser les ressources utilisées

La notion de « lexique pivot » fait référence aux données issues du dictionnaire bilingue qui seront exploitées dans le processus de transfert des unités à traduire de la langue source vers la langue cible.

Dans les travaux fondateurs de Fung et McKeown (1997), le lexique pivot est construit comme étant une sous partie sélective du dictionnaire, notamment pour limiter la polysémie inhérente à l'exploitation de documents journalistiques lorsque l'on prend en compte l'ensemble des entrées du dictionnaire (un mot pouvant disposer de plusieurs traductions). Fung et McKeown (1997) introduisent ainsi la notion de « *seed word* » (*mots amorces*) pour réduire les entrées du dictionnaire en s'appuyant sur les règles suivantes : (i) chaque facette des entrées du dictionnaire doit avoir une fréquence entre 100 et 1 000 dans chaque partie monolingue du corpus, (ii) les entrées du dictionnaire sont réduites aux mots pleins, et plus particulièrement, aux noms, verbes et adjectifs, (iii) les entrées du dictionnaire qui proposent une unique traduction identifiée en langue cible sont favorisées (l'entrée de la langue source pouvant être composée de plusieurs mots). Avec cette technique, le lexique pivot se réduit de manière drastique à 307 mots amorces. Cette stratégie, qui vise à maîtriser la polysémie induite par le dictionnaire, semble plus cruciale en langue générale qu'en langue de spécialité. Le plus souvent, la stratégie adoptée consiste soit à se limiter à une traduction, pour peu de disposer d'un dictionnaire où les traductions sont ordonnées (Chiao et Zweigenbaum, 2002), soit à pondérer chacune des traductions par leur fréquence d'apparition en langue cible (Déjean et Gaussier, 2002; Morin et Daille, 2004, 2006).

À défaut de disposer d'un lexicographe, l'évaluation des méthodes repose sur un lexique de référence. Ce dernier peut être construit à partir de termes extraits du corpus comparable (Déjean et Gaussier, 2002; Morin et Daille, 2004, 2006) ou encore extrait du lexique pivot (Chiao et Zweigenbaum, 2002). Dans les deux cas, il faut bien sûr retirer les termes du lexique de référence du lexique pivot (ce point n'est malheureusement pas toujours précisé clairement[13]). En ce qui nous concerne, les termes des lexiques de référence sont (i) choisis dans le corpus, en nous appuyant sur des nomenclatures existantes, sans être des hapax, et dans la mesure du possible (ii) absents du lexique pivot (car à quoi bon rechercher des traductions déjà recensées dans un dictionnaire sauf à vouloir en vérifier la couverture).

Ces choix nous conduisent souvent à travailler avec des lexiques de référence dont la fréquence des termes est faible (ce qui est aussi une caractéristique forte de notre travail). Comme nous l'avons signalé à de nombreuses reprises, plus un terme du lexique de référence est fréquent, plus son vecteur de contexte sera nourri et par conséquent représentatif de la variété des usages. Inversement, un terme peu fréquent disposera d'un vecteur de contexte peu fourni. Il est bien évident que l'évaluation est plus propice avec des termes fréquents. Ainsi, Déjean et Gaussier (2002) utilisent (i) un lexique de référence composé de 1 800 mots qui comporte des mots de faible fréquence ainsi que des hapax pour un corpus médical anglo/allemand de 100 000 mots et (ii) un lexique de référence de 180 termes simples dont la fréquence des termes est très souvent supérieure à 100 voire à 1000 pour un corpus anglo/allemand relatif

[13](Zweigenbaum et Habert, 2006, p. 38) précise que :« *Dans (Chiao & Zweigenbaum, 2002), qui utilisent la méthode standard, les tests se font successivement sur chaque mot du lexique pivot, que l'on supprime temporairement de ce lexique pour le test (« leave-one-out »).* ».

aux sciences sociales de 8 millions de mots. Dans le premier cas, ils obtiennent avec la méthode directe pour les 10 et 20 premiers candidats une précision de 43 % et 51 %, et dans le second cas une précision de 35 % et 42 %. Ces résultats peuvent sembler *a priori* contredire notre précédente affirmation relative à la fréquence des unités à traduire. Il n'en est rien, puisque la ressource dictionnairique associée au corpus médical, composé de résumés d'articles scientifiques provenant de la base de données MEDLINE, n'est autre que le thesaurus MeSH (Medical Subject Headings). Il s'agit donc d'une ressource particulièrement bien adaptée au corpus. Pour autant cela n'autorise pas à s'affranchir d'un dictionnaire de langue générale comme cela a été démontré par Chiao et Zweigenbaum (2003).

Dans nos travaux, des termes complexes sont aussi présents dans le lexique pivot et souvent plus difficiles à exploiter. Afin de juger des données réellement exploitées dans le processus d'alignement, nous avons compté pour l'expérience avec les ressources [DIAB] le nombre de mots simples et composés (respectivement *Nb. MS* et *Nb. MC*) de l'espace vectoriel[14] qui peuvent être traduits à l'aide du dictionnaire (respectivement *Nb. MST* et *Nb. MCT*) et dont les traductions sont présentes dans le corpus comparable. Le tableau 2.10 présente les résultats obtenus pour chaque langue et pour chaque corpus. Pour les deux corpus, le taux de traduction des mots composés est très faible (autour de 1 %) en comparaison avec celui des mots simples (environ 30 % pour le français et 20 % pour le japonais) qui reste lui aussi faible.

	Français		*Japonais*	
	Nb. MST/ Nb. MS	*Nb. MCT/ Nb. MC*	*Nb. MST/ Nb. MS*	*Nb. MCT/ Nb. MC*
[corpus scientifique]	2 300/7 443	80/7 225	2 614/12 941	78/8 655
[corpus mixte]	3 085/9 888	131/10 110	4 293/9 604	95/11 847

TAB. 2.10 – Éléments lexicaux des corpus comparables traduits à partir du dictionnaire bilingue

Le faible taux de mots composés traduits directement à partir de notre dictionnaire bilingue constitue un handicap majeur dans le processus d'alignement, et plus particulièrement lors de l'étape de transfert d'une unité à traduire de la langue source à la langue cible. Afin de suppléer l'insuffisance du dictionnaire bilingue pour la traduction des mots composés, nous avons essayé d'en élargir la couverture en utilisant une approche par traduction compositionnelle (Janssen, 1996; Melamed, 1997; Grefenstette, 1999). Ainsi, pour chaque mot composé de la langue source absent du dictionnaire et identifié par *ACABIT*, nous traduisons chacun de ses composants. Par exemple, pour le terme français *fatigue chronique*, nous obtenons pour *fatigue* les quatre traductions suivantes : 疲れ, 疲労, 倦怠 et 飽き et pour *chronique* les deux traductions suivantes : 記事番組 et 慢性. Ensuite, nous combinons entre-elles les traductions des différents composants pour obtenir des traductions candidates du mot composé en langue cible. Pour le même exemple, nous obtenons les 8 traductions candidates[15]

[14]Nous rappelons que l'espace vectoriel est constitué d'unités lexicales qui ne sont pas des hapax et qui font référence à des mots, à des termes simples à des termes complexes reconnus par *ACABIT*.

[15]Entre le français et le japonais les constituants sont inversés.

présentées en tableau 2.11. Les traductions retenues sont celles qui font référence à un mot composé existant en langue cible (c'est-à-dire identifié par la version japonaise d'*ACABIT*). Dans notre exemple, le seul mot composé identifié en japonais est 慢性疲労.

chronique	*fatigue*
記事番組	疲れ
慢性	疲れ
記事番組	疲労
慢性	疲労
記事番組	倦怠
慢性	倦怠
記事番組	飽き
慢性	飽き

TAB. 2.11 – Exemple de traduction compositionnelle

Cette approche présente aussi des limites (Baldwin et Tanaka, 2004; Brown et al., 1993; Morin et Daille, 2004) notamment en ce qui concerne la *fertilité*. Par exemple, le terme simple français *hypertension* est traduit en japonais par le terme complexe 高血圧 (où le caractère 高 (*taka*) signifie *haut* et le terme 血圧 (*ketsuatsu*) désigne la *pression sanguine*). Elle diffère aussi de celle utilisée par Robitaille et al. (2006) pour la traduction compositionnelle de termes complexes anglo/japonais. Dans Robitaille et al. (2006), les termes de longueur n (avec $n > 2$) sont préalablement décomposés en toutes les combinaisons de termes de longueur inférieure ou égale à n éléments. Cette approche permet de pouvoir éventuellement traduire directement une sous partie du terme complexe s'il est présent dans le dictionnaire bilingue. Par exemple, pour le terme *syndrome de fatigue chronique*, Robitaille et al. (2006) génèrent les quatre combinaisons suivantes i) [*syndrome de fatigue chronique*], ii) [*syndrome de fatigue*] [*chronique*], iii) [*syndrome*] [*fatigue chronique*] et iv) [*syndrome*] [*fatigue*] [*chronique*]. Par exemple pour la troisième combinaison, si le dictionnaire dispose de la traduction de *fatigue chronique* sa traduction sera alors directement utilisée. Dans ce travail, nous nous limitons à la dernière combinaison[16]. La première combinaison, quant à elle, sera obtenue directement si elle est présente dans le dictionnaire.

Avec cette approche compositionnelle, nous définissons un nouveau dictionnaire comportant 111 traductions de mots composés pour le [corpus scientifique] et 201 traductions pour le [corpus mixte]. En combinant ces entrées avec celles obtenues directement par le dictionnaire, le taux de traduction des mots composés passe de 1 à environ 3 %, ce qui reste globalement très faible. D'une manière générale, nous avons relevé que sur 100 mots composés 32 ne peuvent pas être traduits dans la mesure où l'un des composants est absent du dictionnaire auxquels viennent s'ajouter 17 mots induits par le bruit de la chaîne de traitement (mots composés mal orthographiés ou segmentés, anglais, incomplets ou incohérents). Il n'y a donc que 51 mots composés qui peuvent prétendre à une traduction compositionnelle. Cette difficulté de prise en compte des mots composés a d'ailleurs conduit Déjean et Gaussier (2002) à s'en tenir aux

[16]Environ 90 % des candidats termes fournis par *ACABIT*, après regroupement, ne sont composés que de deux mots pleins.

mots simples.

2.3.4 Mixer les indices de comparabilité

Si la notion de *comparable* fait référence aux caractéristiques partagées des documents d'un corpus comme la période, le domaine, le thème, le support médiatique, le type de discours, etc. (Baayen, 1994), les indices de comparabilité des corpus utilisés ne sont pas toujours précisés. Ils se réduisent souvent à un domaine générique comme le médical (Chiao et Zweigenbaum, 2002) ou la finance (Fung, 1998), ou à un support communicationnel comme les articles journalistiques (Fung, 1998; Sadat et al., 2003). Pour des corpus comparables de langue générale, le second indice largement exploité est celui de la période des journaux utilisés (Fung, 1998; Sadat et al., 2003).

En langue de spécialité, la première caractéristique retenue est celle du domaine. Il s'agit majoritairement du médical (Déjean et Gaussier, 2002; Chiao et Zweigenbaum, 2002; Morin et Daille, 2006) en raison certainement de la facilité d'accès de ces documents, notamment à partir du web. Dans Déjean et Gaussier (2002), le corpus anglais/allemand est construit à partir de résumés d'articles scientifiques issus de la base MEDLINE. Il couvre naturellement des thématiques diverses mais se limite au seul type de discours scientifique et au même genre textuel (les résumés d'articles scientifiques). Dans Chiao et Zweigenbaum (2002), le champ d'investigation est réduit aux « signes et symptômes, états pathologiques ». Les données étant issues de CISMeF pour la partie française et de CliniWeb pour la partie anglaise, il peut s'agir de documents relevant du discours scientifique mais aussi vulgarisé[17] qui font référence à des genres textuels variés, notamment des guides de bonnes pratiques, des sites associatifs, des notes de cours, des QCM... Dans notre cas (Morin et Daille, 2006), nous nous restreignons à la thématique des « maladies liées aux régimes alimentaires » et plus particulièrement au « diabète » et distinguons les documents français/japonais selon qu'ils appartiennent au discours scientifique ou vulgarisé. En ce qui concerne le genre des documents extraits, il s'agit principalement de rapports scientifiques ou vulgarisés avec des différences notables selon la langue du corpus comparable. En japonais, les rapports émanent presque exclusivement d'institutions privées alors que, pour le français, ils émanent de sites gouvernementaux, d'instituts universitaires ou d'institutions publiques (hôpitaux). Étant donné les contextes culturels envisagés dans ce travail, le problème de la comparabilité des textes extraits est aussi à prendre en compte. Ainsi, nous avons constaté une atténuation du caractère distinctif du type du discours pour le japonais puisque que certains termes complexes extraits par *ACABIT* dans la partie française du corpus scientifique ont été identifiés par *ACABIT* dans la partie japonaise du corpus mixte.

Au sens des *dimensions* de Biber (1995), nous pouvons constater que les indices mobilisés en langue de spécialité sont un croisement de la thématique, du type de discours et du genre. À la différence des corpus de langue générale, la période n'est pas prise en compte, certainement de par la difficulté à disposer de cette information. À notre connaissance, il n'existe pas d'études qui mobilisent finement ces trois indices ou qui précisent leur apport respectif. De plus, les résultats que nous avons obtenus pour l'alignement de termes simples semblent

[17] Si le catalogue CISMEF s'adresse en priorité aux professionnels de santé, on y trouve également des informations destinées aux patients et à leurs familles.

indiquer que le type de discours n'est pas un indice significatif puisque les résultats sont meilleurs avec un corpus composé de documents scientifiques et vulgarisés par comparaison avec un corpus composé exclusivement de documents scientifiques. Ce résultat est compréhensible dans la mesure où les termes simples utilisés pour cette évaluation ne sont pas plus caractéristiques du discours scientifique que vulgarisé. Par exemple, le terme *excès* peut tout aussi bien faire référence au discours scientifique dans *excès pondéral* qu'au discours vulgarisé dans *excès de poids*. En revanche, le type de discours semble être un indice pertinent pour l'alignement de termes complexes. À la lumière de ces résultats, qui devront être affinés, il nous semble nécessaire pour l'alignement de termes complexes de croiser au minimum les indices liés à la thématique, au type de discours et à la période. D'autres études sont nécessaires pour préciser l'apport du genre qui induit une certaine circularité avec le type de discours. En ce qui concerne l'alignement de termes simples, c'est la thématique, le genre et la période qui doivent être privilégiés. Dans les deux cas, le contexte culturel inhérent aux langues exploitées ne doit pas induire de décalage marqué dans les dimensions utilisées. L'objectif visé, ne l'oublions pas, est de ne pas bruiter le corpus comparable par des usages différents de termes. Il s'agit donc de contrôler et de croiser ces dimensions pour assurer *a priori* la pertinence des contextes lexicaux extraits (Zweigenbaum et Habert, 2006; Péry-Woodley, 2000).

Cette démarche s'oppose « frontalement » à la position communément adoptée en fouille textuelle — appelée « *Gros, c'est beau* » par Benoît Habert (2000, p. 18)[18] — qui privilégie la quantité des données sur leur qualité. Les raisons souvent invoquées sont, d'une part, la nécessité de disposer de grandes masses de données pour mettre en œuvre les méthodes informatiques et, d'autre part, le manque de méthodes opérationnelles pour automatiser la construction d'un corpus *représentatif* d'un domaine, d'une activité ou encore d'une situation de communication et répondant donc à des critères langagiers précis. En ce sens, notre démarche constitue un aspect novateur à la fouille terminologique multilingue.

2.4 Synthèse

Les acquis engrangés en fouille terminologique multilingue ouvrent de nombreuses perspectives de recherche que cela soit au niveau des méthodes employées que des ressources mobilisées. Nous aurons loisir de développer ces aspects dans le chapitre 5 de ce mémoire.

Dans ce chapitre, nous avons proposé et décrit deux approches possibles pour l'acquisition de terminologie bilingue. En ce qui concerne la méthode par similarité interlangue, elle se propose comme une alternative intéressante à la méthode par traduction directe pour éviter l'appauvrissement du vecteur de contexte lors de son transfert en langue cible. En revanche, elle semble souffrir de quelques difficultés lorsque les termes à traduire ne sont pas suffisamment ancrés conceptuellement dans le texte étudié. Comme cette méthode s'appuie sur les unités similaires à l'unité à traduire, il est nécessaire que ces dernières soient sémantiquement proches de l'unité visée. Un décalage sémantique fortement marqué induira un transfert des traductions proposées vers un autre champ sémantique.

En ce qui concerne l'alignement des termes simples nos résultats semblent indiquer que la prise en compte des termes complexes, quoique pertinente, n'est pas totalement adaptée

[18] Il est fait ici référence à l'article de Marie-Paule Pery-Woodley (1995).

si la ressource n'est pas fortement ajustée au corpus d'étude. En outre, nous avons précisé que le type de discours ne semble pas être un indice de comparabilité à privilégier pour l'alignement de termes simples ce qui est l'inverse pour l'alignement des termes complexes. Les résultats obtenus pour ces derniers sont encourageants mais demeurent encore insuffisants pour envisager une application industrielle effective. Le travail de révision manuelle étant encore à notre sens trop important.

En conclusion de ce chapitre, nous souhaitons simplement revenir sur la notion même de corpus comparable et son positionnement par rapport aux corpus parallèles. Comme nous l'avons déjà indiqué à plusieurs reprises, l'intérêt premier d'un corpus comparable se situe dans la possibilité de disposer d'un matériau textuel qui met en correspondance des documents dans des langues différentes sans être des traductions. Les indices mobilisés pour parvenir à créer un tel corpus s'appuient sur le domaine, la thématique, le couple de langues, le type de discours, le genre, la période. L'objectif étant de pouvoir mettre à jour des lexiques bilingues où le vocabulaire rencontré dans la partie traduite n'est pas influencé par celui de la langue source (c'est-à-dire au plus proche de l'usage). En accord avec cet objectif, il ne peut y avoir de continuum du corpus parallèle au corpus comparable. Un corpus parallèle ne peut constituer un corpus comparable idéal. L'essence même du matériau n'est pas la même, dans un cas, il s'agit d'un matériau artéfactuel, dans l'autre cas, d'un matériau naturel.

Chapitre 3

Reconnaissance de l'écriture manuscrite en ligne

Les travaux présentés dans ce chapitre se situent dans le cadre de la reconnaissance de l'écriture manuscrite en ligne (encore appelée *écriture dynamique)*. Dans ce champ de recherche lié à la reconnaissance de formes, notre contribution se situe au niveau de la modélisation du langage en vue de renforcer un système de reconnaissance de l'écriture manuscrite existant[1]. Nous avons en particulier développé des modèles de langages adaptés à différents types d'écritures. D'une part, il s'agit de pouvoir traiter des documents de langue générale — que nous qualifions ici d'« *écrit standard* » — où le stylo numérique vient remplacer la saisie sur un clavier d'ordinateur. D'autre part, nous cherchons aussi à traiter de l'écriture manuscrite où le stylo numérique s'offre comme une nouvelle méthode de saisie pour l'écriture de SMS (Short Message Service). Il s'agit ici d'un « *écrit déviant* » — par opposition à un écrit standard — qui s'écarte de manière notable des formes de communications écrites traditionnelles et qui nécessite, par conséquent, une modélisation spécifique.

Dans la suite de ce chapitre, nous commençons par présenter en section 3.1 le contexte scientifique et industriel propre à ces travaux. Ensuite, nous décrivons en section 3.2 les modèles probabilistes développés pour la reconnaissance d'un écrit standard et en section 3.3 les techniques adoptées pour la prise en compte d'un écrit déviant. Enfin, la section 3.4 dresse le bilan de nos travaux dans ce champ de recherche.

3.1 Introduction à la reconnaissance de l'écriture manuscrite

Les travaux développés dans ce chapitre se situent dans un contexte technologique particulier, où les documents à traiter sont produits par un scripteur au moyen d'un crayon numérique spécifique. Ce crayon se propose comme une alternative crédible pour remplacer différentes interfaces de saisie. Il peut s'agir du clavier d'un ordinateur pour la saisie de données textuelles, mais aussi du clavier d'un téléphone portable pour l'écriture de SMS.

[1] Il s'agit ici du système *MyScript Builder* de la société française *Vision Objects* – `http://www.visionobjects.com/`

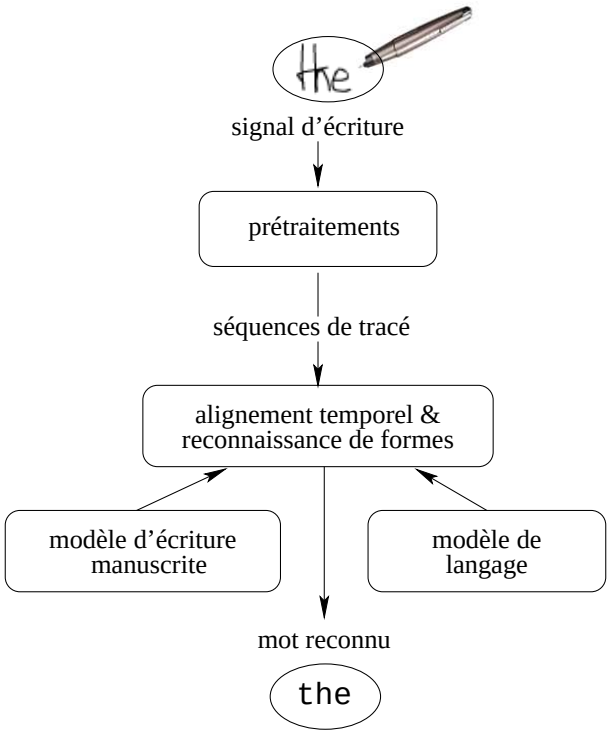


FIG. 3.1 – Architecture générale du système de reconnaissance de l'écriture manuscrite

D'un point de vue pratique, le scripteur dispose d'un stylo numérique spécifique qui enregistre les informations relatives au tracé de son écriture. Grâce à une caméra miniature placée sous la bille, le stylo numérique reconnaît et enregistre sa position à intervalle de temps régulier. Pour se repérer, le stylo doit être couplé avec un papier spécialement pré-imprimé qui est doté d'une grille microscopique. L'écriture est ensuite disponible sous forme d'une séquence de points représentant la trajectoire de l'instrument d'écriture. Il s'agit d'une représentation dite « en ligne » à différencier des représentations « hors ligne » comme celles obtenues sous la forme d'une image après avoir numérisé un document.

Dans notre cas, il s'agit de réaliser la reconnaissance d'une forme décrite par un signal contenant la dynamique de l'écriture par opposition au traitement d'une image où la forme à reconnaître est figée ou statique. Plus précisément, l'objectif est de proposer une séquence de mots correspondant à un signal observé pour lequel nous ne connaissons pas la longueur de chaque forme élémentaire (chaque mot dans la phrase et chaque caractère dans un mot). Il est donc nécessaire d'ajuster par une procédure d'alignement les hypothèses de formes élémentaires au signal observé résultant du tracé manuscrit. La figure 3.1 présente l'architecture générale du système de reconnaissance de l'écriture manuscrite.

À partir du signal fourni par l'instrument d'écriture, une série de prétraitements est réalisée pour fournir en sortie une séquence de points représentant la trajectoire de l'outil d'écriture. Ces prétraitements assurent la normalisation de la représentation de la trace écrite en éliminant le plus possible toutes les variations non significatives concernant l'identité du mot écrit (taille, inclinaison, orientation, vitesse du tracé). Il en résulte une description sous

formes de trames de vecteurs de caractéristiques décrivant localement le tracé. L'analyse de cette trame permet de retrouver l'identité du mot recherché. Dans le cas de l'écriture cursive non contrainte[2], se pose alors le problème de la segmentation du mot en entités de base pour lesquelles des hypothèses de reconnaissance sont envisagées.

La principale difficulté du moteur de reconnaissance est de surmonter la dualité des tâches de segmentation du tracé et de reconnaissance des caractères le composant. Ces deux tâches doivent nécessairement coopérer et sont très dépendantes. Pour circonvenir cette impossibilité de déterminer à l'avance de manière fiable les points de segmentation du tracé en caractères, la solution retenue consiste en une sur-segmentation garantissant l'obtention de segments élémentaires de taille inférieure ou égale à celle du plus petit caractère possible. Ces segments, représentant des graphèmes du tracé, sont alors combinés et chaque combinaison est alors interprétée comme une hypothèse de segmentation (Tay et al., 2001). D'un point de vue technique, le système de reconnaissance *MyScript Builder* combine un réseau de neurones formels, qui joue le rôle d'un reconnaisseur de caractères, avec un modèle de Markov à états cachés (MMC), qui permet de passer de la reconnaissance au niveau des caractères à la reconnaissance au niveau des mots. Nous ne détaillons pas davantage ici les caractéristiques du système de reconnaissance, sachant que ce sujet n'est pas au centre de nos travaux. Pour une présentation détaillée du système, nous renvoyons les lecteurs vers Tay et al. (2001, 2002).

L'objectif est donc de déterminer la séquence de mots la plus probable, étant donné à la fois un signal x correspondant à un tracé manuscrit et les informations fournies par le modèle de langage. Ces différents éléments, à savoir, le signal observé, noté x ; une phrase à reconnaître, notée s et le modèle de langage sont intimement liés par la relation de Bayes :

$$p(s|x) = \frac{p(x|s)p(s)}{p(x)} \tag{1}$$

Dans cette relation, le rôle du modèle de langage est de calculer la probabilité *a priori* $p(s)$ d'une phrase donnée s. Le modèle d'écriture, pour sa part, sait pour une phrase donnée s, évaluer la vraisemblance que le signal observé x résulte de l'écriture de celle-ci, soit $p(x|s)$. Dès lors, le résultat final de la reconnaissance s'obtient en cherchant à maximiser la probabilité *a posteriori* $p(s|x)$. En utilisant le critère *Maximum A Posteriori* (MAP), on retient donc la phrase $s_{reconnue}$ telle que :

$$s_{reconnue} = \arg\max_{s} p(x|s)p(s) \tag{2}$$

La précédente équation suppose possible le calcul de $p(s)$ quelle que soit la phrase s. Il s'agit là de la problématique majeure du modèle de langage que d'être capable d'estimer de façon fiable cette probabilité. L'utilisation d'un modèle de langage statistique dépend de l'estimation de paramètres qui sont calculés à partir de corpus textuels volumineux, et comprend également des techniques de lissage pour gérer les événements inconnus.

Un deuxième point crucial concerne le calcul du terme $p(x|s)$. Ce calcul se fait nécessairement en ligne, au fur et à mesure que le signal x est disponible. De ce fait, il n'est pas possible, d'un point de vue opérationnel, pour chaque phrase possible s, d'évaluer $p(x|s)$. Il

[2]En reconnaissance de l'écriture manuscrite, en première approximation, deux formes d'écritures sont envisagées : l'écriture cursive non contrainte (ou libre) et l'écriture contrainte scripte (où les lettres sont détachées).

faut obligatoirement restreindre l'espace des recherches à un sous-ensemble de phrases admissibles. Pour cela, nous nous sommes orientés sur des techniques de programmation dynamique basées sur des algorithmes de type *Viterbi* et *Beam Search* (Rabiner et Juang, 1993). En fait, l'ensemble des phrases peut être représenté par un graphe orienté où chaque chemin représente une phrase. Il faut alors être capable de ne gérer qu'un nombre limité de chemins en élaguant les moins vraisemblables et en conservant uniquement les chemins ayant le plus de chance d'aboutir. À ce niveau, le rôle du modèle de langage est de participer à ces procédures de choix en ajoutant de l'information *a priori*. Dès lors, il s'agit de privilégier les séquences de mots les plus probables d'un point de vue linguistique. De cette manière, le modèle de langage et le système de reconnaissance coopèrent pour privilégier les chemins à conserver.

L'approche statistique privilégiée ici, pour la reconnaissance d'un écrit standard, assure un renforcement de la cohérence globale du système, puisque les outils utilisés pour la modélisation du langage sont de même nature que ceux utilisés pour le système de reconnaissance proprement dit.

3.2 Reconnaissance d'un écrit standard

Comme nous l'avons indiqué nous cherchons à intégrer à un système de reconnaissance de l'écriture manuscrite en ligne des connaissances linguistiques issues de la langue générale. Plus précisément, il s'agit de construire des modèles probabilistes de langage élaborés lors d'une étape d'entraînement sur des corpus écrits. En outre, les modèles développés doivent répondre à des contraintes industrielles fortes au niveau des temps de traitement et de l'occupation de l'espace mémoire, puisque l'application *MyScript Builder* est destinée à être déployée sur des engins nomades aux capacités matérielles limitées.

3.2.1 Modélisation probabiliste du langage

Dans une modélisation probabiliste, une phrase s peut être représentée sous la forme d'une séquence de L mots w_i, soit $s = w_1...w_i...w_L$. Une phrase particulière peut alors être interprétée comme la réalisation d'un processus stochastique discret. Si l'on suppose que l'évolution de ce processus (le prochain mot) ne dépend que de son état actuel (ce qui est déjà écrit), on s'inscrit alors dans le cadre des processus de Markov. Dans la mesure où l'ensemble des états (le lexique) est fini, on est ramené au cas particulier des chaînes de Markov.

Dans ces conditions, il devient possible à partir de la connaissance des probabilités de transitions entre états de calculer la vraisemblance de toute chaîne particulière, c'est-à-dire dans notre cas d'une phrase spécifique. Il suffit pour cela d'utiliser la relation de Chapman-Kolmogorov. Pour définir le modèle de langage, nous devrons connaître au préalable, grâce à une étape d'entraînent, les probabilités de transitions entre états.

La probabilité d'une phrase s donnée peut alors s'écrire comme étant :

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_L|w_1...w_{L-1}) = \prod_{i=1}^{L} p(w_i|w_1...w_{i-1}) \qquad (3)$$

Ainsi, à chaque position i d'un mot w d'une phrase s, le modèle de langage doit prédire le mot w_i connaissant l'historique $w_1...w_{i-1}$. Par prédiction, nous voulons dire que chaque mot du vocabulaire de taille V du modèle est susceptible d'apparaître et que le modèle affecte à chaque mot w_i une probabilité $p(w_i|w_1...w_{i-1}) > 0$ vérifiant la contrainte de normalisation suivante :

$$\sum_{j=1}^{V} p(w_i = w^j|w_1...w_{i-1}) = 1 \tag{4}$$

(où w^j désigne le $j^{\text{ème}}$ mot du lexique de taille V)

3.2.2 Modèle n-gramme

Lorsque la longueur de l'historique du mot à prédire devient importante, l'estimation de la probabilité conditionnelle $p(w_i|w_1...w_{i-1})$ est difficile à obtenir car le contexte gauche du mot w_i est moins susceptible d'être rencontré lors de l'entraînement. Dans ces conditions une hypothèse classique consiste à réduire l'ordre de la chaîne de Markov en supposant qu'un nouveau mot ne dépend que d'un historique limité aux $n-1$ mots le précédant. L'équation 3 devient alors :

$$p(s) \approx p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_L|w_{L-n+1}...w_{L-1}) \tag{5}$$

$$p(s) \approx \prod_{i=1}^{L} p(w_i|w_{i-n+1}...w_{i-1}) \tag{6}$$

Ainsi, en fixant n à 2, nous obtenons un modèle bi-gramme qui limite le calcul de la probabilité d'un mot au seul mot qui le précède. Pour $n = 3$, le modèle s'appuie sur les deux mots qui précèdent. Dans un modèle n-gramme, seuls les $n-1$ mots sont pris en considération. En ayant recours à des corpus d'entraînement de taille importante, il est possible d'estimer la fréquence des événements recherchés par une simple méthode de comptage :

$$p(w_i|w_{i-n+1}...w_{i-1}) = \frac{c(w_{i-n+1}...w_i)}{c(w_{i-n+1}...w_{i-1})} \tag{7}$$

(où $c(.)$ représente le nombre d'occurrences de la chaîne correspondante)

Dans le cas d'un modèle n-gramme, l'historique est réduit et une approche probabiliste devient envisageable. En pratique, il faut noter que la grande majorité des séquences de mots (même de longueur réduite) ne sera pas observée pendant la phase d'entraînement. Or, le modèle basé sur le maximum de vraisemblance attribue une probabilité nulle à des séquences non observées et donc d'occurrence nulle, même si la séquence est valide dans la langue considérée. Il est donc indispensable d'utiliser des techniques permettant d'estimer des probabilités associées à des séquences non observées (Ney et al., 1994; Kneser et Ney, 1995). La solution consiste à attribuer une partie de la masse totale des probabilités aux

événements inconnus, ceux qui n'auront pas été rencontrés sur le corpus d'entraînement, mais qui vont être rencontrés en situation de reconnaissance. Dans les modèles élaborés, nous avons retenu la technique de redistribution utilisant l'estimateur *absolute discounting backing-off* (Chen et Goodman, 1996, 1998; Manning et Schütze, 2000). À noter que les événements inconnus peuvent faire référence à des séquences non rencontrées comme à des séquences impossibles (Langlois et al., 2003). Quoique nous ne fassions pas de distinction entre ces différentes séquences, la technique proposée par Langlois et al. (2003) pour redistribuer la masse de probabilité attribuée aux séquences impossibles vers celle des autres séquences est très intéressante.

Un modèle n-gramme présente deux limites importantes. D'une part, comme il est peu concevable d'espérer rencontrer l'ensemble des n-grammes possibles dans la base d'entraînement, ceux-ci souffrent d'un manque de robustesse. D'autre part, le nombre de n-grammes devient vite considérable pour un lexique de taille importante dès lors que n augmente. Afin de pallier ce double problème, une méthode consiste à réduire le nombre d'événements observables en regroupant les mots en classes, pour obtenir un modèle dit « n-classe ». En appliquant ce regroupement, le modèle prédit non plus un mot en fonction des $n-1$ mots le précédant, mais une classe de mots en fonction des $n-1$ classes qui la précèdent.

3.2.3 Modèle n-classe

Dans un modèle n-classe, il s'agit d'associer à chaque mot w_i une ou plusieurs classes $g(w_i) = g_k$ avec $k \in [1...K]$ où K représente le nombre de classes souhaitées. Dans le cas où un mot ne peut appartenir qu'à une seule classe, la probabilité conditionnelle $p(w_i|w_{i-n+1}...w_{i-1})$ est donnée par :

$$p(w_i|w_{i-n+1}...w_{i-1}) = p(w_i|g(w_i))p(g(w_i)|g(w_{i-n+1})...g(w_{i-1})) \tag{8}$$

Dans le cas d'un modèle bi-classe, la formule devient alors :

$$p(w_i|w_{i-1}) = p(w_i|g(w_i))p(g(w_i)|g(w_{i-1})) \tag{9}$$

Dans ce type de modèle, le nombre de paramètres requis est $V + K^2$, là où un modèle n-gramme en nécessite V^2. Toute la difficulté de cette approche réside donc dans la définition de la fonction $g()$ qui réalise le regroupement en classes.

Au niveau du regroupement, nous avons plus particulièrement étudié deux critères :

Critère contextuel Les mots qui partagent les mêmes contextes lexicaux sont regroupés au sein d'une même classe (Beaujard et Jardino, 1999). Il s'agit donc ici d'identifier des « affinités du second ordre » (Grefenstette, 1994b) comme nous l'avons évoqué en section 2.1. Ce type de modèle est appelé « modèle n-classe statistique ».

Critère grammatical Les classes sont construites en s'appuyant sur la catégorie morpho-syntaxique des mots (Niesler, 1997). Par exemple, une classe rassemblera l'ensemble des adverbes, une autre celle des pronoms, etc. Ce type de modèle est appelé « modèle n-classe syntaxique ».

En fonction du critère de classification étudié, les classes construites n'ont pas la même signification. Dans le cas d'un modèle n-classe statistique, la classification n'a pas de signification *a priori* explicite, ce qui est l'inverse pour une classification avec un modèle n-classe syntaxique.

3.2.3.1 Modèle n-classe statistique

La principale difficulté d'un modèle n-classe statistique réside dans la construction des classes. Sachant que pour deux classifications différentes le modèle ne présente pas les mêmes performances, il s'agit donc de déterminer la meilleure. En outre, les temps de calcul peuvent devenir rédhibitoires pour peu que les données soient volumineuses (Martin et al., 1998). Les méthodes les plus efficaces procèdent de manière itérative en recherchant à l'itération $n+1$ une classification « meilleure » que celle disponible à l'itération n. Pour déterminer en quoi une classification est meilleure qu'une autre, nous utilisons la mesure de perplexité[3] (notée PP) qui permet d'évaluer la qualité du modèle correspondant aux classes obtenues à l'itération courante. Dans notre cas, la recherche d'un optimum consiste à identifier la classification avec laquelle on obtient une perplexité minimale. Une méthode permettant d'obtenir l'optimum global consisterait à tester l'ensemble des configurations possibles. Mais pour 20 000 mots à diviser en 100 classes, il y a déjà 100^{20000} classifications possibles... Fort heureusement, il existe des méthodes sous optimales de moindre complexité, en particulier celle de Martin et al. (1998) qui s'inspire de l'algorithme des *nuées dynamiques* (*k-means*) de MacQueen (1967).

Dans cette méthode, les classes sont initialisées en rangeant les mots par ordre décroissant de fréquence, puis on affecte les $K-1$ premiers mots dans une classe différente et les $V-K-1$ mots restant dans une dernière classe. Ensuite, à chaque itération, chaque mot est déplacé temporairement dans chaque classe et la perplexité de la classification associée est évaluée[4]. En fonction des résultats obtenus, la meilleure classification est conservée et utilisée à la prochaine itération. Avec cette approche, toutes les solutions de voisinages sont explorées. Pour 20 000 mots à diviser en 100 classes, le nombre de classification étudié est de $20\,000 \times 100 = 2.10^6$ à chaque itération.

3.2.3.2 Modèle n-classe syntaxique

Dans un modèle n-classe syntaxique, les classes sont construites en s'appuyant sur la catégorisation morpho-syntaxique des occurrences de formes. Dans la mesure où plusieurs catégories peuvent être associées à une même forme, un mot peut appartenir à plusieurs classes. Ainsi pour décrire une séquence de mots, plusieurs séquences de classes sont possibles.

Pour ce faire, un lexique morpho-syntaxique indique pour chaque mot l'ensemble des catégories qu'il possède. Nous illustrons en figure 3.2 le graphe orienté associé à la phrase « *Il court vers la victoire* » où chaque chemin représente une séquence possible de classes. Ici, la

[3] La mesure de perplexité est l'inverse de la moyenne géométrique des probabilités conditionnelles $p(w_i|w_{i-n+1}...w_{i-1})$ (Jelinek et al., 1977). Cette mesure représente l'« hésitation moyenne » du modèle de langage sur un texte donné.

[4] À ce niveau, la technique utilisée correspond à celle proposée par Martin et al. (1998).

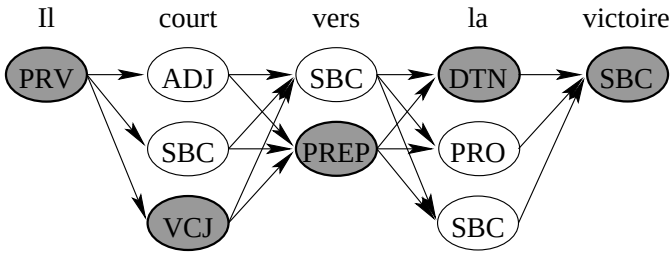


FIG. 3.2 – Exemple de graphe orienté

vraisemblance d'un chemin est calculée suivant le même principe que pour un modèle n-classe statistique (Niesler, 1997).

Dans la mesure où il existe plusieurs chemins valides, la vraisemblance de la séquence de mots est égale à la somme des vraisemblances pour tous les chemins possibles (Bernard, 1994). Pour calculer de manière efficace cette vraisemblance, nous pouvons représenter un modèle n-classe syntaxique sous la forme d'un modèle de Markov à états cachés (MMC) (Rabiner et Juang, 1993). Un MMC est un double processus stochastique. Un premier processus décrit les transitions entre des états invisibles (cachés). Dans notre contexte ces états correspondent aux catégories morpho-syntaxiques. Ils sont dits « cachés » car la valeur de l'état est *a priori* indéterminée. Lorsque ce processus arrive à un état s_i, celui-ci émet une observation. L'ensemble des observations émises par les états est décrit par le second processus. Dans notre cas, les observations sont les mots de la séquence étudiée. Ainsi à chaque séquence de mots est associée un MMC dont la topologie est fonction du lexique morpho-syntaxique exploité. La figure 3.2 présente le MMC associé à la séquence de mots « *Il court vers la victoire* ».

Il existe de nombreux algorithmes associés aux MMC. Nous avons utilisé ici l'algorithme *Forward* qui permet d'établir la vraisemblance d'une séquence d'observations et donc dans notre contexte d'une séquence de mots. De la même manière qu'un modèle n-gramme, un modèle n-classe nécessite une base d'entraînement qui va lui servir à régler ses paramètres. Les différents modèles n-classes développés pour le français ont été entraînés sur des corpus catégorisés à l'aide de *Brill* (Brill, 1994; Lecomte et Paroubek, 1996) associé à l'analyseur flexionnel *Flemm* (Namer, 2000).

Dans le cas d'un modèle bi-classe syntaxique, les paramètres à estimer sont :
- Les probabilités de transition $p(g(w_i)|g(w_{i-1}))$ entre états. Il s'agit ici d'étudier l'enchaînement des catégories morpho-syntaxiques sur l'ensemble du corpus et d'établir la probabilité conditionnelle que la catégorie $g(w_i)$ puisse succéder à la catégorie $g(w_{i-1})$.
- Les probabilités d'observation $p(w_i|g(w_i))$ estimées en s'appuyant sur les associations « mot/catégorie » présentes dans le corpus. Il s'agit d'estimer la probabilité que lorsque le processus se trouve dans l'état $g(w_i)$, l'observation émise soit le mot w_i.

3.2.4 Principaux résultats

Nous présentons ici les principaux résultats obtenus en reconnaissance de l'écriture manuscrite en ligne par l'exploitation d'un modèle de langage. Pour une présentation plus exhaustive, nous renvoyons les lecteurs vers Perraud (2005); Perraud et al. (2005, 2003a).

Dans le cadre de nos expérimentations, deux corpus de textes ont été utilisés pour l'entraînement des différents modèles :

- le corpus [ROMAN], qui comporte 4,2 millions de mots, principalement composé de romans du XIX^{e} et XX^{e} siècle.
- le corpus [LEMONDE], qui comporte 4,1 millions de mots, composé d'articles issus du journal *Le Monde*.

En ce qui concerne l'évaluation des modèles, nous utilisons le corpus [EVAL] qui est principalement composé d'articles issus de journaux en ligne tels que *Libération*, *Le Figaro*... Ce corpus de 1,6 millions de mots ne recouvre que partiellement au niveau lexical les corpus d'entraînement (52 % pour [ROMAN] et 65 % pour [LEMONDE]).

3.2.4.1 Évaluation du modèle bi-classe statistique

La figure 3.3.a présente les résultats obtenus pour un modèle bi-classe statistique où nous faisons varier le nombre de classes suivant les valeurs 10, 50, 100, 500 et 1000. Le dernier point, noté BG, correspond aux résultats obtenus pour un modèle bi-gramme. L'analyse des résultats indique une corrélation entre le nombre de classes et les performances associées. Ainsi, plus le nombre de classes augmente, plus les performances sont intéressantes. Au delà de 500 classes, un phénomène de sur-apprentissage semble apparaître. En outre, les résultats obtenus avec un modèle bi-classe de 500 classes sont sensiblement identiques à ceux obtenus avec un modèle bi-gramme. Ce résultat est d'autant plus intéressant que l'encombrement mémoire associé à un modèle bi-classe est bien moindre que celui associé à un modèle bi-gramme. Dans Perraud et al. (2006), nous avons évalué la taille mémoire nécessaire pour stocker différents modèles de langage. Pour le français, un modèle uni-gramme nécessite 0,33 Mo, un modèle bi-classe 1 Mo et un modèle bi-gramme 27 Mo.

D'un point de vue qualitatif, les classes obtenues ne semblent pas correspondre à une distribution aléatoire. Il est possible d'observer des points communs entre les mots appartenant à une même classe (cf. tableau 3.1). L'explication est peut être liée à la méthode de classification utilisée qui regroupe les mots partageant les mêmes voisins de gauche et de droite. Ces points communs sont d'ordre sémantique (cf. classes 1 et 3) ou syntaxique (cf. classes 2 et 4). Bien sûr la grande majorité des classes ne semblent pas avoir de signification particulière (cf. classe 5). En outre, nous avons noté que les mots vides ayant une forte fréquence (p. ex. *le*, *de*, *un*...) se retrouvent dans des classes distinctes dont ils sont l'unique représentant.

3.2.4.2 Évaluation du modèle bi-classe syntaxique

La mise en place du modèle bi-classe syntaxique peut se faire quant à lui en s'appuyant sur les catégories du discours fournies par l'outil d'étiquetage. Dans notre cas, nous pouvons construire deux classifications différentes en nous appuyant sur les catégories syntaxiques ou sur les catégories morpho-syntaxiques (c'est-à-dire en tenant compte des informations fournies par l'analyseur flexionnel *Flemm* de Namer (2000) notamment le genre et le nombre). Nous créons respectivement ainsi un premier modèle bi-classe syntaxique composé de 57 classes et un second composé de 210 classes.

Les résultats obtenus pour ces deux classifications sont présentés en figure 3.3.b. Nous

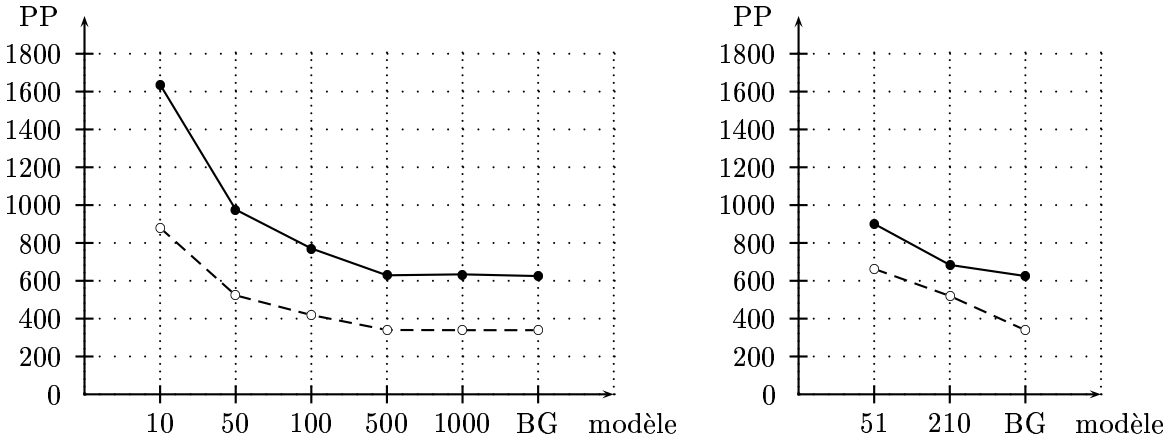
Classe	*Exemple*
1 (pays ou région commençant par une voyelle)	Afghanistan, Alaska, Albanie, Alexanderplatz, Algérie, Alsace, Amazonie, Ameublement, Andalousie, Angleterre, Angola, Antarctique, apesanteur, Argentine, Arménie, Australie, Autriche, Azerbaïdjan, Écosse, Égypte, Équateur, Érythrée, Espagne, Estonie, Éthiopie, euthanasie, hématome, herbe, inavouable, Inde, Indonésie, Iowa, Irak, Iran, Islande, Italie, Occident, Ouganda
2 (adverbe)	éminemment, affreusement, anormalement, désespérément, disco, rock, extrêmement, extraordinairement, faussement, follement, foncièrement, Glenn, hautement, horriblement, incroyablement, inhabituellement, mûrement, mûrissent, passablement, purement, relativement, ridiculement, sucré, terriblement, très, trend, universellement, vertigineusement
3 (mois de l'année)	février, grès, injectant, juillet, juin, mai, mars
4 (participe présent)	écartant, échangeant, écoutant, élargissant, éliminant, élisant, éloignera, émettant, étouffant, évitant, évoquant, abaissant, abandonnant, abolissant, abordant, absorbant, accélérant, acceptant, accordant, accrochant, accueillant, accumulant, accusant, achetant, acquérant, adaptant
5 (divers)	conquerront, déballer, écoute, même, quadruplement, trotter

TAB. 3.1 – Exemples de classes obtenues avec un modèle bi-classe statistique de 1 000 classes pour le corpus [LEMONDE]

pouvons ainsi observer le gain apporté par ces deux modèles qui est plus important avec 210 classes qu'avec 57 classes, certainement en raison de sa plus grande richesse. En revanche ces résultats restent éloignés de ceux obtenus avec le seul modèle bi-gramme. Si nous projetons maintenant les résultats des modèles bi-classes syntaxiques sur les modèles bi-classes statistiques, nous pouvons constater un décalage suivant le corpus d'entraînement utilisé. En effet, les résultats obtenus avec le modèle bi-classe syntaxique pour le corpus d'entraînement [ROMAN] se situent entre les résultats des modèles bi-classes statistiques à 10 et 100 classes pour la classification syntaxique à 57 classes et entre 100 et 500 classes pour la classification syntaxique à 210 classes. En revanche pour le corpus [LEMONDE], nous nous situons entre des modèles bi-classes statistiques à 10 et 50 classes pour la classification à 57 classes et entre 50 et 100 classes pour la classification à 210. Ce phénomène est peut être lié au genre du discours du corpus [EVAL] plus proche du corpus [LEMONDE] que du corpus [ROMAN].

3.2.4.3 Évaluation de la combinaison des modèles bi-classes

Comme nous venons de le constater, les modèles bi-classes permettent de s'approcher de manière très intéressante des résultats obtenus avec un modèle bi-gramme. En particulier un modèle bi-classe statistique de 500 classes permet de suppléer un modèle bi-gramme tout en offrant un plus faible encombrement mémoire. Toujours dans cette optique de maîtrise



(a) modèle bi-classe statistique *(b) modèle bi-classe syntaxique*

FIG. 3.3 – Évaluation des modèles bi-classes statistique et syntaxique ([ROMAN] •——• et [LEMONDE] ○- - -○)

de l'encombrement mémoire pour les meilleures performances, nous nous proposons de préciser le gain obtenu par la combinaison d'un modèle bi-classe statistique avec un modèle bi-classe syntaxique. En effet de nombreux auteurs (Jardino et Adda, 1994; Niesler, 1997; Goodman, 2000) suggèrent que la combinaison de deux modèles est plus performante, en ce qui concerne la perplexité, qu'un seul modèle. Cette combinaison de modèles est réalisée par une combinaison linéaire des probabilités conditionnelles de chaque modèle (Perraud et al., 2003a).

Nous présentons, en figure 3.4, les résultats obtenus en combinant chacune des cinq classifications du modèle bi-classe statistique avec chacune des deux classifications du modèle bi-classe syntaxique[5]. Nous pouvons constater que la combinaison améliore notablement les performances. En particulier, les résultats obtenus avec un modèle bi-gramme sont égalés avec la combinaison de seulement 50 classes statistiques avec 51 classes syntaxiques et dépassés avec la combinaison de 500 classes statistiques avec 51 ou 210 classes syntaxiques. Dans ce dernier cas, l'amélioration obtenue est de l'ordre de 30 %. Cette forte disparité doit être relativisée dans la mesure où le modèle bi-gramme est construit à partir d'un corpus de 4 millions de mots, ce qui reste assez faible pour un tel modèle.

3.2.5 Apports et limites d'un modèle de langage

Les précédents résultats fournissent des indices intéressants en vue de l'intégration d'un modèle de langage dans un système de reconnaissance de l'écriture manuscrite. En particulier, nous avons montré l'intérêt d'utiliser un modèle bi-classe statistique qui permet d'avoisiner les résultats d'un modèle bi-gramme et ce pour un encombrement mémoire réduit. Nous avons aussi montré que l'association d'un modèle bi-classe statistique avec un modèle bi-classe syntaxique dépassait les résultats d'un modèle bi-gramme pour un encombrement mémoire encore

[5]Dans un souci de lisibilité, nous ne présentons que les résultats obtenus pour le corpus [LEMONDE].

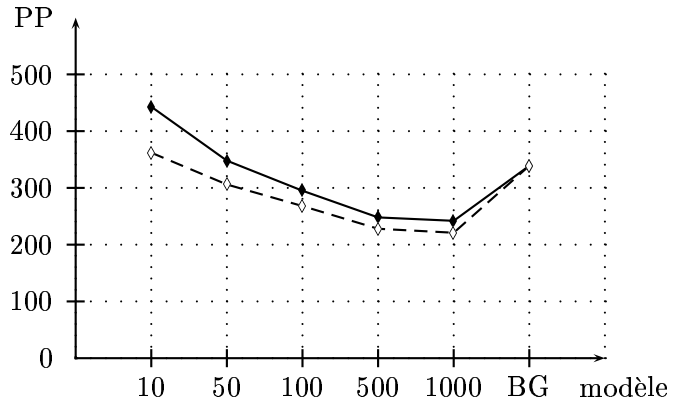


FIG. 3.4 – Évaluation de la combinaison des modèles bi-classes statistique et syntaxique ([LEMONDE] & 51 classes ♦—♦ et [LEMONDE] & 210 classes ◊- - -◊)

acceptable (autour de 3 Mo pour le modèle combiné). Dans Perraud et al. (2003a,b), nous avons aussi étudié le couplage du modèle bi-gramme avec les modèles bi-classes statistique et syntaxique. Les résultats obtenus sont globalement meilleurs que ceux obtenus avec le seul modèle bi-gramme et confirment les résultats indiqués par Jardino (1998) pour l'association d'un modèle tri-classe à un tri-gramme qui dépasse le seul modèle tri-gramme. Ces résultats confirment tout l'intérêt des modèles bi-classes. Pour des corpus d'entraînement plus volumineux, nous pourrions peut être dépasser, avec un modèle bi-classe statistique un modèle bi-gramme comme cela a été réalisé par Jardino (1996) pour trois langues différentes avec un modèle tri-classe par rapport à un modèle tri-gramme.

Les différents modèles de langage développés présentent un apport intéressant une fois couplés avec le système de reconnaissance de l'écriture en ligne. Les tableaux 3.2, 3.3 et 3.4 illustrent de manière concrète cet apport pour différents exemples de phrases à reconnaître. Dans ces exemples choisis, la reconnaissance est effective dès lors que l'on dispose d'un modèle uni-gramme (cf. 3.2) où d'un modèle bi-classe (cf. 3.3). En revanche, la reconnaissance de l'exemple du tableau 3.4 nécessite de disposer au moins d'un modèle tri-classe.

Phrase à reconnaître		
		
Modèle de langage	*Phrase reconnue*	*Taux d'erreur*
aucun	L'écrivain écrit **ln** toutes **inconstances**	2/6
uni-gramme	L'écrivain écrit en toutes circonstances	0/6

TAB. 3.2 – Exemple de reconnaissance : « *L'écrivain écrit en toutes circonstances* »

Phrase à reconnaître		
		
Modèle de langage	*Phrase reconnue*	*Taux d'erreur*
aucun	Son regard reflétait **be** clarté **dean** âme .	3/8
uni-gramme	Son regard reflétait la clarté de **au** âme .	1/8
bi-classe	Son regard reflétait la clarté de son âme .	0/8

TAB. 3.3 – Exemple de reconnaissance : « *Son regard reflétait la clarté de son âme .* »

Phrase à reconnaître		
		
Modèle de langage	*Phrase reconnue*	*Taux d'erreur*
aucun	Plusieurs femmes étaient là **arec** leurs **matis**	2/7
uni-gramme	Plusieurs femmes étaient là **arec** leurs maris	1/7
bi-classe	Plusieurs femmes étaient là **arec** leurs maris	1/7

TAB. 3.4 – Exemple de reconnaissance : « *Plusieurs femmes étaient là avec leurs maris* »

3.3 Reconnaissance d'un écrit déviant

Les travaux présentés dans la précédente section visaient à associer à un système de reconnaissance de l'écriture manuscrite des connaissances linguistiques sur la langue générale. Pour ce faire, l'utilisation d'un modèle de langage probabiliste acquis sur des corpus de textes volumineux permet d'améliorer de manière significative les résultats de reconnaissance de l'écriture. Lorsque les documents à traiter ne relèvent plus de la langue générale mais d'un écrit spécialisé, les modèles développés risquent d'être peu adaptés puisque non caractéristiques des phénomènes rencontrés. C'est dans cette optique que nous nous sommes intéressés à la reconnaissance de Mini-Messages Manuscrits, soit en abrégé MIMEMA.

Ces MIMEMA sont assimilables d'un point de vue linguistique aux SMS qu'un scripteur peut composer à partir du clavier de son téléphone portable (Anis, 2001). D'un point de vue technologique, les MIMEMA ne correspondent pas à un usage existant mais à une évolution possible des messages envoyés sous forme graphique (MMS, Multimedia Messaging Service). Pour envisager cette solution de SMS manuscrits, il est nécessaire de disposer d'un crayon numérique pour leur composition et d'intégrer des fonctionnalités de reconnaissance de l'écriture manuscrite si l'on veut ramener ce message à une simple chaîne de caractères (SMS).

Les SMS, comme les services de discussion en ligne et les forums de discussion, constituent des nouvelles formes de communication écrite (NFCE). Les travaux de recherche relatifs aux

NFCE se sont principalement concentrés sur l'étude linguistique (Mourlhon-Dallies et al., 2004; Véronis et Guimier de Neef, 2006) ou socio-linguistique (Anis, 2003; Piérozak, 2000, 2003) de ces phénomènes. D'un point de vue applicatif, Bove (2005) s'est quant à lui intéressé à la traduction de SMS, saisis au moyen d'un clavier, dans une optique de vocalisation. Dans notre travail, il ne s'agit pas de traduire des MIMEMA, mais de reconnaître une forme manuscrite où le signal d'entrée n'est pas une séquence de caractères électroniques mais un tracé manuscrit. Ces MIMEMA se caractérisent : (i) sur la plan de la reconnaissance de formes par une dégradation importante par rapport aux représentations canoniques, notamment par la perte des espaces inter-mots, le fléchissement de la ligne de base et la fin de mots griffonnée ; et (ii) sur le plan linguistique par une abondance de formes extra-lexicales, une sur-représentation des phrases agrammaticales et un style d'écriture spécifique dans lequel l'orthographe et la ponctuation sont détournées et malmenées.

Il s'agit donc pour nous de déterminer les limites atteignables par un système de reconnaissance de formes par rapport aux différentes contraintes posées lors de la saisie (notamment l'écriture de type pré-casé versus non contrainte), ou bien lors de la reconnaissance (utilisation d'une reconnaissance guidée par le lexique ou sans lexique, utilisation ou non d'une classe de rejet).

3.3.1 Collecte d'un corpus de MIMEMA

Dans le cadre de ce travail, nous devions disposer de données caractéristiques de l'écriture de mini-messages manuscrits. Il faut donc nécessairement acquérir des échantillons d'écritures manuscrites pour évaluer le taux de reconnaissance et développer des ressources linguistiques. Au début de ce travail, le projet « *Faites don de vos sms à la science !;-)*[6] » du CENTAL (Centre de traitement automatique du langage, Université catholique de Louvain, Belgique), visant à constituer un corpus de SMS (Fairon et Paumier, 2006), n'était malheureusement pas disponible. D'autre part, le corpus du français tchaté réalisé par Falaise (2005) nous semblait faire référence à un usage trop lié aux contraintes relatives à la frappe sur le clavier d'un ordinateur. Pour ces différentes raisons, nous avons donc décidé de constituer notre propre corpus de MIMEMA.

Pour réaliser ce corpus, nous avons d'abord collecté, à partir de sites web spécialisés différentes productions « attestées » de SMS. Les principales formes recensées correspondent : i) à de l'écriture rébus (p. ex. `2m1` = `demain`, `kfé` = `café`), ii) à l'utilisation de squelettes consonantiques (p. ex. `slt` = `salut`, `tjs` = `toujours`) ou iii) à la phonétisation de l'écriture (p. ex. `koi` = `quoi`, `comen` = `comment`). Bien sûr ces différents procédés sont souvent combinés entre eux pour former des messages variés. De manière à être le plus proche possible des caractéristiques de l'écriture manuscrite, nous avons délibérément éliminé les formes étroitement liées à l'interface de saisie telles que : i) les messages avec des erreurs caractéristiques de la frappe sur clavier comme l'insertion ou l'interversion ou ii) les messages comportant des étirements graphiques. L'ensemble des formes ainsi collectées est utilisé pour générer différents formulaires papiers. Dans ces derniers, il est proposé au scripteur soit de recopier soit d'écrire librement des SMS avec une écriture en style pré-casé ou libre (cf. figure 3.5). Le fait de disposer de ces deux styles d'écriture permettra de juger de l'effort technique nécessaire

[6] `http://www.smspourlascience.be/`

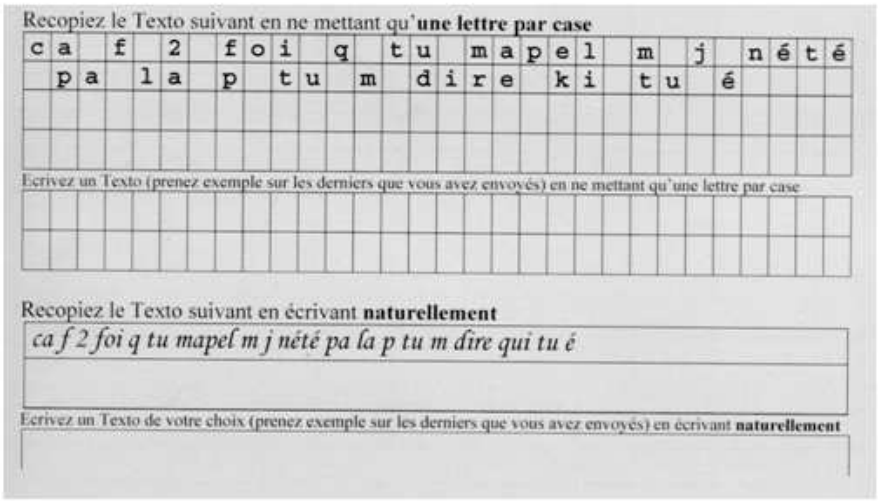


FIG. 3.5 – Extrait du formulaire de collecte de MIMEMA

pour en assurer une reconnaissance acceptable.

Cette collecte a été réalisée auprès de 150 personnes différentes (principalement des lycéens, des étudiants des domaines littéraires et scientifiques). Les messages ainsi collectés ont été manuellement nettoyés (suppression des échantillons non valides), segmentés (segmentation du signal d'écriture par champs du formulaire et par mots) et annotés (association au message de son label texte). Après ces différentes opérations, nous disposons d'un corpus de MIMEMA composé de 1 221 messages pour un total de 11 600 mots ou 38 462 caractères (cf. tableau 3.5).

	Style d'écriture pré-casé	Style d'écriture libre
Contenu imposé	177	174
Contenu libre	493	477
Total	670	551

TAB. 3.5 – Caractéristiques du corpus de MIMEMA collecté

À ce niveau, il est prudent de noter que l'usage du stylo numérique peut introduire un biais dans notre démarche scientifique. En effet, l'utilisation du stylo numérique pour collecter des MIMEMA entraîne l'introduction d'une interface de saisie différente de celle utilisée pour les SMS. Selon les travaux du linguiste français Jacques Anis : « *les conditions matérielles de la communication modèlent fortement la forme linguistique des messages* » (Anis, 2003). Si cette hypothèse — qui s'oppose à l'idée d'un « langage SMS » — était vérifiée, notre protocole de collecte serait « contestable ». Quoique ces aspects, qui relèvent du champ de la sémio-linguistique de l'écrit, dépassent le cadre de ce travail, nous nous devons d'y accorder une attention particulière.

3.3.2 Évaluation avec un système industriel de reconnaissance

De manière à juger de la complexité potentielle de la reconnaissance des MIMEMA, nous commençons par en évaluer la qualité avec un système industriel de reconnaissance de l'écriture manuscrite en ligne.

3.3.2.1 Système de reconnaissance

Nous utilisons à nouveau le système de reconnaissance de l'écriture manuscrite *MyScript Builder*. Mais à la différence des modèles de langage qui étaient directement intégrés dans le système de reconnaissance, nous travaillons ici avec le kit de développement logiciel (SDK) de *MyScript Builder*. Celui-ci permet dans sa version de base de disposer de plusieurs modèles de langage et de construire ses propres ressources notamment des lexiques et expressions régulières. Le moteur de reconnaissance est fourni avec deux ressources linguistiques (Linguistic Knowledge, LK) :

LK-text Cette ressource est adaptée pour du texte de langue générale. Elle comporte un lexique français de grande taille et un modèle de langage (lequel est basé sur les bigrammes au niveau mots). Avec cette ressource, il est néanmoins possible de reconnaître des mots hors-lexique.

LK-free Cette ressource est conseillée pour du texte dont la nature est inconnue. Il y a uniquement un modèle de langage au niveau caractères associé à cette ressource pour faciliter l'identification des mots hors-lexique.

Il est possible d'associer au moteur de reconnaissance aucune ou plusieurs ressources linguistiques.

3.3.2.2 Estimation des performances

Dans la mesure où les messages du corpus de MIMEMA sont étiquetés globalement et non caractère par caractère, le taux de reconnaissance le plus facile à calculer serait une mesure globale donnant le pourcentage de messages bien reconnus. Toutefois, comme une telle mesure ne présente pas une finesse d'analyse suffisante — un message peut être plus ou moins mal reconnu, le rendant néanmoins partiellement compréhensible — nous préférons évaluer les performances de reconnaissance au niveau des caractères des messages. Pour cela, le taux de reconnaissance proposé est dérivé de la distance de Levenshtein, notée D, calculée entre la séquence de mots placée en première position par le moteur de reconnaissance et le label correspondant au message à reconnaître. Le taux de reconnaissance, noté TR, au niveau caractère est défini par la relation suivante :

$$TR = \frac{100 \times (\#label - D)}{\#label} \tag{10}$$

La distance D calcule les coûts d'édition (insertion, substitution et suppression) nécessaires pour passer d'une chaîne à l'autre (Wagner et Fischer, 1974). Afin de ne pas pénaliser deux fois un problème de sur-segmentation d'un caractère, qui le plus souvent va conduire à

une opération de substitution et d'insertion, le coût d'édition est fixé à 1 pour une substitution et à 0 pour une insertion. Le coût d'édition pour une opération de suppression est, quant à lui, de 1. Ainsi, dans le cas où tous les caractères d'un message sont reconnus alors $D = 0$ et $TR = 100\,\%$. Inversement, si aucun caractère n'est reconnu alors $D = \#label$ et $TR = 0\,\%$. Cette formule permet ainsi de circonscrire le taux de reconnaissance entre 0 et 100 %. En revanche, elle ne pénalise pas la reconnaissance de messages où des caractères sont insérés. Le tableau 3.6 est une illustration du calcul de TR pour un exemple simple.

Label	*bjr A 2min*	$\#label = 8$
Résultat de reconnaissance	***lo**jr A **Zmu**in*	$TR = (8-2)/8 = 0,75$

TAB. 3.6 – Exemple de reconnaissance erronée avec une sur-segmentation sur le 'b', une substitution sur le '2', et un caractère 'u' inséré

3.3.2.3 Résultats obtenus avec le système de base

Le tableau 3.7 présente les résultats de reconnaissance obtenus pour les différentes ressources linguistiques à notre disposition et pour des messages manuscrits cursifs ou pré-casés. La première colonne de résultats correspond à une « utilisation normale » du système de reconnaissance sans aucune ressource linguistique, puis avec la ressource LK-text, et enfin avec la ressource LK-free. Les résultats de la seconde colonne utilisent un lexique additionnel optimal comportant l'ensemble des mots présents dans les messages. Les résultats ainsi obtenus définissent une frontière haute pour le système de reconnaissance.

	Sans ressource additionnelle (cursive/pré-casé)	Avec le lexique optimal (cursive /pré-casé)
Aucune ressource	87 % / 94 %	96 % / 96 %
Ressource LK-text	84 % / 90 %	95 % / 96 %
Ressource LK-free	88 % / 95 %	88 % / 95 %

TAB. 3.7 – Taux de reconnaissance des MIMEMA avec le système de base

Nous observons à la lecture de ce tableau que la reconnaissance des caractères pré-casés conduit à des meilleures performances (90 à 95 % de reconnaissance suivant les ressources linguistiques) que la reconnaissance du texte sans contrainte (84 à 87 % de reconnaissance). Ce phénomène n'est bien sûr pas surprenant, la difficulté de segmentation étant bien supérieure dans le cas de textes cursifs. Par ailleurs, l'utilisation de ressources linguistiques standard permet au mieux d'augmenter très légèrement le taux de reconnaissance (passage de 94 à 95 % avec LK-free), ou bien dégrade significativement les performances (passage de 94 à 90 % avec LK-text). Cela confirme l'importance d'associer des ressources appropriées pour reconnaître ce type de messages, et qu'en particulier la ressource LK-text n'est pas vraiment adaptée pour des mini-messages manuscrits. L'adjonction du lexique optimal permet comme le montrent les résultats de la seconde colonne d'améliorer les taux de reconnaissance, et la

seule utilisation de ce lexique donne les meilleurs résultats aussi bien sur le pré-casé que le style libre (96 % / 96 %).

3.3.3 Modélisation du langage relatif au MIMEMA

Nous commençons par décrire les différentes ressources linguistiques développées spécifiquement pour améliorer la reconnaissance des principales formes de MIMEMA recensées, à savoir les squelettes consonantiques, l'écriture rébus, et la phonétisation de l'écriture[7]. Nous intégrons ensuite ces ressources au système de reconnaissance et les évaluons au regard des résultats obtenus sans connaissances spécifiques sur les MIMEMA.

3.3.3.1 Squelettes consonantiques

Un squelette consonantique correspond à l'abréviation d'un mot commun charpenté quasi-exclusivement autour de ses consonnes. Dans ce cas, il faut reconstruire le mot pour pouvoir le lire (p. ex. `dvt=devant`). Ce phénomène peut être modélisé à partir des règles de transformation suivantes :

1. Conservation de la première et de la dernière consonne ainsi que des voyelles situées avant la première et après la dernière (p. ex. pour le mot *indépendance* nous conservons `in...ce`).
2. Suppression des voyelles restantes (p. ex. `indpndnce`).
3. Suppression des consonnes `l`, `r` et `h` en position faible, c'est-à-dire lorsqu'elles sont situées après une consonne en début de syllabe.
4. Suppression des consonnes `n` et `m` en position faible, c'est-à-dire lorsqu'elles sont situées avant une consonne en début de syllabe (p. ex. `indpdce`).

Nous avons utilisé ces règles pour construire un lexique de squelettes consonantiques en nous appuyant sur une liste d'adverbes, adjectifs et noms fréquents extraits d'un corpus français de langue générale. Le lexique ainsi obtenu est composé de 3 244 squelettes consonantiques. Afin d'assurer la complétude de ce lexique, nous avons aussi modélisé les règles de transformation sous la forme d'automates stochastiques à états finis. Ces automates ajoutent aux précédentes règles diverses observations que nous avons réalisées sur les SMS à notre disposition (p. ex. il est fréquent de trouver une sous partie du mot non abrégé comme `bjour` pour `bonjour`). D'une manière générale, ces automates sont moins restrictifs que la stricte application des règles de transformation.

3.3.3.2 Écriture rébus

Les rébus sont généralement construits par un mélange de lettres et de chiffres. Ici, il faut lire chaque symbole du mot mis en évidence par son nom et ne pas lire le son associé au mot (p. ex. `paC` doit se lire `pa-cé` et `non pa-que`). En raison de la grande créativité possible pour

[7] Dans un souci de concision, nous limitons notre propos à la description des phénomènes et non à sa modélisation. Pour une description détaillée, nous renvoyons les lecteurs vers Prochasson (2006); Prochasson et al. (2007a,b).

l'écriture rébus, il n'est pas raisonnable de vouloir construire un lexique dédié (Véronis et Guimier de Neef, 2006). La modélisation de ce phénomène a donc été réduite à la définition d'automates stochastiques à états finis. Ces automates, plus complexes que pour les squelettes consonantiques, reposent sur la modélisation des observations suivantes :

- Possibilité de mélanger lettres, chiffres et symboles (p. ex. `a+` = `a plus`, `2m1` = `demain`...) ;
- Forte prédominance de certains singletons (p. ex. `c` = `ces`, `c'est`... ; `g` = `j'ai` ; `9` = `neuf`...) ;
- Faible probabilité d'avoir deux chiffres ou plus de suite dans la même forme.

3.3.3.3 Phonétisation de l'écriture

La phonétisation de l'écriture est certainement le phénomène le plus complexe à modéliser puisqu'il correspond à des motivations différentes, tout en gardant à l'esprit que la forme produite doit rester compréhensible lorsqu'elle est lue. Il peut s'agir d'une motivation d'abréviation visant à réduire la frappe sur clavier (p. ex. `tro` pour `trop`), d'un simple jeu visant à produire des néographies « amusantes » (p. ex. `bocou` pour `beaucoup`) ou encore d'une simplification ou méconnaissance de l'orthographe, notamment la conjugaison des verbes (p. ex. la transformation des formes `ai`, `aie`, `ait`, `ais`, `é`... par le simple caractère `é`).

Face à la complexité du phénomène et pour ne pas introduire du bruit dans le système, nous avons limité sa modélisation à des règles de transformation conduisant à la production d'un lexique. Par exemple, les règles de transformation pour la phonétisation du mot `cause` produisent des formes comme `kause`, `cose`, `kose`, `koz`... Ces règles ont été appliquées sur les 1 202 mots les plus fréquents d'un corpus français de langue générale pour produire un lexique de 3 171 formes.

3.3.4 Évaluation

Afin d'évaluer l'apport des différentes ressources linguistiques développées, nous avons manuellement classé les messages du corpus MIMEMA correspondant au style d'écriture précasé en quatre catégories : squelettes consonantiques (54 mots/151 caractères), écriture rébus (90 mots/222 caractères), phonétisation de l'écriture (91 mots/327 caractères) et divers. Les résultats obtenus, présentés dans la colonne centrale du tableau 3.8, avec les ressources que nous avons proposées et intégrées au système de reconnaissance, sont comparés avec ceux obtenus avec le système doté des ressources standards (colonne de gauche) et des ressources optimales (colonne de droite) comprenant le lexique des formes à reconnaître[8].

Nous pouvons constater que les ressources linguistiques apportées au système de reconnaissance sont principalement bénéfiques à la reconnaissance des squelettes consonantiques (environ 38 % des erreurs initiales sont corrigées). Dans le cas de l'écriture rébus et de la phonétisation de l'écriture, l'amélioration n'est pas directement visible. Néanmoins si nous étudions les résultats obtenus en combinant nos ressources avec la seule ressource LK-text,

[8]Dans ce tableau, nous n'indiquons que les meilleurs résultats sans préciser les combinaisons de ressources utilisées (c'est-à-dire celles qui sont fournies en standard avec le système de reconnaissance, à savoir : aucune ressource, LK-text ou LK-free).

	Ressources standards	Ressources développées	Ressources optimales
Squelette consonantique	94,7 %	98,0 %	100 %
Ecriture rébus	92,6 %	92,6 %	94,6 %
Phonétisation de l'écriture	94,1 %	94,1 %	99,3 %

TAB. 3.8 – Taux de reconnaissance des MIMEMA avec et sans les ressources développées

nous pouvons constater l'apport global des ressources fournies au système (cf. tableau 3.9). Alors que les résultats du tableau 3.8 ne présentaient dans le meilleur des cas qu'une légère augmentation de la précision, la différence est ici beaucoup plus nette. La ressource LK-text est peu adaptée à la reconnaissance des phénomènes étudiés, mais cette combinaison indique clairement que les modèles de langage conçus apportent une information significative et adaptée au système de reconnaissance. La combinaison de nos propres ressources avec la ressource LK-text corrige presque 62 % des erreurs initiales pour la phonétisation et 75 % des erreurs initiales pour les rébus.

	LK-text	LK-text + ressources développées	LK-text + lexique optimal
Squelette consonantique	66,2 %	98,0 %	98,0 %
Ecriture rébus	69,1 %	92,1 %	94,6 %
Phonétisation de l'écriture	75,2 %	90,5 %	99,0 %

TAB. 3.9 – Taux de reconnaissance des MIMEMA pour la ressource LK-text

3.4 Synthèse

La reconnaissance de l'écriture manuscrite est une branche de la reconnaissance de formes qui pose un tel niveau de difficulté, notamment à cause de la très grande variabilité intra-classe et de la très forte proximité inter-classe des symboles, que la seule perception des formes n'est pas suffisante pour permettre un taux de reconnaissance élevé. Il est donc fondamental de pouvoir appuyer la reconnaissance sur des connaissances issues de la langue.

Dans le cadre de la reconnaissance d'un écrit standard, des résultats satisfaisants peuvent être obtenus par l'adjonction d'un modèle de langage au modèle d'écriture manuscrite. En particulier, nous avons montré qu'un modèle bi-classe statistique permet de concurrencer un modèle bi-gramme et ce pour un encombrement mémoire réduit. Un modèle bi-classe syntaxique ne fait quant à lui qu'approcher les résultats obtenus par un modèle bi-classe statistique. En revanche, la combinaison d'un modèle bi-classe statistique avec un modèle bi-classe syntaxique permet de dépasser les résultats d'un modèle bi-gramme, pour un encombrement mémoire toujours acceptable. Bien sûr, l'utilisation d'un modèle bi-classe syntaxique nécessite l'utilisation d'outils d'étiquetage et de lemmatisation, associés à des traitements préalables

de nettoyage et de segmentation de textes. Ainsi, pour chaque nouvelle langue à traiter, il faut disposer des outils adaptés. Dans un contexte industriel, l'énergie et le savoir-faire nécessaires au passage à une nouvelle langue tend à laisser de côté les modèles syntaxiques pour se limiter aux modèles statistiques plus facilement déployables à grande échelle. C'est ce choix raisonné que nous pouvons observer dans la stratégie de portabilité des modèles de langage du système *Myscript Builder* lorsqu'il s'agit de traiter plus de 70 langues distinctes à grande distance linguistique (p. ex. le français, le finnois, le russe, le chinois, le japonais...).

Le développement de ces modèles de langage nécessite de disposer d'un réservoir conséquent de données textuelles pour l'entraînement des modèles. Le web en est souvent la principale source. Et pour créer des modèles relevant de la langue générale, les journaux en ligne représentent une donnée précieuse. Néanmoins, comme la masse textuelle nécessaire à l'entraînement des modèles doivent être volumineux, la tentation est forte de prendre du texte tout-venant. Cette légitime tentation risque de ne pas être sans conséquence sur les modèles développés. Si nous revenons un instant sur l'exemple de reconnaissance du tableau 3.4, nous pouvons être surpris de la reconnaissance de la forme *arec*, qui désigne tout à la fois un palmier des régions chaudes que son fruit, au détriment de la forme *avec*. Par exemple, le mot *arec* n'est pas présent dans le corpus [LEMONDE], là où la préposition *avec* a une fréquence de 15 000... Bien sûr, il peut aussi s'agir d'une situation où le modèle d'écriture l'emporte sur le modèle de langage. Néanmoins, pour arriver à prendre en compte ce type de phénomène, il est nécessaire de garder à l'esprit qu'un modèle de langage adapté au traitement d'un écrit standard doit s'appuyer sur des documents choisis et non sur la plus grande masse disponible. De même, pour modéliser un écrit relevant d'une langue de spécialité, il faudra s'appuyer sur des écrits caractéristiques du domaine visé et trouver un juste compromis avec des caractéristiques issues de la langue générale. Même si le gain attendu ou obtenu semble parfois faible au regard des ressources humaines à déployer, nous devons être vigilant à ces aspects.

En outre, nous avons montré que l'obtention de tels modèles est une tâche qui devient très délicate dès lors que l'on s'intéresse aux nouvelles formes de communication écrite telles que les SMS. En nous concentrant sur trois phénomènes particuliers affectant la production de MIMEMA, nous avons pu proposer des ressources linguistiques particulières pour les modéliser. Celles-ci sont représentées soit par des lexiques spécifiques, soit par des descriptions sous forme d'automates stochastiques à états finis. Pour l'un de ces phénomènes, les squelettes consonantiques, nous avons pu améliorer de façon très significative les performances du système de reconnaissance en réduisant de 38 % le taux d'erreur au niveau caractères en nous limitant au style pré-casé. Pour l'écriture rébus, l'automate proposé est sûrement trop générique et ne reflète pas suffisamment précisément la réalité des rébus trouvés dans les SMS. De même, le lexique utilisé pour couvrir la phonétisation s'est appuyé sur un corpus de langue générale où le discours n'est sans doute pas le plus adapté au langage des SMS. Une fois encore, la construction de ressources idoines doit primé sur la modélisation des phénomènes et non l'inverse. Il s'agit d'une évidence qui n'est malheureusement pas toujours facile à satisfaire en dehors de toutes contingences.

Chapitre 4

Synergie des approches et des ressources déployées

Les travaux que nous avons réalisés en acquisition lexicale multilingue et en reconnaissance de l'écriture manuscrite en ligne, qui s'inscrivent dans le champ du Traitement Automatique du Langage Naturel (TALN), reposent sur des méthodes informatiques éprouvées et s'appuient sur l'exploitation de ressources textuelles adaptées. Les modélisations linguistiques et statistiques employées mettent en œuvre des approches symboliques et numériques isolées ou combinées. En ce qui concerne les ressources textuelles exploitées, elles relèvent exclusivement du champ de l'écrit, qu'il soit standard, spécialisé ou déviant et dans une perspective monolingue ou multilingue.

Dans ce chapitre, nous souhaitons revenir sur les approches et ressources déployées en mettant l'accent sur leur synergie. La section 4.1 se concentre sur les méthodes employées et montre leur complémentarité. La section 4.2, quant à elle, se focalise sur l'importance de la sélection des ressources employées et sa relation aux tâches visées. Enfin, la section 4.3 conclut ce chapitre.

4.1 Complémentarité des approches numériques et symboliques

Aujourd'hui, nous sommes loin de la vision manichéenne qui a opposé pendant de nombreuses années les adeptes des approches reposant sur une modélisation symbolique du fait linguistique aux militants des approches numériques. Jusqu'au début des années 90, la part belle était faite aux approches symboliques comme l'indique Susan Armstrong-Warwick dans la préface aux deux numéros spéciaux de *Computational Linguistics* (CL) consacrés à l'exploitation des grands corpus : « *When the idea first arose to publish a special issue of CL on using large corpora, the topic was not generally considered to be part of mainstream CL, in spite of an active community working in this field.*[1] » (Armstrong-Warwick, 1993, p. iii).

[1] « *Quand l'idée a germé de réaliser un numéro spécial de CL sur l'exploitation de grands corpus, le sujet n'était généralement pas considéré comme faisant partie des courants dominants de CL, malgré une commu-*

Cette reconnaissance de la communauté de la linguistique informatique pour les approches numériques, associée à une plus grande disponibilité des données textuelles mais aussi à des machines capables de les traiter, a donné un élan formidable au quantitatif. Lequel se retrouve également dans la communauté française qui consacre à son tour un double numéro de la revue *Traitement Automatique des Langues* (TAL) aux *Traitements probabilistes et corpus* en 1995. Une autre avancée majeure visant à rapprocher les approches symboliques et numériques s'opère dans le cadre du Workshop *The Balancing Act : Combining Symbolic and Statistical Approaches to Language* associé à la conférence ACL (*Association for Computational Linguistics*) en 1994. À cette époque si les deux approches semblent acceptées dans la communauté scientifique, leur complémentarité est encore loin d'être effective comme cela est rappelé dans la préface de l'ouvrage regroupant les communications du précédent Workshop : « *To many researchers, the mere notion of combining approaches to the study of language seems anathema.*[2] » (Klavans et Resnik, 1996, p. vii).

4.1.1 Approche symbolique

Dans nos travaux, l'approche symbolique privilégiée est l'*analyse syntaxique surfacique* (*shallow parsing*) qui à la différence d'une analyse syntaxique complète, réalisée à l'aide d'un analyseur, ne cherche pas à vérifier la grammaticalité d'un énoncé au regard d'une grammaire de référence. L'analyse surfacique — encore appelée *analyse partielle* ou *analyse superficielle* — cherche typiquement à analyser un énoncé à travers son découpage en segments élémentaires (*chunks*) comme pour le repérage de syntagmes nominaux, l'extraction de la dépendance sujet... (Abney, 1991). Il s'agit ici d'une analyse qualifiée de *robuste* puisqu'elle vise à toujours donner un résultat pour une phrase donnée et n'est pas limitée à cause de l'incomplétude des lexiques et des grammaires, ou encore de la longueur des phrases à traiter (Abney, 1996). Cette approche, décrite généralement par des automates à états finis, est bien adaptée à l'identification de patrons morpho-syntaxiques (Daille, 2003b) ou lexico-syntaxiques (Morin, 1999).

Cette analyse surfacique était à la base du travail réalisé en extraction de relations sémantiques entre termes à partir de corpus de textes techniques (Morin, 1999). Les relations que nous cherchions à identifier étaient modélisées sous la forme de patrons lexico-syntaxiques simples mais très diversifiés. À cette époque, nous cherchions aussi à mieux comprendre le particularisme des relations extraites à l'aide de patrons lexico-syntaxiques par rapport à ce que peut fournir une approche distributionnelle. En ce sens, nous avions étudié avec Michal Finkelstein-Landau (Université de Bar Ilan, Israël) dans le cadre du projet Franco-Israélien *Term Level Text Mining*[3] la complémentarité d'une approche non supervisée (approche distributionnelle) avec une approche supervisée (approche à base de patrons lexico-syntaxiques) pour l'acquisition de relations entre termes caractéristiques de la relation de fusion d'entreprises (Finkelstein-Landau et Morin, 1999). Les résultats obtenus dans cette étude nous ont

nauté active travaillant dans ce domaine. »

[2] « *Pour beaucoup de chercheurs, la seule idée de combiner des approches pour l'étude de langue semble une abomination.* »

[3] Projet Franco-Israélien *Term Level Text Mining : Representations & Algorithms* (1996-1998) sous la direction de Béatrice Daille (IRIN, Université de Nantes) et Ido Dagan (Université de Bar Ilan, Israël) soutenu par l'AFIRST (Association Franco-Israélienne pour la Recherche Scientifique et Technologique) dans le cadre de l'appel sur les Autoroutes de l'information.

permis de montrer la complémentarité des approches afin d'obtenir une large couverture de la relation visée. En effet, l'approche non supervisée extrait des relations implicites non étiquetées alors que l'approche supervisée extrait des relations bien construites. En outre, les résultats obtenus par une approche non supervisée fournissent une amorce intéressante pour l'acquisition de patrons lexico-syntaxiques.

La principale limite d'une approche surfacique à base de patrons lexico-syntaxiques est liée à la portée des patrons, à savoir la phrase. Or, comme les relations visées ne sauraient se limiter à la phrase, il est nécessaire de mettre en œuvre différentes stratégies pour en élargir la couverture. Une première stratégie consiste éventuellement à s'appuyer sur une analyse distributionnelle comme cela a été rappelé dans le précédent paragraphe. Une autre stratégie consiste à s'appuyer sur les relations extraites par les patrons lexico-syntaxiques pour en inférer de nouvelles par un mécanisme de variation sémantique (Morin et Jacquemin, 1999, 2004). Cette recherche de variantes, réalisée au moyen de *FASTER* (Jacquemin, 2001), permet de considérer une large palette de variations : des variations morpho-sémantiques telles que *contenu en isotope*, une variante de *teneur isotopique*, des variantes syntactico-sémantiques telles que *vins élevés en fûts*, une variante de *vin en barrique*, et des variantes morpho-syntactico-sémantiques telles que *dureté de la viande* une variante de *la résistance et la rigidité de la chair*. Le déploiement de ce mécanisme, pour élargir les relations implicites identifiées par des patrons lexico-syntaxiques caractéristiques de la relation d'hyponymie, permet de découvrir de nouvelles relations éventuellement plus spécialisées, ou encore transférées vers un autre domaine conceptuel et ce, avec un bon niveau de précision.

En ce qui concerne la modélisation du langage relatif aux MIMEMA, c'est aussi l'approche surfacique qui a été privilégiée. En effet, cette approche est particulièrement bien adaptée à la modélisation de phénomènes lexicaux. Ici, notre démarche a consisté à modéliser chaque phénomène visé sous la forme de règles de transformation. Ces règles, qui assurent une bonne expressivité des phénomènes, sont ensuite utilisées pour générer des lexiques ou des automates qui sont directement associés au système de reconnaissance de formes. Les différentes ressources ainsi constituées permettent de suppléer les ressources disponibles qui sont peu adaptées à la prise en compte de l'écriture MIMEMA.

4.1.2 Approche numérique

Dans le cadre de la reconnaissance de l'écriture manuscrite en-ligne, l'adjonction de connaissances linguistiques au modèle d'écriture a été réalisée à l'aide d'une approche numérique. Ainsi en nous appuyant sur un algorithme inspiré des *nuées dynamiques* (*k-means*), nous avons pu aisément construire un modèle bi-classe statistique qui permet de concurrencer un modèle bi-gramme. Le contexte technologique propre à ce travail, notamment en vue de maîtriser l'encombrement mémoire, d'assurer la cohérence globale du système informatique et de faciliter l'intégration de nouvelles langues, a bien sûr fortement guidé nos choix. Néanmoins, deux aspects importants de ce travail méritent une attention particulière. D'une part, les travaux réalisés — qui héritent largement des acquis engrangés en reconnaissance de la parole, tout en étant pionniers en reconnaissance de l'écriture manuscrite — assurent le résultat attendu d'améliorer de manière significative la reconnaissance, en déployant une faible énergie, là où une énergie considérable sera nécessaire pour gagner une poussière de précision...

D'autre part, les classes induites par l'approche bi-classe statistique présentent l'inconvénient de regrouper des entités linguistiques peu homogènes, lesquelles n'ont ni signification *a priori* ni représentant privilégié. Il devient alors délicat de combiner les classes issues d'un modèle bi-classe statistique avec celles issues d'autres classifications.

En ce sens, nous avons essayé de prendre en compte, dans le modèle d'écriture, des connaissances linguistiques plus fines, en introduisant un *a priori* syntaxique lors de l'initialisation des classes du modèle bi-classe statistique (Perraud, 2005). Nous avons en particulier expérimenté l'initialisation des classes de l'algorithme de classification en nous appuyant sur celles fournies par les catégories de l'outil d'étiquetage. La différence entre le résultat ainsi obtenu et celui relevé avec une initialisation classique, est de l'ordre de 1%. Martin et al. (1998) arrivent aussi à la même conclusion. En outre, nous avons constaté avec cette technique que les classes initiales sont complètement éclatées au terme de la classification. En fait, dans le cadre d'une telle combinaison nous introduisons certainement un biais. En effet, nous essayons de coordonner une classification reposant sur un critère grammatical, qui fournit des classes qui ne sont pas fortement séparées les unes des autres (notamment en raison de la polysémie des mots qui peuvent appartenir à plusieurs classes), avec une classification reposant sur un critère contextuel à l'aide de l'algorithme des *k-means*, qui tend à produire des classes bien séparées les unes des autres (Kearns et al., 1997). C'est donc plus ici la nature des classes mélangées, plus ou moins disjointes selon la technique de classification utilisée, qui doit être mise en cause que l'idée même de combinaison.

4.1.3 Approche mixte

Les travaux initiés avec Béatrice Daille en reconnaissance des entités nommées pour le français (Daille et Morin, 2000) ont été poursuivis par Nordine Fourour dans le cadre de sa thèse (Fourour, 2004). Il a en particulier développé le système *NEMESIS* pour la reconnaissance et la classification des entités nommées pour des textes français. Ce travail, qui s'inscrit dans la continuité des approches symboliques pour la reconnaissance des entités nommées (McDonald, 1994; Wacholder et al., 1997), s'appuie sur des lexiques de mots déclencheurs et des règles de réécriture pour la reconnaissance des entités nommées. Fourour (2004) associe à son système un module d'apprentissage automatique basé sur des heuristiques pour la mise à jour des lexiques, et propose une première réflexion pour la désambiguïsation et l'apprentissage de nouvelles règles de réécriture. À la différence des approches statistiques, qui effectuent un apprentissage de contextes et de structures (Mikheev et al., 1999; Bikel et al., 1997), Fourour (2004) préconise un apprentissage supervisé suivant une technique de *bootstrapping*, à la manière des travaux de Riloff et Jones (1999) ou Quasthoff et al. (2002). Un pas supplémentaire aurait pu être franchi dans ce travail en direction des techniques supervisées d'inférence de règles à partir d'exemples afin d'assurer la généricité de l'étape d'apprentissage.

En acquisition lexicale multilingue, la méthode que nous avons proposée correspond à une approche mixte (cf. section 2.2). En effet, nous commençons par identifier en corpus les termes complexes à l'aide d'*ACABIT*, puis nous les filtrons en nous appuyant sur leur fréquence (les hapax ne sont pas conservés) et non sur leur score d'association. Ensuite, le processus d'alignement réalisé à l'aide de la méthode directe ou par similarité interlangue correspond typiquement à une approche numérique.

Cette combinaison entre approches symbolique et numérique qui tire parti des avantages de chacune est intéressante. D'un côté, une approche symbolique permet de s'appuyer sur une description linguistique précise des observables qui contraint le champ des possibles, notamment par l'utilisation de patrons morpho-syntaxiques ou lexico-syntaxiques. D'un autre côté, l'approche numérique assure la robustesse nécessaire au traitement de données volumineuses. En revanche, je suis moins catégorique que Daille (2002, p. 21) qui affirme que : « *Dès lors qu'une description linguistique des éléments est possible, elle doit se faire en amont du filtrage statistique de manière à ce que les patrons recherchés même s'ils sont peu représentatifs par rapport à l'ensemble des événements potentiels du texte soient pris en compte* ». En effet, il est tout à fait possible de s'appuyer en amont sur une approche numérique, comme nous l'avons fait pour amorcer l'acquisition de patrons lexico-syntaxiques (Finkelstein-Landau et Morin, 1999) ou pour combiner des modèles probabilistes (Perraud et al., 2003a). Cette même combinaison est aussi réalisée par Claveau et Sébillot (2004) pour l'extraction de couples N-V qualia. Dans ce travail, l'approche symbolique supervisée (PLI comme technique d'apprentissage) est amorcée par une approche numérique (extraction de cooccurrences des couples N-V qualia) qui permet de suppléer la phase manuelle de construction des exemples inhérente à chaque nouveau corpus. En ce sens, il faut, dans une approche mixte, privilégier la description linguistique qui doit assurer le rôle d'un filtre contraignant sur les événements potentiels et ce en amont ou en aval d'une approche numérique qui doit bien relever les événements observés.

4.2 Coordination des ressources textuelles

Dans nos différents travaux nous sommes amenés à exploiter un grand nombre de ressources textuelles notamment des corpus monolingues ou bilingues, des dictionnaires et des listes terminologiques. Le corpus constitue notre matière première, celle que nous façonnons dans un dessein précis. Un peu à l'image d'un artisan choisissant avec soin les essences les plus nobles pour produire une œuvre d'une facture unique. Cette relation entre la matière exploitée et le résultat obtenu est ainsi rappelée par Péry-Woodley (2000, p. 155) : « *De même que l'efficacité d'un étiqueteur dépend de la coïncidence entre le corpus d'entraînement et le corpus à traiter, la validité des résultats de l'analyse de corpus pour une recherche particulière est fonction de l'adéquation des données aux objectifs de cette recherche* ». En fait, entre l'idée que l'on peut se faire d'une bonne pratique et celle que l'on met en œuvre il y a souvent un certain fossé. Et nous n'avons malheureusement pas échappé à la règle.

À la lumière des acquis engrangés en fouille terminologique multilingue et en reconnaissance de l'écriture manuscrite, nous souhaitons revenir sur les données exploitées et préciser leur relation aux tâches visées.

4.2.1 Représentativité des corpus

Au début des années 80 le *Brown corpus* avec son million de mots, puis dans les années 90 le *British National Corpus* avec ses cent millions de mots ont largement dopé le champ de la fouille textuelle. Ce recours à des ressources volumineuses est souvent justifié par la nécessité de disposer de grandes masses de données pour mettre en œuvre les méthodes numériques (Church et Mercer, 1993). Cette position est particulièrement caractéristique des méthodes

probabilistes, où à défaut de pouvoir caractériser avec précision les productions langagières, la masse est vue comme la solution pour disposer d'une large palette de productions. Habert (2000), empruntant l'expression à Péry-Woodley (1995), qualifie cette position de « Gros, c'est beau ».

En modélisation probabiliste du langage naturel pour la reconnaissance de l'écriture manuscrite, nous avons typiquement adopté cette posture. Nous nous sommes en effet plus concentrés sur le développement des modèles que sur la nature des ressources exploitées. Pour l'entraînement des modèles, nous avons ainsi exploité des données issues de romans du XIX^{e} et XX^{e} siècle et d'articles du journal *Le Monde*. Cet ensemble constituant à notre sens un corpus représentatif de la « langue générale ». Cette notion même de corpus de « langue générale » mérite une attention particulière. Elle semble difficile à concevoir *ex nihilo* et plus aisé à appréhender par distinction à la notion de corpus de « langue spécialisée », qui envisage un lexique dédié, et des regroupements spécifiques de classes de mots, notamment au niveau des structures prédicat-arguments. L'article journalistique comme le roman littéraire sont des genres discursifs distincts qui participent bien sûr à une certaine représentation de la langue. Mais le vocable recherché qui y est employé et la rhétorique parfois singulière du genre romanesque semblent pas très caractéristiques d'une langue générale. On peut aussi s'interroger sur l'homogénéité des discours envisagés pour un genre donné, notamment pour les articles journalistiques qui mettent côte à côte des éditoriaux, des articles, des portraits, des interviews, des brèves... pour des thématiques parfois lointaines comme le sport et la politique... N'est-il pas ainsi illusoire de vouloir définir ce que pourrait être un corpus de langue générale ?

Un autre aspect peu évoqué mérite aussi une attention particulière. Il concerne le « formatage » induit par l'utilisation par une communauté des mêmes corpus. Si l'on ne peut que se réjouir de la facilité d'accès de ces corpus, il faut rester prudent sur la généricité des résultats obtenus qui ne sont valides qu'au regard des données exploitées. Le corpus [LE MONDE] ne saurait être le représentant idéal de la langue pour les différentes raisons évoquées précédemment. Nous sommes malheureusement de ceux qui en font une utilisation « discutable ». Ainsi, certains des lexiques construits pour la modélisation des MIMEMA se sont appuyés sur ce corpus. Il s'agissait pour nous de pouvoir disposer d'un lexique de langue générale afin de modéliser des phénomènes comme la phonétisation ou l'abréviation de l'écriture. Même en restant au niveau lexical, il existe des formes caractéristiques de l'écriture SMS qui ne seront pas rencontrées comme *à plus*, *salut*, *à demain* qui relèvent plus de l'oralité ou d'un écrit dénotant d'une certaine familiarité entre l'émetteur et le récepteur.

Pour répondre à ces problèmes de représentativité et d'homogénéité, le web semble être aujourd'hui la solution à tous ces maux. Néanmoins, pour Kilgarriff et Grefenstette (2003, p. 333) : « *[...] the Web is not representative of anything other than itself [...]*[4] ». Il faut donc être très prudent face à cette *Tour de Babel* qui mélange allégrement langue, domaine, genre, discours... et face à laquelle nous manquons cruellement de méthodes opérationnelles pour séparer le bon grain de l'ivraie — ce qui correspond au « *risque de confusion dans la profusion* » de Habert (2006, p. 263).

[4] « *[...] le Web n'est représentatif de rien d'autre que de lui-même [...]* »

4.2.2 Qualité des données

Le passage du monolingue au bilingue pour l'acquisition de lexiques bilingues à partir de corpus comparables induit des difficultés supplémentaires. La première est liée à l'absence d'un corpus comparable de référence à l'image des corpus parallèles *Hansard* ou *ECI*. Ainsi, les premiers travaux à partir de corpus comparables se sont appuyés sur des corpus construits à partir de textes journalistiques dans des langues différentes. Par exemple, Fung (1998) exploite deux ans du Wall Street Journal et du quotidien chinois Nikkei Financial News. En sus des précédentes réserves liées à l'utilisation des corpus journalistiques, la mise en correspondance de documents produits dans des contextes culturels distincts induit une nouvelle difficulté de comparabilité.

Dans nos premiers travaux à partir de corpus comparables, nous avions construit notre corpus en nous appuyant sur des articles scientifiques issus de la revue *Unasylva* (cf. annexe A). Ce corpus étant composé de documents scientifiques, les contextes culturels de production sont ici moins sensibles car les productions scientifiques sont suffisamment normées d'un point de vue académique. En revanche ce corpus comparable a été construit à partir de documents traduits en sélectionnant des textes qui ne sont pas la traduction l'un de l'autre. Ici, le vocabulaire rencontré dans la partie traduite risque d'être fortement influencé par celui de la langue source. À cette époque, notre démarche était typiquement conduite par la nécessité de disposer de grandes masses de données pour mettre en œuvre nos méthodes numériques visant à pouvoir produire des lexiques bilingues prenant en compte les problèmes de fertilité et non-compositionnalité. Nous n'avions pas encore conscience de la nécessité de croiser les caractéristiques de domaine, de période, de discours... lors de constitution du corpus comparable. Nous nous limitions alors au domaine comme principale, pour ne pas dire unique, caractéristique. Lors de l'exploitation de ce corpus, nous nous sommes rapidement rendu compte que le domaine n'était pas le seul facteur de variation. En particulier, comme ce corpus couvre autant des aspects liés à la gestion et la conservation des plantations, des forêts et des animaux, que des aspects liés au développement socio-économique, au commerce international et à l'environnement, nous nous sommes rapidement heurtés à des problèmes de polysémie. La thématique doit être aussi contrôlée pour un domaine donné.

Est-il alors possible de construire un corpus comparable répondant à des critères langagiers précis tout en conservant une masse de données suffisante ? En posant le problème ainsi, nous faisons l'hypothèse que la qualité des données peut suppléer à la quantité. Cette démarche correspond aussi à une position différente de « more data is better data » (Péry-Woodley, 1995). Elle correspond à ce que Habert (2000) qualifie d'« insécurité dans les grands ensembles ».

Dans le cadre du projet TCAN-DECO, nous avons cherché à répondre à cette question. Pour ce faire, nous avons récolté à partir du web des documents français/japonais relevant du domaine médical pour la thématique « hygiène et santé » et réduite à la sous-thématique des « maladies liées aux régimes alimentaires » et plus particulièrement au « diabète ». Le diabète, fléau des pays fortement industrialisés, fait l'objet d'une importante production langagière en français comme en japonais. En vérifiant manuellement la pertinence des documents à la thématique, deux types de discours, scientifique ou vulgarisé, correspondant à deux paramètres situationnels (Biber, 1993) ont émergé : les textes sont essentiellement écrits par

des spécialistes (médecin, diététicien, infirmier, etc.) mais sont destinés, soit à un public restreint composé lui-même de spécialistes, soit à un public plus large incluant les malades, les populations à risque, etc. Les textes de non-spécialistes à non-spécialistes sont pratiquement inexistants. Ces documents sont ensuite classés selon le type de discours scientifique ou vulgarisé, la seule contrainte étant d'atteindre la taille minimale de 200 000 mots pour chaque type de discours et pour chaque langue de manière à pouvoir mettre en œuvre les méthodes numériques. Pour la classification, nous nous sommes appuyés non seulement sur le genre du document du web (Karlgren et Cutting, 1994; Beauvisage, 2001), mais aussi sur des critères internes reflétant le contexte sociolinguistique de production comme, par exemple, le niveau de style, la personnalisation, la technicité (Krivine et al., 2006; Goeuriot et al., 2005, 2007). Nous avons constaté que pour les deux langues, le genre dominant du web est le rapport mais des différences dans les sous-genres apparaissent : pour le japonais, les rapports émanent presque exclusivement d'institutions privées alors que, pour le français, ils émanent de sites gouvernementaux, d'instituts universitaires ou d'institutions publiques (hôpitaux). La prépondérance du rapport d'institution privée pour le japonais s'explique par le statut privé et concurrentiel de l'hôpital. Les deux types de discours se déclinent dans tous les genres du web : par exemple, l'article de recherche relève du rapport scientifique, les conseils nutritionnels pour l'enfant diabétique du rapport vulgarisé.

À partir de ces documents, nous avons créé deux corpus comparables, l'un composé exclusivement de documents scientifiques, et l'autre de documents scientifiques et vulgarisés. Nous avons ensuite utilisé notre chaîne de fouille terminologique multilingue pour évaluer l'impact de la caractéristique du type de discours du corpus comparable sur la constitution de listes terminologiques bilingues (Morin et al., 2007). Une première expérience réalisée avec une liste de référence composée de termes simples nous a permis de nous positionner convenablement par rapport aux travaux existants en obtenant une précision de 51 % et 60 % pour les 10 et 20 premiers candidats pour la méthode directe. À ce niveau, l'apport de la classification suivant le type de discours n'est pas significatif puisque les résultats sont meilleurs avec le corpus composé de documents scientifiques et vulgarisés, par comparaison au corpus composé exclusivement de documents scientifiques. Ce résultat était attendu dans la mesure où les termes simples utilisés dans cette évaluation ne sont pas plus caractéristiques du discours scientifique que vulgarisé. Par exemple, le terme *excès* peut tout aussi bien faire référence au discours scientifique dans *excès pondéral* qu'au discours vulgarisé dans *excès de poids*. La seconde expérience mettait en jeu une liste de référence composée de termes complexes caractéristiques du discours scientifique. Ici, les meilleurs résultats sont obtenus avec le corpus scientifique (précision de 30 % et 42 % pour les 10 et 20 premiers candidats pour la méthode directe). Plus précisément, il semble que les documents vulgarisés induisent du bruit dans le corpus comparable. Dans ce cas, la prise en compte du type de discours semble bien agir comme un sélectionneur sémantique adéquat. Bien sûr d'autres expériences pour des domaines et genres distincts seront nécessaires pour confirmer et affiner l'importance de cette hypothèse.

4.3 Synthèse

Au risque de passer pour un rétrograde, on l'aura compris je ne suis pas un fanatique des approches numériques et encore moins du tout symbolique. À chaque problème son approche,

pourvu qu'elle soit la plus pertinente possible au regard des contraintes imposées. Bien sûr de nombreuses avancées ont été réalisées en modélisation probabiliste du langage naturel comme cela est rappelé par Michèle Jardino et Marc El-Bèze dans l'introduction au numéro spécial de la revue TAL en 2003 sur ce sujet : « *Au-delà de la vieille opposition entre les approches numériques et les méthodes à base de connaissances, tout le monde s'accorde pour introduire des règles dans les modèles stochastiques ou des probabilités dans les grammaires, dans l'espoir de cumuler les avantages des deux points de vue.* » (Jardino et El-Bèze, 2003, p. 8).

En amont des approches développées, une attention particulière doit être accordée sur la caractérisation des ressources exploitées afin d'en assurer l'adéquation aux tâches visées. Cette caractérisation préalable ne se limite pas au corpus d'étude, mais comprend aussi l'ensemble des ressources textuelles exploitées, notamment les dictionnaires et lexiques de référence. La qualité des données ainsi sélectionnées peut suppléer à leur quantité. Les méthodes informatiques doivent aussi faire l'objet d'une attention particulière car face au faible volume de données disponibles, les approches numériques reposant sur des mesures de récurrence contextuelle ou de distance vectorielle peuvent être mal adaptées. En ce sens, les approches symboliques doivent assurer, en amont ou en aval des approches numériques, un filtre sur les événements observés.

Chapitre 5

Conclusion et perspectives

Les activités de recherche présentées dans ce mémoire s'intéressent au traitement de l'écrit et s'appuient sur des approches numériques et symboliques. Ces activités ont tout d'abord évolué dans un cadre monolingue (notamment en ce qui concerne la structuration de données terminologiques, la reconnaissance des entités nommées et la reconnaissance de l'écriture manuscrite) avant de s'étendre à des textes bilingues (notamment au niveau de l'acquisition de terminologies bilingues). Il s'agit de la première évolution notable de nos travaux. D'un autre côté, nous commençons progressivement à nous éloigner du domaine de l'écrit « standard », qui considère des textes bien formés et normés, pour nous rapprocher du domaine de l'écrit « déviant » et en particulier des nouvelles formes de communication écrite (notamment les SMS). Il s'agit incontestablement de la seconde évolution majeure de nos travaux de recherche. En ce qui concerne les approches utilisées, nous contribuons à montrer la synergie possible entre approches numériques et symboliques. En particulier, nous montrons que les méthodes probabilistes ne sont plus une alternative aux systèmes à base de règles, mais bien complémentaires. Enfin, nos différents travaux nous ont sensibilisé à l'importance de la construction et sélection de ressources adaptées aux tâches visées.

Au terme de ce mémoire, nous concluons en évoquant plusieurs perspectives égrenées tout au long de ce document et qui font l'objet de travaux en cours ou en devenir.

5.1 Analyse conjointe de documents multimédia

La collaboration entre les équipes IVC (IRCCyN – UMR 6597) et TALN (LINA – FRE CNRS 2729) et la société Vision Objects, initiée dans le cadre du développement de modèles de langage pour améliorer un système existant de reconnaissance de l'écriture manuscrite en ligne, a confirmé l'intérêt de traiter de manière conjointe différents média (ici reconnaissance de formes et langage naturel). En particulier, la prise en compte des différentes modalités d'un document multimédia améliore globalement les performances des applications visées. Dans cette direction, deux perspectives de recherche sont envisagées.

5.1.1 Indexation et recherche de documents manuscrits en ligne

Dans le cadre du projet ANR-CIEL (2006-2008) et de l'axe multimédia du CPER-MILES (2007-2009), nous cherchons à concevoir des technologies logicielles adaptées au traitement de documents manuscrits complexes (c'est-à-dire des documents contenant du texte mais aussi des schémas, des tableaux et autres composants de nature bi-dimensionnelle). Nous visons à développer des solutions de reconnaissance globale de documents manuscrits complexes et ainsi à augmenter les possibilités d'applications dans le domaine de la prise de notes appliquée à de multiple processus métiers ou à un usage personnel (enregistrement, conversion, échange, recherche et actions associées).

Dans ces projets, notre contribution porte plus particulièrement sur la définition d'outils de gestion adaptés à la recherche d'informations dans des bases de documents manuscrits. Dans un premier temps, il s'agit de définir un modèle de catégorisation de documents de l'utilisateur (c'est-à-dire permettre à l'utilisateur de définir ses propres catégories et d'y associer quelques documents, afin que des nouveaux documents soient automatiquement classés suivant cette catégorisation). À ce niveau, la recherche de mots-clés, pour les parties textuelles, est bien sûr nécessaire. Mais à la différence des textes électroniques, il faut gérer le bruit induit par les erreurs du moteur de reconnaissance de l'écriture manuscrite et trouver un compromis entre la taille du lexique de recherche des mots-clés et les performances en termes de précision et de rappel. Dans un second temps, il sera nécessaire d'assurer l'indexation des documents suivant la catégorisation en exploitant d'autres éléments que les parties textuelles comme par exemple la présence d'une équation mathématique que l'on aura été capable de détecter sans pour autant forcément chercher à la reconnaître, un style de mise en page... tout en l'adaptant au scripteur. Enfin, il sera nécessaire de concevoir la navigation dans cette base de documents pour en faciliter la recherche. Ce travail s'inscrit, pour partie, dans le cadre de la thèse de Sebastián Peña Saldarriaga (débutée en avril 2007).

5.1.2 Mini-messages manuscrits

Les travaux initiés dans le champ de la reconnaissance de mini-messages manuscrits (MIMEMA) offrent aussi des perspectives intéressantes. La première est liée à la nécessité de concevoir un modèle qui puisse intégrer de manière coopérative les différents phénomènes modélisés et ce sans dégrader la qualité globale de la reconnaissance. En effet, les ressources que nous avons développées pour la reconnaissance des MIMEMA (c'est-à-dire des lexiques spécifiques et des descriptions sous forme d'automates stochastiques à états finis) sont pertinentes lorsque les phénomènes modélisés (les squelettes consonantiques, l'écriture rébus et la phonétisation de l'écriture) sont traités individuellement. Si nous appliquons ces ressources simultanément, la qualité de la reconnaissance est globalement dégradée. Les techniques séquentielles du TAL étant peu adaptées à cette tâche (Véronis et Guimier de Neef, 2006), d'autres approches devront être envisagées.

Une autre perspective liée à l'usage même du stylo numérique pour écrire des MIMEMA mérite une attention particulière. Comme nous l'avons déjà indiqué, l'utilisation du stylo numérique pour écrire des mini-messages manuscrits engendre l'introduction d'une interface de saisie différente de celle utilisée pour les SMS. En particulier, nous devrions vérifier en quoi

les conditions de production d'un message en modèlent la forme linguistique (Anis, 2003). Cette perspective de recherche attrayante, qui relève du champ de la sémio-linguistique de l'écrit, devrait faire naître de nouvelles collaborations.

5.2 Fouille terminologique multilingue

Les aspects multilingues sont centraux dans nos travaux, certainement en raison de mon intérêt pour la fouille textuelle, celle qui est réalisée manuellement pendant de longues heures afin de vérifier la pertinence d'une idée en corpus.

En fouille terminologique multilingue, l'acquisition de lexiques bilingues à partir de corpus comparables est un champ de recherche encore peu étudié. Certainement en raison de la difficulté d'accès aux ressources comparables mais aussi d'un certain renouveau pour la traduction statistique à partir de corpus parallèles. Par exemple, nous pouvons dénombrer dans les actes de l'année 2007 de la conférence ACL (*Association for Computational Linguistics*) 18 communications concernant les corpus parallèles pour 4 communications pour les corpus comparables.

Les travaux que nous poursuivons en acquisition de lexiques bilingues visent exclusivement les domaines spécialisés pour lesquels les terminologies bilingues sont réduites (notamment si l'on ne tient pas compte de l'anglais) ou nécessitent d'être continuellement mises à jour (notamment dans le domaine médical). Nous nous intéressons particulièrement à l'alignement de termes complexes qui représentent des données précieuses pour les systèmes de recherche d'information interlangue ou encore d'aide à la traduction. Dans ce domaine de recherche, trois axes de recherche complémentaires sont proposés.

5.2.1 Comparabilité de corpus

Comme nous l'avons déjà indiqué la comparabilité de corpus est un élément préalable à l'acquisition lexicale bilingue. Elle permet de s'assurer de la conformité du matériau textuel exploité. Pour automatiser la découverte et la construction de corpus comparables, nous devons être capable de maîtriser les deux risques inhérents à l'accessibilité des masses de données textuelles : celui de « *disette dans la profusion* » et celui de « *confusion dans la profusion* » (Habert, 2006, p. 263).

Dans le cadre de la thèse de Lorraine Goeuriot (débutée en septembre 2005), nous cherchons à préciser cette notion de comparabilité à partir de documents issus du web en mettant en œuvre une analyse stylistique et contrastive pour caractériser des documents relevant des discours scientifique ou vulgarisé. En ce sens, une typologie a été élaborée pour caractériser le discours d'un document web à travers les aspects structurel, modal et lexical (Goeuriot et al., 2005). Cette typologie, implémentée à l'aide d'algorithmes de catégorisation de documents, notamment les séparateurs à vastes marges et les arbres de décision, donne globalement des résultats satisfaisants (Goeuriot et al., 2007). Ce travail est un premier jalon qui permet de prendre en compte de manière effective la caractéristique du discours en constitution de corpus comparables. Néanmoins, il convient d'une part de pouvoir décrire plus finement les différents aspects de la typologie, notamment en ce qui concerne les objets visuels présents

dans les documents web. Ce point, qui relève des aspects structurels de la typologie de catégorisation du discours, est aussi un élément fortement discriminant du discours scientifique (utilisation de graphique, diagramme, etc.). D'autre part, il convient de vérifier la généricité de la typologie et des outils associés pour d'autres langues que le français, japonais et russe. Enfin, la catégorisation du discours utilisée ici est binaire. Un document appartient ou non à un discours donné. Cette rupture brutale entre discours scientifique et vulgarisé ne correspond pas toujours à la réalité des discours observés dans les documents du web. C'est l'acte de communication entre émetteur et récepteur pour un support donné qui doit être affiné.

Si ce travail permet de prendre en compte le risque de « *confusion dans la profusion* » (Habert, 2006, p. 263) en s'appuyant sur le type de discours, il ne permet pas de s'assurer d'un volume suffisant d'observables pour mettre en œuvre des méthodes numériques. Pour maîtriser aussi le risque de « *disette dans la profusion* » (Habert, 2006, p. 263) d'autres travaux seront nécessaires. Pour ce faire, il sera nécessaire de préciser ce qu'est un « volume d'observables suffisant ». À ce niveau, la quantité comme la qualité des observables sont intimement liées. L'un pouvant suppléer l'autre. La problématique à résoudre est donc complexe à appréhender.

5.2.2 Alignement lexical bilingue

Nous avons indiqué en section 2.3 un certain nombre de points à améliorer dans la technique d'alignement terminologique. Parmi ces points une attention particulière doit être accordée à l'approche vectorielle qui n'est pas toujours adaptée. Deux aspects semblent plus particulièrement intéressants. D'une part, l'étude d'autres approches vectorielles comme la LSA (*Latent Semantic Analysis*), proposée pour cette tâche par Gaussier et al. (2004), est une piste à étudier même si elle semble *a priori* plus adaptée pour les termes simples que pour les termes complexes, en raison de leur faible degré de polysémie. D'autre part, nous devons aussi nous interroger sur la pertinence de l'approche vectorielle qui prend en compte au même niveau des informations d'ordre morphologique, syntaxique ou sémantique. Une meilleure prise en compte de ces indices doit se faire de concert avec un modèle idoine qui intègre ses indices à des niveaux distincts, et non uniquement dans un même et unique niveau. Si nous avons montré l'intérêt de la prise en compte des termes complexes dans la description des vecteurs de contexte, cette intégration a toujours été réalisée au même niveau que pour les termes simples. De manière à conserver l'information véhiculée par les termes simples, tout en tirant partie de celle induite par les termes complexes, une piste pourrait être de construire deux espaces vectoriels judicieusement assemblés.

Enfin, nous souhaiterions porter aux corpus comparables le concept de cognat propre aux corpus parallèles. La présence et l'identification de ces indices représentent une donnée précieuse pour définir des points d'ancrage forts entre la langue source et la langue cible. Là aussi, la méthode devra s'adapter à la prise en compte de ces données. Ce dernier aspect fera l'objet d'un développement précis dans le cadre de la thèse d'Emmanuel Prochasson (débutée en octobre 2006).

5.2.3 Identification de traduction en corpus

Dans les travaux que nous avons réalisés (cf. section 2.3), le lecteur peu averti pourrait être surpris par la circularité de notre démarche qui nécessite des dictionnaires bilingues généralistes et spécialisés pour en inférer de nouveaux. Cette circularité est une fâcheuse apparence. Les termes complexes pour lesquels nous cherchons à identifier des traductions ne sont pas recensés dans les dictionnaires généralistes ou spécialisés[1]. En outre, ces termes complexes sont sujets à des phénomènes de fertilité, non-compositionnalité et variation. Si un terme complexe est directement traduisible par un dictionnaire ou par compositionnalité, il ne sera pas proposé en entrée de la plate-forme de fouille terminologique. Dans cette chaîne son seul apport, qui n'est pas des moindres, concerne l'étape de transfert d'une unité à traduire en vue de suppléer à l'insuffisance des entrées multi-mots présentes dans les dictionnaires. Dans l'optique d'améliorer cette étape de transfert, en disposant d'un plus grand nombre d'entrées multi-mots, deux pistes peuvent être poursuivies.

En premier lieu, nous pourrions élargir le processus de traduction compositionnelle de manière à prendre en compte les modifications de structures syntaxiques. Dans les travaux réalisés en français/japonais (Morin et Daille, 2006; Morin et al., 2007), nous avons remarqué que des termes équivalents dans les deux langues pouvaient avoir des structures syntaxiques différentes. Par exemple, le terme français *cellule graisseuse* de structure N ADJ est traduit en japonais par le terme 脂肪細胞 de structure N N ou le nom *cellule* est traduit par le nom 細胞 (*saiboo* - *cellule*) et l'adjectif *graisseuse* par le nom 脂肪 (*shiboo* - *graisse*)[2]. Pour ce faire, nous pourrions nous appuyer sur des règles de réécriture (p. ex. la règle 1) pour associer une structure N_1 ADJ à une structure N_1 Prep Art$^?$ $\mathcal{M}(\text{ADJ}, N_2)$ où $\mathcal{M}(\text{ADJ}, N_2)$ désigne l'association entre un adjectif relationnel (p. ex. *graiss-eux*) et le nom à partir duquel l'adjectif a été dérivé (p. ex. *graiss-e*).

$$\begin{array}{l} N_1 \text{ ADJ} \rightarrow N_1 \text{ Prep Art}^? \ \mathcal{M}(\text{ADJ}, N_2) \\ \mathcal{M}(\text{ADJ}, N_2) = [-\text{ique}, -\text{ie}] \\ \mathcal{M}(\text{ADJ}, N_2) = [-\text{ulaire}, -\text{le}] \\ \mathcal{M}(\text{ADJ}, N_2) = [-\text{eux}, \text{e}] \\ \ldots \end{array} \quad (1)$$

En second lieu, nous souhaitons revenir sur l'idée de variation sémantique bilingue précédemment étudiée dans un cadre monolingue (Morin et Jacquemin, 2004). Un phénomène précis nous intéresse plus particulièrement. Il concerne les variations sémantiques liées aux coordinations partageant la même tête. En français, nous pouvons identifier dans le corpus [SYLV] les termes *tracteurs à roues ou à chenilles*, *tracteurs lourds à roues ou à chenilles*, *tracteurs actuels à chenille ou à roues*, *tracteurs à chenilles ou à roues*, *tracteurs à chenilles et à roues* et *tracteurs à chenilles ou à pneus* et en anglais *wheel and crawler tractors*, *caterpillar and wheel tractors* et *caterpillar or crawler type tractor*. Le mécanisme de variation sémantique développé, associé à *ACABIT* et à une hypothèse de semi-compositionnalité sur

[1] Et quand bien même le seraient-ils, nous nous interdisons leur prise en compte dans le lexique pivot (cf. section 2.3.3).

[2] À noter que si le terme *cellule de la graisse* était présent dans notre corpus, il serait regroupé sous le même candidat terme que *cellule graisseuse*. Dans ce cas, nous pourrions obtenir la traduction de *cellule graisseuse* et éviter ce problème de modification de structure syntaxique.

la tête du terme devraient nous permettre de mettre en correspondance de traduction de telles structures. La principale difficulté étant de ne pas associer des termes qui apparaissent dans les mêmes coordinations sans équivalence identifiée en corpus (cf. figure 5.1).

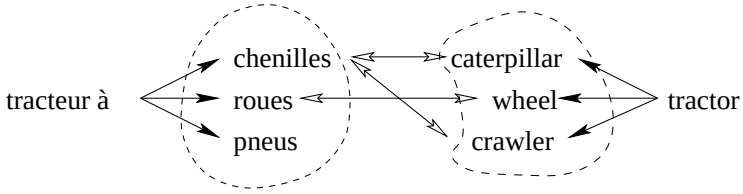


FIG. 5.1 – Exemple de variation bilingue impliquant une coordination

5.3 Pour en finir

Au terme de l'écriture de ce mémoire et de la réflexion qui y est inhérente, je me rends compte du chemin parcouru depuis plus de dix ans dans le *Traitement Automatique de la Langue* (TAL), mais plus encore du chemin restant pour confirmer certaines hypothèses et en vérifier d'autres. Ce mémoire n'est qu'un reflet imparfait de l'ensemble de mes réflexions et des travaux réalisés. Pour donner une vision plus précise de ces aspects, il aurait été nécessaire d'évoquer les idées non poursuivies, les impasses rencontrées ou encore certains travaux aux dimensions technologiques qui trouvent difficilement une place ici. Et pourtant, de même que l'on reconnaît la qualité d'un sportif non seulement à ses victoires mais aussi à ses résultats intermédiaires qui constituent les étapes nécessaires pour arriver au meilleur niveau, la qualité d'un travail scientifique devrait aussi se mesurer à ces étapes intermédiaires. Qu'il me soit donc donné dans l'avenir la possibilité d'en franchir encore de nombreuses. Le TAL ne pouvant se vaincre facilement.

Bibliographie

Abney, S. (1991). Parsing By Chunks. In R. Berwick et C. Tenny, editors, *Principle-Based Parsing*, volume 44, pages 257–278. Kluwer Academic Publishers.

Abney, S. (1996). Part-of-Speech Tagging and Partial Parsing. In S. Young et G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech Processing*, volume 2, pages 118–136. Kluwer Academic Publishers.

Agarwal, R. (1995). *Semantic feature extraction from technical texts with limited human intervention.* Ph.D. thesis, Mississippi State University, MS, USA.

Anis, J. (2001). *Parlez-vous texto ? Guide des nouveaux langages du réseau.* Le Cherche Midi Éditeur, Paris.

Anis, J. (2003). Communication électronique scripturale et formes langagières : chats et SMS. In *Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques*, pages 57–70, Poitiers, France.

Armstrong-Warwick, S. (1993). Preface. *Computational Linguistics*, **19**(1), iii–iv.

Baayen, R. H. (1994). Derivational Productivity and Text Typology. *Journal of Quantitative Linguistics*, **1**(1), 16–34.

Baldwin, T. et Tanaka, T. (2004). Translation by Machine of Complex Nominals : Getting it Right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions : Integrating Processing*, pages 24–31, Barcelona, Spain.

Beaujard, C. et Jardino, M. (1999). Classification de mots non étiquetés par des méthodes statistiques. *Mathématiques Informatique et Sciences humaines*, **147**, 7–23.

Beauvisage, T. (2001). Morphosyntaxe et genres textuels. Exploiter des données morpho-syntaxiques pour l'étude statistique des genres textuels : application au roman policier. *Traitement Automatique des Langues (TAL)*, **42**(2), 579–608.

Bernard, M. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, **20**(2), 155–171.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, **8**(4), 243–257.

Biber, D. (1995). *Dimensions of Register Variation : A Cross-Linguistic Comparison.* Cambridge University Press, Cambridge.

Bikel, D. M., Miller, S., Schwartz, R., et Weischedel, R. (1997). Nymble : a high-performance learning name-finder. In *Procedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington, DC, USA.

Bove, R. (2005). Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS. In *Actes, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL'05)*, pages 625–634, Dourdan, France.

Bowker, L. et Pearson, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora.* Routeledge, London/New York.

Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artifical Intelligence (AAAI'94)*, pages 722–727, Seattle, WA, USA.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., et Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

Cao, Y. et Li, H. (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 127–133, Tapei, Taiwan.

Carl, M. et Langlais, P. (2002). An intelligent Terminology Database as a pre-processor for Statistical Machine Translation. In L.-F. Chien, B. Daille, L. Kageura, et H. Nakagawa, editors, *Proceedings of the COLING 2002 2nd International Workshop on Computational Terminology (COMPUTERM'02)*, pages 15–21, Tapei, Taiwan.

Chen, S. F. et Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 310–318, Santa Cruz, CA, USA.

Chen, S. F. et Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Rapport Technique 10-98, IBM Research.

Chiao, Y.-C. (2004). *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue.* Thèse en Informatique, Université Pierre et Marie Curie, Paris VI.

Chiao, Y.-C. et Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Chiao, Y.-C. et Zweigenbaum, P. (2003). The effect of a general lexicon in corpus-based identification of french-english medical word translations. In R. Baud, M. Fieschi, P. Le Beux,

et P. Ruch, editors, *The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

Church, K. W. et Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, **16**(1), 22–29.

Church, K. W. et Mercer, R. L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, **19**(1), 1–24.

Claveau, V. et Sébillot, P. (2004). Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe. *Traitement Automatique des Langues (TAL)*, **45**(1), 153–182.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse en Informatique, Université de Paris 7, France.

Daille, B. (2002). *Découvertes linguistiques en corpus.* Habilitation à Diriger des Recherches en Informatique, Université de Nantes, France.

Daille, B. (2003a). Conceptual structuring through term variations. In F. Bond, A. Korhonen, D. MacCarthy, et A. Villacicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.

Daille, B. (2003b). Terminology Mining. In M. Pazienza, editor, *Information Extraction in the Web Era*, pages 29–44. Springer.

Daille, B. et Morin, E. (2000). Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *Traitement Automatique des Langues (TAL)*, **41**(3), 601–622.

Daille, B. et Morin, E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Daille, B., Gaussier, E., et Langé, J.-M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, volume I, pages 515–521, Kyoto, Japan.

Daille, B., Enguehard, C., Jacquin, C., Raharinirina, R. L., Ralalaoherivony, B. S., et Lehmann, C. (2000). Traitement automatique de la terminologie en langue malgache. In K. Chibout, J. Mariani, N. Masson, et F. Néel, editors, *Ressources et évaluation en ingénierie des langues*, pages 225–242. Duculot.

Déjean, H. et Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Déjean, H., Gaussier, E., et Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Dufour-Kowalski, S. (2003). *Extraction de terminologies bilingues.* DEA en Informatique, Université de Nantes, France.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1), 61–74.

Fairon, C. et Paumier, S. (2006). A translated corpus of 30,000 French SMS. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC'06)*, pages 615–624, Genève, Suisse.

Falaise, A. (2005). Constitution d'un corpus de français tchaté. In *Actes, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL'05)*, pages 615–624, Dourdan, France.

Fano, R. M. (1961). *Transmission of Information : A statistical Theory of Communications.* MIT Press, Cambridge, MA, USA.

Finkelstein-Landau, M. et Morin, E. (1999). Extracting Semantic Relationships between Terms : Supervised *vs.* Unsupervised Methods. In R. V. Benjamins, D. Fensel, et A. G. Pérez, editors, *Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 71–80, Dagstuhl Castle, Germany.

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*, pages 82–95. Philological Society, Oxford.

Fourour, N. (2004). *Identification et catégorisation automatiques des entités nommées dans les textes français.* Thèse en Informatique, Université de Nantes, France.

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In D. Farwell, L. Gerber, et E. Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Fung, P. et McKeown, K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong, China.

Gaussier, E. et Langé, J.-M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues (TAL)*, **36**(1–2), 133–155.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., et Déjean, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Goeuriot, L., Dubreuil, E., Daille, B., et Morin, E. (2005). Identifying Criteria to Automatically Distinguish between Scientific and Popular Science Registers. In *Proceedings of the 28th SIGIR Workshop on Stylistic Analysis of Text for Information Access*, pages 16–20, Salvador, Brazil.

Goeuriot, L., Grabar, N., et Daille, B. (2007). Caractérisation des discours scientifique et vulgarisé en français, japonais et russe. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 93–102, Toulouse, France.

Goodman, J. (2000). Putting it all together : Language Model Combination. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, pages 1647–1650, Istanbul, Turkey.

Grefenstette, G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, Pays-Bas.

Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publisher, Boston, MA, USA.

Grefenstette, G. (1999). The Word Wide Web as a Ressource for Example-Bases Machine Translation Tasks. In *ASLIB'99 Translating and the Computer 21*, London, United Kingdom.

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In M. Bilger, editor, *Linguistique sur corpus. Études et réflexions*, number 31 in Cahiers de l'université de Perpignan, pages 11–58. Presses Universitaires de Perpignan, Perpignan.

Habert, B. (2006). TAL sur corpus : histoire, acquis, défis. In G. Sabah, editor, *Compréhension des langues et interaction*, Cognition et Traitement de l'Information, chapter 8, pages 249–275. Lavoisier, Paris.

Habert, B., Naulleau, E., et Nazarenko, A. (1996). Symbolic word classification for medium-size corpora. In *Actes, 16th International Conference on Computational Linguistics (COLING'96)*, pages 490–495, Copenhague, Danemark.

Habert, B., Nazarenko, A., et Salem, A. (1997). *Les linguistiques de corpus.* U Linguistique. Armand Colin/Masson, Paris.

Hakusuisha (1989). Dictionnaire des termes techniques et scientifiques français-japonais. Hakusuisha. 4^{e} édition.

Harris, Z. S. (1968). *Mathematical Structures of Language.* Interscience Publishers.

Hindle, D. (1990). Noun classification from predicate argument structures. In *Actes, 28th Annual Meeting of the Association for Computational Linguistics (ACL'90)*, pages 268–275, Berkeley, CA, USA.

Jacquemin, C. (1999). Syntagmatic and Paradigmatic Representations of Term Variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, College Park, MD, USA.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing.* MIT Press, Cambridge, MA, USA.

Janssen, T. M. V. (1996). Compositionality. In J. van Benthem et A. ter Meulen, editors, *Handbook of Logic and Language*, pages 417–473. Elsevier, Amsterdam.

Jardino, M. (1996). Multilingual Stochastic n-gram Class Language Models. In *Proceedings of the 3rd International Colloquium on Grammatical Inference and Applications (ICGI'96)*, pages 161–164, Montpellier, France.

Jardino, M. (1998). Évaluation de modèles de langage à base de trigrammes de classes et de mots, avec le Jeu de Shannon. In *Actes des XXIIèmes Journées d'Etudes sur la Parole (JEP'98)*, pages 363–366, Martigny, Suisse.

Jardino, M. et Adda, G. (1994). Automatic Determination of a Stochastic Bi-Gram Class Language Model. In *Proceedings of the 2nd International Colloquium on Grammatical Inference and Applications (ICGI'94)*, pages 57–65, Alicante, Spain.

Jardino, M. et El-Bèze, M. (2003). Modélisation probabiliste du langage naturel. *Traitement Automatique des Langues (TAL)*, **44**(1), 7–10.

Jelinek, F., Mercer, R. L., Bahl, L. R., et Baker, J. K. (1977). Perplexity - A Measure of Difficulty of Speech Recognition Tasks. In *Proceedings of 94th Meeting of the Acoustic Society of America (ASA'77)*, Miami Beach, FL, USA.

Karlgren, J. et Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, volume II, pages 1071–1075, Kyoto, Japan.

Kearns, M., Mansour, Y., et Ng, A. Y. (1997). An Information-Theoretic Analysis of Hard and Soft Assignment Methods for Clustering. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI'97)*, pages 282–293, Providence, RI, USA.

Kilgarriff, A. et Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, **29**(3), 333–347.

Klavans, J. L. et Resnik, P. (1996). Preface. In J. L. Klavans et P. Resnik, editors, *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, pages vii–xii. MIT Press, Cambridge, MA, USA.

Kneser, R. et Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 181–184.

Krivine, S., Tomimitsu, M., Grabar, N., et Slodzian, M. (2006). Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In *Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles (TALN'06)*, pages 522–531, Leuven, Belgique.

Langé, J.-M., Gaussier, E., et Daille, B. (1997). Bricks and Skeletons : Some Ideas for the Near Future of MAHT. *Machine Translation*, **12**(1–2), 39–51.

Langlois, D., Brun, A., Smaïli, K., et Haton, J.-P. (2003). Événements impossibles en modélisation stochastique du langage. *Traitement Automatique des Langues (TAL)*, **44**(1), 33–61.

Lecomte, J. et Paroubek, P. (1996). Le catégoriseur d'Eric BRILL. Mise en œuvre de la version entraînée pour l'INaLF. Rapport technique, INaLF-CNRS.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In L. LeCam et J. Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Manning, C. D. et Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Martin, S., Liermann, J., et Ney, H. (1998). Algorithms for bigram and trigram word clustering. *Speech Communication*, **24**(1), 19–37.

Matsumoto, Y., Kitauchi, A., Yamashita, T., et Hirano, Y. (1999). Japanese Morphological Analysis System ChaSen 2.0 Users Manual. Rapport technique, Nara Institute of Science and Technology (NAIST).

McDonald, D. D. (1994). Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In B. Boguraev et J. Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, Language, Speech and Communications, chapter 2. MIT Press, Cambridge, MA, USA.

Melamed, I. D. (1997). A Word-to-Word Model of Translational Equivalence. In P. R. Cohen et W. Wahlster, editors, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Madrid, Spain.

Melamed, I. D. (2001). *Empirical Methods for Exploiting Parallel Texts*. MIT Press, Cambridge, MA, USA.

Mikheev, A., Moens, M., et Grover, C. (1999). Named Entity recognition without gazetteers. In *Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Morristown, NJ, USA.

Morin, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse en Informatique, Université de Nantes, France.

Morin, E. (2007). Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité. In *Actes, 7e rencontres Terminologies et Intelligence Artificielle (TIA'07)*, pages 11–20, Sophia Antipolis, France.

Morin, E. et Daille, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, **45**(3), 103–122.

Morin, E. et Daille, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *Traitement Automatique des Langues (TAL)*, **47**(2), 113–136.

Morin, E. et Jacquemin, C. (1999). Projecting Corpus-Based Semantic Links on a Thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 389–396, College Park, MD, USA.

Morin, E. et Jacquemin, C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities (CHUM)*, **38**(4), 363–396.

Morin, E., Dufour-Kowalski, S., et Daille, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables. In *Actes, 11e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN'04)*, pages 309–318, Fès, Maroc.

Morin, E., Daille, B., Takeuchi, K., et Kageura, K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Mourlhon-Dallies, F., Rakotonoelina, F., et Reboul-Touré, S. (2004). *Les discours de l'Internet : nouveaux corpus, nouveaux modèles ?*, volume 8 of *les Carnets de Cediscor*. Presses Sorbonne Nouvelle, Paris.

Namer, F. (2000). Flemm : Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, **41**(2), 523–547.

Ney, H., Essen, U., et Kneser, R. (1994). On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, **8**, 1–38.

Niesler, T. (1997). *Category Based Statistical Language Models*. Ph.D. thesis, University of Cambridge, United Kingdom.

Perraud, F. (2002). *Modélisation du langage naturel basée sur les n-grammes et les n-classes appliquée à la reconnaissance de l'écriture manuscrite en ligne*. DEA en Informatique, Université de Nantes, France.

Perraud, F. (2005). *Modélisation du Langage Naturel Appliquée à la Reconnaissance de l'Écriture Manuscrite En-Ligne*. Thèse en Informatique, Université de Nantes, France.

Perraud, F., Morin, E., Viard-Gaudin, C., et Lallican, P.-M. (2003a). Modèles n-gramme et n-classe pour la reconnaissance de l'écriture manuscrite en ligne. *Traitement Automatique des Langues (TAL)*, **44**(1), 63–92.

Perraud, F., Viard-Gaudin, C., Morin, E., et Lallican, P.-M. (2003b). N-Gram and N-Class Models for On-line Handwriting Recognition. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*, pages 1053–1057, Edinburgh, Scotland.

Perraud, F., Viard-Gaudin, C., Morin, E., et Lallican, P.-M. (2005). Statistical language models for on-line handwriting recognition. *IEICE Transactions on Information and Systems/Document Image Understanding and Digital Document*, **E88-D**(8), 1807–1814.

Perraud, F., Viard-Gaudin, C., Morin, E., et Lallican, P.-M. (2006). Modèles de langages adaptés à la reconnaissance de l'écriture en-ligne. In *Actes du 15e congrès francophone sur la Reconnaissance des Formes et Intelligence Artificielle (RFIA'06)*, page 81, Tours, France.

Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *Traitement Automatique des Langues (TAL)*, **36**(1-2), 213–232.

Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle.* Habilitation à Diriger des Recherches en Linguistique, Université de Toulouse-Le Mirail.

Peters, C. et Picchi, E. (1998). Cross-language information retrieval : A system for comparable corpus querying. In G. Grefenstette, editor, *Cross-language information retrieval*, pages 81–90. Kluwer Academic Publishers.

Piérozak, I. (2000). Les pratiques discursives des internautes. In *Le français moderne*, pages 109–129. Conseil international de la langue française, Paris, France.

Piérozak, I. (2003). Le français tchaté, un objet à géométrie variable ? In M. Marcoccia et B. Fraenkel, editors, *Écrits électroniques : échanges, usages et valeur*, volume 104 of *Langage & Société*, pages 123–144. Maison des sciences de l'homme, Paris, France.

Prochasson, E. (2006). *Reconnaissance de Mini-Messages Manuscrits.* Master en Informatique, Université de Nantes, France.

Prochasson, E., Viard-Gaudin, C., et Morin, E. (2007a). Language Models for Handwritten Short Message Services. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*, pages 83–87, Curitiba, Brazil.

Prochasson, E., Morin, E., et Viard-Gaudin, C. (2007b). Vers la reconnaissance de mini-messages manuscrits. In *Actes du 26e Colloque International sur le Lexique et la Grammaire (LG'07)*, pages 241–248, Bonifacio, France.

Quasthoff, U., Biemann, C., et Wolff, C. (2002). Named Entity Learning and Verification : Expectation Maximization in Large Corpora. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL'02)*, pages 8–14, Tapei, Taiwan.

Rabiner, L. et Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Rajman, M., Besançon, R., et Chappelier, J.-C. (2000). Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues (TAL)*, **41**(2), 549–578.

Rapp, R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Riloff, E. et Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI'99)*, pages 474–479, Orlando, FL, USA.

Robitaille, X., Sasaki, X., Tonoike, M., Sato, S., et Utsuro, S. (2006). Compiling French-Japanese Terminologies from the Web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

Ruge, G. et Schwarz, C. (1991). Term Associations and Computational Linguistics. *Int. Classification*, **18**(1), 19–25.

Sadat, F., Yoshikawa, M., et Uemura, S. (2003). Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval : Hybrid Statistics-based and Linguistics-based Approach. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian languages*, volume 11, pages 57–64, Sappro, Japan.

Salton, G. et Lesk, M. E. (1968). Computer Evaluation of Indexing and Text Processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

Shahzad, I., Ohtake, K., Masuyama, S., et Yamamoto, K. (1999). Identifying Translations of Compound Nouns Using Non-aligned Corpora. In *Proceedings of the Workshop on Multilingual Information Processing and Asian Language processing (MAL'99)*, pages 108–113, Beijing, China.

Smadja, F. (1993). Retrieving Collocations from Text : Xtract. *Computational Linguistics*, **19**(1), 143–177.

TAL (1995). Traitements probabilistes et corpus. *Traitement Automatique des Langues (TAL)*, **36**(1–2).

TAL (2003). Modélisation probabiliste du langage naturel. *Traitement Automatique des Langues (TAL)*, **44**(1).

Tanaka, K. et Iwasaki, H. (1996). Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of the 16th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, volume 2, pages 580–585, Copenhagen, Denmark.

Tanimoto, T. T. (1958). An elementary mathematical theory of classification. Rapport technique, IBM Research.

Tay, Y. H., Marzuki, K., Lallican, P.-M., Knerr, S., et Viard-Gaudin, C. (2001). An Analytical Handwritten Word Recognition System with Word-level Discriminant Training. In *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 726–730, Seattle, WA, USA.

Tay, Y. H., Lallican, P.-M., Knerr, S., et Viard-Gaudin, C. (2002). Un système hybride de reconnaissance de mots manuscrits avec apprentissage global. In *Actes 13e Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA'02)* , pages 771–779, Angers, France.

Véronis, J., editor (2000). *Parallel Text Processing.* Kluwer Academic Publishers.

Véronis, J. et Guimier de Neef, E. (2006). Le traitement des nouvelles formes de communication écrite. In G. Sabah, editor, *Compréhension des langues et interaction*, Cognition et Traitement de l'Information, chapter 8, pages 227–248. Lavoisier, Paris.

Wacholder, N., Ravin, Y., et Choi, M. (1997). Disambiguation of Proper Names in Text. In *Procedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 202–208, Washington, DC, USA.

Wagner, H. (1991). Dictionnaires, bases de données lexicales et lexicographie des langues de spécialité : le traitement des unités complexes. In *Actes, Informatique & Langue Naturelle (ILN'91)*, Nantes, France.

Wagner, R. A. et Fischer, M. J. (1974). The String-to-String Correction Problem. *Association for Computing Machinery*, **21**(1), 168–173.

Zaenen, A. et van den Bosch, A., editors (2007). *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* Association for Computational Linguistics, Prague, Czech Republic.

Zweigenbaum, P. et Habert, B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *Revue de sociolinguistique en ligne GLOTTOPOL*, **8**, 22–44.

Index des auteurs

T

U

V

W

Y

Z

Annexe A

Ressources utilisées en fouille terminologique multilingue

Nous détaillons ici les ressources utilisées en fouille terminologique multilingue pour les expériences liées aux domaines de la sylviculture (noté [SYLV]) et du diabète (noté [DIAB]), à savoir : le corpus comparable, le dictionnaire bilingue et les lexiques de référence.

A.1 Domaine de la sylviculture

A.1.1 Corpus comparable

Le corpus comparable a été construit à partir de la revue *Unasylva* publiée chaque trimestre en anglais, espagnol et français depuis 1947 par la FAO[1]. Cette revue internationale consacrée aux forêts et aux industries forestières couvre autant des aspects liés à la gestion et la conservation des plantations, des forêts et des animaux, que des aspects liés aux développements socio-économiques, au commerce international et à l'environnement. Afin d'obtenir un corpus comparable français/anglais, nous avons sélectionné les textes qui ne sont pas la traduction l'un de l'autre. Nous obtenons ainsi un corpus comparable[2] composé de 2,6 millions de mots pour le français et de 2,3 millions pour l'anglais.

A.1.2 Dictionnaire bilingue

Le dictionnaire bilingue, nécessaire au processus d'alignement, a été construit à partir de ressources disponibles sur le web. Il est composé de 22 300 mots en français avec en moyenne 1,6 traductions par entrée. Il s'agit donc d'un dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine de la sylviculture.

[1] `http://www.fao.org/forestry/foris/webview/forestry2/index.jsp?siteId=2342`

[2] Il convient de préciser, d'une part, que nous utilisons des textes qui ont été manuellement traduits et qui peuvent donc comporter des erreurs humaines de traduction, et d'autre part, que ces textes sont des documents numérisés à partir de la revue papier qui peuvent comporter des erreurs de reconnaissance.

A.1.3 Lexiques de référence

Les lexiques de référence ont été construits à partir de trois ressources terminologiques :

1. Le glossaire bilingue de la terminologie de la sylviculture au Canada du service canadien des forêts[3]. Celui-ci couvre les domaines usuels de la pratique de la sylviculture au Canada. Il est composé de 700 termes spécialisés dont 70 % sont des termes complexes.

2. Le lexique multilingue du projet Eurosilvasur (plate-forme ressource forêt-bois-papier des régions de l'Europe du Sud)[4]. Celui-ci couvre un ensemble de domaines liés à l'exploitation de la forêt : économie et exploitation forestière, transformation du bois, sylviculture, etc. Il est composé de 2 800 termes dont 66 % sont des termes complexes.

3. Le thesaurus multilingue AGROVOC de la FAO[5]. Ce thesaurus, destiné à l'indexation des données entrant dans les systèmes d'informations agricoles, couvre les domaines de l'agriculture, de la pêche, de la sylviculture, de la nutrition, de l'innocuité des produits alimentaires, ainsi que divers sujets connexes comme l'environnement. Il comporte 15 000 descripteurs pour le français dont 47 % sont des termes complexes.

Ces trois ressources terminologiques sont complémentaires dans la mesure où elles proposent des termes plus ou moins spécialisés. Ainsi, le glossaire bilingue est plus spécialisé que le lexique multilingue, qui lui-même est plus spécialisé que le thesaurus multilingue.

À partir de ces ressources, nous avons sélectionné automatiquement 300 termes français, où chaque terme français est au moins présent cinq fois dans le corpus comparable, pour constituer nos trois lexiques de référence :

- [lexique 1] 100 termes simples français dont la traduction, qui n'est pas présente dans notre dictionnaire bilingue, est un terme simple.
- [lexique 2] 100 termes complexes français dont la traduction peut être un terme simple ou complexe. Ces termes ne peuvent pas être traduits directement ou de manière compositionnelle à partir de notre dictionnaire bilingue.
- [lexique 3] 100 termes complexes français dont la traduction est un terme complexe. Ces termes ne peuvent pas être traduits directement mais le sont de manière compositionnelle par leurs composants à partir de notre dictionnaire bilingue.

Ces lexiques de référence sont essentiellement composés de termes peu fréquents (cf. tableau A.1). Deux raisons majeures expliquent ce phénomène. D'une part, les différentes ressources utilisées pour créer les lexiques de référence proposent des termes très spécifiques ou très génériques. D'autre part, le corpus utilisé couvre un grand nombre de domaines liés à la foresterie et ne constitue pas une ressource très spécialisée.

[3] http://nfdp.ccfm.org/silviterm/silvi_f/silvitermintrof.htm
[4] http://www.eurosilvasur.net/francais/lexique.php
[5] http://www.fao.org/agrovoc/

$NB_{occ.}$	$[0, 50[$	$[50, 100]$	$]100, 1000]$	$]1000, \infty[$
[lexique 1]	50	21	18	11
[lexique 2]	54	21	25	0
[lexique 3]	51	18	29	2

TAB. A.1 – Fréquence dans la partie française du corpus comparable des termes français des lexiques de référence

A.2 Domaine du diabète

A.2.1 Corpus comparable

Le corpus comparable a été construit à partir de documents manuellement collectés sur le web[6]. Il s'inscrit dans domaine médical, restreint à la thématique « hygiène et santé » et à la sous-thématique des « régimes alimentaires ». Ce corpus comparable distingue les documents selon le type de discours. Le tableau A.2 présente les principales caractéristiques des documents récoltés, à savoir leur nombre ainsi que le nombre de mots pour chaque langue et pour chaque type de discours[7]. En français comme en japonais, les documents scientifiques sont moins nombreux que les documents vulgarisés mais plus volumineux en terme de nombre de mots.

	Français		*Japonais*	
	Nb. doc.	*Nb. mots*	*Nb. doc.*	*Nb. mots*
Discours scientifique	65	425 781	119	234 857
Discours vulgarisé	183	267 885	419	572 430
Total	248	693 666	538	807 287

TAB. A.2 – Caractéristiques générales des documents français/japonais récoltés à partir du web

À partir des précédents documents, nous avons constitué deux corpus comparables : [corpus scientifique] avec les documents de discours scientifique et [corpus mixte] avec les documents de discours scientifique et vulgarisé.

[6]Pour une description détaillée de la méthode, nous renvoyons les lecteurs à l'article de Morin et Daille (2006).

[7]Pour le japonais, le nombre de mots correspond au nombre d'occurrences des formes lexicales reconnues par Chasen Matsumoto et al. (1999).

A.2.2 Dictionnaire bilingue

Le dictionnaire français/japonais a été construit à partir de quatre ressources disponibles sur le web (notées [dico 1][8], [dico 2][9], [dico 3][10] et [dico 4][11]) et d'un dictionnaire électronique de termes techniques et scientifiques (Hakusuisha, 1989) (noté [dico 5]). En dehors du [dico 4] spécialisé dans le domaine médical, les autres ressources sont des dictionnaires généralistes ou techniques. Le tableau A.3 présente les principales caractéristiques des différents dictionnaires utilisés[12]. La fusion de ces différentes ressources permet de constituer une dictionnaire bilingue composé de 173 156 entrées distinctes pour le français avec en moyenne 2,1 traductions par entrée.

	Nature	*Nb. mots*	*Nb. mots simples*	*Nb. mots composés*	*Nb. traductions par entrée*
[dico 1]	généraliste	9 939	7 414	2 525	1,4
[dico 2]	généraliste	45 042	41 046	3 996	1,6
[dico 3]	généraliste	63 772	45 624	18 148	3,8
[dico 4]	médicale	2 329	1 136	1193	2,2
[dico 5]	technique	65 154	31 269	33 885	1,3
Total		**173 156**	**114 461**	**58 695**	**2,1**

TAB. A.3 – Caractéristiques des différents dictionnaires français/japonais

A.2.3 Lexiques de référence

Deux lexiques de référence ont été construits :
- [lexique_TS] rassemble 100 termes simples (TS) français dont la traduction japonaise est un terme simple. Ces termes n'appartiennent pas au dictionnaire bilingue.
- [lexique_TC] rassemble 60 termes simples ou complexes (TC) avec les caractéristiques suivantes : un terme simple français est traduit par un terme complexe japonais, ou inversement ; un terme complexe français est traduit par un terme complexe japonais. Là encore, ces termes ne peuvent pas être traduits directement ou à l'aide d'un processus de traduction compositionnelle à partir du dictionnaire bilingue.

Les termes complexes du [lexique_TC] doivent répondre à trois contraintes supplémentaires :

1. Ils attestent d'au moins deux occurrences dans le [corpus scientifique] ;
2. Ils ont été proposés par les programmes d'extraction terminologique français et japonais ;

[8] `http://kanji.free.fr/`
[9] `http://quebec-japon.com/lexique/index.php?a=index&d=25`
[10] `http://dico.fj.free.fr/index.php`
[11] `http://quebec-japon.com/lexique/index.php?a=index&d=3`
[12] Pour disposer d'une ressource adaptée au corpus comparable, les entrées du dictionnaire bilingue ont été lemmatisées.

3. Soit le terme français et sa traduction japonaise appartiennent au méta-thesaurus de l'UMLS[13], soit le terme français a été recensé par le *Grand dictionnaire terminologique*[14] dans le domaine de la médecine.

Ces contraintes ne nous ont pas permis d'atteindre la taille de 100 termes simples ou complexes pour le [lexique_TC]. L'intersection entre les candidats termes complexes français extraits, d'une fréquence de deux occurrences au moins et les référentiels terminologiques est de 177 termes. Sur ces 177 termes, seuls 60 d'entre eux ont une traduction en japonais qui correspond à un candidat terme complexe identifié dans le [corpus scientifique].

[13] `http://www.nlm.nih.gov/research/umls`
[14] `http://www.granddictionnaire.com/`