ENRICHISSEMENT TERMINOLOGIQUE EN ANGLAIS FONDÉE SUR DES DICTIONNAIRES GÉNÉRAUX ET SPÉCIALISÉS

Résumé

Nous présentons une méthode d'extraction (enrichissement) terminologique en anglais fondée sur l'utilisation de ressources linguistiques et terminologiques déjà existantes. Nous utilisons les «termes simples significatifs» de chaque domaine, i.e. les substantifs, adjectifs, adverbes et participes apparaissant parmi les termes déjà répertoriés, pour rechercher des nouvelles séquences qui contiennent certains de ces termes simples significatifs, complétés éventuellement par des mots grammaticaux ou néologismes. Cette méthode, indépendante de la taille du corpus traité, permettra de construire un logiciel d'aide à la traduction.

Termes-clés : traitement automatique du langage naturel ; extraction terminologique ; acquisition de terminologie ; dictionnaires électroniques ; traduction assistée par ordinateur.

Introduction

Dans cet article nous présentons un prototype d'outil d'extraction automatique de termes, attaché à une grande base de données terminologique multilingue, LexPro, contenant plusieurs millions de termes pour une trentaine de domaines techniques. Cette ressource terminologique très riche est pour nous le point de départ pour la recherche de termes complexes d'un texte, qui sont soit déjà recensés dans la base, soit construits à partir du même matériau lexical que les termes déjà connus. D'autre part, les lexiques très complets de la langue générale, notamment ceux de l'anglais et du français, élaborés au LADL[1], nous permettent de proposer parmi les nouveaux candidats-termes ceux qui contiennent éventuellement des néologismes[2], des noms propres etc. Les premiers résultats prometteurs obtenus dans le domaine de l'informatique, pour lequel nous disposons d'un lexique de 85000 termes anglais, simples et composés, nous permettent d'envisager l'adaptation de notre extracteur à d'autres langues et domaines de LexPro, ainsi que son utilisation en tant qu'outil d'aide à la traduction.

Au cours de l'extraction, nous nous limitons à la recherche de termes complexes, i.e. contenant au moins deux mots simples (pour la définition du mot simple voir note 5). L'extraction se fait sur un texte étiqueté (partiellement ambigu) par les dictionnaires spécialisés et généraux, et elle est fondée sur la recherche de patrons syntaxiques dans le texte. Nous verrons que déjà des patrons assez simples peuvent donner des bons résultats (très bon rappel, précision relativement bonne) si les dictionnaires utilisés sont suffisamment riches.

Avant que l'extracteur puisse intervenir, une phase importante est celle de la préparation des ressources de LexPro, i.e. le nettoyage des dictionnaires, le classement des termes par catégories grammaticales et par la structure syntaxique, le codage des mots inconnus et la flexion automatique des termes simples et composés. Ce travail, seulement partiellement automatisable, a beaucoup d'importance pour la qualité du résultat final. Ceci peut être vu comme l'inconvénient principal de notre méthode. Remarquons néanmoins qu'une fois la phase préparatoire accomplie, nous pouvons utiliser les mêmes dictionnaires comme une base

très fiable non seulement pour l'extraction terminologique, mais aussi pour d'autres tâches, telles que l'automatisation de la traduction, l'indexation documentaire etc.
Dans la section suivante nous argumentons le choix de notre méthode qui peut être classée comme fortement fondée sur des dictionnaires, et indépendante de la taille des corpus traités. La section 2 explique l'utilité d'un extracteur terminologique en tant qu'outil d'aide à la traduction. La section 3 décrit les formats des dictionnaires utilisés. Dans la section 4 nous montrons les phases du travail de l'extracteur, et la section 5 présente les résultats obtenus en anglais dans le domaine de l'informatique, traité avec un dictionnaire spécialisé de 85000 entrées. Dans la section 6 nous analysons les aspects novateurs de notre approche. Finalement, dans la section 7 nous montrons les perspectives de notre logiciel : les possibilités d'affinement des patrons de recherche, l'adaptation au français et à d'autres langues, la prise en compte des formats non ascii des textes, etc.

1 Pourquoi cette approche ?

De nombreuses sociétés, centres scientifiques et corps administratifs effectuent des travaux visant le recensement et l'unification des terminologies propres à leurs domaines d'activités. Ainsi des dictionnaires spécialisés et des listes terminologiques de tailles souvent très importantes sont accessibles en vente, ou bien fonctionnent comme outils internes, dont la maintenance et l'enrichissement sont parfois confiés à une équipe de terminologues.
D'autre part, des scientifiques en ingénierie linguistique, comme Daille (1994), Bourigault (1994), proposent des outils d'extraction automatique de termes, qui dans la plupart des cas admettent le corpus traité comme le seul point de départ. Ceci est idéal pour le traitement de nouveaux domaines techniques ou pour des utilisateurs ne possédant pas de ressources terminologiques. Néanmoins, ceux qui ont déjà fait un effort de constitution de bases terminologiques ou bien ceux qui utilisent des dictionnaires de domaines bien définis, n'ont pas la possibilité, avec ces logiciels, de réutiliser leurs ressources déjà disponibles.
Un des travaux importants fondés sur un lexique spécialisé existant et permettant l'*enrichissement* terminologique plutôt que l'acquisition initiale, est celui de Jacquemin (1997). Ses résultats étant de très bonne qualité du point de vue de la pertinence des candidats-termes proposés, il est nécessaire d'utiliser un corpus de taille importante (l'auteur travaille sur un texte de 1,6 million de mots) afin d'obtenir un rendement satisfaisant (3300 nouveaux termes à partir d'un lexique de 70 000 entrées).
Pourtant, un très grand corpus du domaine traité n'est pas toujours disponible. En particulier dans le cadre d'aide à la traduction technique, dans lequel nous nous plaçons, les traducteurs ont rarement affaire à des documents qui dépassent 1 mégaoctet de texte. Ils disposent par contre presque toujours d'un ou plusieurs dictionnaires techniques, et éventuellement de lexiques personnels ou fournis par le client. Il nous a donc paru intéressant de proposer un extracteur terminologique qui permette de réutiliser des listes de termes disponibles, et dont la qualité de résultats ne dépende pas de la taille du corpus traité.

2 Extraction terminologique au service d'un traducteur technique

Dans le travail d'un traducteur technique un rôle important est attribué à la constitution d'un *glossaire* du document à traduire. Le traducteur, pas toujours expert du domaine traité, lit le texte en langue source et répertorie tous les termes simples et complexes inconnus ou difficiles à traduire, accompagnés d'exemples de leurs occurrences dans le texte. Cette liste est ensuite envoyée au client qui fournit ses propres traductions ou valide celles proposées par le traducteur. Gouadec (1997) propose une méthodologie très précise de création d'un tel

glossaire, appelé chez lui un *concordancier*, et explique son rôle en tant que garant de l'homogénéité terminologique, ainsi que sa valeur contractuelle entre le traducteur et son client.
La constitution et la validation du glossaire du texte devraient en principe être effectuées avant le début de la phase de traduction. Ceci peut entraîner des délais importants, surtout pour des documents volumineux. C'est ici qu'un programme d'extraction automatique de candidats-termes pourra intervenir. Il analysera le texte et sortira instantanément une liste de candidats-termes, classés selon leurs fréquences d'apparitions, parmi lesquels le traducteur choisira ceux qu'il voudra inclure dans son glossaire.
Dans ce cadre, l'extracteur doit viser le maximum de rappel possible, car, pour la constitution du glossaire, le traducteur ne travaillera plus sur le texte entier, mais sur la liste de candidats. Il n'aura donc aucune possibilité de «rattraper» les termes qui ont échappé à l'extracteur. En même temps, la liste des candidats ne peut pas être excessivement longue (précision relativement bonne), en particulier le temps de sa consultation ne peut pas dépasser celui de la création du lexique «à la main». Remarquons néanmoins la difficulté de définir les notions de rappel et de précision dans notre contexte, liée à la question de ce qui doit être considéré comme un bon terme. Pour un terminologue qui dépouille de grandes quantités de textes, un bon terme est celui avec un statut établi constaté dans différentes sources et chez plusieurs auteurs. En revanche, pour un traducteur, un terme qu'il faut retenir est celui qui pose un problème de traduction dans le texte traité. Un candidat retenu par le traducteur pour le glossaire d'un texte donné peut ne plus l'être pour un autre texte, un autre client ou un autre contrat de traduction.

3 Dictionnaires électroniques généraux et spécialisés

Les points de départ pour la recherche de termes seront pour nous deux grandes ressources linguistiques et terminologiques : les dictionnaires électroniques généraux[3] élaborés selon la méthodologie du LADL, et LexPro, une grande base de données terminologiques multilingue[4] et multidomaine. A l'état actuel, notre outil de recherche de termes n'utilise que la partie anglaise de ces ressources, mais nous envisageons d'étudier son adaptation au français et ensuite à d'autres langues.

3.1 Dictionnaires électroniques du LADL.

Le dictionnaire électronique des mots simples[5] DELAS[6] de l'anglais, contient les formes de base (l'infinitif pour les verbes, le singulier pour les noms etc.) des mots simples avec leurs codes flexionnels. Le *code flexionnel*, décrit la façon d'obtenir toutes les formes fléchies d'un mot simple à partir de sa forme de base. Par exemple l'entrée *loaf,N6* indique le code *N6*, équivalent à l'ensemble de terminaisons *(<E> :s ,1ves :p)* , qui signifient que le singulier est égal à la forme de base (il faut ajouter <E> i.e. une séquence vide à la forme de base) et que le pluriel *loaves* s'obtient en enlevant une lettre de la fin et en rajoutant la terminaison *ves*. Le DELAS anglais fournit une bonne couverture de la langue générale, car il contient à présent plus de 90000 entrées. A partir du DELAS, on obtient automatiquement le dictionnaire des formes fléchies des mots simples, le DELAF, contenant plus de 170000 entrées. A chaque entrée est attribuée une étiquette indiquant sa catégorie (nom, adverbe, adjectif etc.) et éventuellement ses traits morphologiques (nombre, genre, personne etc.). Par exemple, le mot *permits est* décrit par deux lignes suivantes :
[1] permits,permit.N:p

[2] permits,permit.V:P3s

C'est soit le pluriel du nom *permit*, soit la troisième personne du singulier indicatif présent du verbe *to permit.*
Les mots composés[7] sont recensés dans un autre dictionnaire, le DELAC, qui, pour chaque entrée, donne sa catégorie, ses traits flexionnels et la façon dont elle se fléchit, par exemple :

[3] point(point.N1:s) of view,N :s/+N

Cette entrée du DELAC indique, que *point of view* est un nom composé au singulier (N :s) et pouvant se mettre au pluriel (/*+N* signifie qu'il y a un flexion en nombre) par la mise au pluriel de ses *constituants caractéristiques*, i.e. les composants simples pour lesquels le code flexionnel est indiqué, ici *point* (pour les détails de la flexion automatique des composés consulter Chrobot (1998)). Ces informations permettent de générer automatiquement le DELACF, le dictionnaire des mots composés fléchis, comme le montrent les exemples suivants :
[4] point of view,.N :s

[5] points of view,point of view.N :p

Le DELAC anglais, qui est au cours de réalisation, comprend à présent près de 60000 entrées, dont environ 50000 noms composés, 4000 adverbes composés (e.g. *all of a sudden*), 3500 adjectifs composés (e.g. *left-handed*), 300 prépositions composées (e.g. *in front of*) et 100 conjonctions composées (e.g. *as well as*).

3.2 LexPro

La base terminologique LexPro, construite à partir de 120 dictionnaires spécialisés traditionnels, mis sur un support informatique, contient actuellement près de 5 millions de termes en 11 langues, dont environ 2 millions en anglais. L'exhaustivité et la qualité des données varient beaucoup d'un dictionnaire à l'autre et d'un domaine à l'autre. De nombreuses entrées nécessitent la correction orthographique, la mise en évidence de certaines abréviations, la séparation des variantes, ou l'unification du format (effacement des déterminants initiaux, remise au singulier etc.). Très peu d'auteurs de dictionnaires indiquent les propriétés grammaticales de leurs termes, telles que catégorie, genre, nombre, existence du pluriel etc., indispensables pour notre extracteur. C'est pourquoi l'exploitation de notre base demande une phase préparatoire, seulement partiellement automatisable. Elle permet d'obtenir une très haute qualité des données utilisées, qui est la source principale de l'efficacité de notre approche (voir section 6). Cette phase doit comprendre entre autres :

- L'analyse lexicale des termes du LexPro à l'aide du dictionnaire DELAF général décrit ci-dessus, afin de retrouver tous les mots simples inconnus que l'on doit ensuite coder, i.e. fournir pour chacun un code flexionnel comme cela a lieu dans le dictionnaire DELAS (ce codage est manuel).
- La correction des termes mal orthographiés (semi-automatique).
- Pour chaque terme, le marquage de sa catégorie (semi-automatique).
- Pour les termes complexes, le marquage de ses *constituants caractéristiques*, ou *têtes* (semi-automatique) – voir section 3.1.

Tout le contenu de la base LexPro n'est pas pertinent du point de vue de la recherche de termes. Nous n'avons pas besoin de certains champs, comme définitions, précisions, commentaires, sources des données etc. Nous allons donc extraire de la base ce que nous appelons les « formes écrites » : les termes principaux, leurs synonymes, leurs abréviations, leurs antonymes etc., i.e. les « vraies » unités terminologiques telles qu'elles peuvent être trouvées dans des textes. Ensuite nous convertirons les données ainsi obtenues en deux

dictionnaires ressemblants au DELAS et le DELAC du LADL : l'un pour les termes simples et l'autre pour les termes composés.
Nous avons accompli le prétraitement des données décrit ci-dessus pour deux dictionnaires du domaine de l'informatique, celui de De Sollier (1999) et de Hildebert (1998), et nous avons obtenu un DELAS spécialisé de 27000 entrées et un DELAC spécialisé de 57000 entrées, dont voici des exemples :
[6] interrupt,N1+Spec

[7] interrupt,V7+Spec

[8] arithmetic overflow indicator(indicator.N1:s),N+Spec+Comp:s/+N

La flexion automatique de ces deux dictionnaires a engendré le DELAF et le DELACF informatiques de 74000 et 108000 entrées respectivement. Le trait supplémentaire *+Spec* renseigné pour chacun des termes simples et composés permettra, en cours de la recherche de nouveaux termes, de faire la distinction entre les étiquettes provenant des DELAF/DELACF généraux et celles des DELAF/DELACF spécialisés. En revanche, le trait *+Comp* est celui qui différenciera les termes spécialisés composés et simples.

4 Phases de l'extraction

Nous ramenons le problème de l'extraction de termes à celui de la recherche de patrons syntaxiques dans un texte. Le texte est d'abord soumis à l'analyse lexicale qui attribue à chaque unité atomique une ou plusieurs étiquettes syntaxiques provenant des dictionnaires utilisés. Ensuite, dans le corpus ainsi étiqueté, nous cherchons toutes les séquences qui correspondent au patron syntaxique donné. Ces séquences seront les candidats que l'utilisateur va pouvoir valider, i.e. décider s'ils sont ou non des termes.
Le schéma de fonctionnement de l'extraction est montré sur l'illustration I1, où les éléments ovales représentent les différentes phases de l'algorithme, tandis que les éléments rectangulaires correspondent aux entrées et sorties de ces phases (les dictionnaires DELAF et DELACF entourés des rectangles dessinés en gras sont consultés en mode prioritaire – voir section 4.1). Nous pouvons voir que le nombre de données en entrée est le plus important dans l'étape de la recherche des mots simples et composés du texte. En effet, les résultats de cette étape sont décisifs pour l'efficacité de la méthode.
Tous les 4 algorithmes utilisés - l'indexation, l'analyse lexicale des mots simples, l'analyse lexicale des mots composés, et la recherche de patrons - ont été récupérés du système INTEX développé au LADL. Pour la représentation des dictionnaires, aussi bien que pour le dépouillement des corpus, ils emploient un modèle à états finis. Chaque dictionnaire utilisé pour l'étiquetage est converti en un automate à état finis, ce qui permet son consultation en temps linéaire en fonction de la longueur du mot recherché. Les patrons syntaxiques de l'extraction sont, eux aussi, des automates finis. Pour les détails de l'implémentation, consulter Silberztein (1997).

4.1 Etiquetage du texte

Deux phases préliminaires du traitement sont celles de l'identification des items[8] du texte, et de la constitution de l'index qui, pour chaque item, donne la liste de toutes ses occurrences. L'indexation nécessite un temps supplémentaire de traitement, mais elle accélère, surtout pour des corpus volumineux, les autres étapes de l'extraction. L'analyse lexicale s'occupe ensuite de la reconnaissance des mots simples et composés du texte selon 5 dictionnaires DELAF/DELACF et deux niveaux de *priorité*.

Les mots simples sont recherchés dans 3 dictionnaires : le DELAF général et le DELAF spécialisé, décrits dans la section 3, ainsi qu'un dictionnaire des mots grammaticaux, qui a le même format que le DELAF, et qui contient près de 500 prépositions (*about, after, through*…) déterminants (*the, no, one*…), conjonctions (*though, and, or*…), adverbes (*above, almost, yet*…), pronoms (*who, another, few*…) et certains verbes (*are, can, have, may, will*…). A ce petit lexique est attribuée une priorité supérieure aux deux autres dictionnaires DELAF. Cela signifie que chaque mot simple du texte est d'abord recherché parmi les mots grammaticaux, et seulement s'il n'y figure pas, sa recherche est poursuivie dans les DELAF général et spécialisé. Cette mesure a été introduite pour éviter le superflu des séquences incorrectes extraites plus tard par le patron syntaxique : il n'est pas rare qu'un auteur décide de fournir dans son dictionnaire spécialisé les traductions de certains mots grammaticaux, même si elles sont, en principe, universelles. Ceci fait apparaître dans notre DELAF spécialisé des entrées non significatives du domaine. Le dictionnaire prioritaire nous garantira que ces entrées ne participeront par à la recherche de patrons, fondés essentiellement sur l'étiquette *+Spec* comme nous le verrons dans la section suivante.

Il arrive qu'un mot, qui dans la plupart des cas a une fonction grammaticale, puisse avoir dans un langage spécialisé une signification spécifique. Par exemple, en informatique, *or* désigne une opération logique, donc il est, en principe, incorrect d'interdire à l'analyseur lexical l'accès à l'étiquette *or,.N+Spec:s*. Ceci ne permettrait pas la reconnaissance de certains bons candidats-termes informatiques, comme *or gate*. Nous espérons néanmoins, que le silence ainsi introduit est minimal, au moins pour les domaines bien couverts par les dictionnaires LexPro, car les termes « curieux » comme *or gate* figurent déjà dans le DELACF spécialisé, et ils participeront éventuellement à la recherche des candidats plus larges comme *exclusive-or gate.*

Parallèlement à la reconnaissance des mots simples, l'analyseur lexical consulte les deux dictionnaires des mots composés : DELACF général et DELACF spécialisé, décrits dans la section 3. Les deux DELACF ont priorité sur tous les autres dictionnaires. Ceci veut dire, que si une séquence contiguë d'items du texte est reconnue en tant qu'entrée du DELACF spécialisé ou général, elle est dans la suite traitée *en bloc,* i.e. on ne cherche plus à étiqueter ses sous-séquences par les autres dictionnaires, et dans la recherche de patrons elle sera équivalente à un mot simple.

L'analyse lexicale produit deux dictionnaires associés au texte de départ - le dictionnaire des mots simples et celui des mots composés reconnus dans le texte – dont le format est identique à celui des DELAF et DELACF généraux et spécialisés, décrit dans la section 3. Les unités d'un ou plusieurs items reçoivent une ou plusieurs étiquettes grammaticales, pour lesquelles nous n'effectuons aucune désambiguïsation, mis à part l'utilisation des dictionnaires prioritaires. Ainsi un mot se retrouve souvent avec 2, 4 ou 6 étiquettes, dont certaines identiques au trait *Spec* près, par exemple le dictionnaire des mots simples du texte peut contenir des entrées suivantes:

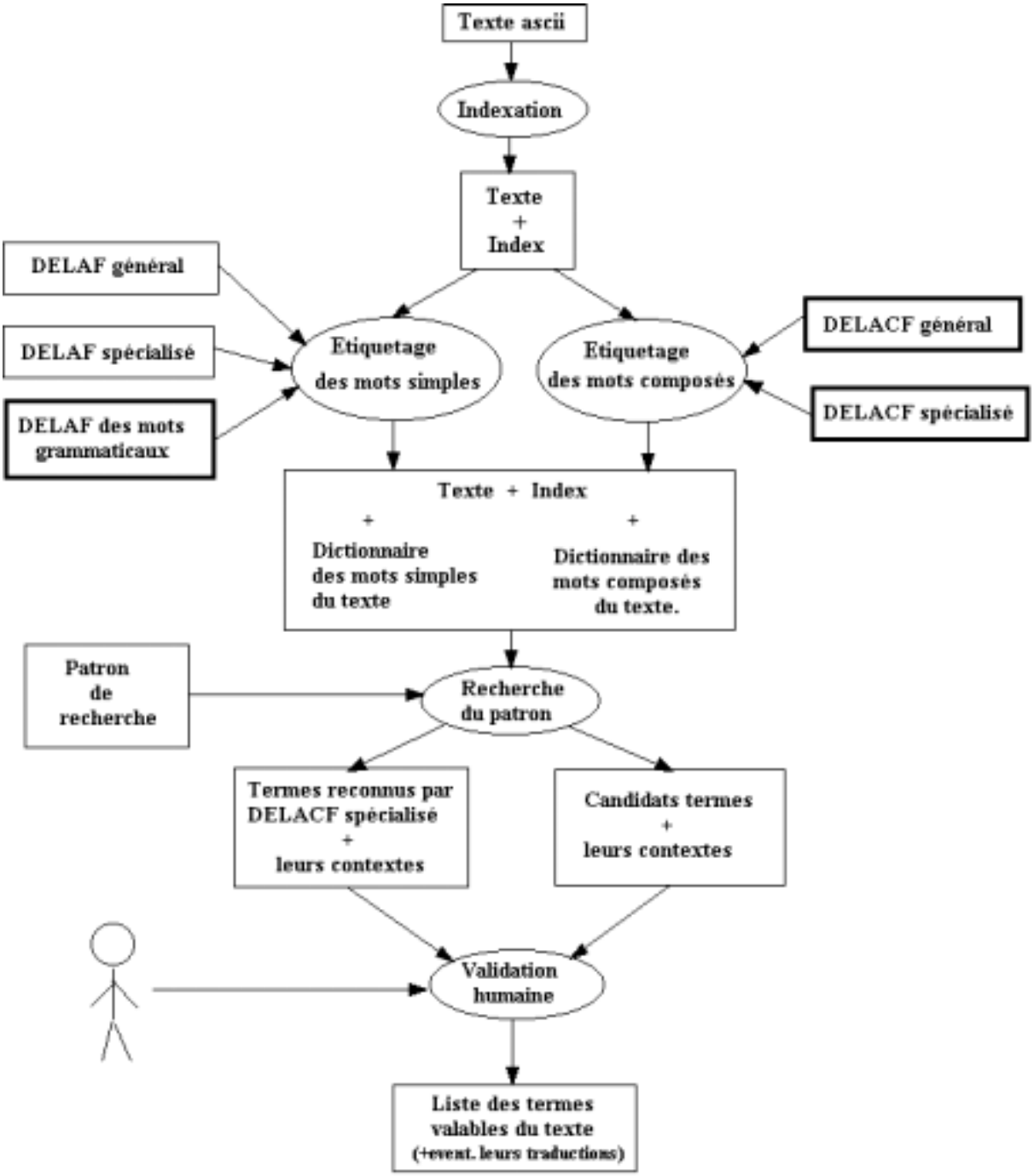
[9] registers, register.N+Spec:p

[10] registers, register.N:p

[11] registers, register.V+Spec:P3s

[12] registers,register.V:P3s

L'ambiguïté entre les étiquettes [9] et [10], ainsi que [11] et [12] ne pose pas de problèmes pour la reconnaissance du patron que nous avons choisi. En revanche, l'attribution par le DELAF spécialisé de catégories différentes pour le même mot (ambiguïté entre [9] et [11]), peut être à l'origine d'un certain nombre de candidats termes incorrects.



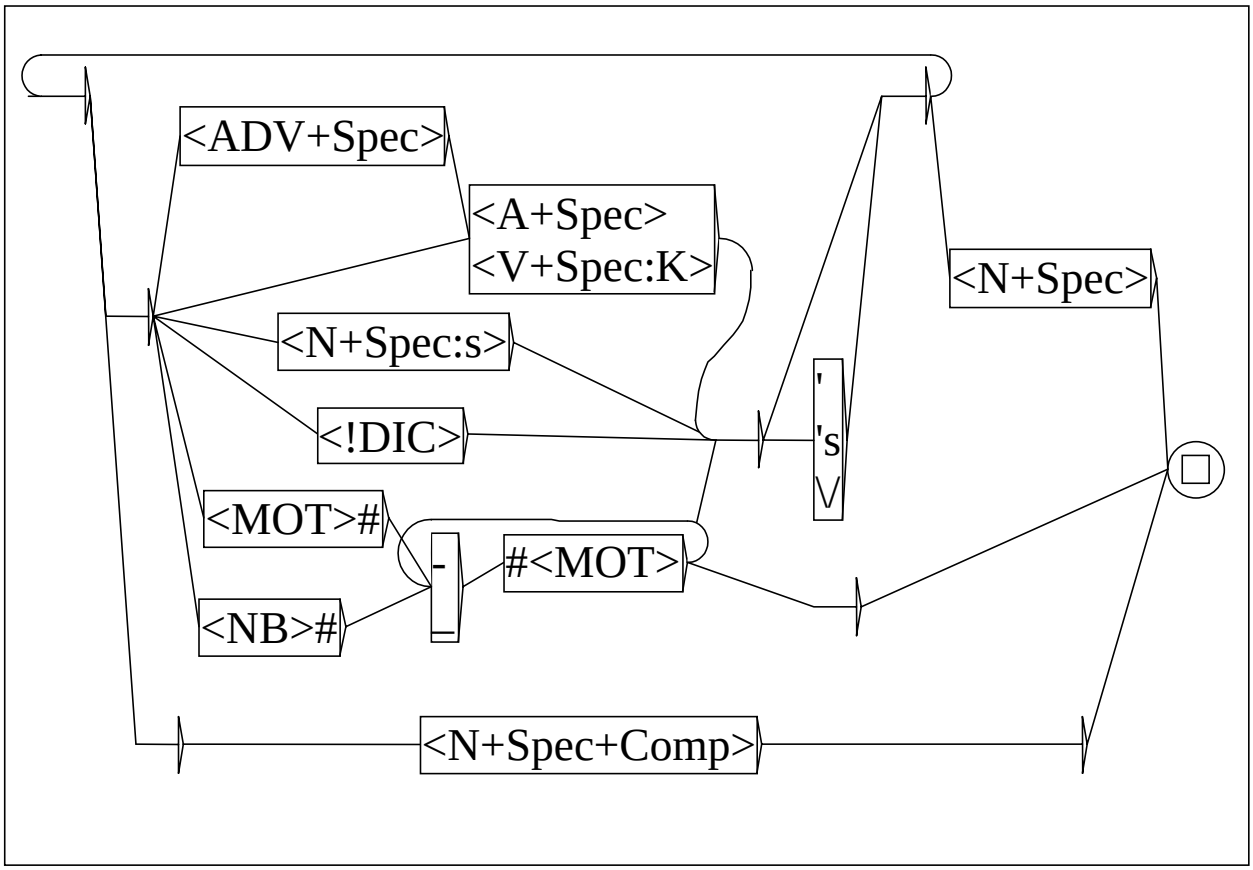
I1 Schéma de fonctionnement de l'extraction.

Les entrées du dictionnaire des mots composés du texte peuvent aussi être ambigües, si elles figurent à la fois dans le DELACF général et dans le DELACF spécialisé, ou bien si un terme complexe a réellement plusieurs emplois avec des catégories différentes. Là aussi seul ce dernier type d'ambiguïté peut influencer les résultats de la recherche du patron syntaxique.

4.2 Recherche de patrons

Nos patrons de recherche se présenteront sous forme d'automates à états finis, dont l'alphabet sera celui des étiquettes grammaticales décrites dans la section 3 et attribuées aux unités du texte dans la phase de l'analyse lexicale. Nous allons toujours chercher à extraire les séquences contiguës et maximales décrites par les patrons, sans nous préoccuper de l'existence de sous-termes ou insertions éventuelles à l'intérieur de ces séquences. Entre

autres, nous n'avons pas pris en compte la possibilité de rechercher des conjonctions, comme *application name and location, flash and SRAM cards,* et d'autres variantes terminologiques qui font objet de l'étude détaillée par Jacquemin (1997).
Jusqu'à présent, nous avons mis au point un seul patron, qui est représenté par le graphe sur l'illustration I2. Un graphe est équivalent à un automate, ce que l'on peut voir si pour chaque nœud du graphe : a) on ajoute un état avant ce nœud ; b) on transforme le nœud en une transition, qui sera étiquettée par le symbole de l'intérieur de ce nœud ; c) le nœud le plus à gauche devient l'état initial ; d) le nœud entouré d'un cercle devient l'état final. La direction des transitions est toujours celle du côté droit d'un nœud vers le côté gauche d'un autre nœud.
Analysons quelques exemples de séquences extraites par les différents chemins du graphe. La branche centrale contenant l'étiquette *<N+Spec :s>* permet de trouver toutes les suites de noms spécialisés. En effet, comme le montrent nos dictionnaires informatiques, les termes complexes du schéma $N_1N_2…N_k$ (avec k>=2), e.g. *access frequency, program reference table, data transmission control unit*, sont de loin les plus nombreux. Nous admettons l'insertion éventuelle de la marque « 's » ou « ' » du génitif, ainsi que du séparateur « / », entre deux noms, pour ne pas manquer les candidats du type *Windows NT User's Guide, server's hostname, matrix' mode, I/O activity*. La contrainte sur le nombre dans l'étiquette *<N+Spec :s>* a été introduite pour éviter le bruit trop important provenant des ambiguïtés entre les verbes à la troisième personne du singulier et les noms au pluriel, comme dans les contextes suivants (les séquences extraites à tort sont soulignées): *the analyzer displays information on the following…, an array supports write caching if it has …, the hit ratio shows cache efficiency and …* Cette contrainte risque d'introduire un certain silence, car il est en principe possible, qu'un terme du type $N_1N_2…N_k$ contienne un nom au pluriel sur une des positions 1…k-1, comme ceci a lieu dans des termes déjà connus, e.g. *active contents type, advanced communications system, american national standards institut.*
Les deux chemins supérieurs du graphe, contenant les étiquettes *<ADV+Spec>*, *<A+Spec>* et *<V+Spec :K>*, assurent la prise en compte des adjectifs, participes passés et adverbes spécialisés à l'intérieur des séquences du type *AN, ANN, NAN, AdvVN etc.*, comme *long return, parallel access array, storage system's physical disks, locally attached arrays.*
La partie du graphe, utilisant les étiquettes *<MOT>* (n'importe quel mot) et *<NB>* (nombre), permet d'extraire les mots liés par le trait d'union ou le soulignement, qui marquent souvent le caractère figé des séquences concernées, à condition qu'il n'y ait pas d'espace autour de ces séparateurs (ceci est exprimé par le signe *#*). Cette branche du patron correspondra à des candidats comme *operating system-related restrictions, DAE-to-DAE interconnection, dual-initiator/dual-bus configuration, mia_output_disable, 512-byte data block.*
Le nœud du graphe étiqueté par le symbole *< !DIC>* est celui qui donne la possibilité de prendre en compte les néologismes, i.e. les mots (communs et propres) non reconnus ni par le DELAF général ni par le DELAF spécialisé. Parmi les exemples de ce type trouvés dans nos corpus se trouvent (les néologismes sont soulignés) : *OpenManage Data Administrator, midplane connectors, nonmirrored write, powerup initialization sequence.*

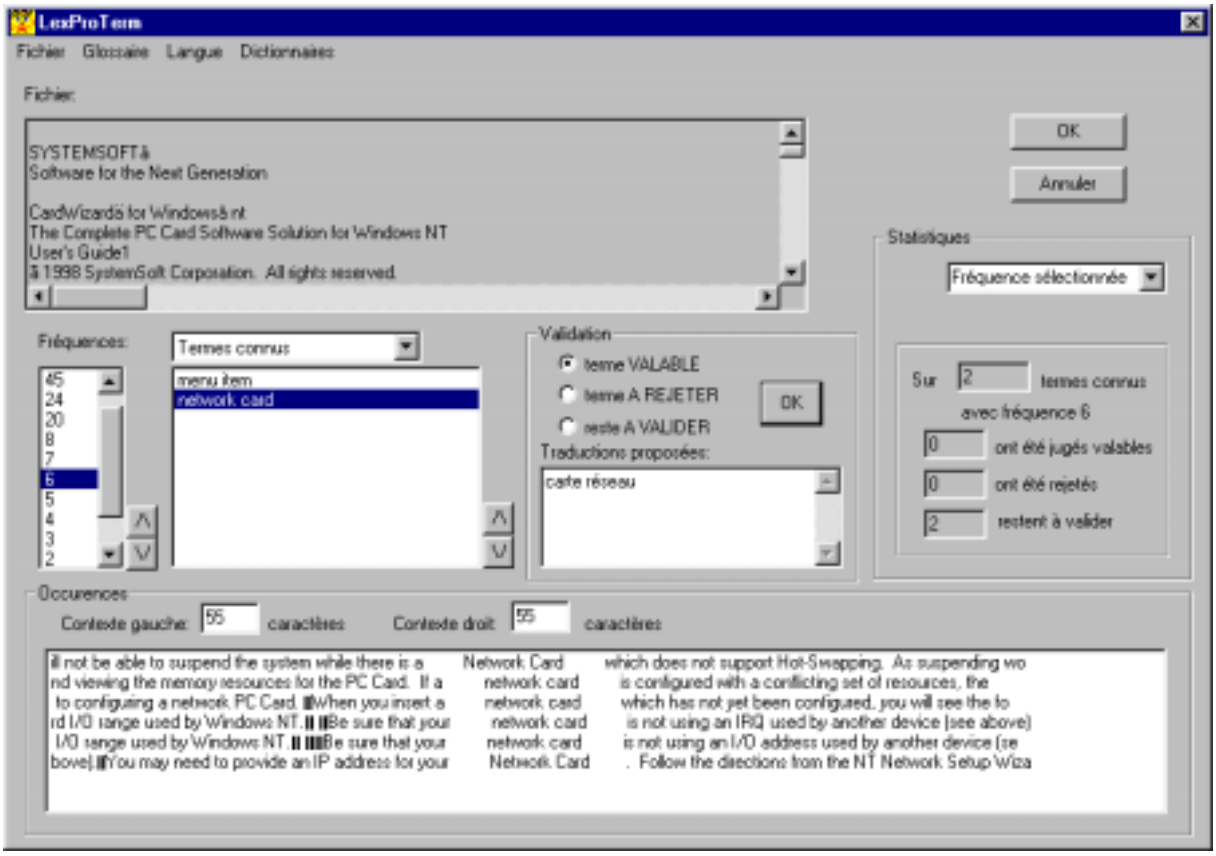


I2 Patron de recherche de nouveaux termes.

Finalement, le chemin inférieur du graphe permettra, grâce aux traits *+Spec+Comp*, d'extraire les noms composés terminologiques reconnus déjà par notre DELACF spécialisé au cours de l'analyse lexicale, qui n'ont pas encore été inclus dans des fréquences plus longues extraites par les autres parties du patron. Ces composés, étant des termes établis du domaine, ont une grande chance d'être retenus par l'utilisateur pour le texte donné.

Remarquons que toutes les étiquettes du graphe contenant le trait *+Spec*, à part celle avec en plus le trait *+Comp*, peuvent correspondre non seulement à des mots simples spécialisés, mais aussi à des composés, ce qui permet la reconnaissance des surcompositions obtenues par ajouts de nouveaux modifieurs ou têtes à des termes déjà connus, comme le montrent les candidats suivants (les composés existants déjà dans le DELACF spécialisé sont soulignés) : *ac power distribution, disk-based application, user free memory*.

La mise au point du graphe ci-dessus a été faite d'une façon expérimentale, par des allers-retours constants entre le patron de recherche et un corpus informatique de 700 kilooctets fourni par un traducteur technique. Nous avons essayé de trouver un juste milieu entre le rappel et la précision introduits, dont nous essayons d'estimer les proportions dans la section 4.3.



I3 Interface de validation.

4.3 Validation

L'illustration I3 présente l'interface de la phase de validation. Cette validation est effectuée par le traducteur qui utilise le logiciel pour créer son glossaire de traduction.
Après l'ouverture du texte, a lieu la phase d'extraction décrite dans la section précédente. Ensuite, les séquences extraites sont regroupées par variantes orthographiques : deux séquences sont considérées comme variantes, si elles sont égales à l'emploi des minuscules et des majuscules près. Pour chaque ensemble de variantes du même candidat, celle qui emploie le moins de majuscules est choisie comme la forme représentative. Toutes les formes représentatives sont triées selon les fréquences de leurs variantes dans le texte et divisées en deux listes : « Termes connus » et « Nouveaux termes », selon si elles figurent ou non dans le DELACF spécialisé (et sont donc des termes connus). La sélection d'un candidat de chaque liste entraîne l'affichage de toutes ses occurrences avec leurs contextes gauche et droit de longueurs réglables. L'utilisateur peut consulter la liste de candidats soit dans l'ordre alphabétique (option « Toutes » sur la liste « Fréquences »), soit selon leurs fréquences. Dans le deuxième cas, seuls les candidats ayant la fréquence sélectionnée s'affichent.
Le rôle principal de l'utilisateur est de valider les candidats-termes en activant l'un des trois boutons du milieu de l'écran (on choisit le bouton intitulé « reste A VALIDER », si l'on n'est pas sûr du statut d'un candidat), et éventuellement de proposer une ou plusieurs traductions pour chaque terme jugé valable. Les statistiques à droite de l'écran indiquent le nombre de candidats déjà retenus ou rejetés et de ceux qui restent à valider. On peut interrompre la validation à tout moment. Typiquement, on ne va examiner que les fréquences les plus élevées, mais cette stratégie n'est pas toujours la bonne : de nombreux candidats corrects n'apparaissent qu'une seule fois, même dans des corpus volumineux.

Les résultats finaux de la validation peuvent être exportés dans un fichier lisible par un logiciel permettant la consultation et la création de bases terminologiques (comme Access etc.) afin de mettre en forme le glossaire de traduction du texte (voir section 2).

5 Premiers résultats

Pour estimer l'efficacité de notre logiciel, nous nous servons des deux critères habituels, ceux de précision et de rappel. La précision est définie comme la proportion de bons termes parmi tous les candidats-termes proposés par l'extracteur. Le rappel signifie la proportion de bons termes proposés par l'extracteur parmi tous les termes existant dans le texte traité.
Nous avons déjà mentionné à la section 2, qu'il était difficile de décider si une séquence est ou non un terme dans le contexte de création du glossaire d'un texte à traduire. Néanmoins, nous avons fait un test pouvant nous donner des premières indications quant à la qualité de notre outil. Il a été réalisé sur un petit corpus de 52 kilooctets (8500 mots) de texte anglais sur le domaine informatique, fourni par un traducteur technique.
Nous avons d'abord effectué un prétraitement du corpus, qui consistait à marquer manuellement toutes les occurrences de termes. Ce choix a dû être parfois arbitraire, car, à part les termes informatiques connus, comme *AC power*, *hard disk*, *power management*, nous avons sélectionné certaines séquences sans statut terminologique établi, mais devant être, à notre avis, traitées comme unités de sens au cours de la traduction, par exemple : *cleanup feature*, *easy-to-read displays*, *non-network function*, *active termination device*, *CardWizard for Windows NT Notify Options screen*, *Card View Display Options*, *notification massage timeout*. Nous prenions en compte toujours la séquence maximale, sans rechercher ses sous-termes éventuels. Le glossaire du texte ainsi obtenu, contenant 839 occurrences[9], a été comparé aux listes de séquences extraites du même texte par notre extracteur. Parmi les 839 termes, 240 ont été reconnus par le DELACF spécialisé, et 450 ont été extraits par le patron syntaxique, ce qui donne le rappel égal à 82%.
Cette valeur, pas encore assez élevée du point de vue de notre application, est due en grande partie aux limites introduites dans le patron de recherche. Les termes non reconnus sont entre autres ceux qui : contiennent des prépositions (46% de cas), comme *PC Card support for Windows NT*; contiennent des noms au pluriel sur des positions non terminales (15%), comme *options menu* ; sont contenus dans des séquences plus longues (20%), comme *PC Card information screen displays* (le bon terme est souligné, *displays* a été extrait à tort). Ce dernier exemple montre le problème très important en anglais d'ambiguïtés entre les noms et les verbes, renforcé encore dans le domaine de l'informatique par le phénomène fréquent de conversion[10] de noms en nouveaux verbes ou de verbes en nouveaux noms. Par exemple, le mot *network*, fonctionnant dans la langue générale en tant que nom, gagne un nouveau sens verbal dans la langue spécialisée: *to network*, avec toutes les formes fléchies associées : *networks*, *networked*, *networking*. Le phénomène inverse est encore plus courant : de nombreux verbes deviennent des noms désignant l'action de ces verbes. Ainsi l'on obtient *an interrupt*, *a merge*, *a reset*, *an assert*, qui peuvent aussi se mettre au pluriel, ambigu avec la troisième personne du singulier des mêmes verbes.
Pour évaluer le taux de précision de l'extraction, il faut comparer le nombre de candidats-termes corrects avec celui de tous les candidats proposés. Puisque l'utilisateur ne consultera chaque candidat qu'une seule fois, nous faisons ce calcul, contrairement à celui du rappel, sur les listes sans doublons. Le nombre de toutes les séquences uniques extraites par le patron est égal à 644 (100 séquences proviennent du DELACF spécialisé). Dans cet ensemble, 339 candidats (dont 87 entrées du DELACF) sont pertinents, donc le taux de précision est égal à 53%. L'utilisateur retiendra donc à-peu-près 1 candidat sur 2, ce qui nous semble raisonnable pour le travail de constitution du glossaire de texte à traduire.

6 Aspects novateurs

L'originalité de notre méthode n'est pas dans les algorithmes employés, car :

- La recherche de patrons dans un texte étiqueté est une technique souvent appliquée dans la tâche d'extraction (par exemple chez Daille (1994) et Auger et al. (1996) en français, ou chez Justeson et Katz (1995) en anglais).
- La méthodologie de construction et d'utilisation des dictionnaires électroniques est celle employée au LADL (voir Courtois et Silberztein (1990)).
- Les principaux programmes informatiques ont été repris du système Intex.
- L'analyse lexicale du texte n'effectue qu'un minimum de désambiguïsation des mots.

Le point fort principal et l'originalité de notre approche est dans le fait de fournir à ces algorithmes des données de très haute qualité et complétude. Nous avons récupéré les résultats des travaux d'experts en lexicographie, terminologie et traduction. Leurs dictionnaires, généraux et spécialisés, étant l'effet de l' « extraction » humaine, sont de très bonne qualité du point de vue de la pertinence des mots et séquences qu'ils contiennent. De plus, nous nous sommes penchés sur la préparation de ces ressources, nécessaire pour le traitement automatique : la correction orthographique, le marquage des catégories et des traits flexionnels, la génération des formes fléchies, etc. Ainsi, nous disposons d'un noyau lexical très fiable que nous pouvons ensuite enrichir par une méthode automatique, standard du point de vue algorithmique, mais originale et efficace grâce à la qualité des ressources.
Les autres aspects novateurs de notre méthode sont à voir dans les points suivants :

1. Application de l'extraction dans le domaine de traduction assistée par ordinateur, qui présente des caractéristiques et exigences particulières, telles que :
   - La nécessité de traiter des textes de tailles très variées, et rarement aussi importantes que les corpus auxquels sont traditionnellement appliqués les outils d'extraction. Cette contrainte exclut l'application efficace de toute méthode d'extraction qui comprend des calculs statistiques, telles que Daille (1994), Justeson et Katz (1995), Nakagawa (1998) et autres, dont un panorama est présenté chez Jacquemin (1997), pp. 24-29 et chez Daille (1994).
   - L'importance pour la qualité de traduction d'un très bon rappel des termes extraits (la même condition est prise en compte par Ladouceur et Cochrane (1996), mais leur article ne précise malheureusement pas les algorithmes employés).
   - La spécificité de la notion du terme valable (il ne doit pas obligatoirement avoir un statut terminologique établi – voir section 2).
2. La variété des ressources utilisées et de leurs rôles dans le procès d'extraction :
   - Le dictionnaire des mots composés terminologiques (le DELACF spécialisé) sépare les séquences qui ont un statut terminologique déjà reconnu.
   - Les dictionnaires des mots simples et composés terminologiques (le DELAF et le DELACF spécialisés) fournissent les étiquettes qui sont à la base du patron de recherche (trait *+Spec*).
   - Les dictionnaires des mots composés généraux et terminologiques (le DELACF général et le DELACF spécialisé) permettent de traiter « en bloc » certaines séquences figées (i.e. nous ne

cherchons pas de nouveaux termes à l'intérieur des mots composés connus).

- La complétude du dictionnaire des mots simples généraux (le DELAF général) permet de considérer les mots simples non reconnus comme néologismes du domaine traité et de les prendre en compte dans le patron de recherche.

3. L'utilisation d'un analyseur lexical qui tient compte des mots composés. Cet aspect est absent ou très limité dans les étiqueteurs employés par les extracteurs existants.
4. L'hypothèse que le matériaux lexical à l'intérieur d'un domaine est relativement stable par rapport à la croissance très importante de la terminologie. Ainsi, nous admettons que la création d'un nouveau terme se fait le plus souvent par une combinaison grammaticalement correcte de termes simples et composés déjà existants. Cette hypothèse est reflétée dans le patron de recherche utilisé et confirmée par les résultats des testes (elle apparaît aussi chez Nakagawa (1998) , mais les mots simples caractéristiques du domaine y sont recherchés non pas dans un dictionnaire mais dans le corpus par une méthode statistique).

7 Perspectives

La méthode d'extraction présentée ci-dessus n'est que le début de notre travail. Il nous reste à mettre en forme tous les dictionnaires LexPro, comme nous le décrivons dans la section 3.2. Beaucoup de ces dictionnaires sont peu volumineux, et donc le nombre de termes simples, sur lesquels est fondée une grande partie du patron de recherche, peut s'avérer trop bas. Dans ce cas, nous pouvons entreprendre, avant de commencer l'extraction, une mesure supplémentaire d'auto-enrichissement de dictionnaires spécialisés par récupération des « termes simples significatifs » de chaque domaine, i.e. des substantifs, adjectifs, adverbes et participes apparaissant parmi les termes complexes déjà répertoriés. Ces nouvelles entrées, soumise ensuite à la flexion, peuvent compléter les DELAF existants.
Il sera aussi nécessaire d'ajouter, au cours de l'extraction, la lemmatisation des candidats-termes, afin de ne plus proposer la même séquence au singulier et au pluriel (e.g. *disk module* et *disk modules*) comme deux candidats indépendants. Remarquons que cette lemmatisation est à faire sur les termes complexes entiers, i.e. seuls leurs constituants caractéristiques doivent être lemmatisés, et non pas tous leurs composants, comme ceci est fait par Daille (1994).
Un problème important qui reste à résoudre est celui des ambiguïtés des mots simples et composés apparues suite à l'étiquetage du texte. Le rattachement à notre système d'un des étiqueteurs disponibles, par exemple de celui de Brill (1994), peut s'avérer difficile à cause de la spécificité des dictionnaires que nous utilisons, entre autres ceux des mots composés. Néanmoins, nous envisageons de tester cette possibilité pour augmenter la précision de notre logiciel.
Nous souhaitons aussi élaborer de nouveaux patrons de recherche. Premièrement, des nouvelles structures syntaxiques, entre autres celles contenant des prépositions, comme *arrangement of slots*, doivent faire l'objet d'une étude détaillée. Deuxièmement, nous voudrions élaborer des méthodes fondées sur l'observation que les listes de termes connus contiennent de nombreuses séries, e.g. *access control group*, *access control list*, *access control machine*, *access control profile*, *access control register*, etc. Il est probable qu'une séquence contenant le même affixe qu'une des séries recensées soit un bon candidat-terme.
Actuellement, notre logiciel ne traite que le texte pur. La prise en compte des formats enrichis, e.g. de l'emploi d'une police spéciale pour certaines parties du document, ainsi que

le traitement des données textuelles inclues dans des tableaux, illustrations etc., sera un affinement important du point de vue d'un traducteur technique.
Finalement, l'adaptation du logiciel à d'autres langues de LexPro est à envisager. Nous sommes consciente que ceci représente un travail considérable, car les patrons de recherche pour une langue seront rarement utilisables dans une autre langue. Nous espérons néanmoins que, le point fort de la méthode étant la taille et la qualité de nos dictionnaires généraux et terminologiques, nous allons pouvoir fournir des résultats intéressants pour les utilisateurs multilingues.

Conclusion

Nous voyons l'un des avantages importants de notre méthode d'extraction de termes dans le fait que ses résultats[11] ne dépendent pas de la taille des documents, sur lesquels elle est effectuée. En effet, si l'on soumet à l'extraction seulement une partie d'un corpus, les candidats termes proposés seront exactement les mêmes que ceux qui dans la même partie ont été trouvés lors de l'extraction sur le corpus entier. Seules les fréquences des candidats (donc leur ordre de présentation) pourront varier, mais ceci n'affecte pas le contenu de la liste[12].
Si l'on dispose d'un dictionnaire couvrant précisément le domaine traité, et si la terminologie disponible dans ce dictionnaire est assez complète et de bonne qualité, les résultats de l'extraction seront riches. Nous espérons qu'avec la taille de LexPro déjà très importante et toujours croissante, notre extracteur terminologique fera ses preuves.

Agata Chrobot,

Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7,

LCI (Langage Communication Informatique), Jouy-en-Josas, France.

Bibliographie

Auger (P.), Drouin (P.), Auger (A.), 1996 : « Filtact : un automate d'extraction des termes complexes », dans Grarson (M.), ed., *Terminologies nouvelles, Banques de terminologie, Actes de la table ronde, Québec, 18 et 19 janvier 1996, N°15*, Bruxelles, pp. 48-51.
Bauer (L.), 1983 : *English Word-Formation*, Cambridge University Press.
Bourigault (D.), 1994 : *LEXTER un Logiciel d'Extraction de Terminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de doctorat en Mathématiques, Informatique Appliquée aux Sciences de l'Homme, École des Hautes Études en Sciences Sociales, Paris.
Brill (E.), 1994 : Supervised part of speach tagger, http://www.cs.jhu.edu/~brill.
Chrobot (A.), 1999 : « Flexion automatique des mots composés », dans Lamiroy (B.), Klein (J.), Peirret (J.-M.), eds., *Cahiers de l'Institut de Linguistique de Louvain. Actes du XVI Colloque Européen sur les lexiques et la grammaires comparés des langues romanes, Louvain-la-Neuve, septembre 1997*, Louvain-la-Neuve.
Courtois (B.), Silberztein (M.) éditeurs, 1990 : *Dictionnaires électroniques, Langue Française 87*, Larousse, Paris.
De Sollier (F.), 1998 : *Dictionnaire Encyclopédique de l'Informatique*, Paris, La Maison du Dictionnaire.

Daille (B.), 1994 : *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat en Informatique Fondamentale, Université Paris 7.
Gouadec (D.), 1997 : *Terminologie et Phraséologie pour Traduire – Le concordancier du Traducteur*, Paris, La Maison du Dictionnaire.
Hildebert, 1999 : *Dictionnaire des Sciences de l'Informatique*, Paris, La Maison du Dictionnaire.
Jacquemin (Ch.), 1997 : *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Habilitation à diriger des recherches en informatique, IRIN, Université de Nantes.
Justeson (J.), Katz (S.), 1995 : « Technical terminology : some linguistic properties and an algorythm for identification in text », dans *Natural Language Engineering, 1(1)*, pp. 9-27.
Ladouceur (J.), Cochrane (G.), 1996 : « Termplus, système d'extraction terminologique », dans Grarson (M.), ed., *Terminologies nouvelles, Banques de terminologie, Actes de la table ronde, Québec, 18 et 19 janvier 1996, N°15*, Bruxelles, pp. 52-56.
Nakagawa (H.), Mori (T.), 1998 : « Nested Collocation and Compound Noun For Term Extraction », dans *Preceedings of COPUTERM, the First Workshop on Computational Terminology, August 15, 1988, University of Montreal.*
Silberztein (M.) , 1997 : *INTEX 3.4. Reference Manual*, LADL, Université Paris 7, Paris.

[1] Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7

[2] Dans ce contexte un néologisme sera pour nous un mot non reconnu ni par un dictionnaire spécialisé, ni par un dictionnaire général.

[3] Il s'agit de bases de données lexicales pour la morphologie flexionnelle de la langue générale. A présent, les dictionnaires généraux du français, de l'anglais, l'allemand, l'italien et l'espagnol sont accessibles.

[4] Anglais, français, russe, allemand, espagnol, portugais, italien, néerlandais, danois, suédois, arabe.

[5] Un mot simple est pour nous une séquence contiguë de lettres de l'alphabet (contenu pour chaque langue dans un fichier à part), délimitée par deux séparateurs : blancs, apostrophes, tirés, points, ou autres caractères de ponctuation. Cette définition est purement orthographique, car e.g. an anglais *air* et *airplane* sont des mots simples, tandis que *air-bed* et *air force* sont des composés.

[6] Pour la description détaillée du format des dictionnaires LADL consulter Courtois et Silberztein (1990).

[7] Un mot composé est, en bref, une séquence contiguë de deux ou plus de mots simples, dont les propriétés sémantiques et/ou syntaxiques ne peuvent pas être déduites de celles de ses constituants.

[8] Un item de texte, tel qu'il est défini dans Intex, correspond à une suite contiguë soit de lettres, soit de séparateurs (caractères non alphabétiques).

[9] Ce nombre se réduit à 417, si l'on compte une seule fois les différentes occurrences et versions orthographiques du même terme.

[10] Pour une discussion sur la conversion en anglais consulter e.g. Bauer (1983).

11 Nous comprenons ici par résultat d'extraction la liste des candidats retenus par le logiciel avant toute intervention humaine.

[12] Nous avons mentionné que, dans le cadre de la traduction, il est important de ne pas négliger les termes d'une fréquence basse d'occurrences. Si néanmoins l'utilisateur choisit de ne valider que les candidats de fréquences élevées, son résultat final, i.e. la liste validée, dépendra évidemment de la taille du corpus.