Dépouillement terminologique assisté par ordinateur de sites Web spécialisés

NANCY BOURASSA
Université de Montréal, Montréal, Canada
nancy.bourassa.1@umontreal.ca

PATRICK DROUIN
Université de Montréal, Montréal, Canada
patrick.drouin@umontreal.ca

RÉSUMÉ
Nous pouvons tous constater quotidiennement que l'importance d'Internet grandit sans cesse. De nombreux sites traitant de domaines spécialisés sont intéressants du point de vue du terminologue, mais n'ont pas encore fait l'objet d'un dépouillement terminologique. Certains domaines, comme celui de la réservation en ligne, dont les réalisations sont majoritairement électroniques, ne peuvent être dépouillés qu'en faisant appel à une batterie d'outils spécialisés. Dans l'élaboration d'un lexique portant sur le domaine de la « réservation en ligne », réalisé conjointement avec l'*Office québécois de la langue française* et *Amex Canada Inc*., nous avons adapté les méthodes de travail traditionnelles afin de pouvoir récupérer et exploiter les données contenues sur la Toile. Dans cet article, nous présenterons d'abord les particularités du domaine à l'étude et la méthodologie utilisée pour mener à bien notre démarche, puis nous analyserons les impacts qu'une telle démarche pourrait avoir sur le travail des terminologues.

ABSTRACT
Every day, we can witness the growing importance of Internet. A substantial amount of Web sites host interesting terminological information dealing with specialised subjects but have yet to be the object of a literature search. The online booking field, to name but one of many subject areas whose realizations are mainly on the Internet, can only be searched by using an array of specialised tools. During the development of a glossary on e-booking, jointly achieved with the *Office québécois de la langue française* and *Amex Canada Inc*., we adapted the traditional work process methods to retrieve and manage data found on the Web. In this article, we will start by presenting the distinctiveness of our field of study and the methodology we used to complete our method. We will then analyse the impact such an approach can have on the terminologists' work.

MOTS-CLÉS/KEYWORDS
terminologie, terminographie, terminologie computationelle, extraction automatique de la terminologie

It is now recognised that the only practical means of processing lexical data is by computer.
(Sager 1990 : 129)

Introduction

La manipulation de données électroniques par le terminologue n'est pas chose nouvelle. En effet, les banques de terminologie font partie du paysage depuis plus d'une quarantaine d'années. Les premiers travaux sur l'informatisation des procédures de travail en terminologie comportaient bien souvent une étape de numérisation des corpus (Auger *et al.* 1991) qui peut désormais être évitée étant donné le nombre sans cesse grandissant

d'ouvrages disponibles en format électronique. Avec l'arrivée sur le marché des logiciels dédiés à l'extraction automatique des termes au début des années 1990, la gestion de corpus en format électronique est elle aussi devenue une réalité du travail terminologique.

Malgré cette prise en charge des corpus électroniques par les outils terminologiques et la diffusion de masse de documents spécialisés dans Internet, de nombreux domaines n'ont pas encore fait l'objet d'un dépouillement terminologique puisque l'élaboration de corpus volumineux pour ces domaines s'avère encore une tâche difficile. En effet, un domaine comme celui de la réservation en ligne, dont les réalisations sont majoritairement électroniques, mais ne se retrouvent pas sous la forme de documents traditionnels, ne peuvent être dépouillés qu'en faisant appel à une batterie d'outils spécialisés.

Dans cet article, nous décrivons un projet, réalisé conjointement avec l'*Office québécois de la langue française* (*OQLF*) et *Amex Canada Inc.*, visant l'élaboration d'une terminologie de la « réservation en ligne ». Nous décrivons succinctement le projet et les particularités du domaine pour ensuite aborder la méthodologie utilisée afin de mener à bien notre recherche. Finalement, nous analysons les impacts qu'une telle approche pourrait avoir sur le travail des terminologues.

Description du projet

Le *Comité de terminologie de la réservation en ligne* (*CTREL*) a été mis sur pied par l'*Office québécois de la langue française* afin de répondre à un besoin d'uniformisation terminologique identifié par la société *Amex Canada Inc*. L'objectif du *CTREL* est de mettre à la disposition des entreprises œuvrant dans le secteur du tourisme une terminologie fiable et représentative de l'usage actuel.

Le premier mandant du *CTREL* consistait à produire soixante fiches terminologiques qui ont été intégrées au *Grand dictionnaire terminologique*. Les fiches feront aussi l'objet d'un signet terminologique qui sera diffusé sur le site de l'*OQLF* dès le mois de novembre 2005.

Particularités du domaine

Le domaine couvert possède quelques propriétés distinctives qui le rendent difficile à manipuler d'un point de vue terminologique : la nature du corpus, l'aspect commercial du domaine, l'importance de la variation terminologique et la forme des termes.

Nos efforts en vue d'assembler un corpus à partir de documentation traditionnelle sur le sujet se sont butés à des problèmes importants. Les seuls documents trouvés portaient sur l'impact des nouvelles technologies sur le domaine de la réservation plutôt que sur la description de cette dernière. Ces documents sont d'un intérêt plus que limité pour le terminologue puisque les notions y sont abordées de façon superficielle. Les réalisations écrites du domaine de la réservation en ligne se retrouvent essentiellement dans Internet, ce qui devrait théoriquement faciliter son analyse à l'aide d'outils informatiques. Cependant, les termes ne se retrouvent pas au sein de documents électroniques traditionnels, ils sont plutôt intégrés aux interfaces des divers systèmes consacrés à la réservation.

La diffusion dans Internet de la terminologie liée au domaine qui nous intéresse n'est donc pas le résultat d'une simple transposition de contenu du support papier au support électronique, contrairement à ce qu'on observe pour bien des domaines. Les interfaces constituent, d'une certaine façon, l'*environnement naturel* de ces termes. Une étape supplémentaire est donc nécessaire en vue d'élaborer le corpus électronique qui servira de point de départ à la présente recherche.

L'aspect commercial du domaine de la réservation en ligne apparaît lui aussi de façon importante dans le corpus assemblé : on y vante répétitivement les mérites de certaines fonctions des outils de réservation en ligne (ORL) alors que d'autres, jugées trop banales ou trop répandues, ne sont tout simplement pas documentées. Les notions ne sont donc pas toutes mises en évidence, contrairement à ce qu'aimerait voir le terminologue. On

constate aussi très rapidement que le point de vue des utilisateurs de tels outils est rarement abordé, alors que celui des transporteurs aériens, surtout l'aspect économique, est présent dans la très grande majorité des textes.

Le domaine de la réservation est relativement récent et on peut considérer qu'il est toujours en pleine émergence. Cet aspect novateur du domaine en influence grandement les termes qui sont soit nouveaux et instables, soit très bien implantés. En effet, on observe une variation importante dans les dénominations attribuées aux notions principales du domaine puisque la terminologie ne s'est toujours pas stabilisée. De plus, les entreprises qui conçoivent les ORL cherchent sans cesse à se démarquer de leur concurrence en innovant et en proposant de nouvelles fonctions aux utilisateurs. Dans la documentation, ici des fichiers d'aide en ligne ou des pages Web, cette course à la nouveauté a pour conséquence la multiplication des dénominations pour une seule et unique notion. Par contre, une bonne part de la terminologie utilisée dans le domaine fait appel à celle du domaine de la réservation traditionnelle. Ce phénomène touche principalement les notions fondamentales ayant trait aux données de réservation, comme l'heure de départ et la destination.

Notre travail a aussi fait ressortir une particularité souvent associée aux outils informatiques, celle d'avoir recours à des termes courts et à des troncatures. Par exemple, on retrouvera régulièrement dans les interfaces le terme *type d'hôtel* alors que la réalité désignée est plus large et recouvre tous les types d'établissement d'hébergement (*hôtel*, *motel*, *auberge*, etc.).

Méthodologie

La méthodologie adoptée dans le cadre de ce projet découle de celle proposée dans L'Homme (2004 : 46). Selon l'auteure, le travail terminologique comporte les sept étapes suivantes :

1) la mise en forme d'un corpus,
2) le repérage des termes,
3) la collecte de données,
4) l'analyse et la synthèse des données,
5) l'encodage des données,
6) l'organisation des données terminologiques,
7) la gestion des données terminologiques.

Les sections qui suivent se concentrent sur la méthodologie adoptée pour les trois premières étapes. La nature essentiellement sémantique des étapes d'analyse et de synthèse des données les rend difficilement informatisables. Puisque la présente recherche s'effectue en partenariat avec l'*OQLF*, les trois dernières étapes s'appuient sur l'infrastructure informatique de l'*OQLF* et son outil de gestion terminologique *SAMI* (*Système d'Alimentation et de Mise à jour Intégré*) dont une application permet de créer et de modifier des fiches terminologiques à distance. Le fonctionnement de cet outil est déjà bien documenté (Bédard et Duchesne 2002) et ne fera pas l'objet d'une description détaillée dans le cadre du présent article.

Mise en forme du corpus

Nous l'avons mentionné précédemment, la terminologie du domaine de la réservation en ligne ne se retrouve que dans les divers portails donnant accès à des ORL. Afin de mettre en place le corpus électronique nécessaire à la présente recherche, nous avons dû faire appel à une technologie permettant de constituer un corpus à partir de ces portails. Nous avons eu recours à un aspirateur de sites Web pour récupérer une copie des ORL[1] dont nous avions besoin. Grâce à ce type de logiciel, il est désormais possible d'obtenir une image locale d'un site Web et de l'ensemble des documents qui le composent. La récupération des sites

Web et le stockage des données sur le poste du terminologue permettent d'obtenir une copie stable des documents du corpus à l'abri des mises à jour fréquentes dans Internet.

Les sites Web sont généralement constitués de documents en format HTML. Cependant, les sites plus élaborés comme les portails dédiés au commerce électronique comportent de nombreux fichiers dans des formats très variés (PDF, RTF, DOC, TXT, etc.). Afin de pouvoir dépouiller ces documents à l'aide d'un extracteur de termes, une étape de normalisation et de conversion de l'ensemble des fichiers vers un format texte facilement exploitable est inévitable. Le volume important de données à manipuler pour ce projet nécessitait la mise en place d'une conversion automatique. Pour ce faire, nous avons eu recours aux logiciels *ABC Amber Text Converter* et *ABC Amber PDF Converter*[2].

Afin de diversifier le type de documents composant le corpus et de ne pas nous limiter strictement au dépouillement de portails, nous avons récupéré quelques articles, en format électronique, portant sur la réservation en ligne. Les versions française et anglaise du *Règlement sur les systèmes informatisés de réservation canadiens* de la *Loi sur l'aéronautique* ont été ajoutées au corpus. En dernier lieu, quelques textes mis à notre disposition par nos partenaires ont été numérisés. Ces documents, malgré le fait qu'ils se trouvaient sur un support traditionnel, étaient à l'origine en format électronique puisqu'ils correspondaient à des saisies d'écran de l'outil utilisé par *Amex Canada Inc.*, *Corporate Travel Online,* et de son fichier d'aide bilingue (anglais/français).

Repérage des termes

Afin de repérer les termes contenus dans notre corpus, nous avons fait appel à deux outils de traitement automatique de la langue (TAL) : un étiqueteur morphosyntaxique et un logiciel d'acquisition automatique de termes. L'étiquetage du corpus a été effectué à l'aide du logiciel *TreeTagger* (Schmid 1994); ce dernier assigne automatiquement à l'ensemble des formes d'un document une partie du discours (*arrive/VER:pres*) et procède à la lemmatisation de ces mêmes formes (*arrive/VER:pres/arriver*). Cette étape est d'autant plus intéressante en vue du repérage des termes qu'elle permet de regrouper les variantes des termes du corpus sous une même entrée (*fournisseur de service, fournisseur de services, fournisseurs de service* et *fournisseurs de services*). Les fréquences ainsi obtenues par l'extracteur sont donc plus représentatives de la réalité puisque les singuliers et les pluriels, ou les verbes conjugués, sont regroupés sous la forme lemmatisée.

L'extraction des termes a été confiée au logiciel *TermoStat* (Drouin 2003). Ce dernier a pour objectif l'identification de candidats termes[3] propres à un domaine. L'outil utilise une technologie hybride qui combine méthodes statistiques et linguistiques en vue de l'extraction de candidats termes simples et complexes. Dans un premier temps, le logiciel recense une première liste de candidats en identifiant dans le corpus des enchaînements qui correspondent généralement aux structures de surface typique des unités terminologiques. Il procède ensuite à une comparaison statistique de la fréquence à laquelle les candidats termes apparaissent dans le corpus dépouillé et la compare avec la fréquence à laquelle ils apparaissent dans un corpus non spécialisé nommé corpus de référence. La comparaison repose sur un test statistique, proposé par Lafon (1980), qui permet de déterminer dans quelle mesure un candidat terme est spécifique au corpus dépouillé. Lorsque la fréquence d'un candidat terme diffère de façon significative entre les deux corpus, *TermoStat* le retient à titre de terme potentiel. Pour le présent projet, nous avons mis à la disposition du logiciel un corpus de référence constitué d'environ 30 millions d'occurrences tirées du journal *Le Monde* pour le français et d'approximativement 7,5 millions d'occurrences du quotidien montréalais *The Gazette* pour l'anglais[4].

La liste des candidats termes proposés par le logiciel comporte des informations additionnelles sur les candidats termes dont la fréquence et les variantes orthographiques. Le regroupement des formes non lemmatisées du candidat terme rend possible la comparaison des variantes orthographiques. Par exemple, pour le terme « outil de réservation en ligne », le logiciel a repéré les trois variantes suivantes : *outil de réservation*

en ligne, *outils de réservation en ligne* et *outils de réservations en ligne*. En plus de permettre de consulter l'ensemble des contextes pour les diverses variantes du terme, ce regroupement effectué par *TermoStat* met en évidence le fait que le pluriel de ce terme pose problème à certains rédacteurs. Le terminologue peut donc déterminer que ce phénomène doit faire l'objet d'une note linguistique sur la fiche terminologique.

Le logiciel attribut aussi un poids à chaque candidat. Ce dernier est établi lors de la comparaison des fréquences. Plus le poids est élevé, plus l'extracteur considère que le candidat terme est caractéristique du corpus[5]. L'interface du logiciel permet au terminologue de consulter tous les contextes d'occurrence d'un candidat à l'aide d'un simple hyperlien. Le terminologue peut ainsi valider l'intérêt de la proposition faite par *TermoStat* directement dans l'outil, sans avoir à retourner au document original.

Comme l'extracteur peut analyser les documents en anglais et en français, il nous a permis de comparer les termes recensés dans les deux langues et de les mettre en parallèle. Afin de conserver toutes les données relevées et de réunir les candidats termes de tous les sites Web et articles que nous avons dépouillés, nous avons consigné les informations propres à chaque terme dans une base de données (*Microsoft Access*). Cette technique nous a permis d'avoir rapidement accès à tous les renseignements nécessaires pour la collecte des données.

Afin d'obtenir un panorama conceptuel du domaine de la réservation en ligne, nous avons eu recours à un autre outil, le gestionnaire d'ontologie *Protégé* (Gennari *et al*. 2003). Ce dernier, distribué gratuitement dans Internet, a été conçu à l'Université de Stanford. Il s'agit d'un outil qui permet d'aller plus loin dans la description de la structure conceptuelle d'un domaine que l'élaboration d'un simple arbre du domaine. En effet, il permet de relier les concepts entre eux à l'aide de relations déterminées par l'utilisateur. Notre choix s'est arrêté sur *Protégé,* car il nous permettait de gérer facilement notre nomenclature sous forme arborescente, d'établir et de modifier les relations entre les concepts et d'obtenir une vue d'ensemble de notre domaine rapidement. De plus, les données peuvent être exportées sous divers formats dont le format HTML.

Collecte des données

La collecte des données (contextes, définition, etc.) relatives à chacun des termes a été principalement effectuée à l'aide du concordancier *WordSmith*. Ce dernier recense toutes les occurrences des mots d'un texte et donne un accès instantané à leurs contextes. De plus, *WordSmith* permet à l'utilisateur d'accéder au contexte en mode plein-texte. D'autres de ses fonctions sont aussi intéressantes pour le terminologue. Par exemple, l'outil permet de trier les concordances en fonction des mots qui précèdent ou qui suivent le mot interrogé. Cette façon d'accéder aux données facilite le dépistage des termes complexes et complète bien le logiciel d'acquisition automatique de termes.

Les possibilités de tri offertes par WordSmith mettent aussi en lumière des collocations liées aux termes de la nomenclature ou des familles de termes. Ainsi, l'analyse des concordances du terme « réservation » avec une fenêtre de deux mots à gauche et d'un mot à droite fait émerger la famille suivante : *outil de réservation en ligne*, *portail de réservation en ligne*, *services de réservation en ligne*, *solution de réservation en ligne* et *système de réservation en ligne*.

Comme dans toute recherche terminologique, la couverture du corpus n'est parfois pas suffisante et le terminologue doit chercher ailleurs les attestions pour certains termes. Conservant notre objectif de maximiser les sources électroniques exploitées pour ce projet, nous nous sommes orientés vers l'utilisation d'une bases de données spécialisées portant sur le monde des affaires et de l'économie : *ABI/INFORM Global* (ProQuest 2005). Les fonctions d'interrogation de cette base de données ne permettent cependant pas la production de concordances et l'accès aux termes se trouve donc limité. En revanche, comme elle met en place un mécanisme d'envoi d'articles à une boîte de courriel, nous avons pu tirer profit de cette fonctionnalité et improviser un concordancier en exploitant le

service de courriel gratuit de la société *Google*, *Gmail* (Google 2005). En effet, grâce au moteur de recherche bien connu de cette entreprise, le terminologue peut rechercher très rapidement et facilement de l'information dans l'ensemble des documents transmis dans sa boîte de réception.

Discussion

Impact sur le travail terminologique : aspects positifs

Au cours de la présente recherche, nous avons adapté les méthodes de travail traditionnelles en terminologie dans le but de prendre en charge un domaine dont l'information se situe presque exclusivement dans Internet. La méthodologie proposée nous a aussi permis d'augmenter la rapidité avec laquelle certaines étapes du travail terminologique sont habituellement effectuées.

Les aspirateurs de sites Web favorisent l'élaboration de corpus électroniques volumineux dans un très court laps de temps. Avec le traitement automatique, le terminologue n'a plus à se soucier de la taille du corpus. Dans le cadre du présent projet, nous avons pu constituer et dépouiller un corpus de plus de 2 200 000 mots, ce qui aurait été impossible de réaliser dans le délai que nous nous sommes fixé, avec les méthodes traditionnelles.

Grâce à un extracteur de termes, la période généralement consacrée à la familiarisation avec le domaine est plus courte. En effet, l'affichage des candidats termes par le logiciel d'acquisition automatique de termes selon leur fréquence ou leur spécificité par rapport au domaine permet de prendre connaissance, par le biais des termes, des notions les plus importantes d'un corpus. Cet accès aux termes pondéré par le logiciel donne, en quelque sorte, un bref résumé de la thématique principale du corpus. Il est ainsi possible de repérer rapidement les textes qui ne portent pas sur le sujet à l'étude et de les écarter du corpus quand ils ont, par exemple, un contenu plus commercial qu'informatif.

L'aspiration des documents du corpus sur le poste du langagier offre un avantage supplémentaire. Grâce à une telle approche, le terminologue obtient une image figée dans le temps du contenu d'un site Web. Cette image se trouve ainsi à l'abri des mises à jour potentielles qui ne permettraient pas au langagier de retrouver les contextes originaux ou les attestations nécessaires à son travail. Le stockage des sites Web évite aussi de nombreuses manipulations des corpus puisqu'il permet un retour rapide et quasi instantané au texte en cas de besoin. Cet aspect passe bien souvent inaperçu, mais le terminologue n'a plus à consigner toutes les informations relatives aux termes à l'étape de la collecte des données puisqu'elles demeurent constamment accessibles.

Impact sur le travail terminologique : aspects négatifs

Bien que la méthodologie utilisée accélère l'initiation au domaine, elle est bien souvent légèrement désordonnée. La lecture d'articles, de manuels ou de monographies sur un sujet précis facilite une entrée en matière logique, alors que la méthode que nous avons utilisée donne tout simplement un aperçu immédiat des notions les plus importantes. Au surplus, une telle entrée en matière ne permet pas de graduellement établir ou découvrir les relations sémantiques entretenues par les notions (générique/spécifique, tout/partie, etc.) sans faire appel aux contextes. Le terminologue doit donc invariablement pousser un peu plus loin sa lecture afin pouvoir se faire une idée plus juste de la structure conceptuelle du domaine.

TermoStat, l'outil d'acquisition automatique de termes que nous avons utilisé, ne donne qu'un accès restreint aux contextes en limitant la consultation à une seule phrase. Cette restriction oblige le terminologue à avoir recours à un autre outil pour accéder à un contexte plus large ou à se référer au texte original afin de s'assurer de la validité d'un terme. Cette difficulté est plus importante lorsque le terme à l'étude apparaît en fin de phrase.

Un tel outil, qui tente de quantifier l'importance des candidats termes pour le corpus, est également tributaire des phénomènes liés à la fréquence des unités recensées. Ainsi, si un candidat est très fréquent dans le corpus dépouillé par rapport au corpus de référence, il apparaîtra en tête de liste et un poids important lui sera accordé. Cette technique conduit donc à l'inclusion de termes *à la mode* dans les résultats proposés par l'extracteur. Le terminologue ne peut éliminer ces suggestions qu'après la consultation de plusieurs contextes ou à la suite de recherches plus approfondies. Il faut cependant noter que cette charge de travail supplémentaire, bien que parfois pénible, force tout de même le terminologue à étudier les divers termes en concurrence et à évaluer leur implantation.

Afin d'éviter de présenter au langagier une liste de candidats termes beaucoup trop longue, les logiciels d'acquisition automatique de termes ont aussi la fâcheuse tendance à laisser de côté certains termes qui seraient jugés intéressants pour le terminologue. Comme il est fastidieux de parcourir manuellement un corpus volumineux, le terminologue ne saura donc jamais que ce terme se retrouve dans le corpus et ne pourra l'attester à moins d'avoir recours à un concordancier et de le rechercher volontairement.

La multitude des types de documents trouvés dans Internet pose aussi un certain problème. Comme nous l'avons mentionné précédemment, une étape d'uniformisation et de conversion s'avère nécessaire si l'on veut manipuler tous les documents constituant les sites Web avec l'extracteur de termes et les divers outils utilisés. Ce faisant, certains éléments d'interface sont inévitablement omis. C'est le cas, par exemple, du texte des menus déroulants, des entrées des listes déroulantes, des étiquettes des formulaires ainsi que du texte des boutons ou des images qui n'ont pas toujours été extraits. Bien que certains logiciels, comme les logiciels de localisation, permettent de récupérer une partie de ces données, d'autres sont tout simplement perdues. Le texte qui se trouve dans les images en mode point, le format le plus fréquent sur les sites Web, ne peut tout simplement pas être pris en charge par ces outils.

Il est important de noter que, dans le cadre d'un projet comme celui décrit ici, la dépendance à l'égard de la performance des outils de traitement automatique de la langue est double. Nous sommes effectivement tributaires du dépouillement automatique et de la qualité des résultats ainsi obtenus. De la même façon, les performances du logiciel d'acquisition automatique de termes sont elles aussi dépendantes de celles de l'étiqueteur morphosyntaxique et de sa capacité à lever les ambiguïtés et à lemmatiser le corpus convenablement.

Conclusion

La collaboration mise en place par notre groupe de recherche et nos partenaires du *CTREL* nous a permis de procéder au dépouillement assisté par ordinateur d'un domaine dont la terminologie se rencontre essentiellement dans Internet. Nous avons mis sur pied une chaîne de travail terminologique flexible grâce à des outils accessibles à très peu de frais ou gratuitement. La méthodologie décrite peut donc être facilement adaptée par les terminologues en substituant des outils semblables à ceux que nous avons utilisés et les résultats obtenus seront comparables. L'approche nous a permis de d'élaborer rapidement environ 60 dossiers terminologiques complexes à partir d'un corpus de taille importante (un peu plus de 2 millions de mots).

Le recours à *TermoStat* pour l'identification des termes nous permet d'envisager diverses pistes pour des recherches futures. En effet, puisque ce dernier met en opposition un corpus de référence et un corpus spécialisé (celui où la terminologie est identifiée), nous pouvons faire varier la nature de ces derniers en fonction de divers objectifs de dépouillement terminologiques. Nous pourrions ainsi envisager de revisiter les sites Web dépouillés pour cette étude de façon périodique et de comparer les anciennes copies du site avec une version fraîchement aspirée. Cette technique permettrait la mise en place d'un système de veille terminologique pour le domaine de la réservation en ligne.

Plutôt que de choisir un axe diachronique décrit au paragraphe précédent, il serait envisageable d'aller de l'avant avec une comparaison de corpus élaborés à partir de sites Web d'origines géographiques différentes. Ainsi, nous serions à même d'identifier les particularités terminologiques d'un pays par rapport à un ou à plusieurs autres pays pour mettre en lumière, de façon semi-automatique, des phénomènes de variation topolectale.

Les avenues de recherche apportées par la linguistique de corpus et par les nouveaux outils issus des recherches en terminologie computationelle sont donc nombreuses. Par contre, les corpus mis sur pied semi-automatiquement à partir d'Internet posent de nouveaux défis dans le contexte des recherches terminologiques, tant du point de vue de leur manipulation, puisqu'ils sont généralement volumineux, que de celui de leur contenu. En effet, le style et la qualité de la langue qu'on y retrouve ne sont pas toujours adéquats pour le traitement terminologique et l'utilisation de ce type de documents pour l'attestation de termes est rarement bien perçue. Il reste donc beaucoup de chemin à parcourir avant que la démarche terminologique et les langagiers soient en mesure d'exploiter à son maximum cette mine d'informations qu'est Internet. En effet, cette source documentaire extraordinaire tarde à être exploitée convenablement et à sa juste valeur.

REMERCIEMENTS

Les auteurs tiennent à remercier Pierrette Vachon-L'Heureux de l'*Office québécois de la langue française* et Caroline Milaire de la société *Amex Canada Inc.* pour leur participation au *CTREL* et leurs conseils tout au long du travail d'élaboration des dossiers terminologiques.

NOTES

1. Dans le cadre de notre projet, nous avons utilisé le logiciel *HTTrack Website Copier* disponible gratuitement dans Internet : http://www.httrack.com.
2. Ces logiciels sont distribués par la compagnie *ABC Amber* : http://www.thebeatlesforever.com/processtext.
3. Puisque les outils informatiques ne peuvent, sans risquer de se tromper, distinguer les unités terminologiques des autres unités, les chercheurs travaillant dans le domaine de l'acquisition automatique des termes adoptent une terminologie prudente et parlent généralement de candidats termes. Cette notion, mise de l'avant par Bourigault (1994) a été reprise par la majorité des chercheurs œuvrant dans le domaine : Frantzi et Ananiadou (1997), Habert et al. (1997), Jacquemin (1997) et Daille (1999), etc.
4. Les auteurs tiennent à remercier André Clas d'avoir mis à leur disposition le corpus *The Gazette* pour ce projet.
5. Cette notion de « poids » des candidats rejoint celle de *termhood* proposée par Kageura et Umino (1996) que l'on pourrait traduire par *potentiel terminologique* des candidats termes.

RÉFÉRENCES

AUGER, P., DROUIN, P. et M.C. L'HOMME (1991): «Automatisation des procédures de travail en terminographie», *META* 36-1, Montréal, Presses de l'Université de Montréal, p. 121-128

BÉDARD, C. et E. DUCHESNE (2002): *Système d'alimentation et de mise à jour intégré de la Banque de terminologie du Québec (SAMI) : Guide de l'utilisateur*, Québec, Office québécois de la langue française.

BOURIGAULT, D. (1994): *Un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*, thèse de doctorat, École des Hautes Études en Sciences Sociales, 352 p.

DAILLE, B. (1999): «Identification des adjectifs relationnels en corpus», *TALN '99,* Cargèse, p. 105-114.

DROUIN, P. (2003): "Term extraction using non-technical corpora as a point of leverage", *Terminology* 9-1, p. 99-117.

FRANTZI, K. et S. ANANIADOU (1997): "Automatic Term Recognition Using Contextual Cues", *Proceedings of the 3rd DELOS Workshop*, 8 p.

GENNARI, J. H. *et al*. (2003): "The evolution of Protégé: an environment for knowledge-based systems development", *International Journal of Human-Computer Studies* 58-1, p. 89-123.
GOOGLE (2005): *Gmail*, http://www.gmail.com, page consultée le 8 mars.
Habert, B., NAZARENKO, A. et A. SALEM (1997) : *Les linguistiques de corpus*, Paris, Armand Colin, 240 p.
JACQUEMIN, C. (1997): *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, HDR en informatique, IRIN, Université de Nantes.
KAGEURA, K. et B. UMINO (1996): "Methods for automatic term recognition: A review", *Terminology* 3-22. p. 259-289.
LAFON, P. (1980): « Sur la variabilité de la fréquence des formes dans un corpus », *MOTS* 1, Presses de la Fondation Nationale des Sciences Politiques, p. 128-165.
L'HOMME, M.C. (2004) : *La terminologie : principes et techniques*, Montréal, Les Presses de l'Université de Montréal.
PROQUEST (2005): *ABI/INFORM : Global*, http://proquest.umi.com/login, page consultée le 4 février.
SAGER, J. C. (1990): *A practical course in terminology processing*, Amsterdam/Philadelphia, John Benjamins Publishing Company.
SCHMID, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Proceedings of the Conference on New Methods in Language Processing, p. 44-49.