Extension des ressources lexicales grâce à un corpus dynamique

Anne Dister

Centre de traitement automatique du langage (CENTAL), Université de Louvain
dister@tedm.ucl.ac.be

Cédrick Fairon

Centre de traitement automatique du langage (CENTAL), Université de Louvain
fairon@tedm.ucl.ac.be

ABSTRACT. A common problem in automatic text analysis is that of unknown words: i.e. those which are not recognized by the system because they are not listed in its dictionaries. These words belong to various categories: misspellings, neologisms, proper names, etc. Moreover, it is often the case that the dictionaries of a given language do not include the lexical variants particular to the different regions of the world where that language is spoken.
In this paper, we analyse a corpus of Québec newspapers and based on the use of GlossaNet we present a simple methodology for extracting unknown words and extending dictionaries.

KEYWORDS: corpus, dynamic corpus, natural language processing, neologisms, electronic dictionaries.

RÉSUMÉ. L'analyse automatique est régulièrement confrontée au problème trivial des mots inconnus, c'est-à-dire non reconnus par le système parce qu'ils n'apparaissent pas dans les dictionnaires de celui-ci. Cette catégorie de mots inconnus est hétérogène : mots mal orthographiés (coquilles ou fautes d'orthographe), néologismes, noms propres, etc. De plus, lorsque le logiciel de traitement automatique est conçu pour analyser des corpus du français de France, les particularités lexicales de textes suisses, sénégalais, belges ou encore québécois figurent elles aussi parmi les mots inconnus.
Dans cet article qui se base sur l'analyse d'un corpus dynamique de presse québécois, nous présentons une manière simple, grâce au système GlossaNet, de repérer ces mots inconnus afin, s'il y a lieu, de les faire figurer dans un dictionnaire.

MOTS-CLÉS : corpus, corpus dynamique, traitement automatique des langues, Unitex, Glossanet, mot inconnu, néologisme, québécisme.

1. Introduction

Le recours aux corpus pour la recherche d'exemples et d'attestations est aujourd'hui une pratique bien ancrée en linguistique : les chercheurs constituent des exempliers en collectant dans de vastes corpus les occurrences des phénomènes linguistiques qu'ils souhaitent étudier. Le recours à ces larges collections de textes comme base de toute analyse a naturellement pour but de fonder les recherches sur une réalité tangible. Des outils informatiques permettent d'automatiser les recherches avec plus ou moins de précision : il s'agit en général de logiciels entrant dans la catégorie des

concordanceurs. Une fois qu'un corpus est passé par le crible du concordanceur et que le chercheur a relevé toutes les occurrences des structures auxquelles il s'intéresse, le corpus est « épuisé » et doit être remplacé pour que la recherche puisse se poursuivre. Reste donc au linguiste à se procurer de nouveaux textes et à recommencer la procédure d'analyse, ce qui peut être coûteux en temps et en efforts. Le système GlossaNet propose une solution originale à ce problème

2. Glossanet et le corpus dynamique

En effet, plutôt que de donner accès à des corpus « fermés », c'est-à-dire établis de manière définitive sur la base de critères stricts, Glossanet donne accès à des corpus « ouverts » qui sont continuellement mis à jour grâce au prélèvement de nouveaux textes sur Internet.

Pour décrire la méthodologie implémentée dans ce système, nous avons repris à Antoinette Renouff [1992] le concept de « corpus dynamique ». Dans cette perspective, le corpus est vu comme un flux de données textuelles plutôt que comme une banque de données plus ou moins fermée. Ce flux apporte continuellement de nouveaux textes et donc de nouveaux exemples. Dans cet article, nous montrerons comment un tel système peut servir à la constitution et à l'extension automatique de ressources linguistiques destinées aux applications de traitement automatique du langage.

2.1 Mise à jour des corpus

Concrètement, le système est spécialisé dans la collecte des textes de presse : il télécharge quotidiennement sur Internet l'édition du jour de plus de 80 journaux dans 9 langues (français, anglais, italien, norvégien, portugais, espagnol, grec, néerlandais et allemands). Les textes récupérés sont ensuite analysés grâce aux programmes du logiciel Unitex[1], un analyseur de corpus qui permet l'application de ressources lexicales sur les textes. Une fois passés au travers de ce processus, ces textes sont prêts pour l'exploitation par GlossaNet. Comme on le voit, il s'agit d'une exploitation essentiellement « verticale » du Web. Des sites ont été définis et sont visités en profondeur. Cela permet, nous semble-t-il, de garantir une certaine homogénéité des corpus, mais ce n'est pas le seul point de vue possible : des expériences qui favorisent la sélection de textes par le biais d'une exploration horizontale ont également été réalisées (cf., par exemple [WALKER 99]).

De leur côté, les utilisateurs accèdent au système par Internet et configurent des « requêtes » sur un ou plusieurs journaux. Ces requêtes peuvent contenir des mots ou des représentations formelles de structures syntaxiques à rechercher. A chaque remise à jour des journaux sélectionnés (c'est-à-dire en général quotidiennement), les requêtes des utilisateurs sont réappliquées automatiquement sur le nouveaux corpus et les éventuels résultats obtenus sont envoyés par courriel sous la forme d'une concordance en HTML.

[1] Unitex est un analyseur de corpus Open Source basé sur Unicode et développé par Sébastien Paumier à l'Université de Marne-la-Vallée (http://www-igm.univ-mlv.fr/~unitex/). Ce système permet d'appliquer des dictionnaires électroniques et des grammaires sur des textes.

était poursuivi pour diffamation dans l' affaire Omar . {S}Par Brigitte VITAL-DUR
o et la préface d'un livre consacré à l' affaire Omar Raddad, le jardinier reconn
un côté plastique, pour la nudité". {S} Affaire Dutroux .{S} L'un des rares à ass
Je me masturbais.{S}" Puis il y a eu l' affaire Dutroux .{S} "Là, on a vu que l'a
oeuvre et a porté à bout de bras l' affaire Dreyfus , sa lutte en faveur des
.{S} " On baignait à ce moment dans les affaires Dutroux et Agusta-Dassault ", e
hier, quatre personnes du ministère des Affaires Etrangères s'occupaient du cond
me Patricia De Jonghe, du ministère des Affaires Etrangères . {S}Ce marché est u
frappe tous les esprits.{S} Surtout, l' affaire Enron n'est pas un cas isolé.{S}
incompatibles avec l'audit légal. {S}L' affaire Enron a montré qu'il y avait un
des mois en incorporant à son chiffre d' affaires des échanges de droits d'accès
de créer une diversion sur le front des affaires du patinage artistique . {S}Eri

ASSISES {S}Pas de prescription dans l' affaire des disparues de l 'Yonne. {S}Pa
aquien {S}Nouvelle perquisition dans l' affaire des emplois fictifs . {S}Par Fab
condamné à deux ans avec sursis dans l' affaire de la gestion de son cabinet. {
s mises en accusation des inculpés de l' affaire Cools : Van der Biest, Taxquet,
s peser lourdement sur l'évolution de l' affaire Cools et sur le procès qui s'ann

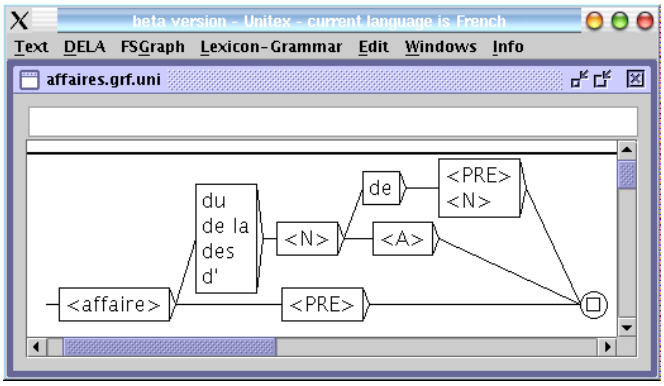
ion d'Alain Van der Biest soit liée à l' affaire Cools , elle ne clarifie rien mai
nternational estime que le passage de l' affaire Semira en chambre du conseil con
nt également pour exiger justice dans l' affaire Sémira Adamu.{S} " Les gendarmes
eser sans aucun doute sur la suite de l’ affaire Cools .{S} RENÉ HAQUIN {S}Dix an
ires datant du début des années '90 : l' affaire des titres volés et un dossier d
e potentiel.{S} Mais l’instruction de l’ affaire Cools est étrange.{S}Il y a sans
nternational estime que le passage de l' affaire Semira en chambre du conseil con.

(Les concordances en HTML contiennent des liens hypertextes qui permettent de retrouver le texte original dans lequel l'occurrence a été trouvée.)

Le grand avantage de ce système est d'automatiser les recherches. Une fois qu'une requête est enregistrée, elle est appliquée chaque jour sur les nouveaux corpus et l'utilisateur reçoit un rapport si de nouveaux exemples ont été trouvés. Dans ce sens, on peut parler de « veille linguistique ». Comme toutes les tâches d'analyse et de mise à jour des textes exécutées automatiquement sur le serveur, le système est extrêmement simple d'utilisation.

2.2 Les requêtes

Les requêtes peuvent être exprimées sous forme d'expressions régulières ou sous forme de graphes. Dans les deux cas, elles peuvent contenir des mots (*avion*, *voitures*, etc.), des symboles grammaticaux (<N>, <V>, <A>, <ADV>, etc.), des symboles formels (<MAJ>, pour un mot en majuscules, <PRE> pour un mot dont la première lettre est une majuscule, etc.) et des lemmes (<être>, représente la forme infinitive et toutes les formes conjuguées du verbe *être*, <beau>, représente toutes les formes fléchies de *beau*). Pour être complet, disons plus simlement que GlossaNet reconnaît toutes les possibilités de recherche du logiciel de traitement de corpus Unitex[2].
Cet usage est rendu possible par l'analyse linguistique du texte qui est réalisée par Unitex. Ce logiciel a recours à une série de dictionnaires électroniques construits selon la méthodologie du LADL (cf. le système DELA [COURTOIS 90] et [GROSS 89]).
Voici deux exemples de requêtes valides :
1) Expression régulière : <être> en <N:s>. Cette expression permet de retrouver dans le corpus sélectionné toutes les occurrences de structures commençant par une forme du verbe *être* suivie de *en* suivi d'un nom au singulier
2) Exemple de graphe



3. Patricularités de GlossaNet

Comme GlossaNet ne nécessite l'installation d'aucun logiciel spécifique (autre que le navigateur Web), le système peut facilement être utilisé avec des étudiants, par exemple pour leur apprendre à formaliser des expressions et à faire des recherches complexes qui impliquent l'utilisation de codes linguistiques (morphologique, sémantique, syntaxiques, etc.). Attendu la nature des corpus auxquels GlossaNet s'intéresse, le système est également utilisé par des personnes intéressées par la

[2] On se reportera donc au manuel de ce logiciel pour une description complète des possibilités de recherche.

recherche d'information ou par la veille d'informations : des personnes qui souhaitent être averties quand un article traitant de tel ou tel sujet est publié. D'une certaine manière, cela permet de faire une « revue de presse » automatique. Comme nous l'avons mentionné, les corpus dynamiques sont également intéressants pour toutes les applications dans lesquelles on constitue des ressources automatiquement ou semi-automatiquement à partir de corpus (par exemple dans des applications telles que celles décrites par [BOGURAEV 96]).

4. Les mots inconnus dans Cyberpresse

Pour ce travail, dans lequel nous nous intéressons aux « mots inconnus », la requête que nous avons formulée à Glossanet est relativement simple : il s'agissait pour le système de nous fournir tous les mots (nous parlons ici de mots simples) présents dans notre corpus et absents du dictionnaire des mots simples delafm du LADL. La requête a la forme suivante : < !DIC>.

Notre corpus est constitué d'extraits de 7 journaux québécois mis en ligne sur le site Cyberpresse (http://www.cyberpresse.ca/) : il s'agit des journaux *Le Soleil, La Presse, Le Nouvelliste, Le Droit, La Tribune, Le Quotidien* et *La Voix de l'Est.* Le volume est approximativement de 300 à 400 mégas de données renouvelées quotidiennement grâce à Glossanet.

Max Silberztein [SILBERZTEIN 95] a calculé que près de 80 % des mots inconnus d'un texte étaient des noms propres. Puisque c'est uniquement cette catégorie des mots inconnus qui nous intéresse ici et afin de ne pas encombrer inutilement l'analyse, nous avons choisi d'appliquer au texte un automate qui analyse automatiquement comme des noms propres les formes non reconnues par le dictionnaire des mots simples et qui commencent par une majuscule[3].

Les premières concordances recueillies étaient surchargées d'occurrences qui occasionnaient du bruit : des suites de lettres séparées par un point (des mots donc, au sens du système Unitex) qui forment les adresses internet, informations très présentes dans notre corpus. Les mots inconnus qui nous intéressaient vraiment étaient perdus dans ce « fatras », qui tout comme les noms propres, surcharge le travail de lecture de la concordance.

Nous avons alors recensé systématiquement pendant plusieurs jours les URL présentes dans le corpus, afin d'en établir la typologie. Un graphe a alors été créé, qui fonctionne comme un filtre et supprime les URL des séquences à analyser par Unitex.
Restaient alors dans nos concordances les « vrais mots inconnus », dont nous allons voir maintenant qu'on peut les regrouper en différentes classes[4].

[3] Dans un travail comparable sur les mots inconnus que nous avions fait sur un corpus de presse statique (par opposition à la notion de corpus dynamique de GlossaNet), nous avions obtenu des résultats comparables à ceux de [SILBERZTEIN 95] : sur les 11 031 formes non reconnues par le dictionnaire des mots simples delafm, 8470 étaient des noms propres, c'est-à-dire 77 %. Voir [DISTER 2000].

[4] Les résultats que nous observons ici sont semblables à ceux présentés dans [DISTER 2000] ; la typologie des mots inconnus est donc relativement semblable. En ce qui concerne les néologismes, les résultats sont bien entendu très différents.

4.1 Les mots mal orthographiés : coquilles et fautes d'orthographe

Un grand nombre de mots de notre corpus ne sont pas reconnus tout simplement parce qu'ils sont mal orthographiés. Ce type de problème apparaît peu ou prou dans tous les types de corpus analysés[5]. Dans *Cyberpresse*, les principales coquilles se répartissent comme suit :
- une lettre manque : *annoné, milions, arrire, réfomes, miliards, corénne, etc.*
- une lettre est en trop : *rapppelle, juilllet, attdendent, afffaires, ballles, papilllons, etc.*
- une lettre est mise pour une autre : *sapré, matérieux, champanzés, padadoxe, etc.*
- deux lettres sont interverties : *soldtas, orgeuilleuse, etc.*
- l'accent manque : *décu, deuxieme, troisieme, ambiguité, etc.*
- un accent est ajouté : *démantélé, écclésiastique, désséchée, auraît, etc.*
- l'espace manque entre deux mots qui sont alors soudés : *aumoins, vlandans, leursresorts, consulterle, etc.*
- un nom propre est écrit sans majuscule (et n'est donc pas retenu par le filtre) : *artique, céline, vigneault, duc d'albe, etc.*

Il y a aussi les erreurs qui s'apparentent plus à des fautes d'orthographe qu'à des coquilles. On peut citer, entre autres :
rattrappent, accomodent, décolage, occurence, égoîste, succédés, tenacité, jettez, rejettez, paléonthologue, cauchemard, etc.

4.2 L'orthographe réformée

On constate, à observer attentivement les mots non reconnus, qu'un certain nombre d'entre eux qui pourraient a priori être considérés comme mal orthographiés participent en fait des rectifications orthographiques proposées par le Conseil supérieur de la langue française en 1990. Cette réforme, qui visait des « aménagements » afin de supprimer de l'orthographe française « un certain nombre d'anomalies », ne s'est pas véritablement imposée, même si certaines tentatives continuent de se manifester pour la promouvoir[6]. Néanmoins, et nous pouvons le constater dans notre corpus, de nombreuses occurrences de mots orthographiés selon les recommandations de la « Réforme » sont quotidiennement utilisées dans les textes. Ainsi, nous trouvons dans *Cyberpresse* des formes qui résultent des propositions suivantes[7] :

- Suppression de l'accent circonflexe sur le *i* et le *u* (On trouve notamment les formes *connait, connaitre, parait, reconnait, surement, etc.*)
- Ajout ou changement d'un accent qui rend la forme graphique conforme à la prononciation. C'est le cas de *sénior* ou encore de *répèteraient.*
  La conjugaison de quelque 240 verbes est affectée par cette rectification[8]. Ainsi, le futur et le conditionnel de verbes comme *régler* n'est plus *réglerai / réglerais* mais *règlerai / règlerais*. Parmi les verbes touchés par cette modification, on peut citer *abréger, accéder, accélérer, acérer, assécher, assiéger, avérer, etc.*
- Suppression de lettres superflues : *asseoir* devient, comme nous l'avons trouvé, *assoir*.
- Certains noms d'origine étrangère sont francisés : *supporteur*.
- Certains mots composés s'écrivent soudés : *basketball, baseball, hanball,* etc.

[5] Pour une étude comparable sur une année du journal *Le Monde*, voir [MATHIEU 98].
[6] Citons par exemple la réédition récente au Seuil du livre dirigé par Jacques Chaurand *Nouvelle histoire de la langue française* ou encore l'ouvrage collectif *Le Français dans tous ses états*.
[7] Nous résumons ici. Pour un panorama complet des « règles » proposées, de leur justification ainsi que des formes concernées voir [GOOSSE 91].
[8] Après flexion, cela donne 2 668 entrées.

Pour résoudre ce problème général des mots non reconnus participant des recommandations de 1990, nous avons créé un dictionnaire électronique de l'orthographe réformée (delafmOR). Pour les mots simples, il comprend dans sa forme actuelle près de 9000 entrées. Il adopte le format et les codes des autres dictionnaires du système Unitex. Une ligne correspond à une entrée lexicale qui se compose comme suit :

- la forme telle qu'on peut la trouver dans le texte ;
- une virgule ;
- le lemme (c'est-à-dire la forme canonique, telle qu'on la trouve en entrée dans un dictionnaire courant) ;
- un point ;
- la catégorie grammaticale ;
- un double point ;
- les informations flexionnelles (genre et nombre pour les noms et les adjectifs, conjugaison pour les verbes).

Voici un extrait du delafmOR :

abime,abime.N:ms
abimes,abime.N:mp
abrègement,abrègement.N:ms
abrègements,abrègement.N:mp
absout,absoudre.V:Kms[9]
absouts,absoudre.V:Kmp
addenda,addenda.N:ms
addendas,addenda.N:mp
affèterie,affèterie.N:fs
affèteries,affèterie.N:fp
affut,affut.N:ms

L'utilisation de notre dictionnaire de l'orthographe réformée, parallèlement au dictionnaire des mots simples, permet de reconnaître tous les mots cités dans cette section.

4.3 La féminisation des noms de métier, fonction, grade ou titre

En 1979 paraissait dans la *Gazette officielle du Québec* un avis recommandant de féminiser les titres. Depuis de nombreuses années, des formes comme *l'auteure, la professeure* ou *la juge* sont courantes en français du Québec. La Suisse en 1988 et la Belgique en 1993 ont également pris des mesures institutionnelles visant à encourager l'utilisation des formes féminines. Dans *Cyberpresse*, comme dans tous les écrits français du Canada, toutes les professions exercées par des femmes sont systématiquement notées sous leur forme féminine. On trouve à de nombreuses reprises les formes, inexistantes en français de France ou de Belgique, *auteure, professeure, amateure, compositeure, metteure en scène, Procureuse générale, réviseuse, ingénieure, gouverneure, ingénieure, écrivaine, entraîneure, sculpteure, entrepreneure, etc.* Et sur le même modèle : *aggresseure, vainqueure,* etc.

[9] V:Kms = verbe au participe passé masculin singulier

En ce qui concerne la féminisation en Communauté française de Belgique, nous avons constitué sous format électronique un « Dictionnaire des noms de profession » qui comprend 4190 formes fléchies[10]. Néanmoins, se basant sur le guide publié par la communauté française de Belgique, il n'inclut pas, à l'heure actuelle, les formes féminisées typiquement québécoises mentionnées plus haut.

4.4 Les mots étrangers

Ces mots figurent souvent dans de titres de films, de livres ou de chansons :

« suite non officielle de Sex, Lies and Videotapes»

Ici, seul le mot *and* est repris dans la concordance puisque les autres mots, commençant par une majuscule, sont retenus par le filtre.

Un grand nombre d'occurrences prend également place dans des discours rapportés :

On me crie «Welcome to Korea, welcome to Korea»

Martin Luther King lance son célèbre : « I had a dream ! »

Une solution « made in Canada », comme le dit Ralph Klein.

Dans certains cas, le mot étranger est employé pour noter une réalité étrangère (la *Casa del Popolo*), ou faire plus « couleur locale » :

Vous ne verrez pas d'indigènes en vedette dans les telenovelas (feuilletons)

Notons la traduction qui suit ici immédiatement l'emploi du terme étranger. Il arrive assez fréquemment que le mot étranger soit traduit en français :

ce qu'on appelle en anglais leur « language of public use»

les abeilles-reines (queen bees) dictent leurs règles

Dans Cyberpresse, la majorité des mots étrangers sont évidemment des mots anglais.

4.5 Les québécismes

Comme on pouvait s'y attendre, un grand nombre de formes non reconnues dans *Cyberpresse* sont en fait des formes propres au français du Québec. On peut citer entre autres : *baloune, bantam, bebitte, bibitte, cégep, cégépien, chialeux, chiac, chum, cenne, chialage, diseux, s'enfarge, s'enfargeant, encarcanne, enfirouaper, fun, frimée, game, guidounes, garrochent, jasette, joualisants, mauditement, patenteux, parlable, pelleteux, placoteux, perchaude, piasse, poutine, placoteux, quêteux, rapaillé, ratoureux, ratoureuse, raplomber, recherchiste, relationniste, robineux, saumonnier, sundae, tabletter, téléroman, télésérie, tounes, tripeux, tripeuse, téteux, québécisants, etc.*

À cette liste, s'ajoutent tous les adjectifs dérivés de noms de lieux : annemontoise, abitibien, abénaquaise, albertaines, bellechassoise, kongais, kongaise, kongaises, latuquois, laurentien, laurentienne, lavalloise, lévisien, magogoise, manitobaines, rimouskois, saskatchewanais, shawiniganais, shawiniganaises, trifluvienne, thetfordois, transgaspésien, etc.

La constitution d'un Dictionnaire électronique des formes simples en français du Québec (DELQUES) est en cours à l'UQAM. Il contient actuellement 87 505 entrées lexicales. Il permet d'étendre la couverture lexicale des corpus québécois.

Néanmoins, la constitution d'un tel dictionnaire ne règle pas le problème trivial des faux-amis. Ainsi, *dépanneur* possède 3 sens en français québécois, sens inconnus du français de France ou de Belgique : « épicerie du coin », « tenancier de cette épicerie » ou encore « grille pour déplacer un véhicule dans la neige ».

[10] Il est téléchargeable à l'adresse suivante : http://ladl.univ-mlv.fr/tools/tools.html#profession

Une autre version du dictionnaire est également en chantier, qui reprend notamment des marques sémantiques et donne, lorsqu'il y a lieu, un équivalent sémantique qui permet de distinguer les faux-amis. En voici un extrait[11], avec notamment les entrées pour *dépanneur* :

- argenté,A32+z3+Q/LAP/BER/GLO/<riche>
- argenteries,N22P+Conc+z2+Q/GLO/<argenterie>
- argents,N2P+Abst+z1+Q/GLO/DUL/BER/<fonds;valeurs>
- argnée,N31+Conc+[Obj]+z3+Q/BER/<crochet servant à remonter les seaux du puits>
- argot,N1+Conc+Pc+z1+Q/BER/<sabot>
- argotage,N1+Abst+z3+Q/BER/=ergotage
- arguer,V3+t+32R3+~cause+z3+Q/GLO/<plaider>
- arguette,N21+Conc+[Obj]+z3+Q/BER<oreiller>
- argument,N1+Abst+z2+Q/BER/GLO/<dispute>
- aria,N1+Abst+z1+Q/DUL/SEU/LAP/DQA/<embarras>
- aria,N1+Abst+z1+Q/DUL/BER/SEU/DQA/GLO/<gâchis;fouillis>
- aria,N1+Conc+z1+Q/DUL/BER/DQA/GLO/<attirail>
- aria,N1+Hum+z1+Q/DQA/<personne qui est un fardeau>
- (…)
- dépanneur,N1+Conc+[Lieu]+z1+Q/DUL/SEU/<épicerie du coin>/=accomodation
- dépanneur,N35+Hum+z2+Q/DUL/SEU/<tenancier de dépanneur>
- dépanneur,N1+Conc+[Obj]+z2+Q/DUL/<grille pour déplacer un véhicule dans la neige>

4.6 Les néologismes

Nous avons rencontré des néologismes qui ne nous semblent pas spécifiques à la « nature québécoise » du corpus. Certains d'entre eux, comme pour certaines occurrences de mots étrangers, apparaissent entre guillemets :
« djihadisée » et *« dédjihadiser »* ; *« pottermaniaques »* et *« pré-pottermaniaque »*, *récréotouristique, atmosphériste, sècheté* (dans *la sècheté de la tragédie*), *folleries, analphabêtise, dinosaurienne* (dans *gestion dinosaurienne*), *taxage, idéateur*, etc.

4.6.1 Les apocopés

De nombreuses formes apocopées apparaissent dans le corpus : *anglo* (pour désigner les anglophones du Québec), *hélico* (hélicoptère), *négos* (négociations), *qualifs* (qualifications), *coloc* (colocataire), *biotechs* (biotechnologies), *condo* (condominium), etc. Bien que beaucoup plus fréquent à l'oral, le phénomène tend à imposer certaines de ses formes à l'écrit, signe sans doute pour celles-ci d'une large diffusion auprès des locuteurs/lecteurs. Recenser toutes ces formes dans un dictionnaire permettrait ici aussi d'étendre la couverture des mots reconnus.

4.6.2 Les préfixes et suffixes productifs

Pour terminer, nous pouvons également citer les néologismes créés grâce à des préfixes productifs, et qui ne sont pas spécifiques au français du Québec :

[11] Conc = Concret ; Abst = Abstrait ; DUL = *Dulong* ; GLO = *Glossaire du parler français au Canada*.

anticorruption, antimotards, antidrogue, antialgues, antibriseurs, antipauvreté, antimondialisation, antiterrorisme, antinauséeuses, etc. ;
codistribution, coanimera, cocréateur, corédacteur, coscénarisé, etc. ;
mégaporcherie, mégaprocès, mégacomplexe, mégatendance, mégasuccès, mégahôpitaux, mégaspectacle, mégaprojets, etc. ;
préproduction, précriminelle, prémaoïste, préhospitaliers, prémâchées, etc. ;
superministère, superhéros, superdivas, supermenteur, supermamies, etc. ;
Citons également en vrac : *intermodale, intermodalité, intergénérationnels, interrives, écotourisme, écotouristique, postréférendaire, surréglementation, rééquipement réhumanise, hypertechniciens, hypersexualisée, hyperefficace, unifamiliale, minientrepôts, minidisque, postmoderniste, cybercommerce, cybercours, cyberculture, cyberespace, télécommunicateur,* etc.
Citons également des néologismes construits sur les suffixes productifs *–phile* ou *-phobe* : *félinophiles, jazzophiles, homophobes, québécophile,* etc.
Ce problème des néologismes est pour nous, dans le cadre de l'enrichissement de nos dictionnaires électroniques, le même que pour les lexicographes : faut-il entrer tous ces mots, peut-être créations d'un jour, dans le dictionnaire ?

5. Conclusion

Quel que soit le corpus qu'il analyse, le chercheur en traitement automatique des langues est confronté à des problèmes invariants : les coquilles et fautes d'orthographe, les termes spécifiques au corpus utilisé et qui ne sont pas présents dans les dictionnaires du système. Nous avons vu les solutions que le système Unitex peut apporter. Deux nouveaux dictionnaires, valables pour toutes les variétés de français, permettent de reconnaître un certain nombres d'occurrences jusque-là ignorées : le dictionnaire des noms de profession et le dictionnaire de l'orthographe réformée. Par ailleurs, pour les corpus québécois, l'application du Dictionnaire électronique des formes simples en français du Québec (DELQUES) améliore considérablement la couverture lexicale.
L'intérêt de l'utilisation d'un corpus dynamique réside de toute évidence dans la mise à jour continuelle, et pratiquement en temps réel, des données que le linguiste (ou le chercheur d'autres disciplines d'ailleurs) a à sa disposition. Dans le cadre d'une analyse des « mots inconnus », l'utilisation de Glossanet nous semble particulièrement appropriée pour les lexicographes, qui pourraient alors mesurer la fréquence d'emploi de certains néologismes dans la presse.

Références

Dictionnaire québécois d'aujourd'hui, sous la direction de Jean-Claude Boulanger, Saint-Laurent (Québec), DICOROBERT INC, 1992.
Dictionnaire historique du français québécois, sous la direction de Claude Poirier, Sainte-Foy, Les Presses de l'Université Laval, 1998.
Guide de féminisation des noms de profession, Communauté française de Belgique.
[BOGURAEV 96] BOGURAEV B. et J. PUSTEJOVSKY dir., *Corpus Processing for Lexical Acquisition*, Cambridge, MA, MIT Press, 1996.
[COURTOIS 90] COURTOIS Blandine, « Dictionnaires électroniques du français », in B. COURTOIS et M. SILBERZTEIN (eds.), *Dictionnaire électronique du français, Langue française* 87, Paris, Larousse, pp. 11-22, 1990.
[DISTER 2000] DISTER Anne, « La presse québécoise vue à travers l'analyse de ses particularités lexicales », dans *Actes des XIVes Journées de Linguistique (23 et 24 mars 2000)*, Québec, Presses de l'Université Laval, 2000.
[FAIRON 98-99] FAIRON Cédrick, "Parsing a Web Site as a Corpus", in C. FAIRON dir., *Analyse lexicale et syntaxique : le système INTEX, Lingvisticae Investigationes* Tome XXII (Volume spécial), Amsterdam/Philadelphie, John Benjamins, 1998-1999.
[FAIRON 2000] FAIRON Cédrick et Blandine COURTOIS, « Extension de la couverture lexicale des dictionnaires électroniques du LADL à l'aide de GlossaNet », In *Actes du Colloque JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles*, Lausanne, 2000.
[GOOSSE 91] GOOSSE André, *La « nouvelle » orthographe*, Duculot, 1991
[LABELLE 94] LABELLE Jacques, *Dictionnaire électronique des formes simples en français du Québec*, DELQUES V1.0, Rapport de recherche n° 9, Montréal, GRFL, UQAM, 1994.
[LABELLE 96] LABELLE Jacques, « Linguistique comparée et dictionnaires électroniques et INTEX », *Actes des premières journées INTEX*, Paris, Université Paris 7, 1996.
[MATHIEU 98] MATHIEU Yvette Yannick, GROSS Gaston et Christophe FOUQUERÉ, « Vers une extraction automatique des néologismes », *Cahiers de Lexicologie*, n° 72, p. 199-208, 1998.
[RENOUF 92] RENOUF Antoinette "A Word in Time : first findings from the investigation of dynamic text", *ICAME Conference*, Nijmegen, 1992.
[SILBERZTEIN 95] SILBERZTEIN, Max, « Dictionnaires électroniques et comptage de mots », *JADT 1995*, Rome, vol.1, p.93-101, 1995.
[WALKER 99] WALKER D., "Taking Snapshots of the Web with a TEI Camera", *Computers and the Humanities* 33 (1/2), New York, Kluwer Academic, 1999