Constitution et étude de corpus spécialisés sur le Web

C. Fairon
Université catholique de Louvain (CENTAL)

Introduction

Les besoins croissants de la linguistique de corpus et du traitement automatique du langage suscitent une demande toujours plus forte en matière de constitution de corpus, que ce soit du point de vue de la taille des corpus recherchés ou de leur spécificité. Si l'on se contente d'observer les chiffres, l'évolution est très marquée. Le *Brown Corpus* établi dans les années 60 comptait un peu plus d'un million de mots ; dans les années 90, le *British National Corpus* (BNC) atteignait les 10 millions de mots ; dans les années 2000, le *Cobuild Corpus* dépassait les 500 millions de mots. Enfin, en 2010, une équipe de Harvard a constitué un « corpus » de plus de 500 milliards de mots extraits de quelque 5 millions de livres digitalisés par Google soit, selon les auteurs, environ 4% de l'ensemble des livres jamais publiés (J.-B. Michel *et al*. 2010)[1]. Les corpus se sont aussi diversifiés : à côté des corpus généraux sont apparus les corpus thématiques, multilingues, multi-genres, multimodaux, etc. et beaucoup ont été enrichis (en tout ou en partie) de différents niveaux d'annotation.

Dans ce contexte, le Web a très rapidement été perçu comme une véritable opportunité pour créer des ressources spécialisées habituellement difficiles à collecter, comme les corpus parallèles (Resnik 1999, Grefenstet 1999) ou les corpus dédiés aux langues « peu dotées » (de Schryver 2002, Scannell 2007). Cet intérêt s'explique aisément : le Web constitue une source immense de données numériques, facilement accessible, en évolution constante, représentant virtuellement toutes les langues écrites (dans des proportions très variées) ainsi qu'un grand nombre de genres, styles et domaines de spécialités. Depuis la fin des années 1990, les travaux visant à exploiter les données langagières issues du Web se sont multipliés et récemment, le développement de cette thématique de recherche a justifié la création d'un groupe d'intérêt au sein de l'*Association for Computational Linguistics*[2] (ACL). Le développement de cette discipline a été et restera largement influencé par l'évolution du Web lui-même ; nous y reviendrons ci-dessous.

Remarquons cependant que malgré la facilité d'accès aux documents (sur le plan technique), ceux-ci n'en sont pas moins protégés par le droit d'auteur et que leur

[1] Soulignons que la notion de « corpus » varie d'un projet à l'autre : le *Brown Corpus* et le BNC sont des corpus équilibrés, le *Cobuild Corpus* est une base ouverte qui évolue au cours du temps en accueillant de nouveaux textes, le corpus composé des livres de Google est multilingue, et rassemble uniquement des textes extraits de livres (contrairement aux autres qui tentent de représenter différentes modalités du langage). L'évolution des chiffres dessine une tendance générale selon laquelle les corpus traités sont au moins dix fois plus grand à chaque décennie. Mais cette évolution comporte également certains désavantages : les grands corpus ne sont plus complètement annotés, ce qui limite les possibilités d'exploitation en linguistique. Dans le cas du corpus de livres de Google, le volume de données est tel que les méthodes d'analyse utilisables sont relativement limitées et parfois insatisfaisantes pour le linguiste.

[2] http://www.sigwac.org.uk/

redistribution sous forme de corpus ne va pas de soi. Pour contourner ce genre de problèmes, certains auteurs ont étudié la possibilité d'utiliser les documents libres de droits que l'on peut trouver sur Internet, par exemple sous licence *Creative Commons*[3]. Brunello (2009) montre dans son étude sur la question que, bien que les documents sous licence *Creative Commons* soient moins diversifiés et en nombre plus limité, ils représentent une source intéressante de données permettant (dans le cadre de son étude) d'obtenir des résultats comparables[4].

Mais peut-on réellement considérer le Web comme un « corpus » ? La notion de « corpus », utilisée en linguistique et en traitement automatique des langues recouvre en fait des réalités extrêmement diversifiées (voir Dister 2007). Selon les auteurs, la définition est tantôt large (attribuant le statut de corpus à toute collection de documents) tantôt très précise. En linguistique de corpus, les définitions sont généralement assez contraintes .Biber explique qu'un corpus « is not simply a collection of texts. Rather, a corpus seeks to **represent** a language or some part of a language » (Biber *et al.* 1998, p. 246; c'est nous qui soulignons). McEnery and Wilson (2001, p. 32; c'est nous qui soulignons) vont dans le même sens, tout en précisant que, même si toute archive de texte peut en théorie être un corpus,« a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a **finite-sized** body of **machine-readable text**, sampled in order to be maximally **representative** of the language variety under consideration. ». Rastier (2004), pour sa part, met en avant la **démarche réflexive** qui mène à la constitution de corpus en insistant sur les unités « textes » qui composent le corpus : « un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. » Si l'on se place comme le suggèrent Kilgarriff et Grefenshted (2003) dans un cadre large et que l'on définit un corpus comme « une collection de documents », alors on répondra sans hésitation : oui, le Web est un corpus et surtout, une excellente **source** de corpus. Cependant, cette dernière définition ne règle pas pour autant les questions fondamentales auxquelles sont confrontés les chercheurs qui exploitent les corpus : représentativité, comparabilité, échantillonnage, etc.

Dans cet article, nous aborderons la question du Web comme source de corpus à travers le cas particuliers du logiciel GlossaNet et nous monterons que l'approche suivie est particulièrement adaptée à la constitution de corpus thématiques.

1. Les outils

Les outils permettant d'utiliser le Web comme source de corpus peuvent être classés en deux grandes catégories : (i) les logiciels qui **considèrent et utilisent le Web** lui-même **comme un corpus** et (ii) les logiciels qui permettent de **constituer des corpus à partir**

---

[3] *Creative Commons* est une organisation qui procure des « free licenses and other legal tools to mark creative work with the freedom the creator wants it to carry, so others can share, remix, use commercially, or any combination thereof. » (D'après http://creativecommons.org/about ; page consultée le 10/12/2010).

[4] L'étude de Brunello (2009) porte plus particulièrement sur la comparaison de deux corpus italiens « généraux » extraits du web à l'aide du logiciel Wacky (basé sur Yahoo). Le premier corpus est composé uniquement de textes référencés par Yahoo comme libres de droits, tandis que le second n'est pas contraint par cette particularité.

de documents sélectionnés sur le Web. Certains auteurs ont pris l'habitude d'opposer le « Web *as* corpus » et le « Web *for* Corpus » pour faire référence à ces deux pratiques (de Schryver 2002).

La **première catégorie** rassemble des outils de consultation du Web qui reposent en général sur l'utilisation de moteurs de recherche généraux (tels que Google, Bing, etc.) qu'ils interrogent avant de présenter les résultats à l'utilisateur sous forme de concordance. C'est le cas de WebCorp[5] (Renouf 2003), de WebConcordancer[6] (Fletcher 2007) ou encore de Corpeus[7], adapté pour sa part à la nature agglutinante de la langue basque (Leturia *et al.* 2007). Cette approche se pratique à peu de frais puisque l'essentiel du traitement est réalisé par le moteur de recherche commercial qui est interrogé. Mais cette approche « low-cost » (pour reprendre le terme de Kilgarriff 2007) souffre de nombreuses limitations, dans la mesure où les moteurs ne proposent pas d'indexation linguistique des textes (analyse morphologique ou en parties du discours), leur syntaxe est relativement limitée, les informations statistiques qu'ils fournissent (nombre de *hits*) concerne les pages et non les occurrences des termes recherchés. Ces arguments amènent Kilgarriff (2007) à la conclusion suivante : « Googleology is bad science ». À nos yeux, cette approche reste cependant utilisable et pertinente quand on en connait les limites et que l'on est capable de tenir compte des biais qu'elle peut entrainer (en particulier en ce qui concerne les fréquences observées).

La **deuxième catégorie** de logiciels rassemble des outils dont l'objectif est d'extraire du Web des documents particuliers répondant à des critères précis. Les projets de ce type sont nombreux comme en témoignent les travaux de Baroni et Bernardini (2006), Duclaye *et al.* (2003) Fletcher (2007), etc. Les logiciels de cette nature doivent généralement disposer de modules capables de prendre en charge **trois fonctions** principales :

1. la **recherche** sur le Web de documents potentiellement pertinents. Pour ce faire, des automates (*crawlers*) sont utilisés pour parcourir le Web en suivant les liens collectés de page en page. Ils peuvent être guidés par une liste de sites à consulter en priorité ou par les résultats retournés par un moteur de recherche ;
2. la **sélection** des documents visités en fonction de leur pertinence par rapport à des critères prédéfinis (par exemple, en fonction du thème, du genre discursif, de la complexité du texte, de sa longueur, etc.). Cette sélection doit se faire automatiquement et repose donc sur des algorithmes de classification.
3. le **filtrage** des documents et leur stockage sous une forme utile pour la tâche envisagée. Cette dernière opération nécessite généralement de supprimer le codage HTML ainsi que les paragraphes et éléments de mise en page non pertinents (*boilerplate*)[8].

---

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

[6] http://webascorpus.org/

[7] http://corpeus.elhuyar.org/

[8] Bien que cette tâche de « filtrage HTML » soit considérée comme une tâche technique et quelque peu triviale, elle s'avère souvent relativement compliquée. Il peut en effet être complexe de déterminer quelle partie de la page mérite d'être considérée comme pertinente et quelle partie ne l'est pas. Il est également nécessaire de filtrer l'HTML (qui ne respecte que très rarement les normes de codage). Un concours comme « Cleaneval » qui vise à mettre en compétition des outils permettant de faire ce type de nettoyage automatiquement rend compte de cette difficulté.

Le logiciel GlossaNet (Fairon *et al.* 2008) que nous allons présenter ci-dessous appartient à cette deuxième catégorie. Il permet de constituer des corpus spécialisés[9] par sélection de documents sur le Web. La particularité de ce système réside dans le fait qu'il est conçu pour observer un nombre prédéfini de sources (nous expliquerons ci-dessous qu'il s'agit de flux RSS) plutôt que de naviguer largement sur le Web comme le font les *crawlers* décrits précédemment.

Des corpus dérivés de flux RSS

La méthodologie qui sous-tend le système GlossaNet consiste à créer des corpus à partir de flux RSS (Fairon 2006). Les flux RSS (*Really Simple Syndication*)[10] appartiennent à un mode de communication qui a été conçu pour échanger facilement des informations d'un site Web à un autre. Ils permettent par exemple à un administrateur de site de publier automatiquement sur son site les nouvelles diffusées par un autre éditeur de contenu (les messages d'un blog, d'un forum, des nouvelles d'actualité, des photos, etc.) et disponibles sous forme de flux RSS. Ces échanges de données sont facilités par l'usage du format XML. Les flux contiennent une série d'items décrits par des champs bien précis et facilement identifiables : titre, auteur, résumé, date de publication, etc. Les flux RSS ne sont pas exclusivement utilisés par les rédacteurs de site Web, puisqu'ils peuvent également être utilisés par le grand public au travers de logiciels spécialisés qui collectent ces flux et les présentent à l'utilisateur (on appelle ces logiciels des « agrégateurs »[11] car ils concentrent dans une interface unique des données provenant de sources diverses. Par exemple, l'utilisateur intéressé par le domaine financier peut « s'inscrire » à des flux particuliers à ce domaine et mis à disposition par différents éditeurs (presse quotidienne, presse spécialisée, institutions financières, entreprises, etc.) et lire ainsi les nouvelles publiées sur ces multiples flux dans une interface unique.

[9] Par *corpus spécialisés*, nous entendons « corpora designed for the purpose of creating a sample of specialized language either by collecting texts of similar content (*e.g.* science, medicine, business, philosphy) or similar text-type or genre (*e.g.* research papers, letters, book chapters) or both (e.g. medical research article or science lectures), or even texts from other types of specialized categories, such as newspaper language or academic language. » (Gavioli 2005, p. 7). Par opposition, les corpus généraux ont pour vocation de représenter la langue dans son ensemble (ou du mois une variété géographique de celle-ci).

[10] http://www.rssboard.org/rss-specification

[11] En général, les navigateurs Web sont capables d'afficher ces flux et proposent parfois des plug-ins qui permettent de s'y abonner. Par défaut, le navigateur Firefox permet de s'abonner à des flux RSS et d'afficher l'actualité dans des menus déroulants.

```
<title>WikiLeaks : Berlusconi, un allié imprévisible, autoritaire et affaibli</title>
    <link>http://www.lemonde.fr/documents-wikileaks/[...].html#xtor=RSS-3208</link>
    <description>
        Les télégrammes américains obtenus par WikiLeaks et révélés par "Le Monde" s'interrogent
        sur la capacité à décider d'un Silvio Berlusconi, évoluant entre "scandales sexuels, enquêtes
        judiciaires,problèmes familiaux et financiers", dans un climat politique "délétère".
    </description>
<pubDate>Thu, 02 Dec 2010 21:32:05 GMT</pubDate>
</item>

<title>Les regroupements d'université vont pouvoir délivrer directement des diplômes</title>
    <link>http://www.lemonde.fr/societe/article/2010/12/02/[...]_3224.html#xtor=RSS-3208</link>
    <description>
        L'Assemblée nationale a voté, mercredi, une proposition de loi autorisant les regroupements
        d'université à délivrer en lieu et place de leurs membres qui y sont favorables, une licence,
        un master ou un doctorat.
    </description>
</item>
```

Figure 1 : Exemple simplifié extrait du flux RSS de « Page Une » disponible sur le site du journal *Le Monde*.

Ce format connait un grand succès, car il permet aux éditeurs de contenu de faire circuler leur information au-delà des frontières de leur site Web et d'attirer de cette manière de nouveaux lecteurs. La plupart des sites professionnels proposent ce genre d'outils (la presse, les entreprises, etc.). Les sites de réseaux sociaux (Twitter, Facebook), les forums et les blogs peuvent également être suivis grâce à des flux RSS. Le succès de ce format a même justifié l'apparition de moteurs de recherche spécialisés dans la recherche de flux RSS[12].

Comme on le voit dans la Figure 1, les flux RSS ne contiennent généralement pas des textes complets, mais une brève description et un lien vers la page sur laquelle se trouve le texte. Même si des études spécifiques pourraient être réalisées sur la nature de ces descriptions, c'est surtout le texte d'origine qui nous intéresse pour constituer les corpus. C'est pourquoi le programme que nous utilisons pour créer les corpus (Fairon 2006) observe une série de flux et à chaque fois qu'il trouve de nouveau *items* sur ces flux, il récupère l'URL de référence et télécharge le document HTML, extrait le texte à l'aide d'un filtre puis l'ajoute au corpus.

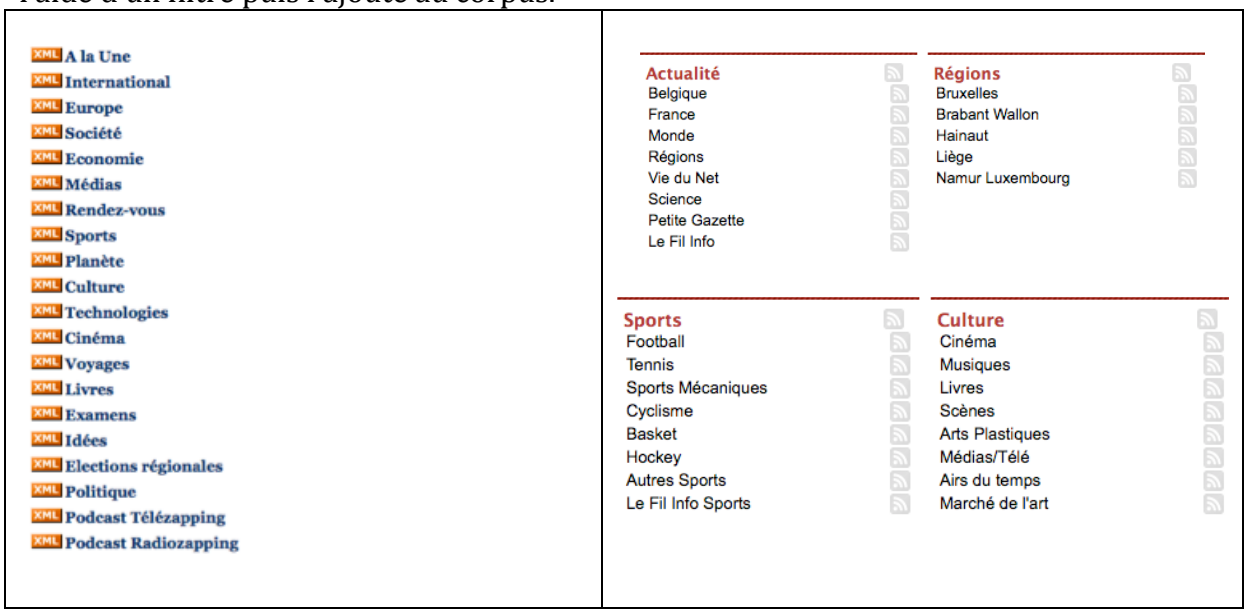


Figure 2 : Extrait des flux RSS proposés par le journal *Le Monde* (gauche) et le journal *Le Soir* (droite)

[12] Cf. par exemple www.jamespot.com ou feedmil.com.

Cette approche basée sur les flux RSS offre plusieurs avantages aux concepteurs de corpus par rapport aux *crawlers* :

a) en terme de **cohérence thématique** ou **discursive** : une pratique très courante consiste à créer des flux thématiquement cohérents (c'est par exemple ce que proposent les éditeurs de nouvelles). Il est donc aisé de constituer des corpus de textes spécialisés dans un domaine particulier (sport, médecine, finance, etc.). Dans certains cas, on peut trouver des flux ayant une cohérence de type discursive : par exemple, on peut soutenir l'idée que les flux rassemblant des éditoriaux de journaux auront un style plutôt argumentatif, etc. ;
b) en terme de **cohérence qualitative** : la constitution de corpus à partir du Web pose souvent des difficultés en raison de la qualité variable des textes disponibles en ligne. Des textes provenant d'un processus éditorial traditionnel côtoient des textes publiés sans contrôle ou reflétant des compétences très variables en matière de maitrise de la langue et il est souvent difficile de faire le tri. Au contraire, au sein d'un flux, le niveau de qualité est plus ou moins constant (sauf dans le cas de flux rassemblant des publications émanant d'un grand nombre d'auteurs, comme les flux provenant de forums, par exemple ;
c) en terme de **méta-informations** : les flux RSS sont des fichiers XML qui respectent un formalisme précis. Il est donc aisé de repérer automatiquement le nom de l'auteur, le titre de l'article et sa date de publication (quand ces données sont présentes).

Un certain nombre de difficultés subsistent cependant, car :

a) la méthode de classification est rarement explicite et varie d'une source à l'autre : est-elle réalisée de manière manuelle ou automatique ; dans le premier cas, est-ce par une seule personne ou un équipe ; les articles sont-ils diffusés dans un seul flux ou peuvent-ils apparaître dans plusieurs flux – dans ce cas un filtrage de doublons sera nécessaire ? Comme on le voit, les questions sont nombreuses ;
b) les catégories utilisées dans les classifications sont différentes pour chaque éditeur de contenu. Il n'y a pas de norme ou de référence standard en la matière, comme le montrent les listes de flux RSS proposées par *Le Monde* et *Le Soir* (cf. figure 2) ;
c) les flux RSS fournissent des liens vers les articles qu'ils référencent, mais ne contiennent pas le texte intégral de l'article. Il est donc nécessai*re de su*ivre *ce lien* et de télécharger l'article à partir de la page Web sur laquelle il se trouve. En outre, certains éditeurs de contenu découpent les articles longs en plusieurs pages Web qui sont accessibles à partir d'un bouton du type « page suivante ». En l'absence de méthode robuste pour traiter ces cas, on risque de collecter des textes incomplets.

L'exploitation linguistique de corpus basé sur des flux RSS

Les RSS ne représentent qu'un mode de diffusion particulier de l'information et leur utilisation se justifie donc avant tout sur le plan pratique. Elle n'entraine donc pas la création d'un nouveau paradigme en linguistique de corpus ni même l'apparition d'un type original de corpus. Cependant, ces flux permettent de répondre à des besoins rencontrés en linguistique et en TAL pour la création de corpus particuliers. Les

quelques exemples d'application qui suivent visent à illustrer cette adéquation sans prétendre à l'exhaustivité :

- constitution de corpus dans un **domaine thématique particulier** à partir de différentes sources (corpus spécialisé représentant le langage médical, de la finance, etc.);
- constitution de **corpus multilingues comparables**[13] à partir des sources publiées simultanément dans plusieurs langues. Par exemple, de nombreuses sources d'information des institutions européennes sont disponibles sous forme de RSS dans plusieurs langues. C'est le cas de Cordis[14], du service de communiqués de presse RAPID[15], etc. ;
- constitutions de **corpus comparables** à partir de sources provenant de **différentes régions** du monde, en vue par exemple de rechercher des d'expressions figées dans des textes français provenant de Belgique, Suisse, Québec, France, etc. Une deuxième exemple est la recherche de Fairon et Singler (2006), qui ont étudié l'usage et les modes d'insertion discursive de l'expression « be like » dans différentes régions anglophones du monde ;
- constitution de **corpus de genres particuliers** : textes argumentatifs (éditoriaux des journaux), textes provenant de blogs ou de forums, textes scientifiques, pédagogiques, etc.

Les corpus dynamiques de GlossaNet

Le logiciel GlossaNet[16] est un service en ligne qui permet aux utilisateurs de (i) **créer un corpus** à partir de différentes sources productives (c'est-à-dire dont le contenu évolue au cours du temps) et (ii) **d'enregistrer une requête** sur ce « corpus dynamique » ainsi défini. La requête appliquée par GlossaNet sur le corpus va produire une concordance qui sera envoyée à l'utilisateur par courrier électronique et qui sera par ailleurs consultable dans l'espace personnel de l'utilisateur sur le site Web de GlossaNet. À intervalles réguliers, le corpus est remis à jour avec les nouveaux documents disponibles dans les sources qui constituent le corpus et la requête est à nouveau appliquée sur le corpus. Les nouveaux résultats sont alors envoyés à l'utilisateur. Dans cette perspective, le corpus n'est pas une base de données fermée et composée d'un nombre fixe de textes. C'est avant tout une source, un flux continu de textes qui se renouvèle au cours du temps. Dans un corpus dynamique[17], la notion de « taille de corpus » est donc directement liée à la notion de temps et il en va de même pour toutes les mesures de fréquence réalisées sur ce corpus.

---

[13] On entend par « corpus comparables », des corpus contenant des textes similaires dans plusieurs langues, sans être pour autant des traductions les uns des autres. Ces textes peuvent être similaires parce qu'ils portent sur le même sujet, parce qu'ils appartiennent au même genre discursif ou parce qu'ils ont été collectés en adoptant un protocole identique.

[14] http://cordis.europa.eu/rss/index.cfm

[15] http://europa.eu/rapid/

[16] Ce service est accessible gratuitement depuis 1999, à l'adresse suivante : http://glossa.fltr.ucl.ac.be

[17] Renouff (2003) a proposé le terme de *Monitor Corpus* pour faire référence à ce type de corpus.

Les utilisateurs de GlossaNet ont donc la possibilité de créer des corpus dynamiques en définissant une liste de flux RSS à observer[18]. Le moteur de GlossaNet va s'abonner à ces flux et récupérer les documents qui y sont listés. À chaque fois que de nouveaux articles seront publiés sur le flux, ils seront récupérés par GlossaNet et ajoutés au corpus.

Introduire une requête

Les corpus récupérés à partir de RSS sont des fichiers textes qui peuvent être traités à l'aide de n'importe quel logiciel de traitement de corpus. Le service en ligne proposé par GlossaNet repose pour sa part sur l'utilisation du logiciel de corpus Open-Source Unitex[19]. Celui-ci est utilisé pour appliquer des ressources linguistiques[20] sur les textes et produire des concordances à partir des requêtes des utilisateurs. Les utilisateurs de GlossaNet peuvent, comme dans Unitex, exprimer leurs requêtes sous forme d'expression régulière[21] ou sous forme de graphe[22] comme suggéré dans l'exemple suivant qui propose deux requêtes équivalentes, capables de repérer des expressions figées du genre *être en rogne, être en nage, être en guerre,* etc. :

<être>en(<A:s>+<E>)<N:s>

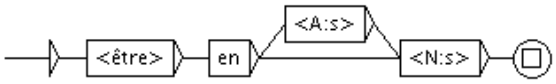


Figure 3 : Exemple de requête utilisable dans GlossaNet

Cette expression permet de recherche n'importe quelle séquence composée d'une forme du verbe <être> suivie de la préposition *en* suivie d'un adjectif singulier (<A:s>) qui est optionnel (+<E>) suivi d'un nom singulier (<N:s>)

Les concordances produites par GlossaNet à partir de ces expressions peuvent être obtenues par email ou consultées dans l'interface en ligne, comme présenté à la figure 4. Un click sur la colonne du milieu permet d'ouvrir la page Web d'origine dans laquelle l'occurrence a été trouvée.

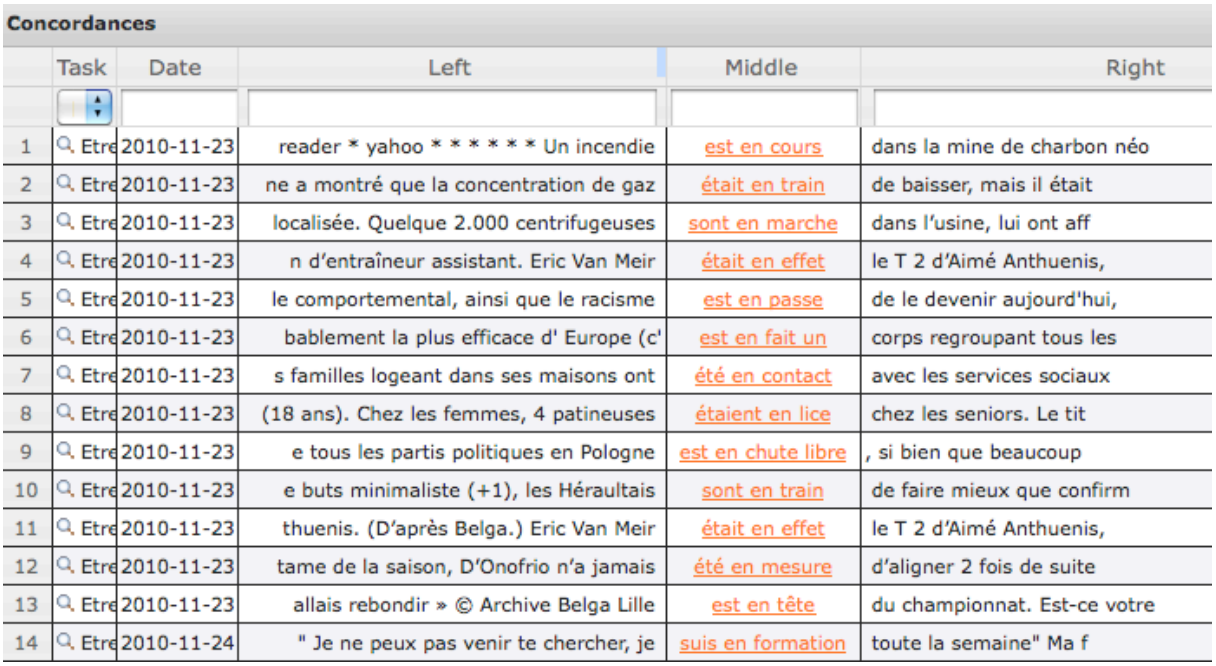
[18] GlossaNet offre une sélection de plus d'un millier de sources RSS (principalement des journaux), mais permet à l'utilisateur d'ajouter n'importe quelle source RSS.

[19] http://www-igm.univ-mlv.fr/~unitex/

[20] Principalement, des dictionnaires électroniques de type Delaf contenant des informations morphologiques et sémantiques ainsi que des grammaires permettant de segmenter le texte en phrase ou de lever certains ambiguïtés.

[21] Le formalisme de ces expressions est décrit dans le manuel en ligne d'Unitex.

[22] Pour créer un graphe, il est nécessaire d'utiliser localement l'éditeur de graphes d'Unitex puis de télécharger le fichier résultant sur le serveur de GlossaNet.



Concordances

	Task	Date	Left	Middle	Right
1	Etre	2010-11-23	reader * yahoo * * * * * * Un incendie	est en cours	dans la mine de charbon néo
2	Etre	2010-11-23	ne a montré que la concentration de gaz	était en train	de baisser, mais il était
3	Etre	2010-11-23	localisée. Quelque 2.000 centrifugeuses	sont en marche	dans l'usine, lui ont aff
4	Etre	2010-11-23	n d'entraîneur assistant. Eric Van Meir	était en effet	le T 2 d'Aimé Anthuenis,
5	Etre	2010-11-23	le comportemental, ainsi que le racisme	est en passe	de le devenir aujourd'hui,
6	Etre	2010-11-23	bablement la plus efficace d' Europe (c'	est en fait un	corps regroupant tous les
7	Etre	2010-11-23	s familles logeant dans ses maisons ont	été en contact	avec les services sociaux
8	Etre	2010-11-23	(18 ans). Chez les femmes, 4 patineuses	étaient en lice	chez les seniors. Le tit
9	Etre	2010-11-23	e tous les partis politiques en Pologne	est en chute libre	, si bien que beaucoup
10	Etre	2010-11-23	e buts minimaliste (+1), les Héraultais	sont en train	de faire mieux que confirm
11	Etre	2010-11-23	thuenis. (D'après Belga.) Eric Van Meir	était en effet	le T 2 d'Aimé Anthuenis,
12	Etre	2010-11-23	tame de la saison, D'Onofrio n'a jamais	été en mesure	d'aligner 2 fois de suite
13	Etre	2010-11-23	allais rebondir » © Archive Belga Lille	est en tête	du championnat. Est-ce votre
14	Etre	2010-11-24	" Je ne peux pas venir te chercher, je	suis en formation	toute la semaine" Ma f

Figure 4. Exemple de concordance présentée dans l'environnement de GlossaNet

En guise de conclusion : quel avenir pour les corpus issus du web ?

La croissance phénoménale du Web ainsi que sa forte diversification ont fait de celui-ci une source de données quasi inépuisable pour la linguistique de corpus, comme nous l'avons illustré à travers l'exemple du logiciel GlossaNet. Il est facile de prédire que l'avenir des corpus issus du Web sera naturellement lié à l'évolution du Web lui-même et cette évolution va incontestablement dans le sens d'une plus grande structuration des documents. On ne parle pas ici de la structure linguistique, mais bien du codage informatique. En effet, les opérations de constitution de corpus à partir du Web nécessitent de pouvoir analyser la structure des documents HTML pour bien comprendre de quoi ils se composent et identifier les parties pertinentes à conserver dans le corpus. Le Web 2.0 a favorisé l'utilisation des outils de gestion et de publication de contenu qui produisent des documents à la structure plus régulière et (parfois) plus explicite que les publications individuelles. Néanmoins, il n'y a pas de standardisation et chaque outil ajoute des métadonnées différentes.

Le projet d'un Web sémantique a pour vocation d'offrir un encodage des documents permettant aux programmes informatiques de traiter le contenu des documents publiés sur le Web. L'information rendue plus facilement accessible permettra aussi de faire évoluer la constitution de corpus. En effet, à l'heure actuelle, la possibilité d'identifier automatiquement des choses aussi simples que l'auteur ou la date de publication d'un texte reste souvent difficile dans la configuration actuelle du Web. Une autre avancée de taille que l'on peut légitimement espérer est qu'il sera un jour possible, grâce à l'étiquetage sémantique, de distinguer les différents sens d'un mot et de sélectionner avec plus de précisions les documents collectés pour constituer des corpus spécialisés grâce à un étiquetage sémantique généralisé et standardisé

Remerciement

GlossaNet est un projet développé par l'équipe du Cental. C'est un plaisir de remercier le Cental et plus particulièrement Kévin Macé, Hubert Naets et Bernadette Dehottay. Par ailleurs, le projet a bénéficié d'un financement de la Région Wallonne dans le cadre du projet First Actispel. Cet article a été partiellement rédigé durant un séjour de recherche au *Stanford Center for Biomedical Informatics Research* (BMIR). Merci enfin à Stéphanie Audrit pour sa lecture et ses remarques.

Références

Baroni Marco, Silvia Bernardini (2006). « WaCky!: working papers on the web as corpus ». Traduzione, lingue e culture. Studi interdisciplinari su traduzione, lingue e culture 6, p. 224.

Baroni Mario, Silvia Bernardini (éd.) (2006). *Wacky! Working papers on theWeb as Corpus*. GEDIT, Bologna.

Biber Douglas, Conrad Susan, Reppen Randi (1998). *Corpus Linguistics – Investigating Language Structure and Use*, Cambridge, Cambridge University Press.

Brunello Marco (2009). « The création of free linguistic corpora from the Web ». In *Proceedings of the Fifth Web as Corpus Workshop (WAC5),* Donastia-San Sebastian. Basque Country, Spain.

De Schryver Gilles-Maurice (2002). « Web for/as Corpus: A Perspective for the African Languages ». *Nordic Journal of African Studies* 11 (2), pp. 266-282.

Dister Anne (2007). De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel. Université catholique de Louvain. Thèse non publiée.

Duclaye Florence, François Yvon, Olivier Collin (2003). « Unsupervised incremental acquisition of a thematic corpus from the web ». In *Proceedings of Natural Language Processing and Knowledge Engineering. IEEE*. (DOI 10.1109/NLPKE.2003.1276006), pp. 752 -757.

Fairon Cédrick (2006). « Corporator: A tool for creating rss- based specialized corpora ». In *Proceedings of the Workshop Web as corpus*, Trento. EACL.

Fairon Cédrick, John V. Singler (2006). « I'am like, 'Hey, it works': Using GlossaNet to find attestations of the quotative (be) like in English-language newspapers ». In A. Renouf and A. Kehoe (eds). *The Changing Face of Corpus Linguistics*, number 55, pages 325–337. Language and computers: Studies in Practical Linguistics, Amsterdam - New York.

Fletcher William (2007). « Implementing a BNCCompareable Web Corpus ». In C. Fairon, H. Naets, A. Kilgariff, G.-M. de Schryver (éds), *Building and Exploring Web Corpora*, volume 4, Louvain-la-Neuve. Cahiers du Cental.

Gavioli Laura (2005). « Exploring corpora for ESP Learning ». Studies in Corpus Linguistics 21. John Benjamins. Amsterdam/Philadelphia. 176 p.

Grefenstette Gregory (1999). The WWWas a resource for example-based MT tasks. In ASLIB Translating and the Computer Conference, London.

Kilgarriff Adam, Grefenstette, Gregory (2003). « Web as Corpus: Introduction to the Special Issue ». Computational Linguistics 29 (3). pp. 333–347.

Leturia I., Gurrutxaga A., Alegria I., Ezeiza (2007). « CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque ». In C. Fairon, H. Naets, A. Kilgariff, G.-M. de Schryver (éds), *Building and Exploring Web Corpora. Proceedings of the 3rd Web as Corpus Workshop*. Cahiers du Cental 4, pp. 69-81.

McEnery Tony, Wilson Andrew (2001). Corpus Linguistics. An Introduction. Second Edition. Edinburgh. Edinburgh University Press.

Michel Jean-Baptiste *et al.* (2010). Quantitative Analysis of Culture Using Millions of Digitalized Books. In Science.

Renouf Antoinette (2003). « WebCorp: providing a renewable energy source for corpus linguistics ». In S. Granger and S. Petch-Tyson (eds), *Extending the scope of corpus-based research: new applications, new challenges*, Rodopi, Amsterdam, pp. 39-58.

Resnik Philip (1999). « Mining the Web for Bilingual Text ». In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Rastier François (2004). « Enjeux épistémologiques de la linguistique de corpus ». In Texto ! [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.

Scannel Kevin (2007). « The Crúbadán Project: Corpus building for under-resourced languages ». In C. Fairon, H. Naets, A. Kilgarriff, G-M de Schryver (éds.), Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve, Belgium, September 2007. Cahiers du Cental 4 (2007), pp. 5-15.