EXTRACTION INFORMATIQUE DE DONNÉES SUR LE WEB

Fabrice ISAAC, Thierry HAMON, Christophe FOUQUERÉ
Université Paris 13
Lorne BOUCHARD, Louisette EMIRKANIAN[1]
Université du Québec à Montréal

Dans le cadre d'une étude linguistique de l'usage des prépositions sur le Web, nous avons conçu un outil de constitution de corpus. Nous présentons ici quelques réflexions sur la spécificité de corpus composé de données extraites du Web, les problèmes techniques rencontrés ainsi que les résultats préliminaires quantitatifs et qualitatifs obtenus.

INTRODUCTION

C'est un fait entendu, Internet est une source presque infinie de ressources, qu'elles soient textuelles, graphiques ou sonores. Qui plus est, ces ressources sont, dans le cadre d'une consultation, libres de droit. Cette source, alliée à des supports de stockage (cédéroms ou disques durs) importants et de faibles coûts, permet la constitution rapide de corpus utilisables par un grand nombre de personnes.

Mais tout n'est pas si simple. La constitution de corpus à partir du Web soulève un certain nombre d'interrogations : le Web est-il vraiment une bonne source de données textuelles, en d'autres termes, peut-elle servir de base à une étude linguistique, sociologique ou à de la veille technologique? Comment récupérer des données, quels sont les outils disponibles, comment traiter ces données, les données récupérées sont-elles – directement – exploitables?

De notre point de vue, un corpus constitué de données extraites du Web est construit par rapport à des contraintes syntaxico-sématiques, *i.e.* en fonction de la présence et de l'ordre d'un certain nombre de mots dans une page. Nous avons donc réalisé un outil capable (i) de donner, à partir d'une requête booléenne, une liste d'URL[2], (ii) de récupérer et normaliser les pages

[1] http://www.emirkanian.louisette@uqam.ca/

[2] Uniform Resource Locator, *i.e.* une adresse unique sur le réseau Internet permettant de retrouver une page.

correspondantes (iii) d'extraire de chacune de ces pages la phrase correspondant à la requête.

La section 2 présente notre point de vue sur la notion de corpus et plus précisément de corpus créés à partir de données du Web. Puis en nous plaçant dans le cadre du projet de reformulation de requêtes, nous présenterons un outil de constitution de corpus et les résultats tant quantitatifs (section 3) que qualitatifs (section 4) lors de son utilisation. Un bilan et des perspectives concluront cet article.

Cet outil a été utilisé dans le cadre d'un projet de coopération France-Québec[3], dont l'objectif est d'évaluer l'apport des connaissances linguistiques à l'amélioration de la précision dans les requêtes sur le Web, que nous décrivons brièvement ci-dessous.

Recherche d'information

Le résultat des recherches sur le Web, à partir d'un moteur de recherche, n'est en général guère satisfaisant. En effet, obtenir un résultat pertinent demande bien souvent une bonne connaissance du domaine de recherche, de l'intuition sur le fonctionnement – caché – du moteur et beaucoup de patience.

Les requêtes simples[4] permettent un rappel[5] important mais une précision faible. Une requête de la forme « mot_1 mot_2 ... mot_n » correspond à la requête booléenne «mot_1 OU mot_2 OU ... OU mot_n ». Cela signifie que le moteur va sélectionner toutes les pages où l'un des mots est présent. Si la requête est plus contrainte (par l'utilisation du connecteur ET), alors la précision est meilleure; de manière pratique, le nombre de pages devient raisonnable, mais le rappel, lui, diminue. On améliore donc la précision avec une requête plus contrainte (Buckland et Gey, 1994). À cet effet, l'utilisation du connecteur NEAR (proche) impose non seulement l'occurrence de deux mots, comme le ET logique, mais aussi leur proximité dans une fenêtre graphique.

---

3 Ce projet de recherche est financé par le ministère des Relations internationales du Québec (MRI) et le ministère des Affaires étrangères de France (MAE).

4 Nous utilisons comme référence le moteur de recherche `Altavista`; tous les moteurs ayant un fonctionnement plus ou moins identique.

5 Le taux de rappel mesure la proportion de textes pertinents sélectionnés par rapport à l'ensemble des documents pertinents de la base. Le taux de précision mesure la proportion de documents sélectionnés jugés pertinents par rapport à l'ensemble des documents sélectionnés par le système.

Description du projet

L'objectif du projet est d'améliorer la précision et le rappel d'une requête dans un cadre linguistique spécifique. Nous nous proposons d'étudier le comportement des noms prédicatifs marquant la localisation ou le déplacement.

À partir d'une classification sémantique des verbes et des noms prédicatifs, notre objectif est de prédire le type de compléments locatifs qu'ils sélectionnent, et dans le cas des syntagmes prépositionnels, leur sens et le type de prépositions qui les introduit. Nous pouvons alors, à partir d'une requête donnée, générer des requêtes équivalentes avec les prépositions adéquates. À partir de la requête `pérégrination au Tibet` nous générons `voyage à Lhassa`, `départ pour le Népal`, `traversée du Tibet`, ... L'ajout des prépositions dans les requêtes constitue une contrainte très sévère[6]. Cependant, une augmentation de la précision fait baisser le rappel[7]. C'est pourquoi, à partir d'une requête, nous générons d'autres requêtes par glissement synonymique des noms en ajoutant d'autres contraintes linguistiques reliées à ceux-ci : les prépositions.

Le Web comme source de corpus

Évaluer la grammaticalité de la combinaison des noms et des prépositions est un exercice difficile. Un corpus de grande taille s'imposait afin de valider nos intuitions. Cependant, les corpus dont nous disposions (*Le Monde, Le Monde Diplomatique*) ne pouvaient pas être les seules ressources dans le cadre de notre étude qui cible des occurrences (rares) de constructions particulières. Notre besoin est double. D'une part, l'étude linguistique doit s'appuyer sur un corpus pertinent et, d'autre part, nos résultats, en ce qui concerne la reformulation de requête, doivent être validés. C'est pourquoi nous avons considéré le Web non seulement comme un objet d'étude mais aussi comme une source potentielle de corpus, charge à nous de réaliser un outil de constitution de ces corpus. L'approche choisie est basée sur l'utilisation des outils existants permettant – théoriquement – d'accéder en tout point du Web : les moteurs de recherche.

6 Notre projet permettra, entre autres, de déterminer si le fait d'introduire une préposition dans la requête permet d'améliorer la précision.

7 Néanmoins, le rappel sur le Web étant en général trop élevé, on ne doit pas craindre de le réduire pour augmenter la précision.

CONSTITUTION D'UN CORPUS À PARTIR DU WEB

Quelles spécificités linguistiques?

Les travaux sur corpus s'intéressent essentiellement aux données écrites ou orales dans leur version électronique. Cependant, les textes produits directement sur support électronique, tels que les pages personnelles accessibles sur l'Internet ou le courrier électronique, constituent également un type de matériel linguistique à part entière (Sinclair et Ball, 1996). Bien que leur statut ne soit pas clairement et unanimement identifié, les textes électroniques possèdent certaines spécificités. Ils se différencient notamment de l'écrit et de l'oral par leurs caractéristiques internes mais aussi par leur rôle social.

Amitay (1999) effectue une étude sur l'anglais des 35 mots les plus fréquents dans le *British National Corpus* (BNC) et dans un corpus de pages web personnelles (*HomeCorpus*), et plus particulièrement des mots présents uniquement dans l'un des deux corpus. L'auteure observe plusieurs traits caractéristiques du corpus de pages internet (i) l'absence de la troisième personne et l'emploi très fréquent de la première et de la deuxième personne (ii) l'absence du verbe *be* au passé et l'utilisation importante du temps présent (iii) l'absence des connecteurs *but* et *which* entre deux phrases. Elle conclut ainsi que la langue utilisée dans les pages personnelles s'apparente à une conversation présentant des faits entre l'auteur et le lecteur. Nous avons, lors de notre expérience, fait le même constat. Par rapport au corpus *Le Monde*, nous recensons 2 à 3 fois moins de prépositions associées à chaque prédicatif dans le corpus constitué à partir du Web. Bien qu'il soit difficile de quantifier un phénomène rare, il semble que le Web soit plus pauvre. Le style du Web est donc entre l'écrit, puisque ce n'est pas de la transcription d'oral, et l'oral, puisque plus pauvre que l'écrit. De plus, l'étude du texte des ancres[8] montre que les termes ne sont pas définis explicitement dans les pages : le contexte et les définitions sont supposés connus. Grâce aux ancres, le rédacteur laisse au lecteur la liberté d'en prendre connaissance.

Cependant, lors de l'analyse des résultats de notre expérience et du parcours de sites Web, nous avons constaté que le contenu des pages accessibles sur l'Internet peut aussi être des transcriptions de textes écrits au

[8] Une ancre est composée de deux balises, c'est-à-dire `<a href=...> texte_ancre</a>`. Amitay (Amitay, 1999) s'intéresse au texte compris entre celles-ci, ici `texte_ancre`.

format électronique (articles de presse, de dépêches, de textes littéraires, etc.) et parfois de transcriptions de documents oraux (exposés ou discours).

Quel corpus construire?

La variété des phénomènes langagiers existant dans les textes accessibles sur l'Internet nécessite donc une étude linguistique particulière. La conception d'un outil produisant un corpus à partir de ce type de données nous amène donc à considérer les critères de constitution d'une telle ressource textuelle.

Sinclair (1996) définit un corpus comme « une collection de données langagières sélectionnées et organisées suivant des critères linguistiques explicites utilisés comme échantillons du langage[9] ». L'auteur note que cette définition offre l'avantage de permettre la constitution d'un corpus non pas à partir de textes mais à partir de fragments de textes. Le corpus se différencie de la collection de textes par l'utilisation de critères linguistiques dans la sélection. Le Web n'est donc pas un corpus. La définition proposée par Sinclair (1996) est suffisamment large pour être interprétée suivant les besoins. Cependant, il est nécessaire de considérer soigneusement plusieurs critères externes afin de disposer de données exploitables. Il s'agit également d'assurer leur représentativité ainsi que leur réutilisabilité vis-à-vis de la tâche finale.

La taille est le premier point à prendre en compte. Péry-Woodley (1995) constate que, dans de nombreux cas dans la littérature, le volume des données doit être important sans qu'il y ait besoin de réflexion *a priori* sur la typologie du corpus obtenu. Il s'agit surtout ici de corpus de textes. À cette approche orientée vers le gigantisme, s'oppose une construction plus raisonnée du corpus (Habert *et al.*, 1997). L'objectif est d'y assurer un équilibre des phénomènes linguistiques. On parle alors de corpus d'échantillons. Nous considérons que le corpus recueilli lors de notre expérience est un corpus d'échantillons. Dans notre cas, la taille du corpus obtenu dépend surtout du nombre de requêtes réalisées et du moteur de recherche employé.

La qualité du corpus est en partie conditionnée par une documentation indiquant son origine, c'est-à-dire les documents utilisés, les objectifs et les responsables de sa construction, mais aussi son histoire au travers des différentes révisions qui ont pu avoir lieu (Habert *et al.*, 1997). La connaissance de l'origine des données permet de s'assurer de leur authenticité

9 « A corpus is a collection of piece of language that are selected and ordered according to explicit linguisitic criteria in order to be used as a sample of the language. »

(Sinclair, 1996) et de permettre une réutilisation ultérieure (Pery-Woodley, 1995). En ce qui concerne la constitution d'un corpus à partir du Web, nous avons choisi de documenter chaque échantillon en indiquant la requête utilisée pour récupérer la page, la date à laquelle a été effectuée l'interrogation du moteur de recherche, ainsi que l'URL de la page.

Enfin, le format des données regroupées est également un critère important. Pour Sinclair (1996), le texte brut composé de caractères ASCII est l'approche la plus simple. Cependant, Habert *et al.* (1997) constatent que la représentation physique de données n'est pas suffisante et n'assure pas une réutilisation et un échange des corpus. La normalisation des corpus passe par une représentation logique indiquant la structure des documents. À ce problème, la TEI (Ide, 1995) apporte une réponse en proposant des recommandations basées sur le langage à balises SGML (Van Herwijnen, 1995). Bien que la DTD HTML soit basée sur SGML, ce langage a été essentiellement conçu pour la présentation des pages Web. Cependant, il n'est pas toujours utilisé de façon standard. Ainsi, il y a mélange entre les balises structurelles et celles de mise en forme (le balises de niveaux de titres – <h$_n$> – sont souvent utilisées comme balises de présentation). Il ne peut donc convenir comme format de normalisation. À partir de ce constat, nous avons choisi de représenter la structure des données puis des échantillons du corpus à l'aide de la DTD CES (Corpus Encoding Standard) proposée par la TEI.

Par ailleurs, Habert *et al.* (1997) soulèvent les problèmes légaux liés à l'utilisation de données textuelles sans toutefois évoquer l'exploitation des données de l'Internet. Les corpus constitués à partir du Web sont particuliers. Contrairement aux autres types de corpus, les problèmes juridiques semblent moins se poser lors de la constitution puisque les informations d'origine sont publiques. Cependant, la diffusion est plus délicate et nous semble même impossible.

Comment exploiter les info rmations du Web?

L'utilisation du Web comme base pour la constitution de ressources textuelles est très récente. Les tentatives d'exploitation de ce type de données sont peu nombreuses.

Dans une perspective de traduction automatique, Resnik (1998a) étudie la possibilité d'utiliser les sites Internet proposant les informations en plusieurs langues pour constituer des corpus parallèles bilingues. Le moteur de recherche d'`Altavista` et des heuristiques lui permettent, dans un premier temps, de retrouver les sites Web proposant des pages dans plusieurs

langues. Ces sites sont ensuite explorés afin de retrouver les pages existant dans plusieurs langues. Cowie *et al*. (1998) proposent un outil d'exploration construisant des corpus multilingues en fonction des codes de langages et des adresses de sites spécifiés par l'utilisateur.

Dans le cadre de la génération automatique de résumés, Radev et McKeown (1997) exploitent les informations sur les forums de discussion et les sites proposant des dépêches de presse afin d'extraire des définitions d'entités nommées. L'objectif est de compléter et d'ajuster la base de connaissances constituée à partir de corpus de dépêches comme le corpus REUTER.

Notre approche diffère des précédentes dans le mode de constitution du corpus ainsi que dans sa perspective d'utilisation.

NOTRE EXPÉRIENCE

Description de l'outil

Un corpus regroupe donc un ensemble de données textuelles dont la sélection se fait en fonction de critères syntaxico-sémantiques.

Il existe deux moyens de constituer un corpus à partir du Web. Soit l'ensemble des données qui vont constituer le corpus sont regroupées sur un ensemble de sites connus (Resnik, 1998a), et dans ce cas on utilise un aspirateur à Web[10], c'est-à-dire un outil qui permet de récupérer un ensemble de pages à partir d'une adresse. Cette méthode est très rapide mais présuppose que les sites aspirés constituent déjà des corpus, ce qui est rarement le cas. La deuxième méthode consiste à utiliser dans un premier temps un moteur de recherche, dont le langage de requête permet d'exprimer des contraintes sur les mots, pour effectuer la sélection des données. On récupère ainsi un certain nombre d'adresses à partir d'une ou de plusieurs requêtes (dont la complexité dépend du moteur). Puis, à partir de ces adresses, il s'agit de récupérer manuellement ou automatiquement les pages correspondantes.

[10] Par exemple *WebMirror* (http://www.maccasoft.com/webmirror/) ou *HTTrack* (http://httrack.free.fr).

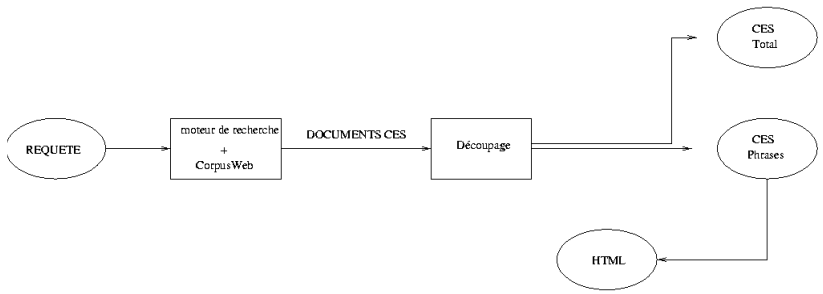

Figure 1 Vue d'ensemble du système

C'est ce principe qu'utilise notre outil. La figure 1 présente une vue d'ensemble du processus de constitution d'un corpus.À partir d'une ou de plusieurs requêtes, exprimées dans un langage propre à un moteur de recherche, un ensemble de pages est récupéré puis transcodé au format CES. Il est possible, soit de conserver le corpus dans cet état pour, par exemple, constituer un corpus thématique, soit d'extraire les phrases contenant les mots de la requête. À cet effet, une exportation vers le format HTML a été réalisée afin de simplifier la consultation. La figure 2 présente un exemple de `sejour*+NEAR+au+NEAR+tibet`, *i.e.* les phrases où les trois mots `sejour`, `au`, `tibet` sont proches. Nous proposons, via un hyperlien, un retour vers la page originelle, pour permettre de replacer la phrase dans son contexte.

Ce point soulève un problème intrinsèque au support Web : l'aspect éphémère des données. En effet, sauf à sauvegarder la totalité des pages concernées, il n'est pas possible de certifier la récupération du contexte originel à partir d'un corpus d'échantillons. L'origine ici n'est donc qu'une information indicative non valide dans le temps.

sejour*+NEAR+au+NEAR+tibet

un séjour au tibet en 1996 et première sensibilisation à la souffrance du peuple tibétain, une

formation en médecine humanitaire et la recherche d'un terrain de stage, enfin la multiplicité
de contacts en France ont orienté mes pas vers ces enfants-là.
Origine

ce recueil d'histoires et de témoignages est le résultat de plusieurs séjours au Tibet ainsi
que dans la communauté tibétaine en exil en Inde et au Népal.
Origine

Figure 2 Consultation au format HTML

L'outil est en fait constitué d'un ensemble modulaire de sous programmes écrits en PERL.

La figure 3 en présente les différents composants.

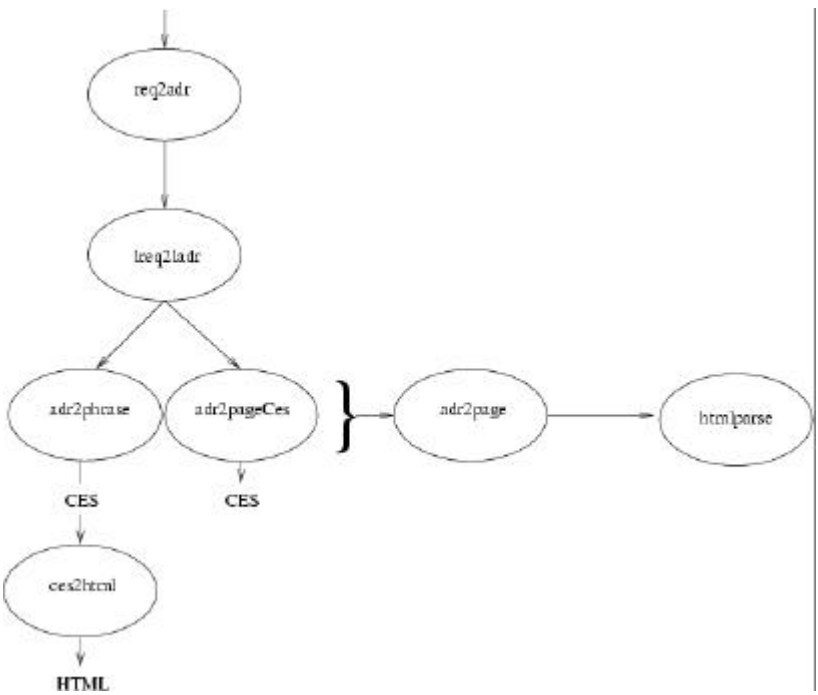

Figure 3 Fonctionnement de *CorpusWeb*

Les requêtes ont le même format que celui utilisé dans l'appel du moteur de recherche :
`www.altavista.com/cgi-bin/query?hl=on&q=`**`sejour+NEAR+au+NEAR+nepal`**`&search.x=51&..`

et correspondent à un moteur de recherche particulier (ici `Altavista`). On utilise ici le connecteur NEAR qui exige une proximité entre les mots[11]. Ce connecteur est donc plus contraignant que le connecteur AND qui ne demande que de trouver les différents mots de la requête dans la page.

Le composant `req2URL` interroge le moteur de recherche et récupère toutes les adresses des pages correspondant aux requêtes. À partir d'une requête, `Altavista` retourne au plus 1 000 URL. Les adresses apparaissant en double sont éliminées.

Le composant `listeURL` permet de factoriser la liste des URL de pages obtenue. Cette étape permet de réduire considérablement le temps de récupération des pages. Puisqu'un URL de pages peut être associé à plusieurs requêtes, il y a de fortes chances que l'URL d'une page soit récupéré plusieurs fois, surtout dans le cas où les requêtes sont proches. Par exemple, la phrase `Mon voyage et mon séjour au NEPAL` sera reconnue par deux requêtes, l'une avec `voyage` et l'autre avec `sejour`. Il ne reste plus, à partir de ces données, qu'à récupérer les phrases qui répondent à la requête (`URL2phrase`) pour constituer un corpus d'échantillons, ou les pages complètes (`URL2Ces`).

Le composant `URL2page` récupère une page à partir de son URL. Il utilise le module `htmlparse` pour transcoder la page d'un format HTML vers un format CES.

Le composant `htmlparse` effectue une analyse syntaxique de la page au format HTML et, lorsque l'analyse a pu être effectuée correctement, réalise un transcodage vers un format CES niveau 1 (*i.e.* balises structurelles élémentaires) avec la gestion des listes et des tableaux. Il est à noter que la permissivité des différents navigateurs encourage un non-respect de la norme HTML dans beaucoup de cas[12].

Les deux composants `URL2phrase` et `URL2Ces` permettent de mettre en forme le retour de `URL2page` en y ajoutant un en-tête documentant le corpus. Dans le cas de `URL2phrase`, on sélectionne les phrases où effectivement les mots de la requête apparaissent dans le bon ordre. En effet, les moteurs de recherche ne tiennent compte ni des ponctuations, ni de l'ordre des mots de la requête.

---

11 Altavista définit ce connecteur sur une fenêtre de longueur égale à dix mots.

12 La mise au point de cet outil a consisté à relâcher les contraintes d'analyse. Lorsque le Web aura migré du HTML à XML, ce travail sera à la fois simplifié et plus productif. En effet, XML est un autre langage à balises dérivé de SGML qui sépare le contenu du style de présentation de celui-ci.

Le composant `ces2html` permet le transcodage du résultat de `URL2phrase` vers un format HTML pour une consultation visuelle du corpus.

Évaluation quantitative

Nous présentons ici les résultats quantitatifs de l'utilisation de notre outil par rapport aux objectifs du projet.

Requêtes

Notre étude a porté à l'origine sur la requête pérégrination au Tibet. À partir des différents composants, par glissement synonymique, sur pérégrination (séjour, trek, voyage, etc.) et sur Tibet (Lhassa, Népal, toit du monde, etc.) et par variation des prépositions, on obtient trois listes de mots. En combinant toutes les possibilités on obtient 2590 requêtes. La génération se faisant ainsi de manière brutale, beaucoup de requêtes ne sont pas grammaticales et ont donc peu de chances d'être attestées, par exemple pérégrination sous Lhassa. Cependant notre objectif ici n'est pas de générer des bonnes requêtes mais de constituer un corpus d'échantillons.

Données recueillies

En réponse à ces 2590 requêtes, on obtient une liste de 15 673 URL de pages. À titre indicatif, sur ces 2590 requêtes, 747 retournent au moins un URL de pages. La factorisation des URL permet de réduire leur nombre à 3 093, ce qui diminue le temps de téléchargement d'un facteur 5. Les composants URL2page et URL2phrase permettent de filtrer les pages en éliminant celles qui ne sont pas analysables syntaxiquement et celles où la position et/ou l'ordre des différents mots ne correspond pas à notre requête (qui est plus précise que celle du moteur de recherche). Nous obtenons finalement 248 phrases pour un total de 10 722 mots.

Temps

Il est très difficile d'évaluer le temps de constitution du corpus. En effet, encore une fois à cause du support Web, beaucoup de facteurs entrent en ligne de compte : l'heure, le lieu, la configuration matérielle et logicielle, la capacité des câbles. Nous donnons néanmoins des valeurs à titre indicatif. Le processus de constitution de la liste des adresses de pages a pris environ 6

heures soit 8 secondes par requête. Le temps de récupération et de traitement des pages, i.e. le transcodage au format CES, est négligeable devant le temps pris par le téléchargement de la page. Le temps moyen est d'environ 20 secondes par page soit environ 17 heures.

ANALYSE DES RÉSULTATS

Problèmes rencontrés lors de la constitution du corpus

En préliminaire à une analyse linguistique des résultats, nous avons observé plusieurs problèmes techniques lors des différentes étapes de la constitution du corpus.

Nous avons tout d'abord éprouvé quelques difficultés dans l'utilisation des requêtes du moteur de recherche `Altavista`. Alors que l'opérateur NEAR devrait retourner les pages contenant les mots spécifiés s'ils sont distants de moins de 10 mots, nous avons constaté que la distance entre les mots pouvait être plus grande[13].

L'indexation des pages composées de plusieurs cadres (frame) a posé également des problèmes lors de la constitution du corpus. L'adresse de la page contenant les mots pertinents pour la requête n'est pas fournie par le moteur de recherche, c'est la page contenant les références aux cadres et notamment l'adresse de cette page qui apparaît dans le résultat de la requête. La page principale peut contenir les mots de la requête uniquement dans le titre non visible par l'utilisateur[14]. Par exemple, pour la requête `trek AND au AND nepal`, une des pages proposées par `Altavista` a pour URL : `http://www.asie.com/asia/trek.htm`. Or la page pertinente et contenant les données textuelles qui nous intéressent est un cadre de cette page, dont l'adresse est `http://www.asie.com/asia/trek.htm`. Cette adresse n'est pas indiquée par le moteur de recherche, seule la page principale apparaît dans la réponse à la requête. Nous n'avons pas considéré ces pages dans cette première expérience. Leur prise en compte nécessite une analyse spécifique.

---

13 En effet, le fonctionnement interne précis des moteurs de recherche n'est pas toujours bien documenté puisqu'il s'agit de secrets industriels.

14 La balise `<TITLE>` permet de spécifier un titre pouvant être utilisé par les navigateurs.

L'outil développé est dépendant de l'indexation des pages Web par le moteur de recherche. Certaines pages indexées n'existent plus lors de l'interrogation du moteur ou ne font plus partie des 1000 premières réponses pertinentes. Il existe un décalage entre le contenu des pages et celui de l'index. Nous avons notamment constaté que d'autres pages ont pu subir des modifications depuis l'indexation et ne contiennent plus les mots recherchés.

Enfin, nous avons dû rejeter les pages utilisant des balises ou des règles HTML trop ambiguës pour être analysées correctement. En effet, la correspondance entre les balises et les règles HTML propres à certains navigateurs et la grammaire SGML de CES s'avère délicate dans ce cas et entraîne des temps de calculs trop importants lors de l'analyse.

Nous avons noté que certaines combinaisons pourtant possibles n'apparaissent pas : aventure à travers le Tibet ou encore marche vers le Tibet. Il est vrai que marche vers le Tibet ne correspondrait pas à une extension de la requête initiale pérégrination au Tibet.

Dans la plupart des pages pertinentes, la distance entre les différents éléments de la requête est très inférieure à 10. Nous constatons ainsi que la distance spécifiée dans l'opérateur NEAR (10 mots) est trop grande pour l'objectif. Par ailleurs, nous obtenons de bonnes pages avec une mauvaise préposition. Par exemple, marche ... de ... Katmandou ou encore départ ... dans ... Tibet.

Nous nous sommes d'abord penchés sur le rôle de l'astérisque. Dans le cas de voyag*, elle permet de récupérer non seulement le nom (au singulier et au pluriel) mais aussi le verbe conjugué ainsi que les dérivés voyageur, -euse. Il en est de même pour march*. Nous obtenons également des pages avec le mot marché; bizarrement, ces quelques pages correspondent tout à fait au type d'informations qu'un utilisateur désirant se rendre au Tibet pourrait rechercher. L'utilisation de l'astérisque, en revanche, n'est pas pertinente avec part* dans la mesure où nous récupérons beaucoup trop de choses inutiles : nous obtenons très peu d'occurrences de partir (conjugué ou à l'infinitif) ou de en partance, par exemple. Au contraire, trek, déplac ou séjour ont une précision proche de 100 %. Cependant, le nombre de phrases n'est pas très élevé.

Étude linguistique superficielle du corpus

Pour évaluer le contenu des résultats, nous nous sommes mis à la place d'un utilisateur qui cherche des renseignements pour se rendre dans cette région du monde, c'est-à-dire des informations telles que celles que pourrait lui donner une agence de voyage. Sur les 248 énoncés obtenus, une vingtaine

contiennent précisément ce type d'informations. Les autres résultats peuvent être classés en 5 groupes : politique, récits de voyage, informations biographiques, religion et histoire et compte rendu.

CONCLUSION ET PERSPECTIVES

Cette première expérience nous a permis de mettre au point une méthodologie de constitution de corpus à partir d'informations accessibles sur le Web. Cependant, l'analyse du corpus montre que les réponses doivent être filtrées plus précisément par des heuristiques prenant en compte certaines erreurs du moteur de recherche. Ce filtrage à l'aide de règles plus complexes doit permettre de réduire le nombre de phrases non pertinentes pour le corpus et la tâche. Nous projetons l'utilisation d'une distance très réduite entre les mots recherchés.

Une autre extension consiste également à analyser les pages contenant des cadres (frames). Il serait ainsi possible de récupérer les pages contenant du texte mais non renvoyées par le moteur de recherche.

De plus, bien qu'il soit possible d'exploiter d'autres moteurs de recherche, il nous semble plus important, d'un point de vue linguistique, de s'intéresser à d'autres modes de communication utilisés sur l'Internet tels que les forums de discussion.

La construction d'un corpus à partir d'informations en ligne est délicate étant donné le manque de connaissances précises des données récupérées. L'utilisation du Web comme source pour des corpus peut être délicate suivant l'objectif. Actuellement, notre outil offre la possibilité de constituer des corpus d'échantillons pour l'étude linguistique de la langue sur l'Internet. Il peut être utile dans le cadre de la construction de corpus de suivi, afin d'étudier les phénomènes langagiers sur le Web ou d'autres modes de communication existant sur l'Internet.

RÉFÉRENCES BIBLIOGRAPHIQUES

AMITAY, E. (1999). Anchors in Context : A corpus analysis of web pages authoring conventions, *In Words on the Web – Computer Mediated Communication.* Lynn Pemberton and Simon Shorville : Intellect Books, UK.

BUCKLAND, M. et F. GEY (1994). The relationship between recall and precision. *Journal of American Society for Information Science*, 45(1), 12-19.

COWIE, J., E. LUDOVIK et R. ZACHARSKI (1998). An autonomous, web-based, multilingual corpus collection tool. *In Proceedings of the International Conference on Natural Language Processing and Induxtrial Applications*.

HABERT B., A. NAZARENKO et A. SALEM (1997). *Les linguistiques de corpus. U Linguistique*. Paris : Armand Colin/Masson.

HERWIJNEN E.V. (1995). *SGML pratique*. Intemational Thomson Publishing. Traduction de Alain Herbuel, Frédéric Orth et Christelle Chaloin.

IDE, N. (1995). *Encoding standards for large text ressources : The text encoding initiative*. T.A.L., 36(1-2), 201-211.

PE'RY-WOODLEY, M.-P. (1995). *Quels corpus pour quels traitements automatiques?*, T.A.L., 36(1-2), 213-232.

RADEV, D.R. et K.R. McKeown (1997). Building a generation knowledge source using internet accessible newswire. *In Proceedings of the 5th Conference on Applied Natural Processing*, Washington, DC.

RESNIK, P. (1998). Parallel strands : A preliminary investigation into mining the web for bilingual text. *In Proceedings of 1998 Conference of the Association for Machine Translation in the Americas*.

SINCLAIR, J. (1996). Preliminary recommendations on corpus typology. EAGLES Document EAGTCWG-CTYP/P.

SINCLAIR, J. et J. BALL (1996). Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P.