Linguistica dei Corpora

Strumenti e applicazioni

A CURA DI
MARIE HÉDIARD

EDIZIONI DELL'UNIVERSITÀ DEGLI STUDI DI CASSINO
2007

Via G. Marconi 10 – Cassino
ISBN 978-88-8317-043-0

Elaborazione: Centro Editoriale d'Ateneo

Distribuzione
Università degli Studi di Cassino
Centro Editoriale di Ateneo
Via Sant'Angelo in Theodice
Campus Folcara – 03043 Cassino (FR)
Acquisto online: http://www.unicas.it/cea
E-mail: editoria@unicas.it
Tel. +39 07762993225 – Fax +39 07762994806

Finito di stampare nel mese di dicembre 2007
presso Grafica Sud S.r.l.
via Nazionale delle Puglie, Km 35.935
80013 Casalnuovo, Napoli

Indice

Sonia Di Vito

Les ressources en français pour la linguistique de corpus

Sintesi. *I risultati dell'analisi contestuale di una lingua dipendono molto dal tipo di dati che vengono esplorati. In particolare, gli studi condotti con la metodologia proposta dalla linguistica dei corpora hanno come punto di partenza i dati contenuti nel corpus; per questo le fonti, gli autori, il pubblico e tutte le altre informazioni relative ai testi che lo compongono sono di fondamentale importanza per uno studioso della lingua che voglia utilizzare tale metodologia. Nell'approccio alla costituzione di un corpus è dunque necessario avere presenti gli obiettivi dell'analisi linguistica e pertanto è importante definire fin dall'inizio i criteri di composizione della risorsa, affinché risponda in pieno alle esigenze del ricercatore che la crea; tali criteri devono inoltre essere ben esplicitati in modo da rendere la risorsa fruibile ad altri studiosi dei fenomeni linguistici. Nel presente articolo analizzerò in un primo momento le caratteristiche e le tipologie di un corpus, nonché i criteri di composizione; presenterò, in seguito, un panorama delle risorse esistenti in francese per la linguistica dei corpora.*

Introduction

Les approches de l'étude de la langue en tant que système étaient, au début du développement de la linguistique en tant que science, plutôt abstraites car, dans la plupart des cas, il s'agissait d'explications et de théorisations liées plus à l'intuition du chercheur qu'à des vraies analyses de la langue en emploi. Toutefois, à un certain moment du développement des théories linguistiques, certains linguistes ont ressenti l'exigence de corroborer leurs théories du langage par des vérifications qui avaient, parfois comme point de départ, d'autres fois comme point d'arrivée, des données réelles dérivant de la langue en emploi.

C'est ainsi que, aux alentours des années 60, un linguiste anglais, J. R. Firth, a mis en évidence l'importance du contexte de production

d'un discours, qu'il soit écrit ou oral, dans l'étude de la langue [Firth 1957]. Il a introduit, en particulier, la notion de *context of situation*, à savoir d'un ensemble de facteurs qui interagissent entre eux et qui constituent la base de toute production linguistique. Ces facteurs [Tognini-Bonelli 2001, 158] sont:

- les participants, personnes et personnalités (et leurs traits caractéristiques pertinents), leurs actions verbales;

- leurs actions non verbales;

- les objets qui entrent en ligne de compte avec le contexte de communication, ainsi que les événements non verbaux et non liés aux personnes[1];

- les effets de l'action verbale.

En suivant un célèbre *dictum* de Wittgenstein [Tognini-Bonelli 2001, 159] «the meaning of words lies in their use», Firth a donc proposé une méthode d'analyse qui ne considère la langue que dans son contexte de situation; de cette assertion initiale découle la nécessité d'intégrer le niveau linguistique et le niveau extralinguistique en prenant en considération des exemples concrets de langue en emploi liés à des contextes de situation déterminés. Nous avons donc deux types de considérations à faire: la première concernant l'importance de la composition des corpus pour avoir à disposition des ensembles très vastes d'exemples concrets et la deuxième relevant de la nécessité d'établir des critères dans la collection des textes qui seront la source de notre analyse.

Au fur et à mesure que la méthodologie d'analyse se précisait et que les nouvelles technologies se développaient, les spécialistes du domaine ont envisagé la nécessité d'une définition plus précise de corpus et de tout ce qui concerne la collecte des données authentiques. Dans le présent article, nous nous proposerons tout d'abord d'analyser les caractéristiques d'un corpus en cherchant à définir ce qu'est un corpus.

[1] Tout ce qui fait partie de l'échange communicatif mais qui ne concerne pas les êtres humains impliqués dans la communication.

Puis nous mettrons en évidence les critères de compilation et énumérerons les différentes typologies de corpus. Dans la dernière partie nous décrirons enfin les ressources en français pour l'analyse linguistique à l'aide des nouvelles technologies.

1. DÉFINITION DE CORPUS

Une définition adéquate de corpus s'avère nécessaire à partir du moment où a eu lieu un certain nombre de malentendus en dehors du domaine de la linguistique de corpus. En effet, un corpus n'est pas une simple collecte de données textuelles sous format papier ou électronique faite au hasard, mais il s'agit d'un ensemble de textes qui sont collectés selon des critères précis et définis de façon claire, afin que d'autres chercheurs puissent l'utiliser sciemment dans leurs études sur la langue.

Il existe trois concepts essentiels qui interviennnent dans la définition de corpus [Pearson 1998, 42]:

- la collection: après avoir décidé son domaine d'investigation et d'étude, un chercheur doit rassembler des textes qui l'aideront dans sa recherche; il peut prendre en considération ce que Sinclair [EAGLES[2] 1996, 2] définit *pieces of language* à savoir, non seulement les textes traditionnels (tout ce qui a un début et une fin et qui peut être encadré dans les genres textuels traditionnels), mais aussi les morceaux de langue qui ne sont pas des textes complets dans le sens traditionnel du mot «texte» mais qui, cependant, rentrent dans la catégorie des exemples concrets de langue en emploi;

- la représentativité: un corpus doit être représentatif dans le sens qu'il doit représenter tout le "système" de la langue qu'on est en train

[2] EAGLES (*Expert Advisory Group on Language Engineering Standards*) est une initiative de la Commission Européenne dans le cadre du programme DG XIII *Linguistic Research and Engineering*, qui a pour objectifs: 1) de donner des modèles standards pour des ressources utilisables pour l'analyse des langues à une très grande échelle (tel que les corpus de l'écrit, de l'oral et les lexiques computationnels); 2) de produire des outils pour utiliser de telles ressources (logiciels de concordance, de codage XML, morphosyntaxique, etc.); 3) d'établir des critères pour évaluer ces ressources et ces outils. Ces informations sont disponibles sur le site http://www.ilc.cnr.it/EAGLES96/intro.html [consulté le 19 novembre 2006].

d'analyser. Bien évidemment, ce facteur dépend de la "portée" du phénomène langagier qu'on veut décrire dans la mesure où, plus ce phénomène est étendu et général (comme par exemple l'emploi des prépositions *par* et *pour* dans la langue française), plus les composantes du corpus doivent être variées (composantes écrites et orales, genres et registres différents, etc.)[3];

- l'échantillon: le corpus ne doit contenir que des exemplaires réels de langue car c'est seulement grâce à l'étude de ces exemplaires qu'on parvient à une description de la langue telle qu'elle est effectivement utilisée.

Les concepts essentiels qui entrent en jeu dans la définition d'un corpus peuvent toutefois être mis en doute par tout linguiste qui n'est pas un spécialiste du domaine car nous assistons à une prolifération de collections de textes qui sont appelées «corpus» mais qui ne respectent pas les caractéristiques essentielles d'un corpus. En effet, parfois on appelle «corpus» un archive de données linguistiques qui ont été collectées en dépit de tout critère, ou encore on parle de corpus pour indiquer un ensemble de textes dont les caractéristiques n'ont pas été clairement explicitées.

Pour éviter ce risque de confusion Sinclair [EAGLES 1996] a indiqué quatre paramètres:

- quantité: la valeur utilisée pour décrire cette caractéristique est "large", à savoir le corpus doit être étendu et varié car seul un corpus composé de nombreuses et différentes composantes peut représenter les différentes facettes de la langue;

- qualité: la valeur utilisée pour décrire cette caractéristique est "authentique" dans la mesure où, comme cela a été explicité à maintes reprises, le matériel qui compose le corpus doit être tiré «from genuine communications of people going about their normal business» [Tognini-Bonelli 2001, 55];

[3] Leech, cité par Kennedy, affirme qu'un corpus est représentatif si les résultats obtenus dans la recherche peuvent être généralisés et être valables pour la langue perçue comme un ensemble homogène ou comme une variété [Kennedy 1998, 62].

- simplicité: la valeur utilisée pour décrire cette caractéristique est "vide" et elle indique le fait qu'un texte ne doit contenir aucune forme d'encodage car, selon Sinclair, les annotations présupposent une théorie préexistante qui ne permettrait pas au chercheur de se baser seulement sur l'évidence des occurrences et des co-occurrences[4];

- documentation: la valeur utilisée pour décrire cette caractéristique est "documenté" car le locuteur doit avoir accès aux informations concernant l'auteur/les auteurs, la date de publication, le but, le public auquel le texte est adressé, etc.; ces détails doivent être contenus dans un document autre que le texte[5].

2. LES CRITÈRES DE COMPILATION D'UN CORPUS

Les spécialistes de la linguistique de corpus ont insisté à plusieurs reprises sur l'importance de suivre des critères formels clairs dans la collecte des composantes d'un corpus, car les données contenues dans l'ensemble des textes influencent énormément les résultats de l'analyse.

Il faut tout d'abord décider le domaine de la recherche, à savoir s'il s'agit d'une recherche sur le système général de la langue ou bien sur un phénomène particulier ainsi que le type de données dont nous avons besoin pour mener notre étude (le genre de texte, la typologie, le registre); les décisions prises à ce propos doivent être claires et explicitées.

Ensuite, nous devons définir la taille du corpus qui correspond non seulement au nombre total des mots, mais aussi au nombre d'exemplaires et au nombre de mots de chaque exemplaire. Il est nécessaire de tenir compte de l'équilibre (*balance*) de ces variables numériques car, comme l'affirme Kennedy, il est important, dans le cas d'un corpus général de langue par exemple, de maintenir une équivalence entre le

[4] Nous partageons l'opinion de Aarts qui est contre ce refus des annotations basé sur le fait qu'il s'agirait seulement d'informations additionnelles; en fait il est important de signaler la présence des annotations aux utilisateurs du corpus et de leur permettre d'avoir accès au texte simple [Aarts 2002, 7].

[5] Avec un étiquetage selon les normes TEI adoptées pour donner ce type d'information.

nombre de textes écrits et oraux, entre les différents genres, entre le nombre de mots dans chaque texte [Kennedy 1998, 62].

Il est enfin fondamental de déterminer la période à laquelle appartiennent les textes: s'il s'agit d'un corpus général il faudrait prendre en considération plusieurs décennies, voire plusieurs siècles, pour avoir un panorama de l'évolution que la langue a subie dans cette période; nous pouvons, au contraire, décider de n'analyser qu'une période particulière de la langue et la choisir en fonction de nos intérêts spécifiques.

Après avoir arrêté les choix mentionnés ci-dessus, il faut classifier chaque texte selon des critères qui sont à la fois linguistiques et extralinguistiques, et qui sont appelés en linguistique de corpus [Pearson 1998, 52] «critères internes» et «critères externes».

2.1. Les critères internes

Ce type de paramètres concerne les caractéristiques internes des textes, notamment leur domaine et leur style. Ce sont des critères qui ne sont pas supportés par une évidence scientifique, il s'agit donc d'une classification, à notre avis, aléatoire car il existe différentes versions de classification sans qu'aucune n'ait de bonnes raisons pour être retenue. Dans le domaine de l'étude de l'anglais il y a deux linguistes qui ont essayé de mettre au point des méthodes objectives pour classifier les textes selon ces critères.

En ce qui concerne le domaine, Phillips a employé une méthodologie distributionnelle visant à relever la macrostructure du texte en analysant la fréquence des collocations et des associations particulières et en menant une comparaison entre la macrostructure de chaque partie du texte [Phillips 1983].

Pour le style c'est Biber qui a proposé une méthodologie se basant sur des analyses statistiques et visant à étudier les caractéristiques similaires des textes ayant la même fonction communicative, pour établir ainsi des paramètres classificatoires objectifs [Biber 1988].

2.2. Les critères externes

Il s'agit en général de critères socioculturels qui concernent les participants, la fonction communicative, l'évènement et le cadre so-

cial dans lequel le texte a été créé:

- l'origine: l'auteur, l'éditeur, le traducteur, l'âge, le sexe, le lieu de résidence et le *background* linguistique des personnes impliquées, éléments qui peuvent affecter le contenu ou la structure linguistique du texte;

- l'état: le mode de transmission, les conditions du mode de transmission, l'apparence du texte;

- le but: les personnes pour qui le texte a été conçu et les objectifs communicatifs.

Ces éléments différenciateurs nous permettent de faire une distinction entre les différents corpus et de les classifier dans des typologies que nous allons décrire.

3. LES DIFFÉRENTS TYPES DE CORPUS

Les linguistes de corpus mettent en évidence plusieurs types de corpus classifiés sur la base des critères internes et externes qui dépendent des choix effectués par le *corpus builder*.

3.1. Le corpus de référence

Selon Sinclair ce genre de corpus doit être composé de matériel homogène avec des sous-corpus (les *sub-corpora*) qui ont le même nombre de textes, le même nombre de mots pour chaque texte et qui sont tirés de plusieurs sources [Sinclair 1991, 17]. Le corpus de référence (*reference corpus*) est à la base de toute recherche qui voudrait analyser le phénomène "langue", à savoir les mécanismes qui caractérisent une certaine langue en emploi. Il est constitué de données qualitativement et quantitativement importantes qui peuvent servir de base pour la constitution de dictionnaires, grammaires et thésaurus. Un exemple de corpus de référence en anglais est la *Bank of English* qui atteint 200 millions de mots et qui se compose de plusieurs sous-corpus.

3.2. Le corpus de suivi[6]

La caractéristique d'un corpus de suivi est le fait d'être continuellement mis à jour par de nouveaux textes qui vont remplacer les textes "vieillis" en termes de date de parution. Ce type de corpus permet d'observer l'évolution de la langue; cela est possible dans la mesure où les textes "anciens" restent disponibles à la consultation et peuvent ainsi être comparés aux textes nouvellement introduits, de façon à permettre une analyse diachronique[7] de la langue. Le corpus de suivi se différencie d'un corpus de référence car il s'agit d'un corpus ouvert et en mouvement qui est modifié périodiquement sans, toutefois, perdre ses anciennes composantes.

3.3. Le corpus spécialisé

Le corpus spécialisé ou corpus spécial est un échantillon de langue "particulière" dans la mesure où la langue utilisée ne peut pas être considérée comme langue quotidienne et standard. Il s'agit, en effet, de textes qui contiennent un grand nombre de caractéristiques inusuelles et qui sont utilisés pour étudier des langues telles que la langue des enfants, des personnes âgées, des parlants non natifs, des personnes qui ont des troubles langagiers, etc.

3.4. Le corpus constitué à des fins spéciales

Cette appellation a été créée et définie par Pearson pour indiquer un type de corpus qui est compilé pour atteindre des objectifs déterminés, comme par exemple l'étude de la langue employée dans les cours universitaires ou encore la langue de la médecine, etc. [Pearson 1998, 48]. Pearson a constitué elle-même un corpus à des fins spéciales pour analyser la structure des définitions dans des textes spécialisés.

[6] Terme adopté dans [Habert *et alii* 1997, 145].

[7] En effet, ce modèle de corpus a ouvert une nouvelle perspective pour les corpus contemporains, celle de la dimension diachronique qui peut être analysée à partir d'un seul corpus.

3.5. Le corpus parallèle

La caractéristique fondamentale d'un corpus parallèle est d'être composé de textes traduits dans une ou plusieurs langues à partir d'une langue source. Il existe trois possibilités de corpus parallèle: 1) il n'y a que deux langues employées, la langue source (A) et la langue cible (B), et les textes contenus sont des traductions différentes dans la langue cible B du texte en langue source A; 2) les langues employées sont toujours deux mais le corpus contient des textes écrits dans les deux langues (A et B) et leurs respectives traductions; 3) les langues employées sont plusieurs et les textes constituent la traductions de textes originairement écrits dans une langue Z [Teubert 1996, 245]. Nous pouvons considérer comme corpus parallèles le recueil des séances de la Communauté Européenne qui sont écrites dans les langues de tous les pays membres, ou bien le *Monde Diplomatique*, mensuel français qui est traduit dans plusieurs langues.

3.6. Le corpus comparable

Un corpus comparable est un corpus mono-, bi- ou multilingue dont les différentes parties peuvent être comparées entre elles dans la mesure où elles se composent de textes qui ont été collectés en suivant les mêmes critères internes et externes. Il ne s'agit pas comme dans le cas du corpus parallèle, de textes traduits, mais de textes qui abordent le même sujet, qui sont écrits dans le même registre de langue ou qui ont été produits dans un même contexte situationnel. Un exemple de corpus comparable est le ICE, le *International Corpus of English* qui se compose de plusieurs variétés d'anglais parlées dans les différents pays anglophones. Un autre exemple, sur lequel nous reviendrons, est celui de C-ORAL-ROM, corpus oral multilingue composé de discours formels et informels qui constituent quatre collections comparables de données en français, espagnol, italien et portugais.

4. RESSOURCES DISPONIBLES EN FRANÇAIS

L'existence d'une si grande typologie de ressources et surtout l'absence en français d'une base de données accessible qui puisse toutes les

rassembler, nous ont poussée à faire l'inventaire et la description des ressources à disposition du chercheur qui se penche sur l'analyse du français. Nous avons fait une subdivision entre ressources écrites, orales et hybrides (ces dernières comprenant en particulier les nouveaux langages qui se développent à partir des nouveaux moyens de communication, tels que les textos ou les tchats).

4.1. Ressources écrites

4.1.1. FRANTEXT

Frantext est une base de données littéraires de la langue française associée à un outil de consultation et d'hypernavigation qui a servi à l'origine de base d'exemples pour la rédaction des articles du *Trésor de la Langue Française*. Il a été mis au point dans le cadre d'une collaboration entre l'Université de Nancy 2, le Centre National de la Recherche Scientifique et l'ATILF (laboratoire *d'Analyse* et *Traitement Informatique de la Langue Française*).

Cette base de données contient 3 737 œuvres littéraires qui couvrent 5 siècles de littérature (du XVIe au XXe siècle) et auxquelles ont été ajoutés des textes appartenant aux domaines des sciences, des arts, des techniques.

Il existe deux versions des textes littéraires: une version catégorisée qui comprend 1 940 œuvres en prose des XIXe et XXe siècles, soit environ 127 millions d'occurrences, qui ont été codées avec une étiquette grammaticale correspondant aux catégories grammaticales (nom, adjectif ...); l'outil de consultation a été mis à jour en insérant, dans les possibilités de recherche, des requêtes portant sur ces mêmes catégories grammaticales. La deuxième version est une version non catégorisée qui se compose de l'intégralité de la base, à savoir tous les textes, environ 210 millions d'occurrences, environ un millier d'auteurs, avec la possibilité d'effectuer des recherches à un niveau simple ou complexe.

Frantext est accessible sur Internet, moyennant un abonnement[8].

[8] On trouve le formulaire pour la demande d'abonnement sur le site http://www.atilf.fr/.

4.1.2. Colisciences

Le projet *Colisciences* est une extension d'un projet ancien, COLIS (*Corpus de Littérature Scientifique* de langue française); né du constat de l'absence quasi-totale de ressources en ligne offrant des corpus francophones traitant des sciences, il a été mis en place par le *Laboratoire Communication et Politique* du CNRS de Paris [Vignaux *et alii* 1999].

Ce corpus, raisonné et annoté, comprend environ 5 500 pages empruntées aux éditions originales des textes de Claude Bernard, Ludwig Büchner, Etienne et Isidore Geoffroy Saint-Hilaire, Armand de Quatrefages, George G. Romanes, Oskar Schmidt. L'objectif est de l'élargir en y ajoutant une édition complète des livres de Claude Bernard et une sélection de ses articles (au total 5 000 pages supplémentaires).

Le corpus comprend aussi des ressources complémentaires, à savoir une courte biographie de l'auteur, une bibliographie de l'auteur et sur lui, une présentation de ses travaux et de leur portée.

Des liens hypertextuels entre les différents parcours d'exploration et de lecture permettront de trouver aisément les informations et de saisir les idées contenues dans les textes par d'autres moyens que par la lecture linéaire qui caractérise la lecture d'un livre classique.

Ce corpus a été constitué dans le but d'atteindre quatre objectifs:

- culturels et patrimoniaux dans la mesure où l'on a collecté, numérisé et rendu disponible un grand corpus des ouvrages et des articles des principaux biologistes et naturalistes du XIX^{e} siècle en langue française;

- intellectuels car, à travers les textes qui composent le corpus, on peut retracer le développement de la biologie, articulée en plusieurs grands domaines;

- scientifiques car l'architecture du site se base sur les réflexions autour du concept d'hypertexte qui favorise la construction et le maintien en place de liens sémantiques entre les objets repérables dans l'ensemble des documents;

- cognitifs et pédagogiques dans la mesure où le corpus permet de répondre aux questions soulevées par la nouvelle modalité de lecture (comment on lit un texte électronique, comment on peut l'utiliser

dans le processus d'acquisition des connaissances et comment on peut maîtriser l'hypertextualité qui le caractérise).

Ce corpus est conçu plus pour l'étude des processus d'appropriation et d'utilisation des hypertextes que pour l'étude proprement linguistique de ses contenus. Toutefois, il pourrait être utile pour une analyse comparative entre la langue employée dans des textes scientifiques du XIX[e] siècle et du XX[e], de même que pour l'étude de l'évolution du lexique, etc..

Ce corpus est en accès libre à la page http://colisciences.in2p3.fr/.

4.1.3. Le corpus du Témiscouata

Ce corpus est né à partir d'un projet réalisé à l'Université du Québec à Rimouski[9]; il s'agit d'une collection de lettres rédigées entre 1930-1936 par des colons ou des futurs colons, généralement peu lettrés, et envoyées à l'abbé Léo-Pierre Bernier, missionnaire colonisateur[10]. Le contenu de ces lettres concerne des questions d'administration quotidienne, des plaintes pour les mœurs des voisins, des réclamations pour les services (église, école, moulin), des requêtes auprès de l'abbé, du gouvernement, du député. En parallèle, il y a d'autres lettres qui témoignent les échanges de l'abbé Léo-Pierre avec d'autres acteurs de la colonisation comme des agents des terres, des curés de paroisse, des religieuses, des gardes-malade; le contenu de ces lettres illustre très bien les coutumes d'une époque difficile à cause du chômage.

Du point de vue linguistique, ce corpus a été étudié à partir de la constatation que les colons, malgré leur connaissance approximative de l'orthographe, connaissaient et maîtrisaient certaines règles du style épistolaire et de la rhétorique de la requête et de la persuasion. Ainsi a-t-on constaté l'existence d'une "tension" entre la langue populaire, locale et individuelle, qui contenait des expressions régionales et des tournures

[9] Les liens qui donnent accès à ces informations sont http://www.ling.uqam.ca/forum/corpus/messages/3.html et http://www.ling.uqam.ca/corpus/faribault/temis.htm [consultés le 19 juillet 2006].

[10] En effet ce corpus a été extrait du *Fonds Léo-Pierre Bernier*, appartenant au service des Archives de l'Université du Québec.

syntaxiques populaires, et l'écrit relevé avec ses formes prestigieuses.

Toutes les lettres sont manuscrites et pour rendre le corpus accessible on a dû produire une version normalisée, dans la mesure où le français utilisé n'était pas normé. Le type de normalisation qui a été effectué a consisté en:

- la mise en page des textes;
- la ponctuation;
- l'usage des majuscules;
- l'usage des diacritiques (les accents et le tréma sur les voyelles, la cédille);
- la coupure des mots (avec l'ajout ou le retrait de l'apostrophe et du trait d'union).

Toutefois, même s'il y a eu un aménagement typographique pour rendre la lecture plus agréable, l'intervention a toujours été prudente pour ne pas compromettre les analyses linguistiques ultérieures.

Un autre problème rencontré, relevant du manque de scolarisation des colons, est celui de la substitution de certaines formes graphiques à d'autres (comme *ses* pour *c'est*), ce qui a comporté des problèmes de reconnaissance automatique et d'étiquetage des formes. Certaines substitutions sont récurrentes, ce qui permet d'éviter un traitement occurrence par occurrence.

Il existe une version accessible en mode NOMINO qui fournit l'ensemble des lettres que nous pouvons lire de façon traditionnelle. En effet, nous avons accès à une page qui est subdivisée en deux parties: la première classifie les documents selon différents dossiers (par exemple "Colons du plan Vautrin", "Lettres des colons", "Biencourt", "La-Nativité-de-la-B.-V.-M.", "Hygiène", "Infirmières"); la deuxième nous fournit un index des lettres divisées par catégories d'auteurs (comme par exemple "agents-du-gouvernement", "colons-et-simples-particuliers", "membres-des-communautés-religieuses" etc.).

Le traitement du corpus se fait également à l'aide d'un corcordancier à partir de la page http://www.ling.uqam.ca/forum/corpus/messages/3.html en cliquant sur le lien "en mode SATO": nous pouvons

mener une recherche précise à l'aide du concordancier qui affiche les résultats de la recherche du mot-clé sous forme de concordances et avec leur co-texte immédiat, sans prendre en considération la totalité du texte. Il existe différentes recherches possibles sur les contextes d'un ou de plusieurs mots-clé:

- contextes de longueur fixe:
- contextes de phrases: dans ce menu nous pouvons rechercher une liste de mots aussi bien que des expressions figées;
- contextes de paragraphe.

Il existe aussi un lien au "lexique du corpus" qui vise à analyser les mots utilisés dans le texte avec leur fréquence d'utilisation; la liste des mots peut comprendre la ponctuation, les nombres, les abréviations, les noms propres, etc.

Un extrait de la version normalisée du corpus est accessible en ligne à l'adresse http://www.ling.uqam.ca/corpus/faribault/temis.htm, en cliquant sur le lien "Accès au corpus".

4.1.4. Le corpus BAF

Le corpus BAF (*bi-textuel anglais-français*) est une base de données qui se compose d'un ensemble de documents bi-textes anglais et français, qui sont la traduction l'un de l'autre[11]. Ce corpus a été constitué par l'équipe de traduction assistée par ordinateur du CITI, à l'Université de Montréal. La plus grande partie du corpus est constituée de textes de nature institutionnelle; il y a aussi quelques articles scientifiques de même qu'une œuvre littéraire, pour un total de 400 000 mots dans chaque langue.

Ce corpus comprend, en particulier:

- le journal des débats de la chambre des Communes (Canada) séance du 14 mars 1994 (55 625 mots en anglais et 58 994 mots en français);
- un procès à la cour suprême du Canada (*Terrence Wayne Burlin-*

[11] Informations et téléchargement sur le site http://rali.iro.umontreal.ca/Ressources/BAF/ [consulté le 03 octobre 2006].

gham c. sa majesté la Reine- avec 31 092 mots en anglais et 33 170 mots en français);

- des reportages de l'ONU:

1) *241st and 242nd Reports of the Committee on Freedom of Association* de 1985 – 145 346 mots en anglais et 153 061 mots en français;

2) *Report of the Secretary General on the Work of the Organisation* de 1993 – 48 407 mots en anglais et 54 869 mots en français);

- des articles scientifiques concernant les technologies de communication et de l'information, des analyses de traduction traditionnelle et automatique, le changement technologique et l'organisation du travail, la cohérence automatique, la réaccentuation automatique de textes français, pour un total de 49 941 mots en anglais et 53 735 mots en français;

- le texte littéraire *De la terre à la lune* de Jules Verne (40 161 mots en anglais et 53 181 mots en français);

- un document technique *ScanWorX User's Guide* de la Xerox Corporation (39 328 mots en anglais et 46 828 mots en français).

Les phrases de ces documents ont été alignées: pour chaque segment de texte dans une langue, nous avons le correspondant dans l'autre langue.

La version 1.1 du corpus est disponible gratuitement en trois formats, à la page http://rali.iro.umontreal.ca/Ressources/BAF/:

- dans le format UNIX: baf-1.1.tar.gz, fichier créé à l'aide des utilitaires UNIX *tar* et *gzip*;

- dans le format Zip: baf.1.1.zip, fichier créé à l'aide du programme *zip*;

- dans le format "texte" qui nous permet de télécharger un à un les différents fichiers qui constituent le corpus en cliquant sur les liens qui apparaissent dans la description du corpus.

4.1.5. FRIDA

Le projet FRIDA[12] (*French Interlanguage Database*) a été lancé en 1998 par le *Centre for English Corpus Linguistics* de Sylviane Granger de l'Université de Louvain la Neuve en Belgique, dans le but de rassembler un corpus de français langue étrangère. Cette base de données contient actuellement 200 000 mots mais les chercheurs qui y travaillent espèrent bientôt atteindre les 450 000 mots.

Le but ultime de l'utilisation du corpus Frida est celui de créer un correcteur automatique adapté aux utilisateurs non francophones: en effet, il est possible de créer un outil multimédia d'apprentissage du français en se basant sur les erreurs des apprenants relevées par Frida.

Les textes contenus dans Frida ont des caractéristiques précises: tout d'abord il s'agit de textes qui ont été produits par des apprenants de niveau intermédiaire, qui ont une longueur de 100 à 1 000 mots par texte et qui sont des compositions libres (textes descriptifs, argumentatifs, narratifs, courrier, journal). La langue maternelle des producteurs n'a pas d'importance dans la mesure où les chercheurs qui travaillent à ce projet souhaitent rassembler des écrits d'apprenants de langues maternelles très diverses.

Les textes qui sont envoyés au Centre de recherches sont minutieusement analysés et étiquetés une fois que toutes les erreurs ont été répertoriées. Cet étiquetage est fait sous format XML, car l'emploi de cette norme permet d'avoir un système d'étiquetage plus flexible; il existe trois balises pour chaque erreur, à savoir chaque erreur est classifiée selon trois catégories: le domaine d'erreurs (grammaire, lexique, orthographe,...), le type d'erreur (genre, nombre, etc.) et la catégorie grammaticale (nom, adjectif, etc.).

En raison du fait que le corpus n'a pas encore été achevé, il n'est pas encore disponible[13].

[12] Informations obtenues à partir du site www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Frida/frida.htm [consulté le 15 novembre 2006].

[13] En effet les chercheurs souhaitent élargir et diversifier le corpus grâce à l'aide précieuse des étudiants de français langue étrangère de niveau intermédiaire qui peuvent envoyer un message contenant un extrait de leurs productions en français ainsi que leur profil, à l'adresse granger@lige.ucl.ac.be (bien évidemment les textes doivent respecter des critères qui sont explicités à la page www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Frida/frida.htm).

4.1.6. GLOSSANET

Une autre base de données qui est disponible en ligne à l'adresse http://glossa.fltr.ucl.ac.be[14] est celle de *Glossanet*, un concordancier en ligne qui permet aux usagers de faire une recherche sur des mots ou des séquences de mots dans un ensemble de journaux dont les éditions quotidiennes sont disponibles sur la toile et qui sont sélectionnés au début de l'opération. Ce genre de corpus est disponible pour plusieurs langues (français, anglais, espagnol, italien, portugais, finlandais, allemand, hollandais, norvégien, grec, slovaque et russe).

Quand le concordancier trouve, dans le corpus sélectionné, plusieurs occurrences des mots-clé, la liste des concordances est envoyée automatiquement à l'usager sur son adresse électronique: l'occurrence est présentée dans son co-texte qui correspond à 40 caractères à droite et 40 caractères à gauche.

Glossanet nous permet d'effectuer des recherches complexes dans l'ensemble de notre corpus en utilisant des symboles, des codes ou des opérateurs.

Nous pouvons accéder, par exemple, à l'ensemble des formes d'un lemme en utilisant les signes < > qui permettent de trouver toutes les formes par exemple du présent de l'indicatif du verbe *avoir* en tapant *<avoir>*.

En utilisant certains codes, nous pouvons avoir accès à une recherche plus approfondie qui tient compte des éléments grammaticaux, morphologiques et sémantiques et qui nous permet de mener une recherche générale (par exemple le code <V:P3s> nous permet de rechercher les occurrences de tous les verbes de l'indicatif présent à la troisième personne du singulier).

A l'aide d'opérateurs nous pouvons effectuer également la recherche d'un ensemble de mots. Par exemple, le signe + nous permet de retrouver dans le corpus la liste des mots qui sont utilisés dans une expression régulière: *<je+tu+il+elle+nous+vous+ils+elles> <être> <allé+allée+allés+allées>* nous donnera les séquences suivantes *je suis allé, nous étions allées, ils sont allés*, etc.

[14] [consulté le 06 octobre 2006].

La personne qui visite le site pour la première fois doit s'enregistrer pour avoir un nom d'utilisateur et un mot de passe. Il peut à tout moment modifier le type de recherche qu'il veut effectuer, à savoir il peut chercher de nouvelles expressions, ou bien il peut modifier les composantes du corpus. L'usager peut choisir aussi de recevoir les résultats de sa requête directement sur son mail, chaque jour ou chaque semaine.

Une autre possibilité c'est d'obtenir tout de suite les occurrences des mots recherchés; cela implique que l'opération soit effectuée en ligne, à savoir les résultats seront immédiatement affichés sur l'écran et le moteur de recherche ne poursuivra pas l'analyse des journaux choisis dans les jours ou les semaines à venir. *Glossanet* est un outil très intéressant mais qui implique des délais assez longs pour la collecte des données.

4.1.7. WEBCORP

Webcorp[15] est un outil qui permet de chercher des exemples d'emplois des mots à partir du Web pour des analyses linguistiques. L'exigence de créer un tel outil est née de la constatation que les corpus, dans la plupart des cas, sont fixes, à savoir qu'ils ne peuvent pas être mis à jour et qu'ils ne reflètent donc pas les changements de la langue [Kehoe & Renouf 2002]. Cet outil considère le *World Wide Web* comme un corpus, une grande collection de textes d'où les linguistes peuvent extraire des faits de langue. Ce système opère des recherches à partir des moteurs de recherche existant sur le Net.

Dans les options de base de WebCorp l'utilisateur peut spécifier plusieurs choix:

- le moteur de recherche qu'il faut utiliser;
- la longueur du co-texte à gauche et à droite du mot-clé;
- la recherche du mot exact (*case sensitive*) ou d'un mot qui contient les caractères tapés dans la case *search term* (*case insensitive*);
- *l'output* des résultats (qui peuvent être envoyés sur le courrier électronique ou affichés à l'écran).

[15] À l'adresse http://www.webcorp.org.uk/ [consulté le 21 novembre 2006].

On peut, en outre, choisir le format des résultats qui peut varier selon les besoins de l'utilisateur: on peut avoir des formats HTML standards, des textes simples et des formats de KWIC (*Key Word In Context*). L'utilisateur peut également choisir si le système doit mentionner les adresses internet des pages où le mot-clé a été trouvé.

Par ailleurs, l'utilisateur peut choisir de restreindre la recherche à toutes les pages d'un site ou à tous les sites d'un domaine donné. Par exemple, en choisissant l'option *Newspaper Domains* l'utilisateur peut restreindre sa recherche aux sites Web d'un ensemble défini de journaux (anglais, français, etc.).

WebCorp donne d'autres informations qui concernent la longueur des pages où le mot-clé a été trouvé, le nombre des formes (les *tokens*) et le nombre des lemmes (les *types*). Cet outil permet aussi d'avoir une liste de mots ordonnés alphabétiquement ou sur la base de leur fréquence, à laquelle on peut accéder en cliquant sur le lien qui apparaît dans le texte.

WebCorp permet d'avoir, enfin, des informations sur les collo-cations du mot-clé à gauche et à droite, qui sont reliées aux sens particuliers du mot. Les informations sur les collocations, sauvegardées dans un fichier de défaut contenant 4 mots à gauche et à droite, sont affichées dans un tableau et les *key phrases* sont mises en évidence.

WebCorp prévoit l'emploi de caractères spéciaux pour la recherche des modèles linguistiques; ces caractères sont:

- l'astérisque (*) qui, à la fin d'un mot, permet de rechercher les formes d'un lemme donné (en tapant *ét** le système recherchera les formes *était, été, étaient*, etc.); placé devant une série de lettres il permet de chercher tous les mots qui se terminent par la série donnée (en tapant **aient* le système recherchera les formes *mangeaient, riaient, allaient*, etc.); placé entre deux mots il permet de rechercher toutes les expressions qui commencent par le premier mot et qui se terminent par le dernier (en tapant *le * homme* WebCorp recherchera les expressions commençant par *le* et se terminant par *homme*);

- les crochets ([ ]) et la barre verticale (|) qui indiquent au système de rechercher les caractères ou les mots qui y sont contenus (en tapant *le [navire|bateau] coula* le système cherchera les expressions *le navire coula* et *le bateau coula*).

Une fois les résultats obtenus, l'utilisateur peut les modifier de différentes façons:

- en éliminant les concordances qui ne sont pas pertinentes à son analyse;
- en les ordonnant alphabétiquement selon leur contexte à gauche ou à droite;
- en les ordonnant selon leur date de parution.

Cet aperçu sur le fonctionnement de WebCorp nous a montré comment ce système peut affiner le processus de recherche sur le Web, en combinant la nécessité d'extraire des exemples de langue en emploi dans le réseau internet et de les afficher de la façon la plus avantageuse pour une analyse linguistique.

4.1.8. Le JRC-Acquis

Il s'agit d'un corpus parallèle qui contient des documents de la Communauté Européenne, dont la nature est principalement légale. C'est la plus grande ressource de ce genre qui existe: en effet, elle comprend des textes en plus de 20 langues (les pays membres de l'Union Européenne); plus de 8 000 documents, correspondant à 9 millions de mots par langue; des informations sur l'alignement des paragraphes aussi bien que le codage en XLM suivant les normes de la TEI[16]. Les documents ont été groupés (à la main) selon la classification de L'Eurovoc Thesaurus[17].

En ce qui concerne sa nature, le *JRC-Acquis* contient l'ensemble des obligations et des droits communs auxquels sont liés les Pays Membres. En particulier, il comprend les principes et les objectifs politiques des traités; la législation de l'Union, les déclarations et les résolutions, les accords internationaux et les actes et les objectifs communs.

[16] TEI: *Text Encoding Initiative*, http://www.tei-c.org/.

[17] Il s'agit d'un système de classification sémantique contenant plus de 6000 classes organisées hiérarchiquement sur la base de différents sujets. Les textes rassemblés dans le corpus concernent des domaines tels que l'économie, la santé, les technologies, le droit, l'agriculture, la politique, etc.. [Eurovoc 1995].

Ce n'est pas une ressource fermée dans la mesure où elle est en évolution constante (on y ajoute les documents traduits dans les langues des nouveaux pays qui sont entrés dans la Communauté Européenne, aussi bien que de nouveaux documents stipulés et rédigés au cours des activités de l'Union).

Cette ressource est complètement gratuite du moment que le *EC's Office for Official Pubblications* a donné la permission de la rendre publique pour des fins de recherche[18].

4.1.9. Le Corpus Carmel

Le corpus Carmel a été réalisé dans le cadre du projet Carmel, financé par le CNC, qui a réuni quatre partenaires: SINEQUA (éditeur de logiciel), le LIDILEM (*Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles de l'Université Stendhal Grenoble 3*), l'ACCE (*Académie Civilisation Cultures Européennes*) et le LIA (*Laboratoire Informatique Avignon*). Leur but était celui de constituer un corpus multilingue aligné, associé à des outils d'exploitation adaptés à ce genre de ressource.

Il s'agit d'une collection de récits de voyages, qui datent presque tous de la même période (fin du XIX^{e} siècle) et qui ont été traduits en quatre langues européennes: français, anglais, espagnol et italien[19]. Il contient 36 œuvres originales écrites par 24 auteurs de 6 nationalités différentes, pour un total de 10 millions de mots environ. La cohérence de ce corpus est due au fait que les différents auteurs ont tenu un journal de bord rigoureux pour décrire leurs aventures autour du monde.

La ressource a été collectée en suivant trois axes [Richard *et alii*, 2006]:

- tout d'abord, on a appliqué un alignement multilingue qui consiste dans la mise en correspondance des parties qui coïncident dans les différents textes, telles que paragraphes, phrases, lexèmes, etc.;

[18] Le corpus peut être téléchargé à partir de l'adresse http://wt.jrc/It/Acquis.

[19] Malheureusement on ne peut pas avoir accès à la totalité des traductions des originaux car seuls les textes libres de droit sont disponibles.

- ensuite, on a réalisé automatiquement un étiquetage sémantique d'une quinzaine de lexèmes en anglais et d'une trentaine en français, afin d'obtenir une désambiguïsation sémantique plus précise;

- enfin, on a opéré une annotation au niveau thématique, sur la base d'un thésaurus de 13 thèmes, qui avaient été fixés au début du travail et qui n'ont pas été modifiés pendant la réalisation du corpus, et de 64 sous-thèmes dont le choix s'est affiné au cours de l'acquisition des textes.

L'utilisateur peut effectuer deux types d'observation, l'une basée sur la recherche de mots-clés, comme avec les concordanciers classiques, l'autre en examinant les étiquettes morphosyntaxiques croisées avec les donnés thématiques ou sémantiques [Kraif *et alii* 2006].

Cette ressource est disponible gratuitement[20] en ligne à l'adresse http://projetcarmel.org.

4.2. Ressources orales

Après avoir décrit les ressources collectées pour l'écrit, nous allons maintenant présenter l'ensemble des corpus disponibles pour l'étude de l'oral.

4.2.1. Le corpus ELICOP

Dans le cadre d'un projet international, approuvé et subventionné par les Fonds de Recherche Scientifique[21], les Universités de Louvain la Neuve (Belgique) et d'Amsterdam ont repris des corpus élaborés dans le cadre de projets de recherche précédents, les ont enrichis et ont créé ce nouveau corpus, *Etude Linguistique de la Communication Parlée* (ELICOP).

Ce nouveau corpus se compose de deux parties, le corpus ELILAP et le corpus LANCOM, dont nous allons parler plus en détail.

[20] Seuls les textes libres de droits peuvent être téléchargés.

[21] Informations sur le site http://bach.arts.kuleuven.be/elicop.htm [consulté le 12 janvier 2006].

4.2.1.1. Le corpus ELILAP

La constitution du corpus ELILAP (*Etude Linguistique de la Langue Parlée*) remonte à la fin des années 60, quand des sociologues et des professeurs de français britanniques ont décidé de constituer une banque de données pour les analyses du français parlé, et se termine en 1976 quand la dernière composante a été incluse dans le projet. Il se compose de trois parties.

La première composante de ce corpus a été l'ESLO, *l'Etude SocioLinguistique sur Orléans*. Il s'agit d'un corpus d'environ 325 heures d'enregistrements effectués à partir d'interviews à des habitants d'Orléans, choisis de façon aléatoire à partir de la liste de l'Institut National de Statistique et d'Etudes Economiques. Ces enregistrements reflètent des situations de communication formelle qui présentent toutefois des variations du degré de spontanéité en fonction de la familiarité que les locuteurs ont avec le sujet de l'interview.

La deuxième composante d'ELILAP est le *Livre parlé de Tours*: il se compose de 120 heures d'enregistrements effectués à partir d'entretiens avec 193 témoins tourangeaux choisis de façon aléatoire qui exercent des professions variées et appartiennent aux deux sexes et à différentes tranches d'age. Les entretiens portaient sur 13 thèmes différents, en particulier sur le vécu des témoins, de façon à susciter des réponses spontanées, malgré le caractère formel de la situation de communication.

La troisième composante, la moins étendue du corpus ELILAP, est la *Voix d'Auvergne*, corpus qui contient deux types de situation: des tables rondes, reflétant une situation de communication très formelle, et des entretiens en face à face, se déroulant dans un contexte formel qui devient petit à petit plus spontané.

Les chercheurs ont visé trois objectifs:

- un objectif linguistique, pour avoir une base de données à partir de laquelle observer la langue parlée;
- des objectifs didactiques, dans la mesure où les résultats des recherches sur le français parlé peuvent modifier et renouveler la didactique du français en l'orientant dans le sens de la communication orale;
- un objectif socio-culturel, afin d'approfondir différents aspects de la vie sociale et culturelle des Français.

4.2.1.2. Le corpus LANCOM

Le corpus LANCOM a été constitué aux débuts des années 90 dans le but d'avoir une ressource de français parlé pour l'observation des savoir-faire communicationnels en français. Ce corpus revêt deux types d'intérêt: d'une part les chercheurs qui l'ont conçu voulaient offrir aux enseignants et aux concepteurs de manuels, un observatoire qui pouvait guider leurs choix; d'autre part, ils voulaient donner aux linguistes un observatoire pour décrire la langue de la communication.

Le corpus LANCOM se compose de trois éléments:

- des jeux de rôle organisés dans les classes de FLE en Belgique néerlandophone;
- des jeux de rôle analogues qui ont été joués dans des classes de FLM en Belgique francophone;
- les mêmes jeux de rôle joués par des francophones volontaires de la région de Lille et de Douai.

Actuellement ce corpus se compose d'environ 39 heures d'enregistrements qui constituent un corpus comparé, dans la mesure où les données apportées par les trois types d'enregistrements peuvent être alignées et analysées de façon comparative.

En outre, nous sommes en présence d'un corpus ouvert, notamment grâce au fait qu'il est constamment enrichi par de nouveaux enregistrements et de nouvelles transcriptions apportés par des étudiants en langues romanes de l'Université catholique de Louvain.

Les données recueillies ont été en partie transcrites et codées, pour un total de 18 heures.

Les scénarios organisés sont très divers: les rencontres dans la rue, l'agence de voyage, l'office du tourisme, la contravention, l'invitation, le baby-sitting. Parmi ces scénarios il existe certains qui se rapprochent beaucoup de la vie des élèves: dans ce cas-là la spontanéité des dialogues augmente. De tous les enregistrements et scénarios, les chercheurs ont retenu seulement ceux qui ont été jugés plus utiles pour les recherches envisagées.

4.2.1.3. Problèmes et solutions

Ces deux corpus sont divergents de par leur nature et requièrent des types d'annotations différents dans la mesure où le premier est un corpus oral et il a donc un système de renvois basé sur la numération des lignes ou des interventions et sur l'identité du locuteur. Le deuxième est, en revanche, un corpus qui vise l'étude et l'annotation des erreurs de langue chez des apprenants de français langue étrangère.

Il s'est agi alors de résoudre ces problèmes qui concernaient la normalisation des deux corpus en élaborant un format uniforme, et d'enrichir les corpus par un ultérieur étiquetage grammatical. Tous les éléments qui constituent l'étiquetage du corpus, quelle que soit leur nature, ont la forme d'une balise au format SGML insérée dans des crochets pointus (< >) [Mertens 2002, 2-4].

Une fois étiqueté, le corpus ELICOP a été mis à la disposition des utilisateurs sur le réseau Internet à l'adresse http://bach.arts.kuleuven.ac.be/elicop, à partir du lien "Consultation des corpus".

4.2.2. Le corpus VALIBEL

Le corpus VALIBEL[22] (*Variétés Linguistiques du français de Belgique*) a été constitué à partir de l'année 1987 et il comprend 22 sous-corpus qui ont été collectés pour la plupart par des étudiants de l'Université catholique de Louvain, dans le cadre de leur mémoire de fin d'études. Il s'agit d'entretiens à 533 informateurs, dont le lieu d'origine était Bruxelles ou la Wallonie, entretiens qui ont été enregistrés sur support audio, parfois aussi sur support vidéo.

La plupart des corpus ont été donc réalisés à l'occasion de recherches orientées vers des thématiques précises telles que des enquêtes sur l'accent, sur les représentations linguistiques, sur la liaison, même si ces sous-corpus peuvent être exploités à des fins diverses. Chaque entretien est accompagné de fiches d'identification qui précisent les caractéristiques sociologiques de chaque locuteur et les caractéristiques de l'interaction elle-même.

[22] On peut retrouver ces informations à l'adresse http://valibel.fltr.ucl.ac/be/corpus.htm consultée le 12 janvier 2006.

Tous les entretiens ont été transcrits orthographiquement et leur numérisation est en cours, afin de permettre une consultation aisée des corpus alignés.

Ces corpus ne sont pas accessibles en ligne actuellement mais certaines bandes peuvent être consultées sur demande en prenant contact avec francard@frwa.ucl.ac.be.

4.2.3. C-ORAL-ROM

Le C-ORAL-ROM (*Corpus for Spoken Romance Languages*)[23] est une ressource produite dans le cadre d'un projet de l'Union Européenne qui a commencé en 2001 et qui a eu une durée de trois ans et demi. Il s'agit d'un corpus oral multilingue composé de discours formels et informels qui constituent quatre collections comparables de données en français, espagnol, italien et portugais. Chaque partie du corpus a environ 300 000 mots pour un total de 1 200 000 mots[24].

L'objectif de cette collection est celui de représenter une variété d'actes langagiers actualisés dans la vie quotidienne, pour étudier les structures syntaxiques et prosodiques d'un point de vue qualitatif et quantitatif. Chaque composante a une histoire de constitution différente et intéressante et, pour être acceptée comme élément du corpus, il a fallu l'adapter aux critères et aux consignes dictées par les idéateurs du projet, afin d'avoir un corpus homogène.

Nous nous limiterons à analyser plus en détail la composante française qui a été produite par le Laboratoire DELIC (*Description linguistique Informatisée sur Corpus*, dirigé par J. Véronis) de l'Université de Provence [Cresti & Moneglia 2005, XI]. Le corpus français se composait à l'origine du CORPAIX (*Corpus de Français Parlé*) collecté depuis 1978 à Aix-en-Provence qui contenait trois types de textes avec une cassette des enregistrements originels. Ce corpus présentait des difficultés d'ordre légal car les chercheurs qui ont fait

[23] Voir également Moneglia & Cresti dans le présent recueil, 00-00.

[24] Le corpus dans cette dimension a été publié par Benjamins [Cresti & Moneglia 2005]; en réalité, les quatres corpus sont beaucoup plus importants et ils sont distribués sans restrictions par ELDA (*European Language Distribution Agency*).

les enregistrements n'ont pas demandé aux personnes interviewées la permission d'utiliser les données personnelles. En effet, on n'a pu utiliser que 18 textes sur un total de 570.

Les données de la deuxième composante ont été enregistrées entre 1999 et 2000, le *Corpus de Référence de Français Parlé*, pour lequel on a demandé systématiquement le consentement légal. Le corpus se base sur des interviews classées selon l'âge, le niveau d'éducation et le type de situation (privée, professionnelle, publique). Cependant, sur 134 éléments faisant partie de cette composante, seuls 31 ont pu être utilisés dans le projet C-ORAL-ROM, car les critères de constitution utilisés pour le CRFP n'étaient pas complètement compatibles avec ceux du projet.

Les chercheurs ont donc dû créer un corpus additionnel qui constitue, en réalité, la majorité des données du C-ORAL-ROM français (115 textes pour 70% de la totalité du corpus).

Le corpus se compose de plusieurs textes qui font partie de différents domaines et qui concernent différents genres de la langue parlée. Tout d'abord, nous avons une subdivision entre langue formelle et informelle. Dans la langue formelle nous pouvons analyser la langue des médias, la langue parlée au téléphone (interaction homme-machine ou conversations privées) ou dans des contextes naturels (conférences, débats publiques, discours politiques, enseignement, explications professionnelles, etc.). Dans la langue informelle nous trouvons une subdivision entre la langue parlée dans le privé et dans le public, ces deux catégories étant subdivisées ultérieurement en conversations, dialogues et monologues.

Le corpus C-ORAL-ROM est disponible aux éditions Benjamins [Cresti & Moneglia 2005].

4.3. Ressources entre l'écrit et l'oral

Avec l'avènement de nouveaux moyens de communication, on assiste au développement de nouvelles façons de communiquer basées sur des systèmes langagiers qui se détachent sensiblement de la langue utilisée dans la communication standard. L'intérêt donc que ces nouvelles formes suscitent dans la communauté linguistique est

tellement grand que certains chercheurs ont décidé de collecter des ressources pour l'étude de nouveaux langages, notamment les messages textos et le tchat, ressources que nous allons décrire dans les paragraphes suivants.

4.3.1. Corpus de français tchaté

A partir de 2004, dans le cadre d'un stage de deuxième année de master de recherche à l'Université de Grenoble 1, Achille Falaise, doctorant en informatique, a collecté un corpus de français tchaté afin d'évaluer les difficultés posées par la langue du tchat à son traitement automatique [Falaise 2005]. Ce corpus est constitué d'enregistrements des sessions de discussion, disponibles sur le site http://www.botstats.com[25], qui sont archivés pour une durée maximale de trois mois. L'auteur du corpus a choisi de collecter les enregistrements des canaux qui ont été rendus accessibles à tous[26] par leurs créateurs. Une fois ces enregistrements collectés sous format HTML, ils ont été «convertis automatiquement en XML grâce à des expressions régulières, afin de pouvoir être exploités» [Falaise 2005].

La quantité des données collectées à partir des enregistrements est remarquable: en effet, avec 4 192 033 messages, soit 23 millions de mots environ, on dispose de la plus grande ressource représentative de la langue utilisée dans un tchat[27]. D'un point de vue qualitatif, le corpus comporte une variété assez grande de canaux où les sujets de discussion peuvent aller du tchat généraliste au tchat spécifique (tel que celui concernant des problèmes de programmation), ou encore au tchat relatif à l'actualité.

[25] Les canaux de tchat sont des serveurs où il existe des plateformes pour l'échange de messages instantanés, pour le partage de photos ou de documents, pour participer à des vidéo-tchats et à des forums. Le site publie les résultats du service web Botstats dédié aux canaux de tchat.

[26] En effet, chaque créateur de canaux de discussion sur un tchat, peut décider de publier les enregistrements des discussions ou bien de ne les rendre consultables qu'aux utilisateurs enregistrés.

[27] L'autre ressource de langue tchatée que l'auteur a pu consulter est le corpus d'italien tchaté constitué par Eulogos (http://www.intratext.com/X/ITA0192.HTM) qui compte 800 000 mots environ.

Pour que le corpus soit accessible à tous ceux qui s'intéressent aux analyses de la langue du tchat, il faut effectuer des modifications sur le corpus pour ce qui concerne les canaux, dans la mesure où chaque canal comporte un sujet différent, mais aussi des «spécificités pragmatiques» [Falaise 2005]. En outre, l'auteur envisage de compléter le corpus avec une annotation lexicale plus fine, pour identifier les particularités graphiques qui lui sont propres.

Une version démo du corpus est disponible en ligne à l'adresse http://www-clips.imag.fr/geta/User/achille.falaise/corpustchat/.

4.3.2. Corpus de textos[28]

C'est le *Center for natural language processing* de l'Université de Louvain en Belgique qui a collecté un corpus de 75 000 messages écrits sur des portables par 2 400 auteurs de 12 à 65 ans. Une partie de ce corpus a été publiée par les Presses Universitaires de Louvain [Fairon *et alii* 2006].

Il s'agit d'un des plus grands corpus de SMS existant à l'heure actuelle qui contient des messages collectés en Belgique francophone dans le cadre du projet *Faites don de vos SMS à la science* sous l'égide de deux centres de recherche de l'Université de Louvain, le CENTAL (*Centre de traitement automatique du langage*) et le CELEXROM (*Centre d'études sur le lexique roman*)

Cette ressource est intéressante d'un double point de vue:

- sur le plan des profils sociolinguistiques: on a accès aux informations concernant le profil des auteurs des messages, à savoir l'âge, le sexe, la provenance, le travail, la langue maternelle et la deuxième langue, etc. Il est possible à tout moment d'accéder aux informations qui permettent de sélectionner un sous-corpus sur la base des détails sociolinguistiques donnés, qui deviennent ainsi des critères de recherche;

- sur le plan des transcriptions: les messages ont été transcrits dans un français normalisé de façon à pouvoir chercher les va-

[28] Ces informations sont disponibles à l'adresse http://www.smspourlascience.be [consulté le 1er septembre 2006].

riantes d'un mot à l'intérieur du corpus, grâce au logiciel d'interrogation qui visualise, en outre, les textos originaux où sont présentes ces variantes graphiques.

Conclusion: pour un usage informatique des ressources langagières du français

Comme nous l'avons mis en évidence dans les paragraphes précédents, il existe de nombreuses ressources pour l'analyse de la langue française, ressources collectées dans des contextes de recherche variés et pour diverses finalités. Ces ressources sont disponibles parfois gratuitement, parfois moyennant un abonnement, dans la plupart des cas elles le sont sur support électronique. Il n'existe cependant pas de ressource qui les rassemble toutes comme on peut le constater pour l'anglais (nous pensons au *British National Corpus* ou à la *Bank of English*, qui sont des corpus de référence sur lesquels on peut mener des analyses pour la description générale de la langue).

Des tentatives de constitution de grands corpus français ont été menées dans les dernières années, notamment avec le projet ANANAS[29] pour la création d'une base de données linguistiques en français contenant de textes libres de droits et annotés (selon un découpage structurel et un étiquetage morphosyntaxique), accessible en ligne et conçue pour l'étude des anaphores. Mais on a dû constater une série de problèmes du moment que les annotations des différentes composantes avaient été effectuées sur la base de schémas très différents, que les phénomènes annotés étaient hétérogènes et que seul un corpus présentait une annotation multiniveaux systématique.

Une fois ces problèmes constatés, du fait que l'annotation des différents corpus en partant quasiment de zéro est très coûteuse, une équipe de chercheurs de l'ATILF-CNRS et LORIA a commencé à travailler à la constitution d'un ensemble de corpus français comprenant des textes libres de droits et diffusables en ligne, ainsi qu'à leur annotation multiniveaux (structurelle, morpho-syntaxique, syntaxique, coré-

[29] http://www.inalf.fr/atilf/ananas.

férence et anaphores) suivant des normalisations standards (TEI). En outre, cette ressource est conçue comme une ressource évolutive dans la mesure où peuvent s'y ajouter d'autres parties validées par un comité d'édition. Elle couvre une variété de genres tels que les textes littéraires, journalistiques, scientifiques, constats d'accidents, recueils de dialogues téléphoniques, monologues[30].

Cette ressource, cependant, est encore loin de constituer un corpus de référence pour la langue française, mais elle le deviendra, nous espérons dans un avenir proche, fournissant ainsi une base de données exploitables par la communauté scientifique. Pour parvenir à ce but, une collaboration plus étroite entre les différentes équipes doit être envisagée, en pensant à l'intérêt commun de disposer d'un outil précieux pour l'analyse de la langue française.

BIBLIOGRAPHIE

[Aarts 2002] Aarts J., «Does corpus linguistics exist? Some old and new issues», in Breivik L. E. & Hasselgren A.(ed. by), *From the COLT's mouth ... to others'*, Amsterdam, Rodophi, 2002, 1-17.

[Biber 1988] Biber D., *Variation across speech and writing*, Cambridge, Cambridge University Press, 1988.

[Cresti & Moneglia 2005] Cresti E. & Moneglia M. (ed. by), *C-ORAL-ROM Corpus for Spoken Romance Languages*, Amsterdam/Philadelphia, John Benjamins Publishing, 2005.

Derrock M. & Flament-Boistrancourt D., «Le corpus LANCOM: bilan et perspectives», *ITL-Review of Applied Linguistics*, 111-112, 1996, 1-36.

[EAGLES 1996] EAGLES, *Preliminary recommendations on corpus typology*, disponible sur le site http://www.ilc.cnr.it/EAGLES96/intro.html [consulté le 19 novembre 2006].

[Eurovoc 1995] Eurovoc, «Thesaurus Eurovoc – Volume 2: Subject-Oriented Version. Ed. 3/English Language», in *Annex to the index of the Official Journal of the EC*, Luxembourg, Office for Official Publications of the European Communities,

[30] L'ensemble complet des parties du corpus disponibles se trouve à la page http://freebank.loria.fr/corpus.php, [consulté le 23 janvier 2007].

1995, disponible sur le site http://europa.eu./eurovoc, [consulté le 21 novembre 2006].

Fairon C., «Parsing a Web site as a corpus», in Fairon C. (éd. par), *Analyse lexicale et syntaxique: Le système INTEX, Linguisticae Investigationes*, 22, 1999, 327-340.

[Fairon *et alii* 2006] Fairon C., Clein J. & Paumier S., *Le Corpus SMS pour la science. Base de données de 30 000 sms et logiciels de consultation, CD-Rom*, Louvain La Neuve, Presses Universitaires de Louvain, Cahiers du Cental, 2006.

Fairon C., Clein J. & Paumier S., «A translated corpus of French SMS», in *Proceedings of the fifth International Conference on Language Resources and Evaluation, LREC* 2006, Genoa.

[Falaise 2005] Falaise A., «Constitution d'un corpus de français tchaté», in *Actes de RECITAL 2005*, Dourdan, disponible en ligne à l'adresse http://www-clips.imag.fr/geta/User/achille.falaise/publis/AF-Article_RECITAL_2005.pdf, [consulté le 22 février 2007].

[Firth 1957] Firth J. R., *Papers in Linguistics 1934-51*, Oxford, Oxford University Press, 1957.

[Habert *et alii* 1997] Habert B., Nazarenko A. & Salem A., *Les linguistiques de corpus*, Paris, Armand Colin/Masson, 1997.

[Kehoe & Renouf 2002] Kehoe A. & Renouf A., «WebCorp: Applying the Web to Linguistics and Linguistics to the Web», in *Proceedings of the WWW2002 conference*, Honolulu, 2002, disponible sur le site http://www2002.org/CDROM/poster/67/, [consulté le 21 novembre 2006].

[Kennedy 1998] Kennedy G., *An Introduction to Corpus Linguistics*, London, Longman, 1998.

[Kraif *et alii* 2006] Kraif O., El-Bèze M., Meyer R. & Richard C., «Le corpus Carmel: un corpus multilingue de récits de voyage», in *Actes de TALN-RECITAL 2006, Atélier Technolangue*, Louvain la Neuve, disponible sur le site http://test.sinequa.com/carmel/publis/Carmel-TALN-2007.pdf [consulté le 16 décembre 2006].

[Mertens 2002] Mertens P., «Le corpus de français parlé ELICOP: consultation et exploitation», in Binon J., Desmet P., Elen J., Mertens P. & Sercu L. (éd. par), *Tableaux Vivants*, Louvain la Neuve, Presses Universitaires, 2002, disponible à l'adresse http://bach.arts.kuleuven.be/pmertens/papers/elicop.pdf, [consulté le 19 juillet 2006].

[Pearson 1998] Pearson J., *Terms in Context. Studies in Corpus Linguistics*, Amsterdam / Philadelphia, John Benjamins, 1998.

[Phillips 1983] Phillips M.K., *Lexical Macrostructure in science text*, PhD Thesis, University of Birmingham, 1983.

[Richard *et alii* 2006] Richard C., Meyer R. & El-Bèze M., «Projet CARMEL: récits de voyages», in *Actes de TALN-RECITAL 2006, Atelier Technolangue*, Louvain la Neuve, disponible sur le site http://test.sinequa.com/carmel/publis/Carmel-TaLC-2006.pdf, [consulté le 16 décembre 2006].

[Sinclair 1991] Sinclair J., *Corpus, concordance and collocation*, Oxford, Oxford University Press, 1991.

Steinberger R., Pouliquen B., Widiger A., Ignat C. Erjavec T., Tufi D. & Varga D., «The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages», in *Prooceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Genova, disponible sur le site http://langtech.jrc.it/Documents/0605_LREC_JRC-Acquis_Steinberger-et-al.pdf, [consulté le 22 février 2007].

[Teubert 1996] Teubert, W., «Comparable or Parallel Corpora?», *International Journal of Lexicography*, 9/3, 1996, 238-264.

[Tognini-Bonelli 2001] Tognini-Bonelli E., *Linguistics at Work*, Amsterdam-Philadelphia, John Benjamins, 2001.

[Vignaux *et alii* 1999] Vignaux G., Attali A., Augier M., Jardin P., Piotrowski D. & Silberstein M., «Le projet Colisciences», *MEDIATEC*, CNRS, disponible sur le site www.colisciences.net/pdf/projet.pdf, [consulté le 19 juillet 2006].