Les grands corpus du français moderne

Inka Wissner

Inka Wissner

Les grands corpus du français moderne : des outils pour étudier le lexique diatopiquement marqué ?[1]

Abstract

In modern French dialectology, the critical usage of text data for the appropriate description of diatopic (regional) elements of national or regional varieties of French – though crucial – is hardly commented in scientific discourse. The paper offers an overview on the main corpora that are, or could be used for the lexical study of diatopicisms in contemporary French in the northern part of the francophone world, whether it be traditional, mostly literary corpora, journalistic texts, recent web corpora, or 'oral' corpora. The author examines their suitability for identifying diatopicisms with their spatial distribution in a francophone perspective, presents current methods of text analysis, and points out various problems related in particular to text location, regional annotation, discourse genres and treated subjects, but also to formal compatibility, text size and practical access. The article therewith underlines the need, in a close future, of comparable data for the different varieties in the francophone world in order to enhance the description of diatopic variation in modern French.

1. Introduction

La lexicologie française dispose depuis le tournant des XX^{e}/XXI^{e} siècles de grands dictionnaires différentiels de régionalismes, ou *diatopismes* (comme DSR, DHFQ, DRF et DictBelg) outre une base de données lexicographiques panfrancophone (BDLP) – conçus comme compléments des dictionnaires généraux. Leur préparation ayant impliqué la constitution et l'exploitation de bases de données textuelles, plusieurs questions

[1] Notre travail a bénéficié du regard critique de l'éditeur et de nos collègues France Lagueunière et Christophe Benzitoun, spécialistes de la variation diatopique pour l'une, et de corpus 'oraux' pour l'autre, ainsi que d'Alain Polguère pour le projet RLF, présenté ci-dessous.

s'imposent pour l'étude du lexique du français contemporain sous l'angle de sa variation dans l'espace, c'est-à-dire dans sa dimension diatopique :

> – Quels sont les grands corpus existants et exploités en ce domaine ?
> – Quels sont les problèmes liés à leur consultation, en termes pratiques d'accès aux bases, des données et métadonnées qu'ils fournissent, et des types de recherche qui sont possibles ?
> – Quels sont les besoins en lexicologie variationnelle pour une meilleure identification et description des diatopismes du français contemporain à travers la francophonie ?

Pour tenter de répondre à ces questions, on rappellera d'abord quels sont les corpus traditionnels de la lexicographie différentielle. On s'interrogera ensuite sur la possibilité et les besoins d'intégrer d'autres corpus, existants ou en préparation, qui pourraient permettre de mieux identifier les diatopismes dans une optique panfrancophone. On se concentrera surtout sur les plus grands corpus du français moderne de la francophonie du Nord, où l'on dispose déjà de corpus d'une certaine envergure.

La question du rôle des corpus est largement discutée dans de nombreux domaines de la linguistique contemporaine, et prend une ampleur particulière dans les travaux sur les réalisations médiatiques orales. Dans le champ disciplinaire de l'étude de la variation diatopique du français, bien établi en philologie romane depuis un peu plus de deux décennies (cf. Rézeau 2007), l'apport des corpus est toutefois peu thématisé. Les publications des spécialistes qui ont largement fait avancer les recherches en ce champ d'étude portent essentiellement sur l'exploitation de ressources lexicographiques. La problématique des corpus est notamment abordée du point de vue de leur constitution par des équipes en Belgique et au Québec et, du point de vue de leur exploitation – avec des optiques différentes – dans les articles de Queffélec (1997) et de Thibault (2007), surtout.

On a pu confirmer en lexicographie différentielle que l'exploitation de corpus est « de plus en plus sentie comme indispensable par la plupart des lexicographes et marque un renouveau de la méthodologie lexicographique » (Vézina 1998 : 228). Les corpus comme ensembles textuels complètent ainsi les matériaux traditionnels de la lexicologie variationnelle – qui sont surtout lexicographiques et très variables d'une aire de la francophonie à l'autre. L'accès croissant à des ensembles textuels permet en outre d'envisager à moyen terme d'identifier un diatopisme non pas (exclusivement) par rapport aux ressources métalinguistiques – surtout

des dictionnaires généraux, jouant le rôle de point de comparaison (cf. Poirier 2005 : 497) – mais aussi par rapport aux réalisations discursives effectives dans les diverses variétés diatopiques du français, par l'intermédiaire d'une analyse comparative des corpus.

Leur analyse présuppose bien entendu d'être familiarisé avec la situation sociolinguistique dans les aires de la francophonie dont on souhaite rendre compte. Nous mettons ici de côté la discussion de l'apport d'informateurs et de lexicographes locuteurs des différentes variétés diatopiques pour nous concentrer sur le problème de l'exploitation de corpus. Les réflexions proposées se situent dans le cadre de l'évaluation des possibilités d'une prise en compte de la variation diatopique au sein d'un projet dictionnairique qui porte sur le français contemporain, le Réseau Lexical du Français, préparé depuis l'été 2011 au CNRS (ATILF) sous la direction scientifique d'A. Polguère (cf. Lux-Pogodalla & Polguère 2011).

Si le terme de *corpus* désigne un ensemble d'éléments sur lequel se fonde l'étude du phénomène linguistique, il est utilisé en lexicographie différentielle pour renvoyer à un ensemble de ressources (méta-)linguistiques – d'où la notion de 'corpus d'exclusion' (cf. Francard 2001 : 228–230) – mais aussi à des ensembles de réalisations discursives. Ces derniers peuvent avoir été constitués à des fins d'analyse spécifiques, dans le respect de critères de sélection particuliers, comme la base FRANTEXT, mais aussi à d'autres objectifs que la recherche linguistique, comme EUROPRESSE[2]. Les corpus peuvent accueillir des textes dans leur intégralité (comme SCIENTEXT), ou rassembler des extraits d'un nombre limité de propositions phrastiques, pour ce qui est des corpus de citations (comme le FLI, au Québec).

La plupart des corpus existants du français n'ont pas été conçus pour l'analyse de diatopismes, et leur constitution ne s'appuie souvent pas sur une conception claire et actualisée de cette dimension de la langue – d'où résulte toute une série de difficultés d'exploitation pour les analyses à visée diatopique.

Rappelons ici que nous considérons que le français représente une langue historiquement standardisée (bien distincte de variétés historiquement liées comme des créoles à base lexicale française). Elle dispose d'un système à plusieurs modalités, dont la variabilité est

[2] Pour les sigles des corpus, banques de données, plateformes et portails, voir la bibliographie et l'annexe.

inhérente. En accord avec les avancées de la recherche en linguistique variationnelle des langues romanes, nous considérons aussi que la variation diatopique, dans l'espace – qualitative et quantitative – est enchevêtrée avec d'autres facteurs diasystémiques, qu'ils soient diastratiques (socioculturels), diaphasiques (situationnels), mais aussi diachroniques (temporels), diacodiques (dépendant du canal communicatif), ainsi que diamésiques, selon la conception des énoncés (cf. Gadet [2]2007 : 23, 47–49).

Si nous envisageons l'étude du français dans une optique francophone, la francophonie – au sens linguistique – est ici considérée comme l'ensemble de l'espace où le français est langue d'usage, et où il est souvent langue officielle – que ce soit en Europe, en Amérique du Nord et du Sud, en Afrique subsaharienne et au Maghreb, ou encore dans l'Océan Indien, l'Extrême Orient et le Pacifique. Nous ne nous limitons pas à l'étude du français comme langue 'maternelle', terme qui renvoie à un concept inadapté selon de nombreux chercheurs, notamment en dehors de la Métropole (Gadet et al. 2009 : 152*sq*.). Dans l'espace francophone, l'usage de la langue varie en effet fortement selon une diversité de facteurs, y compris la situation sociolinguistique de chaque communauté – et donc les réseaux sociaux et les modes d'acquisition du français – mais aussi sa fréquence d'emploi et son statut (dominant ou dominé) par rapport à des variétés en contact, ou les contraintes systémiques internes de chaque variété diatopique (cf. Gadet & Jones 2008 : 244*sq*.).

En se focalisant sur les variétés diatopiques les mieux dotées en matière de corpus, dans la francophonie du Nord (v. ci-dessous), la présente étude porte toutefois avant tout sur des aires où le français a un statut de langue dominante, représente une langue par tradition, et est un outil de communication quotidien.

2. Des sources traditionnelles

Pour étudier la dimension diatopique du lexique du français contemporain, on dispose de corpus traditionnels qui portent sur les variétés les mieux décrites, c'est-à-dire le français en Belgique (Wallonie, Bruxelles), en France hexagonale, au Québec, et en Suisse romande.

2.1 Des corpus constitués dans le cadre de projets dictionnairiques

Parmi les corpus traditionnels, on dispose avant tout de FRANTEXT – « incontestablement le plus grand corpus de français » (Gadet 2007 : 20), c'est-à-dire parmi les collections ordonnées de textes français, monolingues, publics, et accessibles à la communauté scientifique. Constitué pour l'étude du français général (ou de ce qui est considéré comme 'le français tout court'), il a donné lieu à la publication du dictionnaire de référence de français sur corpus, le TLF.

S'y joignent des corpus inspirés de FRANTEXT, élaborés pour l'étude de variétés diatopiques spécifiques dans le cadre de la préparation des dictionnaires différentiels de diatopismes en Europe et en Amérique du Nord, depuis le dernier tiers du XX^{e} siècle. Il s'agit tout d'abord de QUEBETEXT de Québec, qui contient une petite partie des textes qui ont été dépouillés pour la préparation du dictionnaire DHFQ, et en tant que source textuelle brute complète un très riche corpus de citations (FLI), outre la base métalinguistique ILQ. On dispose également de corpus pour chacune des trois grandes variétés diatopiques en Europe : BELTEXT (cf. Delcourt et al. 1993) élaboré pour la Belgique à Liège, Mons et Louvain-la-Neuve (ayant donné lieu à l'ouvrage de Delcourt 1998–1999), SUISTEXT, conçu à Neuchâtel pour le français en Suisse (DSR), puis REGION, élaboré pour l'étude de diatopismes en France à Nancy pour la publication du DRF. À ces bases traditionnelles s'ajoutent pour l'Amérique la base BDTS (cf. Cajolet-Laganière et al. 2008, Masson et al. 2007) et MCVF (ci-dessous), outre un petit corpus pour un espace entre l'Amérique du Nord et du Sud, les Antilles (ZOBEL).

Toutefois, l'envergure des bases varie énormément, allant de la totalité de l'œuvre de quatorze écrivains romands dans la base suisse, et de 7500 citations de 220 ouvrages littéraires dans la base REGION (ou des onze romans d'un écrivain dans ZOBEL), à 248 millions de mots dans FRANTEXT – en passant par plus de 52 millions de mots dans la récente base BDTS. Si dans d'autres domaines de la linguistique de corpus, il est habituel de faire des comparaisons d'usages à partir de tranches égales (p.ex. mille ou dix mille mots), on ne dispose pas de la possibilité technique de constituer des sous-corpus par nombre de mots pour les corpus traditionnels, ni du nombre de mots pour toutes ces bases (taille p.ex. non communiquée pour BELTEXT). Une telle comparaison serait aussi problématique vu l'hétérogénéité chronologique des corpus, qui portent sur le français d'époques différentes – du XIX^{e} au XXI^{e} siècles (REGION, SUISTEXT,

ZOBEL), du XVI^e^ au XXI^e^ siècles (FRANTEXT et QUEBETEXT) ou du XI^e^ au XIX^e^ siècles (MCVF) – ou exclusivement des années 1960 à nos jours (BDTS).

Pour ce qui est de l'aspect énonciativo-discursif, on sait que de tels corpus recueillent essentiellement du discours à dominante littéraire, élaboré, écrit. Les plus petites bases sont exclusivement littéraires (REGION, SUISTEXT, ZOBEL), d'autres plus diversifiés, tous exclusivement écrits, et relèvent par exemple des domaines administratifs, juridiques, ou scientifiques. La base québécoise BDTS a même une ambition de 'représentativité des usages', qui doit se comprendre au sens de l'intégration de plusieurs types de textes. Si la question de la représentativité se pose pour tout corpus (p.ex. Kleiber 1978 : 65), on sait qu'elle part d'une notion complexe et délicate (v. Cappeau & Gadet 2007). Qu'est-ce qui est représentatif par rapport à quoi – en fonction de quels critères ?

En ce qui concerne la dimension géolinguistique, la base FRANTEXT porte implicitement surtout sur la langue française dans la Métropole. Parmi les écrivains hexagonaux, certains utilisent inévitablement des diatopismes eux aussi, en particulier (quoique non exclusivement) les auteurs qui sont réputés régionalistes, comme H. Vincenot ou R. Bazin. En outre, ce corpus contient aussi – surtout depuis récemment – des textes d'ailleurs, comme ceux dus à des écrivains québécois (comme Guèvremont), belges, réunionnais, ou antillais, tels que Chamoiseau ou Zobel. Les autres corpus, puisque constitués pour l'analyse du français dans sa variation diatopique, fournissent a priori des textes ancrés dans des aires spécifiques. Ils ne sont pas pour autant diatopiquement homogènes, le problème de l'hétérogénéité se posant pour tout discours. À titre d'exemple, le corpus québécois BDTS contient aussi des textes métalinguistiques qui portent sur le traitement lexicographique de diatopismes d'ailleurs (Cajolet-Laganière et al. 2008 : 22*sq.*).

Ces divers types d'hétérogénéité ne sont pas gênants en tant que tels, si l'on procède à une analyse philologique de chaque attestation d'une unité sous étude, et de chaque corpus individuel, comme ceci est habituel pour l'exploitation des dictionnaires (cf. Wissner 2010 : 104–149). De telles divergences deviennent toutefois problématiques dans le cadre de requêtes automatiques et fréquentielles, pour l'analyse d'unités en grand nombre ou de diatopismes qui affichent un nombre très élevé d'attestations, ainsi que pour une exploitation comparative de corpus pour identifier des diatopismes. Celle-ci présuppose en effet une certaine comparabilité selon

plusieurs critères – y compris quantitatifs, chronologiques, discursifs, et géolinguistiques.

Il y a également toute une série de problèmes qui sont liés à l'annotation des textes, à l'accès aux métadonnées (lorsqu'il y en a), et aux types de requêtes par critères que permettent les logiciels d'exploitation des bases qui accueillent les corpus. Les possibilités d'exploitation des corpus de textes intégraux et de citations sont différentes pour des recherches de type énonciativo-pragmatique et syntaxique surtout, quoiqu'à une moindre mesure pour des analyses diatopiques qui portent sur le lexique. En effet, les logiciels d'exploitation qui donnent accès aux corpus contenant l'intégralité de textes sous la forme de bases de données permettent de visualiser non pas ces ensembles, mais seulement des extraits textuels. La taille de ces extraits varie en outre d'une base à l'autre ; à titre d'exemple, pour FRANTEXT, elle est de l'ordre de 700 signes depuis 2012 (300 auparavant). On n'oubliera pas non plus qu'on n'accède pas, ontologiquement parlant, au même discours si on lit un texte comme un roman, en tant qu'ouvrage imprimé destiné à être lu et à divertir et intriguer le public (plus qu'à informer), ou si on y accède par l'intermédiaire d'un corpus pour des fins de recherche.

2.2 L'appartenance géographique des locuteurs : un attribut textuel traditionnel

L'identification géographique de locuteurs est un des paramètres classiques de la sociologie, et l'annotation des énoncés rassemblés dans des corpus selon la région de rattachement des locuteurs est un attribut textuel traditionnel. Il est très peu explicité en lexicographie française, contrairement à la lexicographie anglophone (Atkins et al. 1992 ; Atkins & Rundell 2008 : 89) – avec des paramètres comme le type de source (p.ex. 'informatif'), la catégorie de domaine (du type 'sport et loisir') ou le sous-domaine (*ib*.). Outre l'origine géographique des énonciateurs des textes, les lexicographes anglophones jugent également utile de préciser leur statut de locuteur maternel (« natif ») (*ib*.). Toutefois, la variabilité diatopique d'une langue n'est pas tant liée à son usage en tant que langue 'maternelle' (ou 'officielle'), qu'aux types de fonctions que lui réserve la communauté sociolinguistique (v. 1).

D'une manière générale, l'identification géographique de locuteurs s'appuie traditionnellement sur leur lieu de naissance et/ou de résidence, mais aussi sur les régions où ils ont vécu et qui les ont marqués

linguistiquement. Au cas où l'utilisateur de corpus (intéressé par la diatopie ou non) souhaiterait connaître les rattachements multiples des locuteurs des énoncés qu'il analyse, les corpus ne fournissent le plus souvent aucun balisage ou retiennent seulement un rattachement principal – qui, à lui seul, n'est alors pas nécessairement pertinent. En même temps, l'annotation diatopique de corpus implique des choix qui influent nécessairement sur l'interprétation des données. Les catégories géolinguistiques doivent en outre s'appuyer sur une répartition à jour des aires linguistiques du français – et non pas sur des limites politiques ou géographiques, ni sur des aires d'autres variétés de langue en contact. Or, si l'on prend l'exemple de la France, les catégories géolinguistiques dans la base REGION sont de type politique (p.ex. 'Maine-et-Loire'). Pour une catégorisation à jour, on pourrait désormais s'appuyer sur les quatre aires lexicales principales qui ont été décrites à partir du DRF : une aire occidentale et une aire orientale (passant dans le Sud par l'embouchure du Rhône), puis une aire septentrionale et une aire méridionale, qui inclut le Bordelais à l'ouest et le Lyonnais à l'est (cf. Wissner 2010 : 29) – donc schématiquement 'Nord-ouest', 'Sud-ouest', 'Nord-est' et 'Sud-est'.

Le balisage géolinguistique du discours est bien fourni dans les corpus qui visent explicitement l'étude de diatopismes, comme ce petit corpus REGION, mais aussi dans de nombreux corpus oraux (v. 3.3), dans des manuscrits anciens ou le récent corpus MCVF, et dans les plus grands corpus de référence d'autres langues européennes (v. 4). Au contraire, dans les grands corpus de français comme FRANTEXT, ce type de balisage n'a pas été entrepris – d'où l'impossibilité, pour l'utilisateur, de constituer des sous-ensembles diatopiquement pertinents. Pourtant, le critère géographique est nécessaire pour une interprétation adéquate des données – et ce pour toute étude linguistique, mais aussi pour d'autres disciplines, comme l'histoire ou la sociologie. Dans le cadre d'analyses à visée diatopique, le recours à de tels corpus implique que chaque utilisateur décrypte le discours pour dépister les diatopismes. Ainsi, dans FRANTEXT – que le lexicologue diatopicien n'utilise pas seulement pour identifier ce qui relève de l'usage général (v. Thibault 2007 pour les possibilités de son exploitation) – il faut alors recourir aux procédés traditionnels que sont, schématiquement,

> – l'identification des locuteurs à l'aide de biographies et d'autres travaux critiques, et
> – l'analyse discursive du cotexte immédiat et du contexte plus large.

On trouve alors des indices d'une possible diatopicité si une forme recherchée figure surtout dans le discours de locuteurs qui sont identifiables comme étant marqués par une variété diatopique d'une région donnée, tout en évaluant les liens historiques de cette variété avec d'autres variétés de français. L'analyse du cotexte et de la situation d'énonciation plus large permet, quant à elle, de tirer profit des mises en relief métalinguistiques dans le discours, comme les commentaires métalinguistiques, et d'autres indices qui peuvent expliciter la valeur géolinguistique d'un énoncé étudié – que ce soit par son attribution au discours d'un locuteur, par la nature du référent, ou par la localisation d'un récit. Bien entendu, l'attribution d'un énoncé par un locuteur à une communauté linguistique est à évaluer avec prudence : il s'agit de simples indices.

Dans le cadre d'analyses à visée diatopique, l'annotation géolinguistique des grands corpus faciliterait considérablement les requêtes automatiques pour des analyses qualitatives et quantitatives. Elle ne permet toutefois pas de faire l'économie d'une analyse philologique sérieuse du discours, en présence de l'hétérogénéité et de la polyphonie du discours – où un diatopisme peut aussi être utilisé par un locuteur exogène, ou être attribué aux propos d'autrui. En effet, en présence de ces caractéristiques inhérentes à tout discours, l'annotation géolinguistique du discours est à doubler d'un balisage des passages de discours rapporté. Ceci concerne autant les unités du discours tels que les discours direct ou indirect, que les extraits qui sont présentés comme cités à l'aide de mises en relief métalinguistiques. De tels passages renseignent en outre indirectement sur la conceptualisation que se fait l'énonciateur de discours d'autrui, passés ou futurs, effectifs ou possibles, qu'ils soient associés à l'oral ou non.

Ce type de balisage, bien fourni pour certains corpus de textes d'états anciens du français, n'a toutefois pas été proposé pour les corpus traditionnels présentés ci-dessus sauf pour la plus petite de ces bases, ZOBEL, où le texte est annoté selon des critères à la fois géographiques, discursifs et métadiscursifs : selon les unités de discours, les notes et les indices typographiques (italique, gras), et la langue des séquences (passages en créole, en anglais, etc.).

L'annotation (géolinguistique) du discours et de ses unités discursives est en effet coûteuse en temps, même sur des corpus écrits (raison pour laquelle ladite base n'a pas été élargie pour devenir un véritable corpus antillais). En outre, le balisage du discours rapporté est problématique : son identification s'appuie en partie sur une interprétation, surtout dans le cas

de superpositions de voix – difficulté qui est particulièrement aigüe dans les textes d'enregistrements de réalisations orales. Si certaines séquences qui sont présentées comme rapportées sont aisées à identifier, et donc à annoter – essentiellement le *discours cité direct* – on rencontre des cas de polyphonie hautement complexe lorsque l'attribution de la responsabilité aux différents énonciateurs impliqués est difficile, voire impossible, comme dans le *discours indirect libre*. Ce sont les avancées de la linguistique à l'intersection de l'analyse du discours et du traitement informatique de la langue qui faciliteront à moyen terme l'annotation du discours cité.

2.3 L'accès aux données et aux métadonnées

Pour être exploitable, le critère géographique doit bien entendu être accessible lors de la consultation du corpus à l'aide du logiciel d'exploitation, et figurer parmi les options de constitution de sous-corpus. Or, le balisage diatopique de bases comme REGION, pour la France, est seulement visible lors d'une consultation manuelle, et certaines bases comme SUISTEXT ou ZOBEL sont actuellement consultables seulement en format texte dans des fichiers séparés (tout comme TCOF ; v. 3.3). Parmi les corpus traditionnels, seulement FRANTEXT, QUEBETEXT et MCVF sont consultables avec des logiciels d'exploitation performants, et seul le dernier permet un accès direct à l'ensemble des textes qu'il contient ; il est aussi le premier à permettre de créer des sous-corpus pour mener des recherches dans des textes relevant d'aires spécifiques dans la francophonie.

Si les corpus BELTEXT, QUEBETEXT, REGION et SUISTEXT relèvent bien, en théorie, des sources de référence traditionnelles pour les lexicologues diatopiciens, c'est surtout pour des problèmes de diffusion qu'ils ne sont pas exploités de façon systématique pour étudier des diatopismes. Traditionnellement, les corpus étaient en effet constitués par des équipes pour usage interne ; pour consulter des corpus ouverts à la communauté scientifique, les chercheurs devaient alors se déplacer. Ainsi, le corpus REGION est consultable sur place uniquement, à Nancy, tandis que les corpus BELTEXT et SUISTEXT ne sont toujours pas rendus accessibles : pour le second, la demande de droits d'auteurs pour une diffusion libre n'a pas abouti. Au contraire, d'autres bases sont consultables sur Internet – librement, pour QUEBETEXT, qui ne contient que des textes libres de droits (complété du Fichier lexical FLI) ; sous abonnement pour FRANTEXT ; et sur demande auprès des auteurs, pour la récente base MCVF.

La consultation de corpus en ligne est en effet devenue un standard en linguistique moderne même si les modalités d'accès varient, et certains corpus récents sont conçus comme des bases de données dynamiques sur Internet – qui permettent aussi la gestion, l'archivage et la consultation des métadonnées. Ainsi, la diffusion en ligne est annoncée pour la BDTS, mais est effective surtout pour des bases autres que les corpus traditionnels à dominante littéraire, comme des ensembles journalistiques tel qu'EUROPRESSE, des corpus tirés du web, ou le corpus oral belge VALIBEL (v. 3).

3. Quelles sources (nouvelles) pour l'étude lexicale de diatopismes dans une approche comparative, panfrancophone ?

À côté de ces corpus de référence traditionnels, constitués de discours essentiellement écrits, littéraires, et élaborés, le lexicologue diatopicien peut s'appuyer sur d'autres genres de discours s'il veut tirer des conclusions valables sur l'usage au sein d'une variété diatopique donnée (pour la notion de genre, encore largement débattue, v. p.ex. Adam & Heidmann 2006). En linguistique variationnelle, on souhaiterait en outre disposer d'un ensemble de données discursives qui soit à la fois équilibré et panfrancophone (Dister et al. 2008 : 296–298 pour l'oral). Pour l'étude du français dans sa dimension diatopique, c'est ce qui permettrait d'identifier et de décrire des diatopismes à partir de données textuelles (v. 1), dans une approche comparative, proprement francophone.

Les critères à retenir pour l'établissement de corpus sont largement discutés dans les travaux de spécialistes d'enregistrements de réalisations discursives orales (p.ex. Baude 2006 ; Benzitoun & Cappeau 2010). Pour l'étude du lexique diatopiquement marqué, les principaux critères à stabiliser dans des corpus qui portent sur différentes variétés diatopiques du français me semblent être de type géolinguistique et chronologique, mais aussi thématique et discursif – deux catégories qui semblent plus pertinentes que la distinction oral/écrit, selon des analyses lexicales ciblées de spécialistes de corpus 'oraux' (p.ex. Benzitoun & Cappeau 2010 : 1390 et 1396). La comparabilité des corpus présuppose aussi une certaine homogénéité des méthodes et des problématiques visées lors de leur constitution, ainsi que des choix d'annotation et de logiciels de consultation – ces facteurs influant directement sur les possibilités d'exploitation des corpus.

Si l'on retient l'aspect discursif, en visant un certain équilibre des genres, on pourrait compléter les corpus traditionnels de corpus récents – comme POLITEXT, pour le discours politique français du XX^{e} siècle (donnant accès à dix millions de mots), ou encore SCIENTEXT, pour le discours scientifique, avec plusieurs millions de mots de textes de huit disciplines (parmi lesquelles la médecine, la linguistique et l'électronique). Ces bases sont toutefois franco-centrées, et non conçues pour l'analyse de diatopismes, même si elles peuvent apporter des renseignements utiles. On se concentrera ici sur des corpus qui existent pour plusieurs variétés, et sur des genres discursifs qui se prêtent à l'établissement de corpus pour différentes variétés diatopiques de français dans la francophonie tout en permettant d'atténuer le penchant des ressources traditionnelles vers le discours littéraire.

3.1 Les corpus journalistiques

Ce sont tout d'abord les corpus journalistiques qui sont déjà exploités avec profit en lexicographie différentielle. Les journaux ne sont pas communément considérés comme des 'viviers' de diatopismes, leur emploi y étant en principe peu intense, par rapport aux romans régionalistes traditionnels. On y en trouve toutefois bel et bien, notamment dans les chroniques de gastronomie, de loisirs ou de tourisme (cf. Thibault 2000 : 554 au sujet du *Monde*).

Ces corpus, d'abord rendus disponibles sous la forme de cédéroms puis aussi sous forme de bases de données, ne sont toutefois pas conçus pour des analyses linguistiques poussées : les cédéroms *Le Monde* et *Le Monde diplomatique*, comme les sites commerciaux en ligne EUROPRESSE ou EUREKA, ne permettent pas de faire des requêtes de lemmes ou de locutions, ni d'établir des sous-corpus. À l'exception d'EUREKA, ils ne fournissent pas non plus de métadonnées géolinguistiques, imposant au diatopicien de recourir aux procédés d'identification géolinguistique traditionnels (v. 2.2) – souvent plus difficile lorsqu'il s'agit d'énoncés de journalistes, puisqu'on ne dispose pas souvent de renseignements suffisants à leur sujet. Si le corpus EUREKA est consultable par langue, région, date de publication et domaine thématique, la requête par suites de mots est en partie inopérante – semblant fournir les attestations de la première entité graphique saisie – et les critères d'établissement des catégories géographiques sont de nature politique, et non pas linguistique (v. 2.2).

De façon générale, on observe toujours, pour l'utilisation de tels types d'outils, l'absence de concordances et de fonctions de recherche statistique, l'aspect « trop limité » des recherches de co-occurrences, et la « difficulté de désambiguïser rapidement et efficacement les résultats obtenus », mais aussi dans certains (comme EUROPRESSE) la redondance des résultats, en raison de la reprise de dépêches, qui posent problème en particulier pour des études fréquentielles (Thibault 2007 : 479). Pour ce qui est des archives de journaux – lorsqu'elles sont accessibles en ligne – elles sont dotées de logiciels d'exploitation encore moins adaptés aux besoins du linguiste.

Le discours journalistique est toutefois aussi désormais mis à profit dans plusieurs corpus qui sont conçus et annotés pour des recherches linguistiques, tels que le corpus hexagonal CERF, qui comporte dix tranches d'un million de mots de la presse nationale et autant de la presse régionale. Toutefois, le corpus de presse le plus grand de la langue française (complété d'un corpus littéraire d'Afrique noire) est actuellement la *Kölner romanistische Korpusdatenbank*, en finalisation à Cologne, dont la diffusion libre est prévue. Ce méga-corpus (80 millions de mots environ) conçu pour des recherches linguistiques – qualitatives et quantitatives, y compris de collocations – accueille des extraits de journaux publiés dans les années 2000, répartis en parts égales entre la presse francophone européenne, surtout de France (*Le Figaro*, *L'Est Républicain* et *Sud-Ouest*), et la presse nationale des pays francophones d'Afrique, surtout d'Afrique noire (v. Annexe).

Le discours journalistique étant en effet relativement facile d'accès grâce à l'informatisation, il est en principe possible, et souhaitable, de constituer un corpus journalistique avec des données comparables pour les différentes zones de l'espace francophone. En France, la plateforme CNRTL met actuellement à disposition un corpus journalistique régional de France – des extraits de trois années des éditions régionales et locales de *L'Est Républicain* (1999, 2002 et 2003) – conçu lui aussi pour des analyses linguistiques. Pour établir un corpus journalistique équilibré du français en France se pose alors le problème de l'équilibre géolinguistique, chronologique (et donc partiellement thématique), et quantitatif. En s'appuyant sur une répartition à jour des aires linguistiques, il faudrait disposer de corpus pour les quatre grandes aires en France, grosso modo le Nord-ouest, Nord-est, Sud-ouest, et le Sud-est (v. 2.2). En se focalisant sur les quotidiens, on pourrait alors retenir (comme pendant de l'*Est Républicain*) la presse régionale *Ouest-France* pour l'Ouest septentrional, qui a une très large couverture (Bretagne, Normandie, Pays de la Loire, qui

comprend la Vendée mais non les Charentes), pour la France méridionale surtout occidentale *La Dépêche du Midi*, et le journal du Sud-est : *Le Dauphiné Libéré*, couvrant les régions Rhône-Alpes et Provence-Alpes-Côte d'Azur (incluant donc le Lyonnais). Bien entendu, ces quatre quotidiens ne couvrent pas toutes les régions de France, et on garderait une certaine non-adéquation dans les résultats puisque la couverture de chacun d'entre elles ne se recoupe pas exactement avec la répartition des aires lexicales.

3.2 Les corpus construits à partir du web

Vu les difficultés générales d'accès aux corpus de référence traditionnels de la lexicographie différentielle, et vu les problèmes d'exploitation efficace des corpus journalistiques actuellement disponibles pour l'analyse de diatopismes, de nombreux lexicologues diatopiciens recourent de plus en plus à GALLICA et Google Recherche de Livres (GRL) – qui donnent accès à des textes numérisés – mais aussi au web lui-même, le plus souvent par l'intermédiaire du moteur de recherche Google.

On sait toutefois que « récupérer du corpus sur le net soulève des écueils » (Cappeau & Gadet 2007 : 108), que la taille des corpus ne résout pas tous les problèmes, et que l'utilisation de tels corpus est discutable. Le statut du web comme corpus est ainsi largement débattu surtout en linguistique computationnelle, et du point de vue lexicographique gagne peut-être plutôt à être conçu comme une source de textes, à partir de laquelle peuvent être établis des corpus (Atkins & Rundell 2008 : 78). Par ailleurs, si le web a pour avantage d'être a priori exploitable pour tout chercheur ayant accès à Internet, le corpus évolue très vite, et les données fournies par le moteur de recherche varient d'un ordinateur à l'autre – en fonction des habitudes de requêtes de ses utilisateurs.

L'exploitation critique du web ou des bases GALLICA et GRL présuppose bien entendu de prendre les mesures de prudence nécessaires dans l'analyse philologique des textes, en analysant exclusivement les énoncés dont la localisation – à l'aide des procédés classiques (v. 2.2) – est suffisamment probable, et en tenant notamment compte du contexte de production des textes que l'on exploite, mais aussi des implications ontologiques et techniques de leur consultation via le Net. Les difficultés d'exploitation des bases évolutives GALLICA et GRL, largement discutées ailleurs, relèvent entre autres de problèmes de traçabilité des formes dans les textes (notamment anciens), dont certaines ne sont pas reconnues par les

logiciels de traitement utilisés, et donc inaccessibles pour l'utilisateur ; d'autres problèmes sont liés à la fiabilité des métadonnées mises à disposition, comme la datation des textes (GRL retenant p.ex. souvent, pour les journaux, l'année de publication du premier fascicule). Il est alors de mise de critiquer, croiser et contrôler les données recueillies (Brochard 2012 pour différentes techniques).

GRL et GALLICA permettent de créer des sous-corpus selon des critères comme l'époque de publication, mais ne tiennent pas compte du paramètre diatopique – tout comme le moteur de recherche Google, pour le web. Si le groupe de Google vise à enregistrer un maximum de renseignements (y compris géographiques) sur les utilisateurs du web à partir de données fournies par les usagers eux-mêmes (comm. pers. de Mairi Mc Laughlin, Berkeley University, le 07/09/2011), se pose non seulement le problème de la fiabilité de ces métadonnées, mais aussi celui des catégorisations géolinguistiques, et des appartenances multiples des locuteurs (v. 2.2).

Cependant, aussi bien GALLICA, GRL que le web en général ont pour grand avantage de donner accès à des textes de diverses régions de la francophonie et d'époques différentes. Leur exploitation semble alors justifiée si l'on y cherche justement ce qui est introuvable ailleurs : notamment un ensemble textuel de très grande taille, mais aussi et, avant tout, des usages mal attestés dans les sources traditionnelles. Le web contient en outre des types de discours propres (comme les blogs), ainsi que des textes variés en termes de sujets abordés et de situations d'énonciation qui y sont représentées – formelles et informelles, privées, publiques autant que professionnelles (Atkins & Rundell 2008 : 80 pour le web en anglais).

Il existe en outre depuis peu des méga-corpus construits à partir du web pour plusieurs langues européennes, dans le cadre du projet italien *WaCky wide web*, y compris pour le français : le FRWAC, basé sur les URL en « .fr », d'une taille d'environ 1,6 milliard de mots-occurrences. Les extraits de cette base – constituée de l'image du web à un moment donné (en 2009 ? ; date non précisée) – ne sont pas nécessairement accessibles ultérieurement par l'intermédiaire d'un moteur de recherche comme Google. Par rapport au web, ce corpus est plus restreint, ce qui évite de noyer l'utilisateur sous des quantités de résultats souvent faramineuses, et est non-évolutif – et donc consultable par d'autres chercheurs pour des fins de vérification, ou des études comparatives.

Toutefois, comme le web lui-même, le FRWAC inclut des redondances (aussi 3.1), et est hétérogène dans le sens où les domaines (URL)

accueillent des sites de tout ordre, y compris en d'autres langues – outre les passages en discours cité (v. 2.2). En outre, s'agissant d'un corpus constitué par interrogation du web avec des paires de mots à l'aide d'un logiciel de traitement informatique (cf. Sharoff 2006), la recherche de lemmes avec le moteur de recherche permet d'accéder seulement aux formes dans le corpus qui sont reconnues par les dictionnaires sur lesquels se sont appuyés les auteurs du corpus. Le FRWAC est aussi désormais disponible sous une version conviviale catégorisée par degré de normativité orthographique et grammaticale, conçue et exploitée en lexicographie française générale, pour le projet du Réseau Lexical du Français en cours à l'ATILF (v. 1). Reste à voir si ce corpus sera exploité de façon plus générale, comme le web, y compris en lexicologie variationnelle.

Vu la taille de FRWAC et sa nature hétérogène, il est quasiment impossible de l'annoter (et indexer) selon les paramètres traditionnels qui sont pertinents pour une exploitation critique, en particulier les critères énonciativo-discursifs (genres, thèmes) et le facteur diatopique (rattachement géographique des locuteurs). Les auteurs de FRWAC ont certes essayé de construire des corpus « relativement homogènes » (Baroni et al. 2009 : 15), en se limitant par exemple au domaine « .uk » pour l'anglais britannique, afin d'exclure

> « les problèmes théoriques et méthodologiques concernant l'inclusion ou l'exclusion de variétés où une langue a un statut officiel mais non maternel » (*ib.*).

Toutefois, la distinction entre des variétés à statut officiel *vs* 'maternel' n'est pas nécessairement pertinente (v. 1 et 2.2), et le choix exclusif d'un seul domaine ne permet pas d'assurer l'homogénéité géolinguistique du corpus, puisque les données peuvent ne pas relever de la variété de rattachement officielle du site où elles figurent.

Pour des analyses lexicales à visée diatopique, il serait possible de recourir au corpus francophone I-FR de Leeds, constitué par la même équipe que FRWAC et selon les mêmes procédures, mais contenant environ 200 millions de mots, de tous les domaines francophones. Il aurait pour avantage d'avoir été annoté thématiquement, selon des critères qui comprennent des renseignements géographiques – même s'il ne s'agit pas de catégories proprement géolinguistiques. L'un des objectifs pour l'analyse du lexique diatopiquement marqué étant de disposer de corpus comparables pour différentes zones de la francophonie, il serait aussi possible d'établir assez rapidement des corpus à côté de FRWAC (pour la France) – par exemple pour la Belgique (domaine .be), la Suisse (.ch), le

Québec (.qc.ca), voire le Canada (.ca), ou pour d'autres provinces (comme .on.ca pour l'Ontario). Ceci permettrait de disposer d'ensembles de données plus importants pour une exploitation à visée comparative. Se pose toutefois la question de savoir si les tendances d'usages quantitatives et qualitatives dans les corpus respectifs qui seraient observables peuvent fournir des indices valables, vu l'hétérogénéité géographique des données textuelles qu'ils contiennent.

En outre, comme dans le web, l'identification géolinguistique des énoncés auxquels donnent accès autant l'I-FR que le FRWAC est en général plus délicate que dans le cadre du discours romanesque et même journalistique – et souvent impossible. Le diatopicien peut certes glaner des indices au sein de l'adresse URL ou dans les textes en présence de commentaires métalinguistiques, de marqueurs typographiques et d'autres dispositifs de citations ; toutefois, une analyse philologique poussée y est impossible.

Pour l'étude sémantique de la phraséologie, même des corpus soigneusement annotés comme le BNC pour l'anglais britannique (rassemblant 100 millions de mots) seraient « too small to reveal the meaning-based patterning of collocations » (comme l'angl. *deep sense* ou *real sense*) même dans une perspective sans visée diatopique, et sont à utiliser en même temps que le Web (Wierzbicka 2009 : 125). En effet, même le méga-corpus qu'est le FRWAC ne fournit en fin de compte que peu d'attestations de diatopismes du français, comme de l'expression verbale *tomber en amour* (français de référence *tomber amoureux*) (dix attestations), tout en fournissant cependant certains indices pour la localisation des énoncés (Wissner à paraître).

3.3 Les corpus 'oraux'

En présence du changement de paradigme en linguistique contemporaine vers une conceptualisation de la langue avec ses réalisations non seulement à l'écrit, mais aussi à l'oral, nombreux sont les linguistes qui considèrent qu'il est nécessaire d'exploiter des « collections ordonnées d'enregistrements de productions linguistiques orales et multimodales », communément appelées *corpus oraux* (déf. d'apr. Baude 2006). Les questions portant sur l'exploitation et la constitution de corpus oraux sont

ainsi largement discutées par les spécialistes du domaine (p.ex. *ib.* ; Cappeau & Gadet 2007 ; Gadet et al. 2009)[3].

Il faut ajouter que l'oralité des productions langagières enregistrées et transcrites dans de tels corpus est de nature diacodique, dépendant du médium de communication, et non pas diamésique (v. 1). Les corpus oraux donnent accès non pas à une 'langue' orale spontanée et authentique, mais à des enregistrements et transcriptions d'énoncés produits oralement, qui ont des modalités fonctionnelles et communicatives diverses.

Malgré « la grande vague de l'oral » (Blanche-Benveniste & Jeanjean 1986 : 43–46), et malgré l'intérêt de phonéticiens, pragmaticiens et morphosyntacticiens pour les corpus oraux, ces derniers sont très insuffisamment exploités en lexicographie française – autant dans les années 1990 (Queffélec 1997 : 353), que de nos jours. Cet état de choses s'explique surtout par la primauté traditionnelle de l'écrit (autant en termes diacodique que diamésique), et par le faible rendement de corpus oraux : le coût élevé pour des enregistrements, et le besoin pour les lexicographes de corpus quantitativement importants, sachant que l'exploitation de corpus oraux (de petite taille) pour l'étude de variétés de français comme en Belgique et en Afrique s'est avérée très coûteuse et peu rentable (*ib.* : 353–356). De fait, « la quasi-totalité des corpus oraux de français » ont été « constitués pour des buts autres que lexicologiques » (Queffélec 1997 : 356). Pourtant, le recours à des corpus oraux, y compris de conversations spontanées, est usuel en lexicographie anglophone (Atkins & Rundell 2008 : 77). Pour la description de diatopismes du français, il est fondamental non seulement lorsque les pratiques langagières se manifestent très majoritairement à l'oral, comme dans la plupart des régions francophones en Afrique (Queffélec 1997 : 365*sq*.), mais aussi pour tenir compte de l'usage sous ses diverses facettes en général.

Il est vrai que l'équipe québécoise préparant leur dictionnaire différentiel DHFQ avait déjà utilisé les corpus oraux de Sherbrooke et de Montréal avec des données recueillies dans les années 1960/70 : le *Corpus de l'Estrie*, *Centre-Sud* et *Sankoff-Cedegren* (DFQPrés 1985 : XVII [Poirier])[4]. En effet, c'est surtout depuis les années 1960/70 au Québec, et

[3] On exclut de la discussion des corpus 'oraux' le méga-corpus bilingue Hansard, disponible en ligne, qui fournit des données pour le français au Canada des débats parlementaires canadiens, mais est constitué des comptes-rendus de ces débats, dont une partie est traduite de l'anglais.

[4] Les deux premiers rassemblent chacun autour de 100 000 mots, le dernier plus d'un million (v. le site du TLFQ pour un descriptif de ces corpus : http://www.tlfq.ulaval.ca).

depuis les années 1980 en Afrique Noire puis au Levant et dans l'Océan Indien qu'ont été réalisés des enregistrements, tous concentrés sur une variété ou une comparaison de deux variétés (Dister et al. 2008 : 296) – avant l'arrivée du PFC (v. ci-dessous). Toutefois, les premiers corpus oraux à objectifs général et sociolinguistique comme le *Corpus Sankoff-Cedegren* ou le corpus hexagonal *Français Fondamental* de 1953, désormais effacé (cf. Baude 2006 : 26), sont difficilement comparables aux corpus de la nouvelle génération – les possibilités d'enregistrement, de transcription, d'annotation et d'exploitation à l'aide de logiciels étant passées par une importante (r)évolution informatique.

Parmi la multitude de petits corpus oraux en Europe relevés par Cappeau & Seijido (2005), seize corpus explicitent le rattachement géographique des textes et/ou des locuteurs enregistrés (pour ce qui est des corpus dépassant une taille correspondant à approximativement 200 000 mots) : trois pour la Suisse, autant pour la Belgique, et huit pour la France continentale (Orléans, Tours, Auvergne ; Alsace ; Perpignan – conçu pour l'étude de régionalismes ; Aquitaine, Toulouse et Grenoble). En France – où J.-M. Debaisieux évaluait la taille de l'ensemble des données orales disponibles en 2005 à environ quatre ou cinq millions de mots (d'apr. André & Canut 2010) – aucun des corpus oraux rendus disponibles à la communauté scientifique ne vise toutefois spécifiquement à permettre des recherches de diatopismes. En outre, toutes ces données sont éparpillées (difficultés d'accès), et pour une exploitation comparative à visée diatopique, ils sont trop hétérogènes. En effet, à l'heure actuelle,

> « de nombreuses contraintes pèsent sur la linguistique française sur corpus, notamment au niveau de l'existence et de la disponibilité des données et des outils » (Benzitoun & Cappeau 2010 : 1384)

En outre, le français ne dispose toujours pas d'un « corpus de référence » (*ib.* : 1385) – même si le CERF, le CRFP (cf. DELIC 2004) et le PFC (cf. Durand & Lyche 2003, Durand et al. 2009) s'y portent candidats. Contrairement aux corpus écrits, les corpus oraux posent en outre tous des problèmes de transcription, et donc d'accessibilité des données enregistrées. L'urgence est ainsi actuellement de faire converger les pratiques, comme au niveau des annotations et des transcriptions, et de systématiser les métadonnées (v. Bruxelles et al. 2009 : 13 et Baude 2006), puis de rendre les corpus disponibles conformes aux standards

internationaux, et accessibles[5]. C'est aussi pour améliorer la description du lexique diatopiquement marqué qu'on plaide depuis longtemps pour une homogénéisation des méthodes d'enregistrement et des normes de transcription, pour une mise en commun des corpus oraux, et pour leur diffusion auprès de la communauté scientifique (Queffélec 1997 : 364–366). Toutefois, une codification minimale qui serait commune pour toutes les visées de recherche (même au sein de la linguistique) n'existe pas, « même si le mirage en est régulièrement soulevé » (Cappeau & Gadet 2007 : 107). L'exploitation efficace de corpus oraux pour l'analyse du lexique diatopiquement marqué implique ainsi de retenir des corpus en transcription orthographique standard.

Même si *big is not necessarily beautiful* (v. 3.2), dans le domaine des corpus oraux, la taille d'un corpus est un critère de valeur pour permettre des analyses qualitativement fructueuses – à côté de la 'rareté' des données qu'ils accueillent, et des types de recherche qu'ils permettent grâce aux possibilités de sélection de sous-corpus (p.ex. Dister et al. 2009 : 121 et 123). Un corpus oral « de taille significative » comporterait deux millions de mots, un « vaste corpus » plusieurs millions de mots (Benzitoun & Cappeau 2010 : 1383 et 1384).

Dans les conditions actuelles où les corpus oraux de français sont de la grandeur d'un million de mots, ceux-ci fournissent déjà une base suffisante pour mener des études lexicales qualitatives et fréquentielles valables (Benzitoun & Cappeau 2010 : 1396). D'autres spécialistes de corpus au contraire estiment que vu leur taille, « il n'est guère possible de faire des recherches lexicales, ni d'établir des statistiques fiables sur les usages » (Baude 2006 : 29) – même du français général[6]. L'apport des

[5] Ainsi pour le français hexagonal, un regroupement de corpus oraux est en train de se faire à l'ATILF par Ch. Benzitoun pour usage au sein du laboratoire. Il accueille des transcriptions d'enregistrements de plusieurs villes en France, totalisant plus de 2,3 millions de mots (dont un million environ sont actuellement exploitables). L'archive intègre les données de l'équipe nancéenne (TCOF) et des autres projets d'envergure qui sont compatibles avec ce dernier, donnant accès à des enregistrements de dialogues pour la plupart : le Corpus de Français Parlé Parisien (CFPP2000) (v. 3.3), CORPAIX (version 2000), le Corpus de Référence du Français Parlé (CRFP, cf. DELIC 2004) – corpus de plusieurs villes en France rassemblant plus de 400 000 mots – et enfin la partie des discussions libres de PFC-France (Benzitoun/Cappeau 2010 : 1385sq.).

[6] En effet, même des verbes courants (du français général) comme *causer* n'affichent qu'une fréquence très limitée dans des corpus d'un million de mots, comme le corpus oral CORPAIX (cinq attestations) ou la tranche journalistique du corpus de textes écrits

corpus oraux actuel est d'autant plus limité pour l'étude de diatopismes (par rapport aux larges corpus écrits traditionnels).

Actuellement, les banques de données les plus consistantes en Europe sont PFC, VALIBEL (ci-dessous) et CLAPI (Dister et al. 2008 : 296), ainsi que désormais TCOF (ci-dessous) et ESLO – le plus grand des corpus oraux hexagonaux, avec une taille approchant les sept millions de mots selon des approximations, mais dont la diffusion annoncée n'est pas effective à ce jour. Si CLAPI forme un ensemble textuel de taille significative – la quarantaine de corpus qu'il contient ont chacun une taille allant jusqu'à 200 000 mots (d'apr. Benzitoun & Cappeau 2010 : 1398 note 2) – cet ensemble n'est pas librement accessible et, comme il est constitué pour l'analyse de l'interaction (cf. Bert et al. 2010), il n'est pas adapté pour des recherches lexicales dans l'optique diatopique. À l'heure actuelle, on compte seulement trois corpus oraux de français contemporain qui pourraient selon nous être exploités avec profit pour l'analyse du lexique diatopiquement marqué dans une optique panfrancophone, en répondant aux conditions fondamentales suivantes :

> – être disponibles en transcription orthographique standard pour permettre des recherches lexicales automatisées, avec un moteur de recherche permettant la requête de lemmes, voire de co-occurrences, dans un seul fichier, avec un concordancier permettant de trier les résultats de requête par contexte gauche/droite pour isoler les sens,
> – fournir l'annotation du discours selon le critère géolinguistique en s'appuyant sur des catégories linguistiquement pertinentes, et rendre ce critère disponible lors de recherches au sein du corpus, et enfin
> – être rendus accessibles à la communauté scientifique, en ligne.

On ne dispose pas actuellement de plusieurs corpus qui satisfassent à ces divers critères, mais plusieurs corpus se portent candidats. Ainsi, pour le français continental, on trouve le corpus TCOF, complété du Corpus de Français Parlé Parisien (CFPP2000, cf. Branca-Rosoff et al. 2012), qui sont accessibles à la communauté scientifique et accueillent des transcriptions d'enregistrements avec des locuteurs hexagonaux. Ayant distingué tous les locuteurs enregistrés dans les métadonnées d'après leur 'appartenance régionale dominante' (André & Canut 2010), ils permettent à l'utilisateur d'identifier les énoncés selon ce critère géolinguistique et de créer des sous-corpus selon ce critère pour faire des requêtes dans tous les énoncés

---

CERF (38 attestations) – sachant que ce décalage est aussi lié aux genres et thèmes des deux corpus exploités (Benzitoun/Cappeau 2010 : 1384).

de locuteurs d'une région donnée. Le premier, TCOF, qui contient des parties d'échanges adulte/enfant et entre adultes (surtout des entretiens, des conversations et des réunions de travail), est un corpus essentiellement lorrain dans le sens où les locuteurs enregistrés sont pour la plupart originaires du Nord-est (Lorraine). Cet ensemble totalise approximativement quatre millions de mots (dont un demi-million sont exploitables en ligne, sans restriction, actuellement sous la forme de fichiers séparés). Le CFPP2000, pour sa part, recueille des énoncés de locuteurs originaires de la région parisienne, et rassemble un ensemble de 400 000 mots environ.

Contrairement aux plus grands corpus franco-français, le corpus de français en Belgique, VALIBEL, constitué à partir d'enquêtes sociolinguistiques et d'entretiens (cf. Dister et al. 2009 : 117), a été explicitement conçu pour l'analyse d'une variété diatopique spécifique, et a permis la préparation du DictBelg (2010). Parmi les corpus portant sur une variété diatopique spécifique du français, il s'agit du corpus oral moderne le plus vaste, avec environ 4 millions de mots. Il est accessible en ligne sur demande, sauf pour les tous derniers enregistrements qui n'ont pas encore donné lieu à des publications pour un objet de recherche précis (cf. Dister et al. 2009 : 126*sq.*). S'il s'agit d'un corpus unique, il n'est pas pour autant systématiquement exploité en lexicographie différentielle pour l'étude de diatopismes en dehors de la Belgique. Inspiré des corpus du Groupe Aixois de Recherche en Syntaxe GARS en France et de la sociolinguistique québécoise, VALIBEL inspire à son tour d'autres corpus dans la francophonie, comme VALIRUN, pour le français et le créole à l'île de la Réunion[7].

Les autres variétés diatopiques du français dans la francophonie du Nord ou du Sud ne disposent à l'heure actuelle pas, à ma connaissance, de corpus oraux d'une taille équivalente qui soient rendus accessibles à la communauté scientifique. Pour l'Amérique du Nord, on dispose tout de

[7] D'autres corpus portent exclusivement sur des créoles à base lexicale française, mais affichent des alternances codiques entre créole et français, comme le corpus de Ludwig et al. (2001). En Europe, un projet parallèle à VALIBEL, VALISUISSE pour le français en Suisse (annoncé sur le site de VALIBEL), a été initié en 2004 par Anne Grobet (Genève) mais n'a pas pu aboutir, faute de moyens (comm. pers. de l'auteure du 24/10/11). On attend donc avec impatience le très récent projet OFROM, en préparation à Neuchâtel.

même déjà du *Corpus de français parlé au Québec* (CFPQ) de Sherbrooke, constitué d'un demi-million de mots environ[8].

Le troisième corpus qui est actuellement exploitable pour une analyse du lexique dans une optique diatopique est le premier des projets panfrancophones qui visent explicitement à rassembler des données comparables pour plusieurs variétés diatopiques de la francophonie, recueillies selon une même méthode : le corpus du projet Phonologie du Français Contemporain (PFC). Celui-ci visant d'abord une exploitation phonologique, puis tardivement des analyses à visée syntaxique et sociolinguistique, ce sont ses entretiens qui sont exploitables avec profit pour une analyse lexicale. L'ensemble textuel actuellement disponible correspond à un million de mots transcrits (comm. pers. de B. Laks du 25/09/11), accessibles en ligne sur demande. Comme pour VALIBEL et TCOF, le discours a été entièrement annoté selon le rattachement géolinguistique des locuteurs dont il recueille les énoncés transcrits. L'interface d'exploitation n'est pas conçue pour des recherches de co-occurrences, mais permet la création de sous-corpus, y compris selon le critère diatopique. Même si le PFC reste un corpus ouvert, ses transcriptions seront à moyen terme consultables pour tous 'les français' qu'il couvre à l'échelle francophone, dans trente-trois zones géographiques.

À plus long terme, les données orales du projet PFC pourront en outre être complétées par des données comparables du projet CFA, pour l'Afrique et l'Océan Indien, dont la méthodologie d'enquête s'appuie sur celle de PFC (cf. Dister et al. 2008). Deux autres grands projets internationaux en cours visent également à permettre des comparaisons d'usage dans la francophonie : le Corpus International et Écologique de la Langue Française (CIEL-F) – initiative franco-belgo-allemande qui rassemble des données pour le français de vastes aires, comme en Suisse, en Acadie et en Égypte (cf. *ib* ; Gadet et al. 2009, 2012) – et le projet *Dynamiques des français périphériques*, lancé en 2008 au laboratoire MoDyCo, qui porte en partie sur les mêmes aires que CIEL-F, mais aussi sur des zones comme la Tunisie et les Îles anglo-normandes.

[8] Le projet international d'Ottawa intitulé *Le français à la mesure d'un continent* (2011–2018), mené sous la direction de F. Martineau, visant à établir un corpus de discours non-conventionnel, écrit et oral (en particulier pour mener des analyses en syntaxe), fournira pour sa part des données pour le français dans l'Ontario, ainsi qu'a priori des données comparables pour des variétés diatopiques ailleurs dans la francophonie.

Si l'usage de corpus oraux tend à être moins 'rentable' que celui de corpus écrits, les corpus oraux sont importants pour une analyse plus juste des diatopismes du français, y compris de son lexique – vu les caractéristiques énonciativo-discursives des énoncés transcrits, et vu que le travail d'identification diatopique des locuteurs y est déjà souvent fourni.

4. Perspectives

À l'heure actuelle, c'est surtout par manque de moyens que l'analyse lexicale des diatopismes s'appuie sur des données hétérogènes, qui sont encore en grande partie lexicographiques. En complément des ressources traditionnelles, l'utilisation de grands corpus 'écrits' et 'oraux' est pourtant nécessaire pour s'appuyer sur des données discursives des diverses variétés diatopiques du français dans la francophonie – sous condition de s'astreindre à l'analyse critique de chaque corpus.

Peut-on raisonnablement imaginer la mise en place d'une interface commune qui soit incontournable pour toute étude lexicologique à visée diatopique ? Un outil partagé pourrait se présenter sous la forme d'un portail commun qui centraliserait la description des corpus, et fournirait des hyperliens vers les sites où leur exploitation serait possible. Il est vrai que le souhait de la constitution d'une « bibliothèque de données de corpus » multifonctionnelle, ouverte et partageable, quoique « dans l'air du temps », est irréaliste (Cappeau & Gadet 2007 : 108 et 107–109), mais un *Inventaire des corpus de français hors de France* est bel et bien en préparation (Gadet à paraître). On pourra regrouper les corpus qui sont exploitables selon des visées de recherche comparables, comme ceux qui visent l'analyse de la phonologie ; de l'interaction ; ou de la morphosyntaxe et du lexique. Il existe déjà pour d'autres langues de grands corpus de référence qui sont exploitables pour des recherches lexicales et morphosyntaxiques, annotés diatopiquement aussi bien que morphosyntaxiquement, et équilibrés selon plusieurs critères (y compris discursif) : notamment pour l'anglais britannique, avec le BNC, et pour l'espagnol de l'Hispania, avec le CREA (constitués de 100 million de mots pour l'un, 160 pour l'autre) – outre des corpus de plusieurs centaines de millions de mots pour l'anglais contemporain (tels OEC ou COCA).

Vu l'organisation essentiellement étatique des plateformes existantes qui hébergent des corpus multifonctionnels du français, exploitables à l'aide de logiciels compatibles – le *Réseau des corpus lexicaux québécois*, au Canada, ou en France le SLDR (anciennement CRDO-Aix, pour

'l'oral') et le CNRTL (pour l'écrit, où se trouve toutefois aussi le TCOF) – c'est dans le cadre du projet Open Resources and TOols for LANGuage (ORTOLANG) que l'on peut instaurer à moyen terme une infrastructure commune, tout au moins pour la France. Le projet (Aix, Nancy, Orléans/Tours, Paris 2012–2019) vise en effet à l'élaboration d'une infrastructure en réseau et de ressources et d'outils, pour rassembler, diffuser et archiver les corpus écrits, oraux et patrimoniaux du français et des autres langues de France (cf. http://www.cnrtl.fr/ortolang/). Il se réalisera dans le cadre de deux consortiums linguistiques (IRCOM et 'Corpus Écrits'), mis en place en 2011 (cf. http://www.corpus-ir.fr/).

Pour mener des études comparatives à une échelle proprement panfrancophone, il faudra certes attendre qu'il existe des ensembles textuels suffisamment larges et comparables pour les diverses variétés diatopiques du français de la francophonie, et qui suffisent aux critères formulés ici. Si tout corpus est nécessairement restreint, un ensemble textuel de référence pour la lexicologie variationnelle contenant des réalisations écrites et orales de variétés diatopiques de la francophonie du Nord mais aussi du Sud serait un pas important. Un tel ensemble contribuerait considérablement à l'identification de diatopismes en tant que tels, et à en retracer les caractéristiques diverses – tout particulièrement des emplois 'rares' et des locutions, et des particularismes qui se distinguent justement par une différence de statut ou de fréquence. Il permettrait aussi de dépasser deux raccourcis majeurs : l'explication de la variation du français par les contacts de langue (du type 'germanisme' ou 'anglicisme'), et l'étiquetage géographique hâtif d'un emploi donné (du type 'belgicisme' ou 'québécisme'), par omission d'une étude sociolinguistique et aréologique sérieuse. De nouveaux corpus adaptés à l'analyse lexicale à visée diatopique ou un éventuel corpus de référence pour la lexicologie variationnelle renseigneraient non seulement sur les diatopismes du français contemporain, mais aussi sur le fonctionnement de la langue, vu la diversité des situations sociolinguistiques au sein de la francophonie.

Références

Adam, Jean-Michel & Heidmann, Ute (2006) Six propositions pour l'étude de la généricité. *La Licorne* 79 : 21–34.

André, Virginie & Canut, Emmanuelle (2010) Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques : linguistique, littérature, didactique* 147/148 : 35–51.

Atkins, Sue ; Clear, Jeremy & Ostler, Nicholas (1992) Corpus design criteria. *Literary & Linguistic Computing* 7/1 : 1–16.

—— & Rundell, Michael (2008) *The Oxford Guide to Practical Lexicography* (coll. Oxford Linguistics). Oxford : Oxford University Press.

Baroni, Marco ; Bernardini, Silvia ; Ferraresi, Adriano & Zanchetta, Eros (2009) The WaCky wide web : A collection of very large linguisticially processed web-crawled corpora. *Language Resources and Evaluation* 43/2 : 209–226.

Baude, O. (éd.) (2006) *Corpus oraux, guide des bonnes pratiques*. Paris : Éditions du CNRS/Orléans : PU d'Orléans.

BDLP. Base de données lexicographiques panfrancophone constituée de vingt bases nationales, hébergée par le TLFQ, Laval (Québec) : AUF/TLFQ, en accès libre sur Internet <http://www.tlfq.ulaval.ca/bdlp>, aussi via le CNRTL, rubrique « Portail lexical » <www.cnrtl.fr>.

Benzitoun, Christophe & Cappeau, Paul (2010) Description sur corpus. Quelques réflexions autour des données et des instruments pour le français (parlé) à travers la description de *cause* et *causer*. In Franck Neveu ; Valelia Muni Toke ; Thomas Klingler ; Jacques Durand ; Lorenza Mondada & Sophie Prévost (éds), *Congrès Mondial de Linguistique Française – CMLF 2010 [La Nouvelle-Orléans, 12–15 juillet 2010]*, pp. 1383–1398. Paris : Institut de Linguistique Française. Publié en ligne le 12 Juillet 2010 <http://dx.doi.org/10.1051/cmlf/2010237> (20/06/2011).

Bert, Michel ; Bruxelles, Sylvie ; Étienne, Carole ; Mondada, Lorenza ; Teston, Sandra ; Traverso, Véronique ; Jouin-Chardon, Émilie & Justine, Lascar (2010) Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). *Pratiques : linguistique, littérature, didactique* 147/148 : 17–34.

Blanche-Benveniste, Claire & Jeanjean, Colette (1986) *Le français parlé, transcription et édition*. Paris : Didier-Érudition.

BNC. *British National Corpus*. Corpus textuel de l'anglais britannique de la fin du XX^{e} siècle d'environ 100 millions de mots dont 10 millions de mots de textes 'oraux' (de conversations spontanées et autres enregistrements), d'une grande variété de sources et de genres, annoté linguistiquement et doté de métadonnées contextuelles selon les recommandations de TEI ; dernière version (de 2007) distribuée librement en XML (sous licence).

Branca-Rosoff, Sonia ; Fleury, Serge ; Lefeuvre, Florence & Pires, Mat (2012) *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*, 23 pp. Paris : Université Sorbonne Nouvelle Paris 3. Publié en ligne <http://cfpp2000.univ-paris3.fr/> (24/07/2012).

Brochard, Marie-José (2012) Les ressources numériques en lexicologie historique. In Dörr, Stephen & Thomas Städtler (éds), Ki bien voldreit raisun entendre. *Mélanges en l'honneur du 70^{e} anniversaire de Frankwalt Möhren*, pp. 27–42. Strasbourg : ELiPhi.

Bruxelles, Sylvie ; Mondada, Lorenza ; Simon, Anne Catherine & Traverso, Véronique (2009 [2007]) Grands corpus de français parlé. Bilan historique et perspectives de recherches. *Cahiers de Linguistique* 33/2 : 1–14.

Cajolet-Laganière, Hélène ; Labrecque, Geneviève ; Martel, Pierre ; Masson, Chantal-Édith ; Mercier, Louis & Théoret, Michel (2008) Dictionnaires usuels du français

et Banque de Données Textuelles de Sherbrooke (BDTS) : convergence et divergence des nomenclatures. In Brigitte Horiot (éd.), *Français du Canada, Français de France.* Actes du septième Colloque international de Lyon, du 16 au 18 juin 2003 (coll. Canadiana romanica 22), pp. 9–28. Tübingen : Niemeyer.

Cappeau, Paul & Gadet, Françoise (2007) L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale. *Revue Française de Linguistique Appliquée* 12/1 : 99–110.

—— & Seijido, Magali (2005) *Inventaire des corpus oraux en langue française*, document avec l'inventaire version 1.1 en annexe, accessible en ligne <www.dglflf.culture.gouv.fr>.

COCA. *Corpus of Contemporary American English*. La plus grande base textuelle de l'anglais américain contemporain, en accès libre, constitué de 425 millions de textes (de 1990 à 2011) et d'un large spectre de textes et de genres.

CREA. *Corpus de Referencia del Español Actual*. Corpus de référence de l'espagnol actuel du tournant des XX^e^ et XXI^e^ siècles constitué par la Real Academia Española, rassemblant environ 160 millions de mots (état de 2005) de types de textes divers annoté selon les genres et thèmes abordés, de toutes les variétés hispanophones (Argentine, Bolivie, Chili, Colombie, Costa Rica, Cuba, Le Salvador, Équateur, Espagne, États-Unis, Guatemala, Honduras, Mexique, Nicaragua, Panama, Paraguay, Pérou, Philippines, Porto Rico, République Dominicaine, Uruguay, Venezuela) – de 50% de textes espagnols, 50% hispano-américains, et de 90% de textes écrits, pour moitié journalistiques, et 10% de 'textes oraux', de 1975 au présent – en accès libre en ligne <http://www.rae.es>.

Delcourt, Christian (1998–1999) *Dictionnaire du français de Belgique*, vol. I–II. Bruxelles : Le Cri.

—— ; Francard, Michel & Moreau, Marie-Louise (1993) Une banque de données textuelles sur la langue française en Belgique. In Danièle Latin ; Ambroise Queffélec & Jean Tabi-Manga (éds), *Inventaire des usages de la francophonie : nomenclatures et méthodologies* (coll. Actualité scientifique), pp. 313–331. Paris : J. Libbey-Eurotext.

Delic (collectif) (2004) Présentation du Corpus de Référence du Français Parlé. *Recherches sur le français parlé*, 18 : 11–42.

DFQPrés. Poirier, Cl. (éd.) (1985) *Dictionnaire du français québécois. Volume de présentation*. Sainte-Foy (Québec) : PU Laval.

DHFQ. Poirier, Cl. (éd.) (1998) *Dictionnaire historique du français québécois. Monographies lexicographiques de québécismes*. Sainte-Foy (Québec) : PU Laval.

DictBelg. Francard, Michel ; Geron, Geneviève ; Wilmet, Régine & Wirth, Aude (2010) *Dictionnaire des belgicismes*. Bruxelles : De Boeck-Duculot.

Dister, Anne ; Francard, Michel ; Hambye, Philippe & Simon, Anne Catherine (2009 [2007]) Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989–2009). *Cahiers de Linguistique* 33/2 : 113–129.

—— ; Gadet, Françoise ; Ludwig, Ralph ; Lyche, Chantal ; Mondada, Lorenza ; Pfänder, Stephan ; Simon, Anne Catherine & Skattum, Ingse (2008) Deux nouveaux corpus internationaux du français : CIEL-F (Corpus International et

Écologique de la Langue Française) et CFA (Français contemporain en Afrique et dans l'Océan Indien). *Revue de Linguistique Romane* 72 : 295–314.

DRF. Rézeau, P. (éd.) (2001) *Dictionnaire des régionalismes de France (DRF).* Bruxelles : De Boeck-Duculot.

DSR. Thibault, André (^{2}2004) *Dictionnaire suisse romand* [1997]. Genève : Zoé.

Durand, Jacques & Lyche, Chantal (2003) Le projet Phonologie du Français Contemporain (PFC) et sa méthodologie. In Elisabeth Delais-Roussarie & Jacques Durand (éds), *Corpus et variation en phonologie du français : méthodes et analyses*, pp. 212–276. Toulouse : PU du Mirail.

—— ; Laks, Bernard & Lyche, Chantal (2009) Le projet PFC : une source de données primaires structurées. In *id*. (éd.), *Phonologie, variation et accents du français*, pp. 19–61. Paris : Hermès.

Francard, Michel (2001) Le français de référence : formes, normes et identité. In *id*. (éd.), *Le français de référence. Constructions et appropriations d'un concept*, pp. 223–240. Louvain-la-Neuve : Institut de linguistique de Louvain.

Gadet, Françoise (à paraître) *Inventaire des corpus de français hors de France*. Site de la DGLFLF, <www.dglf.culture.gouv.fr>.

—— (^{2}2007) *La Variation sociale en français* (coll. L'Essentiel français) [2003]. Paris : Ophrys.

—— & Jones, Mari (2008) Variation, contact and convergence in French spoken outside France. *Journal of Language Contact*, série THEMA, n° 2, 238–248.

—— ; Ludwig, Ralph ; Mondada, Lorenza ; Pfänder, Stephan & Simon, Anne Catherine (2012) Un grand corpus de français parlé : le CIEL-F. Choix épistémologiques et réalisations empiriques. *Revue française de linguistique appliquée* 17/1 : 39–54.

—— ; Ludwig, Ralph & Pfänder, Stefan (2009) Francophonie et typologie des situations. *Cahiers de linguistique* 34/1 : 143–162.

ILQ. *Index lexicologique québécois* : Inventaire des mots du français québécois ayant fait l'objet d'un commentaire ou d'une étude depuis 1750 jusqu'à nos jours. Fonds documentaire réalisé de 1979 à 1986 dans le cadre de la préparation du DHFQ sous la direction de Claude Poirier et de Louis Mercier, consultable sous forme informatique dans sa version actuelle reprogrammée en 2003 par Jean-François Smith, dans la rubrique « Fonds documentaires » du site du TLFQ, en accès libre <http://www.tlfq.ulaval.ca/ilq/>.

Kleiber, Georges (1978) *Le Mot « ire » en ancien français (XI^{e}–XIII^{e} siècles). Essai d'analyse sémantique* (coll. Bibliothèque française et romane). Paris : Klincksieck.

Ludwig, R. ; Telchid, S. & Bruneau-Ludwig, F. (éds) (2001) *Corpus créole. Textes oraux dominicains, guadeloupéens, guyanais, haïtiens, mauriciens et seychellois : enregistrements, transcriptions et traductions* (coll. Kreolische Bibliothek). 2 Cédérom, Hamburg : H. Buske.

Lux-Pogodalla, Veronika & Polguère, Alain (2011) Construction of a French Lexical Network : Methodological issues. In Benoît Sagot (éd.), *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011 (Ljubljana, Slovenia, August 1–5 2011)*, pp. 55–62. Disponible en ligne <http://alpage.inria.fr/~sagot/woler2011/WoLeR2011/Program_%26_Proceedings.html> (25/01/12).

Masson, Chantal-Édith ; Cajolet-Laganière, Hélène & Martel, Pierre (2007) La BDTS-concordances : un outil d'enrichissement de la pratique lexicographique. *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 764–775. Disponible en ligne <http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_073.pdf>.

OEC. *Oxford English Corpus*. Corpus textuel de la langue anglaise le plus vaste de son genre avec plus de deux milliards de mots en format XML, tirés de textes de genres divers, annotés selon leurs types de texte, le type de variété d'anglais et l'identité des locuteurs, utilisé dans le cadre du *Oxford English Dictionary* et par le projet de recherche linguistique des Presses universitaires d'Oxford.

Poirier, Claude (2005) La dynamique du français à travers l'espace francophone à la lumière de la base de données lexicographiques panfrancophone. *Revue de Linguistique Romane* 69 : 483–516.

Queffélec, Ambroise (1997) Le corpus textuel oral. Constitution, traitement et exploitation lexicographique. In Claude Frey & Danièle Latin (éds), *Le Corpus lexicographique : méthodes de constitution et de gestion* (coll. Champs linguistiques), pp. 353–368. Louvain-la-Neuve : Duculot.

Rézeau, Pierre (2007) Des variétés dialectales gallo-romanes aux variétés régionales du français : la constitution d'un champ disciplinaire. In David Trotter (éd.), *Actes du XXIVe Congrès international de linguistique et de philologie romanes* (Aberystwyth, 1–6 août 2004), vol. IV, pp. 263–275. Tübingen : Niemeyer.

Sharoff, Serge (2006) Creating general-purpose corpora using automated search engine queries. In Marco Baroni & Silvia Bernardini (éds), *WaCky ! Working papers on the Web as Corpus*, pp. 63–98. Gedit : Bologna. Publié en ligne <http://wackybook.sslmit.unibo.it/> (25/01/2012).

Thibault, André (2000) Trois nouveaux dictionnaires différentiels de français : présentation et considérations méthodologiques. In Annick Englebert ; Michel Pierrard ; Laurence Rosier ; Dan Van Raemdonck (éds), *Des mots aux dictionnaires* (XXIIe ACILPR, Bruxelles, 23–29 juillet 1998), vol. IV, pp. 551–561. Tübingen : Niemeyer.

—— (2007) Banques de données textuelles, régionalismes de fréquence et régionalismes négatifs. In David Trotter (éd.), *Actes du XXIVe Congrès international de linguistique et de philologie romanes* (Aberystwyth, 1–6 août 2004), vol. I, pp. 467–480. Tübingen : Niemeyer.

TLF (1971–1994). *Trésor de la langue française. Dictionnaire de la langue du XIXe et XXe siècle (1789–1960)*, vol. I–XVI, publié sous la direction de P. Imbs (vol. I–X). Paris : CNRS Éditions, et de B. Quemada (vol. XI–XVI). Paris : Gallimard. Dictionnaire en accès libre avec ses mises à jours sur le site du laboratoire ATILF, CNRS/Université de Lorraine <http://atilf.atilf.fr/tlf.htm>.

Vézina, Robert (1998) Compte-rendu de : DSR. *Langues et Linguistique* 24 : 225–230.

Wierzbicka, Anna (2009) Exploring English phraseology with two tools : NSM semantic methodology and Google, *Journal of English Linguistics* 37/2 : 101–129.

Wissner, Inka (2010) *Les diatopismes du français en Vendée et leur utilisation dans la littérature : l'œuvre contemporaine d'Yves Viollier*, Thèse : Bonn/Paris :

Philosophische Fakultät der Universität Bonn, Abteilung für Romanistik / Université Paris-Sorbonne Paris IV, ÉD Concepts et langages. Disponible comme ressource électronique à l'URN urn:nbn:de:hbz:5-24001.

—— (à paraître [2013]) *Tombé en amour avec les corpus ?* What text corpora add to the study of expressions in modern French. In Henry Tyne ; André, Virginie ; Boulton, Alex et al. (éds), *Ecological and Data-Driven Perspectives in French Language Studies*. Cambridge : Cambridge Scholars Publishing.

Annexe

Pour une meilleure vue d'ensemble, on présentera les corpus, bases de données et plateformes en ligne pour l'étude du français contemporain dans une section à part. On a choisi de fournir un maximum de renseignements qui sont pertinents pour l'étude du lexique diatopiquement marqué. Toutefois, les informations qu'on a pu obtenir varient d'une ressource à l'autre, celles-ci étant tributaires des priorités des différents auteurs des corpus, et des renseignements qui sont mis à la disposition de la communauté scientifique (description en date du 22/08/2012).

I. Plateformes et portails

CLARIN. *Common Language Ressources and Technology Infrastructure of the Institute for Language and Speech Processing*. Infrastructure de recherche européenne synthétisant la description de corpus de nombreuses équipes internationales pour diverses langues du monde (890 corpus) et fournissant les hyperliens aux sites respectifs ; contenant 55 corpus [surtout] écrits hétérogènes de langue française (dont quinze uniquement français), parmi lesquels FRANTEXT, GALLICA, Les Voisins du Monde, Morphalou et Ananas (état 09/09/2011), conçus pour des recherches variées, et contenant des données allant du XVI^{e} au XX^{e} siècles <http://www.clarin.eu/view_resources>.

CNRTL. Portail du CNRS fédérant un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue, donnant accès à des corpus de français écrit, notamment littéraire et journalistique, et à un corpus oral transcrit et homogénéisé à Nancy (TCOF), Nancy : ATILF, CNRS/Université de Lorraine, en accès libre <www.cnrtl.fr>.

Réseau des corpus lexicaux québécois. Plateforme mise en place par le Secrétariat à la politique linguistique SPL du Québec, présentant

quinze corpus de français québécois exploitables en ligne, et fournissant une interface d'interrogation par mot-clé (sans recherche de co-occurrences) ainsi que les liens aux sites permettant d'accéder aux corpus provenant de cinq universités québécoises, dont FLI, ILQ, BDLP et QUEBETEXT de Laval, LEXIQUM de Montréal, Corpus du Témiscouata de Rimouski, et BDTS. En accès libre en ligne <http://www.spl.gouv.qc.ca/languefrancaise/corpuslexicaux/description/>.

SLDR. *Speech & Language Data Repository* (anciennement CRDO-Aix) : Banque de données parole et langage archivées au CINES et distribuées par le CC-IN2P3, constituée sous la coordination scientifique de Philippe Blache et Daniel Hirst visant à pérenniser et partager des corpus de parole et à les enrichir à l'aide de transcriptions et d'annotations, en accès libre sur Internet (Open Archives, CLARIN), sous condition (certaines catégories d'utilisateurs, après acceptation d'une licence non-commerciale, selon le Code du patrimoine), hébergée par l'Université de Provence <http://sldr.org>.

II. Corpus et bases de données

BDTS. Banque de données textuelles de Sherbrooke, exploitable avec l'outil BDTS-concordances, recueillant plus de 52 millions de mots tirés de quelque 15 000 'textes représentatifs des différents usages du français en usage au Québec' couvrant les années 1960 à 2000 – textes littéraires, journalistiques, didactiques, et spécialisés (notamment techniques, scientifiques, sociopolitiques, administratifs, juridiques et culturels), et un sous-corpus oral d'environ 2 millions de mots (transcriptions d'enquêtes sociolinguistiques orales et de discours plus formels comme des téléromans et des textes radiophoniques et télévisés) ; corpus préparé dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ) pour servir de base à l'élaboration de la nomenclature du dictionnaire du projet de dictionnaire du français québécois de l'équipe FRANQUS en préparation sous la direction éditoriale d'Hélène Cajolet-Laganière et Pierre Martel, Sherbrooke (Québec) : Université de Sherbrooke. En accès limité aux collaborateurs, diffusion prévue sur le Site de l'Université où un sous-ensemble contenant plus de 16 millions de mots est accessible <http://www.usherbrooke.ca/catifq/accueil/>, rubrique « Recherches ».

BELTEXT. Base de données textuelles sur la langue française en Belgique et sur la littérature belge de langue française constituée de corpus écrits et oraux, transcrits en orthographe conventionnelle et accompagnée des enregistrements originaux, permettant des recherches littéraires et linguistiques pour des exploitations multiples (lexicale, morphologique, syntaxique, pragmatique, etc.) ; contribution de la Communauté Wallonie-Bruxelles au projet international BDLP, dans le cadre de la préparation d'un dictionnaire de belgicismes par une équipe interuniversitaire dirigée à l'origine par Delcourt (Liège), Francard (Louvain-la-Neuve) et Moreau (Mons) ; renseignements non obtenus au sujet des possibilités de consultation auprès de Christian Delcourt.

CERF. Corpus Évolutif de Référence du Français. Corpus de l'équipe DELIC (sous la direction de Jean Véronis) comportant 9 millions de mots, répartis en neuf tranches (à 1 million de mots chacune) et équilibrés discursivement entre neuf types de discours tirés de l'écrit (récupérés en partie sur Internet), y compris de magazines, textes scientifiques et de la presse nationale et régionale, ainsi qu'une tranche orale (CORPAIX) d'environ 1 million de mots ; corpus présenté en ligne <http://sites.univ-provence.fr/delic/corpus/index.html>.

CFA (*en cours*). Corpus de Français contemporain en Afrique et dans l'Océan Indien, projet d'une équipe internationale menée sous la direction d'Ingse Skattum initié en 2006 à Oslo pour compléter le projet PFC, portant sur huit zones (Burkina Faso, Cameroun, Centrafrique, Côte d'Ivoire, La Réunion, Mali, Maurice, Sénégal), recueillant des productions formelles et des entretiens semi-directifs (comme le PFC), des entretiens sur les usages et attitudes des locuteurs, et des conversations libres selon les contextes locaux, avec douze locuteurs par point d'enquête ; ensemble de données comparables pour les variétés considérées, complété de données de genres variés (cours magistral, prêche, débats télé, etc.) qui sera soumis à des outils d'indexation phonologique et syntaxique pour permettre des analyses phonologiques, syntaxiques et sociolinguistiques, transcrit, étiqueté, codé et aligné texte/son/métadonnées ; diffusion en accès libre sur Internet prévue.

CFPP2000. Discours sur la ville. Corpus de Français Parlé Parisien des années 2000. Corpus composé d'un ensemble d'interviews sur les quartiers de Paris et de la proche banlieue (Sonia Branca-Rosoff,

Serge Fleury et Florence Lefeuvre). Corpus en accès libre en ligne, sous licence <http://cfpp2000.univ-paris3.fr/>.

CFPQ. Corpus de français parlé au Québec préparé depuis 2006 dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ) de l'Université de Sherbrooke sous la responsabilité de Gaétane Dostie ; corpus multimodal visant à 'refléter le français québécois' en usage dans les années 2000 qui rassemble des enregistrements effectués sur support audiovisuel de plus de 45 heures et les transcriptions alignées, réalisées à l'aide du logiciel Transana (382.181 mots le 24/05/12, statistique non mise à jour) ; support audiovisuel consultable sur place uniquement ; transcriptions en accès libre sur le Site de l'Université de Sherbrooke <http://pages.usherbrooke.ca/cfpq/corpus.php>.

CIEL-F (*en cours*). Corpus International et Écologique de la Langue Française : données audio et vidéo établies dans des situations comparables dans différents pays choisis selon une typologie des aires et situations d'usages (Acadie, Algérie, Antilles, Belgique [Bruxelles-Brabant Wallon], Cameroun, Côte d'Ivoire, Égypte [Le Caire], France, La Réunion, Liège, Maurice, Québec, Ontario, Sénégal, Suisse) ; initiative franco-belgo-allemande (Lyon, Paris, Freiburg, Halle, Louvain) menée sous la responsabilité de Lorenza Mondada (Lyon) et Stefan Pfänder (Freiburg), en partenariat avec des équipes internationales ; diffusion libre sur Internet prévue via les interfaces [moca] et CLAPI <www.ciel-f.org>, copyright 2008–2012.

CLAPI. Base de données orales contenant 46 corpus, constitués d'enregistrements audio ou vidéo (non média) pour permettre l'analyse de l'interaction (Lyon 2) <http://clapi.univ-lyon2.fr/>.

CORPAIX. Corpus d'enregistrements d'oral réalisés par le Groupe Aixois de Recherche en Syntaxe (GARS), dont la version de 2000 (CORPAIX2000) comporte environ un million de mots , cf. <http://sites.univ-provence.fr/delic/corpus/index.html>.

CRFP. Corpus de Référence du Français Parlé. Corpus présentant des situations variées avec une majorité d'entretiens, dont les enregistrements ont été effectués dans diverses villes de l'Hexagone et selon trois situations de parole (privée, publique et professionnelle) (CRFP-1). Corpus d'environ 440 000 mots transcrits par des transcripteurs multiples selon les Conventions de transcription orthographiques (conventions DELIC), constitué sous la responsabilité de Jean Véronis (Aix-en-Provence) en vue d'études

linguistiques, notamment en syntaxe <http://sites.univ-provence.fr/delic/crfp/> ; sa diffusion par la Délégation DGLFLF est prévue.

Dynamiques des français périphériques (*en préparation*). Projet inter-composantes du laboratoire MoDyCo (Modèles, Dynamiques, Corpus) lancé en 2008 sous la direction de Françoise Gadet et Colette Noyau en partenariat avec des équipes de trois continents, visant la mise en réseau de chercheurs et de laboratoires pour rassembler des données pertinentes pour des comparaisons d'usage dans les différentes variétés diatopiques du français (extraction et mise en forme de corpus disponibles ou collecte coordonnée de données spécifiques) ; les travaux et/ou corpus des partenaires actuels portent sur le français aux Îles anglo-normandes, en Acadie, Ontario et Louisiane, à l'Île Maurice et à la Réunion, au Cameroun, au Gabon, au Sénégal, au Togo, en Côte d'Ivoire, et en Tunisie ; présentation prévue sur le site *Aspects acquisitionnels et sociolinguistiques des dynamiques des français* (ATRADY).

ESLO. Enquête Sociolinguistique à Orléans. Corpus oral constitué au Laboratoire Ligérien de Linguistique (LLL, ex-CORAL) de l'université d'Orléans en collaboration avec d'autres laboratoires, constitué à partir de 700 heures d'enregistrements, dont 300 heures de 1968 à 1971 (ESLO-1, à visée didactique) et 400 heures d'enregistrements comparables dans les modalités de collecte, depuis les années 2000 (ESLO-2, à visée variationniste) ; accès libre sur Internet annoncé pour 01/07/2012, non effectif en date du 21/08/2012 <http://www.univ-orleans.fr/eslo/>.

L'Est Républicain, corpus constitué d'articles de toutes les éditions régionales et locales du quotidien régional de l'est de la France de 1999, 2002 et 2003, consultable via le CNRTL en format XML-TEI P5 en trois fichiers séparés par année <http://www.cnrtl.fr/corpus/estrepublicain/> ; corpus en projet d'extension pour rassembler les textes des années 2005 à 2010.

EUREKA. Corpus de presse multilingue (anciennement *Biblio Branchée*) depuis les années 1980 surtout (sans indication exacte) couvrant l'actualité internationale, nationale, régionale et locale en douze langues dont le français, permettant des requêtes thématiques par mots-clés et en principe par suites de mots (option en partie inopérable), régulièrement enrichi, rassemblant actuellement 6023 sources référencées (état du 06/01/12, sans indication de taille),

consultables par langue, région, date de publication, domaine thématique et parties de texte par une interface affichant les 300 premiers documents ; corpus comportant des parts égales de textes de la presse généraliste et spécialisée – surtout des journaux (comme *Le Monde* et *Le Monde diplomatique*, pour la France), publications spécialisées, fils de presse, émissions télévisées et radiophoniques transcrites, blogues triés sur le volet, sites Web référencés, rapports). Accès sous licence, Copyright CEDROM-SNi inc. 2012 <http://www.eureka.cc/Default.aspx>.

EUROPRESSE. Presse d'information francophone et anglophone en texte intégral qui réunit le texte des publications suivantes : *AFP Général*, dep. 19/03/01 ; *L'Express*, dep. 07/01/93 ; *L'Humanité*, dep. 16/11/1999 ; *La Croix*, dep. 01/09/95 ; *Le Figaro*, dep. 31/10/1996 ; *Le Monde*, dep. 01/01/1997 ; *Le Monde diplomatique*, dep. 01/01/1984 ; *Le Parisien*, dep. 02/05/1998 ; *Le Point*, dep. 07/01/1995 ; *Les Échos*, dep. 02/01/1991 ; *Libération* dep. 02/01/1995 ; accès payant aux archives sur Internet <www.europresse.com>.

FLI. Fichier lexical informatisé. Base de données du français québécois du XVI^e^ au XX^e^ siècles comportant 400 000 fiches avec une ou plusieurs citations chacune, essentiellement d'emplois québécois (état 02/2010), tirées d'un corpus de citations papier constitué pour l'essentiel entre 1975 et 1990 en vue du DHFQ à l'Université de Laval (Québec), contenant plus de 1,2 millions de fiches manuscrites de sources diverses : récits anciens, documents d'archives manuscrits ou imprimés, documents administratifs, journaux et magazines, littérature et textes de création, études et textes spécialisés, manuscrits de radioromans et de téléromans, enregistrements oraux, relevés d'enquêtes sur le terrain. En accès libre sur Internet avec deux modes de recherche, « dans les entrées » et « dans les citations » <http://www.tlfq.ulaval.ca/fichier/>.

FRANTEXT. Base de données textuelles de la littérature française conçue dans le cadre de la préparation du TLF à l'ATILF, CNRS/Université de Lorraine ; corpus à dominante littéraire constitué de quelque 248 millions de mots, de 4.084 références de textes français (état du 09/09/11) du XVI^e^ au XXI^e^ siècles – appartenant aux domaines des sciences, des arts, de la littérature, et des techniques – consultable au format TEI avec le logiciel Stella en ligne sur abonnement, permettant de visualiser des extraits textuels de 700 signes

<http://www.frantext.fr/> ; un corpus réunissant 500 œuvres de la littérature française du XVIII^e^ au XX^e^ siècles libres de droits est en accès libre <http://www.cnrtl.fr/corpus/frantext/>.

FRWAC. Corpus textuel d'une tranche du Web (domaine « .fr ») d'environ 1,6 milliards de mots, construit dans le cadre du projet WaCky Wide Web (Trente/Bologne) par interrogation du web avec des paires de mots et par filtrage et nettoyage des pages, indexé avec CorpusWorkBench, étiqueté et lemmatisé avec TreeTagger. En accès libre sur demande <http://wacky.sslmit.unibo.it/doku.php?id=download> ; également exploitable sous une version catégorisée par degré de normativité orthographique et grammaticale, conçue à l'ATILF, CNRS/Université de Lorraine (Nancy) en collaboration avec Druide Informatique Inc. (Montréal). Sa mise à disposition à la communauté scientifique est envisagée par l'intermédiaire des chercheurs à l'origine du projet WaCky Wide Web, sauf si une diffusion via le site CNRTL est accordée et juridiquement sûre ; pourra y figurer au moins un lien vers la base.

GALLICA. Bibliothèque numérique de la Bibliothèque nationale de France (Paris) : collections de manuscrits occidentaux et orientaux d'époques diverses numérisées depuis 2007, en accès libre sur Internet <http://gallica.bnf.fr/>.

GRL. Google Recherche de Livres : Base de données textuelles scannées et traitées par Google, en accès libre sur Internet <http://books.google.fr/>, consultée en mode de Recherche « Livres entiers ou en aperçu limité » [données à vérifier sur l'original].

HANSARD. Corpus en français au Canada des *Débats de la Chambre des communes* constitué des comptes rendus *in extenso* des débats ayant lieu à la Chambre et en comité plénier du Parlement du Canada, depuis le 35^e^ Parlement de janvier 1994 au jour d'aujourd'hui (41^e^ Parlement) ; textes préparés à partir des enregistrements sonores des délibérations ainsi que des renseignements fournis par le personnel du Service des comptes rendus en poste sur le parquet de la Chambre. Corpus publiés en français et anglais et doté d'un référencement minimal, saisi, enregistré et diffusé à l'aide du logiciel PRISME. En partie en accès libre sur Internet (diffusion de l'ensemble prévue ; taille inconnue), avec une possibilité de requête en ligne par mots-clés et séquences exactes et exclusion d'autres mots-clés <http://www.parl.gc.ca/HouseChamberBusiness/ChamberSittings.asp>.

I-FR. Corpus francophone de Leeds tiré de 50 000 pages Internet du Web francophone en 2006 sans limitation de domaines, d'environ 200 millions de mots, établi au *Centre for Translation Studies* de l'Université de Leeds avec un traitement de recherche automatisé par interrogation du web avec des paires de mots, indexation avec CorpusWorkBench, filtrage et nettoyage des pages, et étiquetage/lemmatisation avec TreeTagger (selon la même procédure que FRWAC), avec annotation automatique par thèmes selon des groupes thématiques (*thematic clusters* ; sans vérification manuelle), qui comprennent des informations de type géographique (comme 'congolais', 'Liban' ou 'Québec') ; corpus consultable via un logiciel d'exploitation permettant des recherches de concordances et de collocations. En accès libre sur Internet (Corpus.Leeds.ac.uk/internet.html).

Kölner romanistische Korpusdatenbank. Corpus écrit préparé par Sascha Diwersy sous la responsabilité de Peter Blumenthal (Cologne) constitué de textes journalistiques des années 2000 (totalisant environ 80 millions de mots), pour moitié environ d'Europe, surtout de la France continentale (presse régionale et nationale : *Le Figaro, L'Est Républicain, Sud-Ouest*), et pour plus de la moitié de presse nationale d'Afrique, en particulier du Cameroun (*Cameroon Tribune, Mutations*), du Sénégal (*Le Soleil*) et de Côte d'Ivoire (*Fraternité Matin*) ; corpus complété d'environ 100 romans publiés depuis 1950 d'auteurs de tous les états francophones d'Afrique noire, en particulier du Cameroun, du Sénégal et de Côte d'Ivoire ; corpus annoté conçu pour des recherches qualitatives et quantitatives de collocations ; plus de 5 millions de mots lemmatisés et syntaxisés ; consultable sur place à l'Université de Cologne, diffusion libre prévue pour 2012.

Le Monde (2003). *Le Monde : L'histoire au jour le jour 1939–2002*, cédérom, Coedrion : Le Monde, Emme et IDM.

Le Monde diplomatique (2011). *Le Monde Diplomatique. 43 années d'archives sur DVD-ROM (1968–2010)*, nouvelle édition. Ressource informatique contenant plus de 40 000 documents, de l'édition française (1968–2010) ainsi que de cinq des éditions étrangères en version originale et sous forme de traductions, Coedrion : Le Monde, Emme et IDM.

MCVF. Modéliser le changement : les voies du français. Corpus constitué de textes en grande partie intégraux de genres discursifs divers tels que des correspondances, romans et documents administratifs, du

français de diverses variétés diatopiques (notamment d'Amérique du Nord), 'structuré de façon dialectale, sociale et historique' et couvrant quatre périodes historiques : ancien, et moyen français, français du XVI^{e} siècle, et français classique (France et Nouvelle-France) (XVII^{e} et XVIII^{e} siècle) (documents 'sources' du XI^{e} au XIX^{e} siècle) ; corpus constitué sous la direction de France Martineau dans le cadre des « Grands Travaux de Recherché concertée » de l'Université d'Ottawa, copyright 2009 ; corpus en format XML-TEI permettant des requêtes sur des mots et des co-occurrences de mots à l'aide d'un concordancier en accès libre sur demande en cédérom et en ligne <http://www.voies.uottawa.ca/corpus_pg_fr.html>.

PFC (*en finalisation*). Phonologie du Français Contemporain : usages, variétés et structure. Base de données ouverte préparée dans le cadre d'un projet international sous la direction de Jacques Durand, Bernard Laks et Chantal Lyche (MoDyCo), recueillies en 60 points d'enquête dans l'ensemble de la francophonie avec dix locuteurs échantillonnés par point selon une méthodologie commune comportant une lecture de mots, une lecture de texte, une conversation soutenue et un dialogue informel, représentant un total de 900 heures d'enregistrements transcrits, dont un ensemble d'environ un million de mots consultable par la communauté scientifique sur demande auprès d'Atanas Tchobanov (état 25/09/2011) <http://www.projet-pfc.net>, copyright 2004–2008.

POLITEXT. Base de données construite à l'Université de Nice rassemblant cinq cents discours des grands hommes politiques français (de Jaurès à Jospin, de Poincaré à Chirac) couvrant le XX^{e} siècle et correspondant à 10 millions d'occurrences, disponible en format word (actuellement indisponible sur Internet) <http://www.unice.fr/ILF-CNRS/politext/>.

QUEBETEXT. Base de données textuelles du Trésor de la langue française au Québec. Contribution québécoise au projet international du « Trésor des vocabulaires francophones » (TVF) – qui visait à créer un ensemble de fonds textuels compatibles avec la base FRANTEXT – élaborée depuis les années 1990 dans le cadre des travaux du TLFQ de l'Université de Laval, Québec pour la préparation du DHFQ, réunissant des textes de genres discursifs différents, à dominante littéraire, d'auteurs québécois des XIX^{e} et XX^{e} siècles (textes publiés depuis 1837, les années 1960 étant les mieux représentées). Corpus consultable à l'aide du logiciel TACT, permettant des recherches

d'occurrences et de co-occurrences de mots selon divers paramètres ; accès individuel libre à quatre corpus de texte intégral libres de droits sur Internet : « Littérature québécoise (1837 à 1919) », « Textes sur l'anglicisme (1826–1930) », « Témoignages des voyageurs (1651–1899) », « Préfaces de répertoires lexicaux (1841–1957) » <www.tlfq.ulaval.ca/quebetext/>.

REGION. Banque de données de régionalismes du français hexagonal conçue dans le cadre de la préparation du DRF à Nancy, réunissant 7500 contextes de 220 ouvrages dus à 156 auteurs de France métropolitaine (le Centre-ouest exclu, sauf *Sentiers d'eau*, Mathé 1978), annoté par région d'appartenance des écrivains ; corpus consultable sur place au centre du FEW, ATILF, CNRS/Université de Lorraine.

SCIENTEXT. Base de données textuelles d'écrits scientifiques, totalisant 4,8 millions de mots de 219 textes en français, de plusieurs disciplines (surtout linguistique et science de l'éducation) (version 1.3.9, état du 01/06/2012) ; conçue à l'Université Stendhal (laboratoire LIDILEM, Grenoble) en collaboration avec les Universités Bretagne Sud (LI-CORN) et de Savoie (LLS), pour permettre l'étude des caractéristiques linguistiques des textes scientifiques à travers des structures phraséologiques et des marques lexicales et pour des requêtes sémantiques et syntaxiques ; corpus annoté syntaxiquement (*Treetagger*) et structurellement (TEI Lite) avec isolation des parties textuelles, consultable en trois modes – sémantique, simple guidé (par lemme, catégorie et relations syntaxiques), et complexe (par expressions régulières). En accès libre sur Internet <http://scientext.msh-alpes.fr>.

SUISTEXT. Base de données textuelles sur la littérature suisse romande contemporaine de langue française, contenant la totalité de l'œuvre de quatorze écrivains romands contemporains du XX^{e} siècle ; contribution suisse au projet international du TVF élaborée dans le cadre de la préparation du DSR sous la responsabilité de Pierre Knecht à l'Université de Neuchâtel ; corpus numérisé de 129 textes consultable par les collaborateurs du Centre de dialectologie et d'étude du français régional à l'Université de Neuchâtel.

TCOF. Traitement de Corpus Oraux du Français. Base de données de corpus oraux de français contemporain hexagonal, avec alignement texte/son (avec le logiciel *JTrans* ou transcriber, format .wav) constitué depuis 2005 à l'ATILF, CNRS/Université de Lorraine, comportant des

enregistrements de données orales numérisées et transcrites en TRS d'échanges adultes/enfants et entre adultes (surtout des entretiens, conversations et réunions de travail), de locuteurs français originaires de Lorraine pour la plupart ; sur approximativement quatre millions de mots enregistrés, environ un demi-million sont consultables avec un concordancier en accès libre (sous forme de fichiers séparés) via la plateforme CNRTL <http://www.cnrtl.fr/corpus/tcof/>.

VALIBEL. Banque de données textuelles du français oral en Belgique constituée par le Centre de recherche VALIBEL à Louvain-la-Neuve (1989–2009) visant à documenter la variation linguistique en Belgique francophone (Wallonie et Bruxelles), rassemblant un ensemble d'environ 4 millions de mots transcrits et alignés (gérée par le logiciel Praat) dans une base de données dynamique et évolutive en ligne, diffusé (sauf les tous derniers enregistrements) à l'aide du système [moca] sur demande avec mot de passe, pour une durée limitée <http://www.uclouvain.be/valibel-corpus.html>.

VALIRUN. Banque de données textuelles orales du français et créole réunionnais établi sous la responsabilité de Gudrun Ledegen, transcrites selon des conventions convergentes au corpus VALIBEL, totalisant environ 1 million de mots ; diffusion libre sur Internet prévue.

ZOBEL. Banque de données textuelles de littérature antillaise francophone comprenant l'œuvre en prose de l'écrivain martiniquais Joseph Zobel établie par André Thibault – onze ouvrages publiés de 1946 à 2002 (en onze fichiers .rtf) – totalisant 2044 pages, numérisées à l'aide d'Omnipage SE de Scansoft et annotées en XML par Patrick Drouin, consultable sur demande auprès des auteurs.

Coordonnées :

Inka Wissner
CNRS, ATILF
44 Avenue de la Libération
B.P. 30687
F-54063 Nancy Cedex
France
e-mail: inka(dot)wissner(at)atilf(dot)fr