Des corpus pour les français hors de France
Présentation de l'inventaire[1]

Plan de la présentation

Introduction

En des temps où les corpus ont acquis une importance considérable dans la pratique courante du travail des linguistes (mais aussi pour d'autres fins qui ne relèvent pas toutes de la connaissance scientifique), il apparaît indispensable de faire le point sur les corpus de français existants, au-delà de ceux qui concernent seulement l'Hexagone. Il peut en effet être intéressant pour plusieurs sortes d'experts de savoir où et comment situer les ressources disponibles, pour le français hexagonal bien entendu, mais aussi pour les français[2] de l'ensemble de la francophonie. Et peut-être même encore plus pour les français hors de

[1] Tous mes remerciements à tous ceux qui m'ont aidée tout au long de ce travail, en particulier pour établir de nouveaux contacts avec de potentiels informateurs sur des corpus de francophonie, de façon plus large que ce que me permettaient mes propres connaissances ou mes rencontres ponctuelles. Tout particulièrement, merci à Hélène Blondeau, Béatrice Akissi Boutin, André Dugas, France Martineau, Katja Ploog - sans compter nombre de répondants au questionnaire, qui ont fait des suggestions ayant souvent conduit à des pistes pertinentes. Merci pour finir à Paul Cappeau, qui m'a constamment accompagnée et soutenue à toutes les étapes du travail, et à Nicoletta Michelis, qui n'a pas plaint son temps, et a gentiment fait beaucoup plus que ce qui était attendu d'elle.

[2] On remarquera que j'utilise le terme au pluriel, ce qui se voit moins pour le terme français *français* que pour l'anglais *Englishes*. C'est une option que l'on peut discuter, mais qui me semble rendre justice à la grande diversité des français à travers le monde, et marginaliser toute velléité d'idéologie du standard.

France, dans la mesure où ils sont en partie méconnus de la majorité des linguistes, même des spécialistes de français, et dans la mesure où les travaux les décrivant se trouvent localisés dans des supports de publication éparpillés à travers le monde.

Pour le français hexagonal et pour les français européens, on dispose depuis 2005 de la recension de Cappeau & Seijedo, dont l'utilité évidente se mesure au nombre de fois où elle est citée, dans des publications diverses. Un inventaire plus récent vient aussi d'être effectué par l'IRCOM : http://ircom.corpus-ir.fr/wiki/doku.php?id=wiki:enquete.

Nous avons essayé ici de constituer un document du même type, en l'étendant à l'ensemble de la francophonie.

1. L'intérêt d'un inventaire des corpus de français hors de France

Il est évidemment toujours intéressant et utile de disposer du plus grand nombre possible de corpus pour une langue (et que cela se sache, en particulier à travers la publicité faite autour de leur existence, et des possibilités d'y recourir). De ce point de vue, on ne peut que se réjouir qu'existent des ouvrages comme Dittmar 2002 (qui parle des corpus à travers les transcriptions, selon des objectifs généralisants), ainsi que des réflexions mettant régulièrement les synthèses à jour, comme le fait la *Revue Française de Linguistique Appliquée* (voir *RFLA* 1996, 2007, 2012, traitant du français mais aussi d'autres langues). Je reviendrai en fin de partie 3 sur le fait qu'à mon avis ces réflexions sont loin d'être en nombre suffisant et suffisamment approfondies.

Mais une telle information, utile partout, devient encore plus cruciale pour les langues pluricentriques, comme l'anglais - sans contexte actuellement la plus pluricentrique de toutes les langues. La pluricentricité est aussi une caractéristique du français ; mais de fait, elle se manifeste surtout dans le réel des situations de francophonie, nettement moins dans les représentations que s'en font les locuteurs – comme en général les descripteurs d'ailleurs. Ce caractère a d'ailleurs pu faire l'objet de polémiques, comme l'évoquait déjà Lüdi en 1992[3]. La définition d'une langue pluricentrique, c'est l'existence d'une diversité de situations et d'aires (et donc de formes) à travers le monde, effet d'héritages historiques complexes et diversifiés, qui conduisent à des contacts relevant de différents types, et aboutissant à des statuts divers (majoritaire *vs* minoritaires *vs* isolats, au minimum) : bref, ce que l'on peut synthétiser sous l'idée de diversification des écologies. Tel est bien le cas pour la francophonie (voir par exemple Gadet, Ludwig & Pfänder 2008).

[3] Comme je lui demandais pourquoi il n'y avait pas de point d'interrogation final à son titre, Georges Lüdi (c. p.) me confirme que c'est bien ce qu'il aurait souhaité, mais que l'article est finalement paru sans.

De tels inventaires des corpus ont été établis sous diverses formes depuis déjà longtemps pour l'anglais, langue pour laquelle il existe de nombreuses publications portant un titre du genre *The World Englishes* (voir, parmi de nombreux autres, Kortmann *et al*. 2004 avec la problématique des « angloversals », ou déjà Cheshire *Ed.* 1991 - et Schneider 2007 pour une synthèse), de même d'ailleurs que des atlas établis à partir des comportements variables de certains traits ou phénomènes phoniques ou grammaticaux (voir Szrmecsany & Kortmann, 2009, par exemple). On est encore très loin que les travaux sur le français soient près à suivre la même voie.

C'est pourtant cette grande diversité des situations dans lesquelles le français est impliqué à travers le monde qui donne à cette langue une telle importance pour un point de vue de linguiste. Beaucoup plus que son nombre de locuteurs, qui linguistiquement n'est signe de rien, et qui somme toute demeure relativement modeste – une fourchette réaliste le situe entre 90 et 110 millions[4], ce qui place le français, selon les évaluations, à un rang entre la 11e et la 13e place des langues les plus parlées du monde (à supposer qu'une évaluation purement numérique revête réellement du sens).

Une meilleure connaissance des ressources disponibles (plus ou moins disponibles, comme le montre l'inventaire) apparaît aussi comme une condition nécessaire pour améliorer les modalités d'établissement de futurs nouveaux corpus (en particulier, pour tenter de remédier aux éventuels manques, insuffisances ou défauts flagrants).

2. Facteurs favorables et difficultés pour la réalisation de l'inventaire

L'établissement de cet inventaire des corpus de français hors de France a rencontré à la fois des conditions favorables, et quelques obstacles.

Parmi les conditions favorables, il y avait le fait de pouvoir bénéficier de l'expérience de prédécesseurs, en reprenant et adaptant le schéma de Cappeau & Seijedo 2005. Nous avons pu en effet repartir de l'organisation de leur questionnaire, en l'adaptant aux corpus de français hors de France, et en tenant compte du temps écoulé depuis qu'il avait été construit.

Un autre atout non négligeable a été l'appui de la DGLFLF, grâce au soutien enthousiaste d'Olivier Baude. C'est ce qui a rendu possible l'embauche de Nicoletta Michelis, jeune docteur de Paris Ouest, laquelle s'est chargée avec beaucoup de précision et d'intelligence de

[4] Il y a beaucoup d'autres évaluations qui circulent ici et là, mais on peut les regarder comme partisanes, militantes, ou tout simplement erronées. La plus grande difficulté est de parvenir à définir le terme « francophone », et de savoir où on ferait passer une frontière à partir de laquelle on renoncerait à ce statut.

la réalisation de la base de données[5]. Enfin, mon implication déjà ancienne dans l'étude de la francophonie hors de France m'a facilité le contact avec des chercheurs relevant de 16 nationalités différentes, dont les travaux se trouvent disséminés dans des publications qui ne sont pas toujours d'accès immédiat, ou qui souvent relèvent de la littérature grise (de même pour les thèses, qui ne sont pas toujours publiées, ni même répertoriées).

Les difficultés rencontrées, quant à elles, étaient en grande partie attendues, et elles relèvent d'autres facteurs, davantage liés à la matière traitée.

2.1. Difficultés liées à la configuration de la francophonie

La première difficulté concerne la situation éclatée de la francophonie[6] : au-delà de l'Europe, qui avait déjà été largement prise en compte dans l'inventaire de Cappeau-Sejeido, il fallait pouvoir s'adresser aussi bien à des spécialistes de français d'Afrique qu'à des spécialistes de français d'Amérique (les deux principales localisations du français hors Europe). Or, ces deux populations de linguistes ne se confrontent à peu près jamais. Les chercheurs ayant des interrogations générales sur la variabilité du français, au sens le plus large (voir Chaudenson, Mougeon & Béniak 1993, Gadet 2007 et 2011, ou l'introduction de Drescher & Neumann-Holzschuh 2010, qui font de ces interrogations un objectif essentiel de leur projet), sont encore loin d'être une perspective généralisée, ou même répandue. La plupart des travaux sont spécialisés sur une ou plusieurs aire(s)), et les « africanistes » et les « américanistes » constituent deux ensembles de linguistes tellement en distribution complémentaire parmi les spécialistes du français qu'ils ne publient pas dans les mêmes lieux, ne fréquentent pas les mêmes colloques… Bref, ils ne se croisent à peu près jamais. Il en va d'ailleurs de même pour les thèses, qui sont rarement réalisées sous la direction des mêmes directeurs (c'est, par nécessité, moins le cas dans des pays où le français est objet d'un enseignement en tant que langue étrangère, un titulaire de chaire pouvant se trouver amené à accepter un sujet qui ne relève pas prioritairement de son terrain de recherche).

Cette rareté de confrontation peut être opposée encore une fois à la situation de l'anglophonie, où existe depuis longtemps une tradition de comparaison entre variétés de l'anglais, comme l'atteste l'existence de revues dont certaines ont déjà une certaine ancienneté, comme *English*

[5] Merci aussi à Atanas Tchobanov, du laboratoire MoDyCo, qui a généreusement et savamment appuyé Nicoletta Michelis pour le formatage initial de la base de données.

[6] Ce n'est pas que l'anglophonie, l'hispanophonie ou la lusophonie (ensemble des entités où sont parlées les autres langues intercontinentales d'origine européenne, disséminées par la colonisation) soient tellement moins éclatées sur une carte du monde, mais du moins connaissent-elles de grands blocs (Amérique du nord pour la première, Amérique du sud pour la deuxième, Brésil pour la troisième). Or, l'existence d'un grand bloc ne se manifesterait dans la francophonie que dans une aire où le français n'est que rarement langue maternelle : l'Afrique de l'Ouest. Voir Gadet 2011 pour un bilan en forme de synthèse.

World-Wide, *English Today*, ou *World Englishes*, toutes dévolues à la comparaison entre les différents anglais[7]. Quant à la possibilité d'une revue qui serait entièrement consacrée à des réflexions sur les français hors de France, sur le modèle de revues comme les trois sur les anglais citées ci-dessus, je ne parviens même pas à imaginer que quelqu'un y songe[8].

D'autres « difficultés » de la francophonie sont davantage spécifiques à un ou à quelques terrain(s) : fallait-il faire figurer les territoires français d'outremer parmi les « français hors de France » ? S'ils sont administrativement français, ils sont souvent linguistiquement plus proches des français des zones créolophones (voir la relation entre la Réunion et Maurice, par exemple). Entre les arguments administratifs et les arguments scientifiques, nous avons penché sans état d'âme du second côté et nous avons, de façon *ad hoc* encore une fois, décidé de privilégier les aspects linguistiques (d'où la présence dans cet inventaire de corpus sur la Guadeloupe, la Martinique, la Guyane, la Nouvelle-Calédonie et la Réunion). Nous en avons toutefois fait une rubrique à part, sous le titre "France d'outremer".

2.2. Les difficultés d'un inventaire

Les autres difficultés rencontrées sont plus générales, et Paul Cappeau et Magali Sejeido s'étaient déjà heurtés à au moins une partie d'entre elles lors de leur inventaire des corpus de français européens :

- Fallait-il s'arrêter aux travaux dont le recueil de données se réclame explicitement du terme « corpus » ?

- Fallait-il s'arrêter aux travaux qui ont recueilli des corpus oraux, en considérant que l'inventaire des corpus écrits relèverait d'une autre problématique ? Cela aurait été du même coup rigidifier la frontière entre oral et écrit (et les corpus que l'on peut constituer sur cette base), à l'encontre de réflexions comme celles de France Martineau, qui recherche des traces de rapport à l'oralité dans des écrits ordinaires, et qui montre à quel point le truchement de l'écrit peut présenter de l'intérêt pour la compréhension même de l'oralité. Voir Martineau 2012 et son corpus de « Français familier ancien », que nous avons inclus ici au titre de l'objectif de recherche de traces d'oralité, ce que nous n'avons pas fait pour des corpus écrits sans visée de rapport à l'oral.

---

[7] Pour les relevés de corpus d'autres langues, on peut se reporter au premier numéro de la *RFLA* portant sur les corpus (1996), qui offre des bilans (évidemment datés), concernant des corpus d'anglais (Stig Johanson), d'espagnol et de portugais (Mireille Bilger), d'italien (Miriam Voghera) et d'allemand (Norbert Dittmar).

[8] Et l'on constate encore une fois la coupure entre d'un côté la revue *Le français en Afrique*, et d'un autre côté des revues canadiennes publiant surtout sur les français du Canada et d'Amérique du nord (voir aussi le colloque récurrent tous les deux ans, *Les français d'ici*, à entendre comme "les français d'Amérique du nord").

- Fallait-il dès lors privilégier les travaux qui se réclament de certaines problématiques, en particulier de la sociolinguistique ? Et du même coup exclure ceux qui se réclament davantage d'autres discipline? La question s'est posée pour la dialectologie (où les chercheurs ne recourent traditionnellement pas à la dénomination de « corpus » pour qualifier leur ensemble de données)[9]; pour l'acquisition (voir les corpus d'Yvan Rose sur Terre-Neuve, ou de Colette Noyau sur différents pays d'Afrique); pour la didactique (en particulier les très nombreux corpus d'interactions en salle de classe); ou encore pour l'analyse de conversation – du moins si on n'en fait pas un sous-domaine de la sociolinguistique (voir les corpus de Gaétane Dostie ou de Diane Vincent pour le Québec).

Tous ces questionnements attestent à quel point les frontières sont loin d'être étanches, pour ne pas dire qu'elles sont parfaitement arbitraires. Voici un exemple de difficultés auxquelles peuvent conduire ces obstacles, choisi à dessein parmi les travaux les plus anciens, non répertoriés dans l'inventaire. Emile Seutin (1975) n'avait nullement l'idée de « constituer un corpus » en documentant le parler de l'Ile-aux-Coudres (une ile sur le Saint-Laurent à côté de Québec). Cependant, bien des aspects de son travail, qu'il considère comme relevant de la dialectologie en rendant hommage à Louis Remacle, pourraient entrer dans une telle rubrique, dont la possibilité d'exploiter les fréquences. Ses données, qui pourraient paraître un peu hétéroclites à l'aune des exigences actuelles, comportent « plus de 500.000 mots » transcrits (dont 130.000 sont exploités informatiquement), « plus de deux cents heures d'enregistrement » (p. 23). Les informateurs sollicités sont « plus d'une centaine » à avoir été enregistrés « peu ou prou », ce qui constitue 1/17^{e} environ des 1700 habitants de la petite île québécoise. La question de la représentativité (qui d'ailleurs n'est pas forcément ni toujours la meilleure question à se poser[10]), se trouve ainsi largement remise en perspective.

Dernier type de question difficile sur laquelle il fallait trancher : il existe aujourd'hui plusieurs projets de grande ampleur, visant une comparabilité pan-francophone, au-delà de monographies sur une situation linguistique donnée ou l'approfondissement ethnographique d'un terrain déterminé. La plupart de ces grands corpus étant ouverts, ils sont toujours aujourd'hui en cours d'enrichissement, avec l'objectif de comparaisons soit au niveau mondial soit à l'échelle d'une partie du monde (en général un continent). Tel est le cas, au

[9] Du moins ne s'agit-il pas de corpus outillés, au sens où on l'entendrait de nos jours.

[10] Tous les termes mis entre guillemets dans ce paragraphe sont de Seutin. Comme il le dit dans son introduction : « chaque nouvelle bande enregistrée nous apporte des surprises et nous pose de nouveaux problèmes » (p. 23), ce qui ouvre des questions sur les limites numériques à envisager pour un ensemble de données - ou corpus. Ces réflexions peuvent fragiliser l'idée même de corpus, pour qui voudrait à toute force en appeler à la représentativité des données. C'est un point sur lequel nous reviendrons plus bas.

moins, de *PFC* (*Phonologie du Français Contemporain* – voir le site), de *CIEL_F* (*Corpus International Ecologique de la Langue Française* – voir le site, ainsi que Gadet *et al.* 2012), de *CFA* (*Contemporary French in Africa and the Indian Ocean* pour l'Afrique et l'Océan Indien – voir le site), ainsi que de *FRAN* (*FRançais en Amérique du nord* – voir le site, ainsi que Gadet & Martineau 2012). Fallait-il faire figurer une par une toutes les enquêtes relevant de ces projets ? Ou bien faire figurer une seule fois le(s) concepteur(s) du projet, compte tenu de ce que la méthodologie est largement reconduite d'un terrain à l'autre (de façon très précise, comme pour *PFC*, ou de façon seulement indicative de consignes générales, laissant davantage d'initiative aux équipes locales, comme pour *CIEL_F* ou pour *FRAN*) ? Cette question n'est pas à sous-estimer, étant donné ce que ces grands projets déterminent quant à la façon de concevoir la recherche en sciences du langage[11].

Pour aucune des questions ainsi soulevées, nous n'avons fait le choix d'une politique unique, stable et systématique, sauf peut-être pour exclure volontairement les textes donnant lieu à analyse de discours. Nous avons en général préféré prendre des décisions locales, au coup par coup. Nous ne cherchons pas à dissimuler qu'elles sont influencées, certes d'abord par des considérations scientifiques et pratiques, mais aussi par des options personnelles, pour chacun des corpus concernés, quant à leur constitution comme dans le mode d'exploitation qu'ils ont connus. Par exemple, nous avons fait figurer quelques corpus de *PFC* recueillis dans des zones assez peu documentées par ailleurs (voir par exemple Douglas Walker sur l'Alberta).

3. Les corpus de français hors de France, parmi les corpus de français

Il y a quelques caractéristiques récurrentes des corpus de français (en France et hors de France), qui sont assez bien connues désormais. Aussi ne ferons-nous ici que les rappeler de façon synthétique, en renvoyant chaque fois que possible à des publications disponibles.

3.1. Un retard historique pour les corpus sur le français, et la France à la traîne

A un niveau général tout d'abord, une première caractéristique est le retard qui a été historiquement pris par les corpus de français : ce que l'on voit aisément en faisant une comparaison avec les bilans pour l'anglais, l'allemand, l'italien ou l'espagnol (comme l'a montré entre autres le premier numéro de la *RFLA* portant sur les corpus - 1996, ou l'ouvrage

[11] Avant ces grands projets institutionnels, les corpus individuels, nécessairement de taille plus modeste, donnent lieu à un éparpillement qui rend un inventaire indispensable. Les grands projets ont eu pour effet un regroupement, au moins des méthodologies si ce n'est des objectifs théoriques, et chaque projet comporte son propre inventaire.

de Dittmar 2002 pour l'allemand)[12]. Les corpus de français hors de France ont été constitués avant, de façon plus rapide, et dans au moins une partie des cas (davantage en Amérique qu'en Afrique), avec plus de moyens institutionnels et financiers que pour les corpus constitués en France même. En effet, la collecte de corpus a commencé dès la fin des années 60 en Amérique du nord (d'abord au Canada, à la fois sur Montréal et sur l'Ontario), alors qu'en France, ces recueils ne commencent qu'à la toute fin des années 70, avec le travail, d'abord largement personnel et « artisanal », de Claire Blanche-Benveniste et de l'équipe du GARS, autour de la revue *Recherches sur le français parlé* (qui paraît à partir de 1977[13] - et se saborde en 2004). C'est un peu plus tard, au début des années 80, que commencent les recueils en Afrique, effectués d'abord davantage par des chercheurs étrangers (surtout français) que par des locaux, eux-mêmes essentiellement mobilisés par les projets d'inventaire des particularités lexicales, qui exigent moins l'appui sur des corpus au sens où nous l'entendons ici (ainsi, on n'a en général pas affaire à des relevés d'énoncés oraux, mais plutôt à des dépouillements de la presse – voir par exemple Lafage 2002 et 2003, qui y a ajouté de nombreuses sources de documentations directes et indirectes, y compris le questionnement de natifs).

On ne peut qu'avancer des hypothèses pour tenter de comprendre un tel retard français par rapport aux autres traditions des grandes langues internationales. Je ferai l'hypothèse que les causes sont sans doute plutôt négatives du côté de la France : les Français eux-mêmes ont longtemps été trop figés dans leur normativisme, affiché ou plus diffus, ainsi que dans leur tradition littéraire pour prendre vraiment au sérieux la langue parlée, surtout ordinaire. L'accueil mitigé réservé à ce qui constitue historiquement le premier corpus de français de France, le *Français fondamental* constitué dans un objectif didactique, peut probablement s'interpréter en partie comme le franchissement d'un tabou culturel quant à l'étude de la

[12] Inutile de préciser que les corpus de français hors de France (du moins ceux rendus publics) se soucient de respecter les « bonnes pratiques » (voir Baude 2006), tout autant qu'en France même, et tout autant qu'ailleurs. Les règles à ce sujet sont même souvent plus draconiennes qu'en France, en particulier en Amérique du nord, comme l'atteste la nécessité, dans toutes les universités canadiennes, de passer devant un comité d'éthique pour soumettre un projet de constitution de corpus – je peux faire état de l'exemple de l'Université de Moncton au Nouveau-Brunswick, en remerciant Annette Boudreau de m'avoir donné communication des documents de son propre comité d'éthique.

[13] A partir de la fin des années 90, l'équipe du GARS a pu bénéficier d'un soutien de la DGLFLF pour la réalisation du *CRFP* (*Corpus de Référence du Français Parlé*, constitué à l'initiative de Mireille Bilger, sur différentes villes de l'Hexagone), qui malheureusement n'est toujours pas disponible sur internet, malgré l'intérêt qu'il y aurait à disposer publiquement de ses 440.000 mots transcrits, diversifiés géographiquement ainsi que dans une organisation en trois rubriques : « parole privée », « parole publique » et « parole professionnelle ».

langue[14]. Et il demeure très instructif, quant au rapport des Français à leur langue parlée ordinaire, de remarquer que ce sont des Britanniques qui ont, à partir des années 50, constitué l'autre grand corpus historique de français hexagonal: le « Corpus d'Orléans », ou plus précisément *Etude Sociolinguistique Sur Orléans*, auquel des chercheurs d'Orléans autour de Gabriel Bergounioux et d'Olivier Baude sont actuellement en train de donner un prolongement, par un deuxième volet tenant compte des nouvelles réflexions sur les corpus – on va donc disposer d'un *ESLO2* qui permettra la comparaison avec *ESLO1*.

Et si c'est ailleurs qu'en France qu'ont d'abord été constitués des corpus de francophonie, cela me semble pour des raisons bien compréhensibles concernant le reste de la francophonie, et en particulier le Canada : à quelle autre ressource qu'à des corpus (oraux comme écrits, d'ailleurs) peut-on recourir pour documenter quelque chose d'un peu précis sur des parlers ne disposant pas de tradition de description grammaticale spécifique, et étant, comme toutes les versions orales ordinaires des langues, objet d'instabilité et de variation (voir Drescher & Neumann-Holzschuh 2010, ou Brasseur & Falkert 2005 pour l'Amérique du nord) ? Sans doute y a-t-il aussi eu une influence de traditions linguistiques plus axées sur le terrain et le travail empirique (la linguistique américaine pour le Canada).

Ce sont donc des Canadiens (Québécois et Ontariens) qui ont fait les premières récoltes de quelque ampleur et avec quelque systématicité, dès la fin des années 60, avec la conséquence imprévue, du fait de l'absence d'autres corpus, qu'ont d'abord été considérés comme spécifiquement canadiens des traits dont on découvrira par la suite qu'une bonne partie d'entre eux concerne tous les vernaculaires du français (Gadet 2011).

3.2. Peu de préoccupations d'affichage et de publicité

A première vue, ce n'est pas une quelconque propriété de la francophonie qui devrait entraîner l'apparente absence de préoccupations pour la notoriété et la diffusion des corpus, prolongée de façon plus récente dans une diffusion toujours restreinte sur la toile. C'est pourtant bien ce qui s'est passé, aussi bien en France (voir *CorpAix*, le *CRFP*, *ESLO1*, tous particulièrement mal diffusés) qu'à l'étranger.

Ainsi, les Canadiens, qui ont été pionniers pour la récolte de corpus de francophonie autour des années 70 (en particulier sous l'impulsion de Gillian et David Sankoff à Montréal - voir Sankoff *et al*. 1976, de Raymond Mougeon et son équipe en Ontario – voir un historique dans

[14] Il demeure un peu mystérieux quant à l'histoire des idées linguistiques que ce soit un linguiste qui n'avait lui-même rien de normatif, Marcel Cohen, qui a signé une charge très polémique contre le projet du français fondamental, avec son ouvrage de 1956 (dont il dit sans davantage de précision qu'il a été écrit avec « un groupe de linguistes »). Il est vraisemblable qu'il s'agit surtout de motivations politiques.

Mougeon & Béniak 1991, de Shana Poplack à Ottawa-Hull – voir son article de 1989[15]), ne participaient d'aucun circuit de diffusion sur la toile. Le premier corpus à avoir été rendu public sur le web est le tout récent *CFPQ* (mis en ligne en mai 2009), dont la réalisation est due à l'impulsion de Gaétane Dostie à Sherbrooke, avec des visées plutôt conversationnalistes. De même pour l'Afrique, la revue *Le français en Afrique* ne présente qu'occasionnellement des corpus (et à ma connaissance n'a jamais présenté un inventaire), son objectif primordial étant davantage de les analyser.

La plupart des chercheurs n'ont accès à ces corpus que par le seul truchement de ce qui figure dans des articles publiés, ou bien parce qu'ils connaissent l'auteur du corpus ou le directeur de la thèse pour laquelle a été récolté le corpus. Ces corpus demeurent donc en grande partie hors de portée de la majorité des chercheurs, et ils finissent par risquer de tomber dans l'oubli ou même par se détériorer, faute de volonté explicite et concertée de conservation et de valorisation.[16]

3.3. Excursus sur les corpus (oraux) en général

L'expression d'un étonnement pour terminer sur ce point : comment se fait-il que l'on rencontre, toutes langues confondues, si peu de réflexions globales sur l'activité de recueil de corpus, et sur ce qu'ils permettent de nouveau quant à la façon de « faire de la linguistique », ou bien, au contraire, les illusions qu'ils laissent passer d'une révolution radicale ? On ne peut pas éviter de rappeler ici les *a prioris* sur la répartition des tâches en sciences du langage : constituer des corpus est une activité qui demeure peu valorisée dans un CV. Ce que l'évaluation de la recherche laisse ainsi entendre au moins implicitement, c'est qu'il s'agit d'une étape sans enjeux théoriques, sur laquelle il n'y a pas lieu de s'appesantir, et que l'on peut abandonner aux tâcherons ou aux petites mains de la recherche afin de passer au plus vite aux activités plus prestigieuses que sont l'analyse, la diffusion et la comparaison des résultats. Les corpus apparaissent ainsi à ce point regardés comme des donnés dont nul n'a le besoin de savoir de quelle façon ils ont été recueillis, que ceux-là même qui déplorent qu'on ne dispose

[15] Mais le présent inventaire fait aussi état, quand la documentation disponible le permettait, d'autres corpus des années 60 et 70, comme ceux de Gesner, de Ryan ou le fond Létourneaux. Merci à Raphaële Wiesmath et à Sandrine Hallion-Brès pour avoir fourni de nombreux détails très précieux sur ces corpus, lors de leurs riches réponses au questionnaire. J'ai aussi pu prendre largement appui, pour les débuts de l'histoire, sur l'article de Boisvert & Laurendeau 1988.

[16] Un exemple sur les aléas de destinée des enregistrements, concernant le corpus du *Français fondamental* : selon Rivenc 2006, les disques qui ont servi de base à la constitution du corpus ont été détruits à la demande de Georges Gougenheim, au vu de leur piètre qualité sonore (ou, le soupçon en vient, de ce qu'ils attestaient sur la langue parlée). Merci à Daniel Coste de m'avoir fait connaître cet article. Le contre-exemple, beaucoup plus rare, est le sauvetage du premier *ESLO* par des chercheurs d'Orléans.

pas de davantage de corpus se penchent en général bien peu sur leurs qualités et sur les modalités de leur recueil. Or, de nos jours, ce n'est pas la masse de données à disposition qui est en cause (exigence assez facile à remplir, même si tout type de données n'est pas forcément aussi aisé d'accès), ce sont leurs qualités intrinsèques, au niveau du recueil comme au niveau de la transcription et des différentes étapes de révision, en tous cas avant même de commencer l'analyse[17].

4. Quelques constats préliminaires sur les corpus présentés dans l'inventaire

L'inventaire montre qu'il existe de fortes disparités entre les corpus dont il est fait état ici, elles aussi plus ou moins prévisibles compte tenu de leur extrême diversité.

4.1. Les aspects matériels

- La première est tellement évidente qu'elle ne présente guère d'intérêt: il s'agit de la disparité des ampleurs.

- Disparités dans les motivations scientifiques des initiatives (objectifs avant tout sociolinguistiques, ou bien syntaxiques, phonétiques ou phonologiques, lexicaux, discursifs et pragmatiques…), au-delà du fait qu'une bonne partie de l'homogénéité constatée n'est qu'un effet, tout à fait artificiel, de la sélection des répondants qui a été pratiquée.

- Disparités dans les façons de dénommer les corpus (par le nom de l'auteur, le lieu de recueil, les objectifs du projet, parfois l'adjonction de la date – en particulier dans les cas de retour sur un terrain[18]…) – parfois il y a pour un même corpus alternance entre les façons de les dénommer, ce qui ne facilite pas leur recensement.

- Disparités dans les modes de conservation, qui va souvent de pair avec le degré de précision? Evidemment, cet aspect se module selon les époques (voir ce qui a été dit à la note 16), les exigences ne faisant qu'augmenter.

[17] Voir les réflexions d'un numéro de la revue *Verbum* (2010 affiché 2008), numéro qui toutefois se concentre davantage sur les étapes d'exploitation que sur celles de constitution des corpus, comme c'est aussi le cas de *Langages* 2008 et de *LIDIL* 2005. Ces publications contrastent ainsi avec les trois numéros de la *RFLA* déjà évoqués, qui s'efforcent de prendre en compte toutes les étapes du processus. Elles ne reviennent aucunement sur des réflexions comme celle de Mondada 1998, qui montre à quel point le fonctionnement du chercheur sur le terrain participe de la configuration des données (avec d'évidents effets sur les corpus).

[18] Un exemple très illustratif est la destinée du corpus Sankoff-Cedergren de Montréal, initialement constitué, parmi les tout premiers, dès la fin des années 60 (voir Sankoff *et al.* 1976), puis repris une première fois en 1984 (voir Thibault & Vincent 1990), et une deuxième fois en 1995, les deux recueils ultérieurs ayant visé à retracer le plus grand nombre possible d'informateurs initiaux – ce qui permet une réflexion sur les trajectoires linguistiques des interviewés (voir Blondeau, Sankoff & Charity, 2002). Il est référé à ces corpus sous les noms de Montréal 71 (mais aussi Sankoff-Cedergren), Montréal 84 et Montréal 95. Voir Vincent 2009 pour le rappel de cette (déjà longue) histoire, qui montre que ce n'est pas parce qu'on reconduit un terrain qu'on n'introduit pas des innovations (en l'occurrence, l'introduction d'objectifs écologiques dans une méthodologie variationniste).

- Disparités dans les modalités de transcription : en général, transcription orthographique – mais qui peut ou non laisser place à des aménagements, plus ou moins étendus selon les cas ; quelquefois transcription phonétique – cette dernière option étant d'ailleurs rare dès lors que le projet manifeste quelque ampleur[19] ; plus récemment, recours ou non à des logiciels d'aide à la transcription, le plus fréquemment Praat ou Transcriber ;
- Disparités dans les façons d'évaluer leur volume (nombre de sous-corpus ou d'informateurs, nombre d'heures d'enregistrement, nombre de mots transcrits, combinaison de ces mesures d'évaluation…) ; on observe avec étonnement qu'un nombre non négligeable de répondants ne documente pas cette question, sans qu'il soit possible de savoir pourquoi.
- Enfin, disparités dans ce qui est dit des modalités d'accessibilité : au-delà de ce qui a été dit plus haut quant à la rareté d'accessibilité sur la toile, les cas d'accessibilité directe et ouverte sont loin de constituer la majorité, même si les choses changent peu à peu.

Les répondants n'ont pas toujours eu la patience de se conformer à la grille qui leur était soumise, et malgré quelques relances, nous ne sommes pas parvenus à des caractérisations complètes et unifiées, ce qui n'a pas non plus été une surprise, compte tenu du nombre de concepteurs de corpus sollicités. Le présent inventaire fait en effet état de 145 corpus, et comme certains répondants sont les maîtres d'œuvre de jusqu'à 5 corpus différents (c'est le cas de Ruth King, qui a constitué 3 corpus en Acadie et 2 ailleurs au Canada, ou de Gisèle Prignitz avec ses 5 corpus du Burkina-Faso), le nombre de concepteurs n'est "que" de 93 relevant d'institutions de 16 pays différents, les plus nombreux étant les Canadiens (qui travaillent à peu près sans exception sur le Canada).

Dès le lancement de l'opération (en 2007 !), je n'avais aucun doute sur le fait que l'exhaustivité était un horizon mythique, inatteignable. Comme on le craignait, il s'est avéré en particulier difficile de documenter des « petits » corpus, dont on n'entend généralement parler que soit par hasard, soit lors de la lecture d'articles qui les évoquent ou en tirent des exemples, généralement sans donner trop de détails[20] – sans compter les mémoires, souvent encore moins diffusés que les thèses. Il faut aussi rappeler l'importance des époques de

[19] Une exception notoire, le corpus de Katja Ploog recueilli à Abidjan, entièrement transcrit en phonétique. Son option fondamentale se comprend dans la mesure où il s'agit d'enregistrements d'enfants des rues, au français particulièrement fragile et instable, qui soulève d'énormes problèmes de transcription. Mais étant donnée l'ampleur du corpus, il est facile d'imaginer le nombre d'heures qu'a exigé la transcription intégrale…

[20] Un exemple, croisé récemment : l'article Fonseca-Greber 2007 fait état d'un corpus de 8 h 30 de « Conversational Swiss French » (français de Suisse romande), comportant 117.000 mots - ce qui n'est pas négligeable, surtout pour quelqu'un qui a travaillé seule (certains des corpus qui figurent dans cette base de données n'en comportent guère plus, parfois moins). L'auteur du corpus figure malheureusement parmi ceux qui n'ont pas donné suite à plusieurs relances.

constitution de ces corpus. Avant le début des années 90 et l'explosion des ordinateurs individuels, les corpus étaient saisis sur des machines à écrire de bureau, et il faudrait aujourd'hui entièrement refaire les saisies - et numériser tout ce qui a été enregistré sur des K7 ou des bandes susceptibles de se dégrader. Il ne semble pas que beaucoup d'institutions de recherche en voient clairement la nécessité.

4.2. Quelques remarques qui se dégagent

Voici maintenant quelques remarques sur ce qui ressort des réponses recueillies. Que ce type de vaste récapitulation soit aujourd'hui à l'ordre du jour, je n'en prendrai pour indice que le fait que France Martineau est actuellement en train d'établir un inventaire du même type pour les français d'Amérique du nord (autour du corpus *FRAN* et du projet GTRC *Le français à la mesure d'un continent*).

- Les dates des recueils, régulières, ne semblent manifester ni ralentissement ni accélération (comme on aurait pu s'y attendre – sans doute un peu naïvement - du fait de l'amélioration des moyens techniques). Les retours (d'un chercheur ou d'une équipe) sur un même terrain demeurent particulièrement rares : on peut évoquer les exemples de la postérité du corpus Sankoff-Cedregren à Montréal, la démarche récurrente de Raymond Mougeon en Ontario, ou les retours de Ruth King à Terre-Neuve.
- Certaines aires s'avèrent largement quadrillées, voire labourées, contrairement à d'autres qui demeurent peu exploitées. Outre le fait qu'on décrit toujours mieux un terrain déjà bien balisé, les raisons de la fréquence de fréquentation relèvent de motifs variés. Ainsi, pour la Belgique, c'est certainement une volonté scientifique du Centre VaLiBel, dont le travail a été constant et soutenu depuis 1989. A Montréal, on peut davantage évoquer des visées politiques, du moins pour le début de l'aventure, des financements publics ayant été octroyés dans les années 60 et 70, afin de disposer des bases pour établir un modèle du « français standard d'ici ». L'Acadie a fourni 32 corpus répertoriés ici (compte non tenu des « comparaisons »), et on est tenté de mettre ce chiffre élevé en relation avec l'importance qu'y revêtent les questions de langue. En Afrique, les motivations de la concentration sont encore d'un autre ordre, une francophonie orale pouvant être documentée étant souvent limitée à des zones bien précises, essentiellement urbaines (voir la domination des recueils à Abidjan pour la Côte d'Ivoire).
- Plus le corpus est gros, plus il requiert des infrastructures institutionnelles, dont Poplack 1989 a donné un bon aperçu. Désormais, les « grands projets de corpus » ont pris le relais, avec d'importants financements publics (voir *PFC*, *CIEL_F*, ou *FRAN* pour les français hors de France).

- Les matériaux enregistrés consistent pour l'essentiel en interviews (dites plus ou moins informelles, selon la tradition initiée par Labov de tenter de « mettre l'interviewé à son aise »). Celles-ci continuent à constituer la majeure partie des enregistrements (beaucoup plus fréquemment interviews individuelles qu'interviews de dyades ou interviews collectives).
- Il y a assez peu de parole spontanée (ou naturelle, ou écologique), ce qui n'est pas une surprise, étant donné la difficulté à les recueillir et tout autant à les transcrire.
- De façon plus étonnante, il apparaît peu de recours aux ressources des médias, et en particulier aux enregistrements de radios. Or, elles peuvent constituer une ressource non négligeable pour des chercheurs ayant des difficultés pour accéder à un terrain[21]. Cappeau & Sejeido 2005 faisaient d'ailleurs déjà le même constat de relative rareté des corpus de radios. Cette observation conduit à s'interroger sur ce qui amène des chercheurs à préférer le caractère socialement sollicité des interviews au naturel des émissions de radio, qui peuvent souvent offrir une large palette de diversité. C'est d'ailleurs l'une des situations retenues par le protocole de *CIEL_F* (voir par exemple l'intérêt communicatif des *phone-in*)[22].
- Contrairement à ce à quoi on pouvait s'attendre, c'est loin d'être le cas que la majorité des corpus aient été établis dans l'objectif de rédaction d'un mémoire ou d'une thèse, qui constituent certainement un aiguillon puissant pour se lancer dans le chemin aride de la réalisation d'un corpus.
- De nombreux chercheurs sont sensibles à l'urgence de documenter des attestations pour des variétés « en danger » (par exemple, les corpus de Ruth King à Terre-Neuve et à l'Ile du Prince Edward, ou ceux de Karen Flikeid ou de Julia Hennemann en Nouvelle-Ecosse). Qu'il s'agisse de risque de disparition imminente, ou d'évolution rapide d'un parler, dans certains lieux, il est important de disposer d'une profondeur de documentation sur des situations instables ou mouvantes, comme la Côte d'Ivoire où les conditions du *français populaire ivoirien* devenu aujourd'hui le *nouchi* évoluent très rapidement : les corpus recueillis à la fin des années 70 – comme celui de Jean-Louis Hattiger attesté dans sa thèse de 1981, regardé comme le premier véritable corpus sur Abidjan (voire sur l'Afrique) – ne correspondent plus à des façons de parler actuelles : cependant, il s'agit de documents irremplaçables pour retracer l'historique d'une évolution.

[21] Voir par exemple le corpus de Cristina Petras, une jeune Roumaine qui a constitué le corpus de sa thèse sur la base d'émissions de radios communautaires grâce à internet, avant même de parvenir à se déplacer personnellement en Nouvelle-Ecosse (voir Petras 2008).

[22] On peut avoir un avant-gout de ce que sera le projet complet quand il sera mis en ligne, avec une collection d'extraits de une minute chacun, sur http://www.ciel-f.org/vitrine.

Conclusion

L'entrecroisement de ces différents facteurs laisse prévoir ce qui apparaîtra malheureusement comme difficilement contournable dans ce travail : son caractère bricolé et intrinsèquement lacunaire, outre qu'il est par nature voué à être très rapidement obsolète.

L'enquête avait initialement été lancée en 2007, à l'occasion d'un article paru depuis longtemps dans la *Revue Française de Linguistique Appliquée* (Cappeau & Gadet 2007). Vu le temps écoulé depuis le début de l'enquête, nous avons tenté de revenir vers les concepteurs des corpus, et de faire des mises à jour chaque fois que cela a été possible. Nous souhaiterions évidemment avoir amorcé là un processus de mise au point continu, et que la base de données puisse régulièrement être mise à jour, ce qui dépend en grande partie de la bonne volonté des concepteurs de corpus, ainsi que de celle des futurs concepteurs.

Qu'il y ait là des spécificités récurrentes des études de linguistique française, cela a récemment été rappelé par Blanche-Benveniste 2010, dans son introduction et dans sa conclusion un peu désabusées. Toutefois, il nous paraît évident que, même imparfait comme il l'est actuellement, cet inventaire permettra d'ouvrir une brèche dans ce qui apparaît au pire comme de l'ignorance (la majorité des linguistes français ou francisants ne savent que peu de choses sur les français hors de France – au-delà des stéréotypes), au mieux comme du désintérêt (les linguistes ne s'intéressent guère plus aux français hors de France qu'ils ne se sont historiquement intéressés au français tel qu'il est réellement parlé). On peut ainsi espérer que le fait de connaître l'existence de ces corpus, voire d'y avoir facilement accès (dans des cas que l'on espère de plus en plus nombreux), fera rapidement évoluer l'état de la documentation sur le français.

Compte tenu de cette extrême diversité, il semble qu'aujourd'hui, au contraire des rêves initiaux de pouvoir parvenir un jour à une harmonisation des disparités, on admette mieux qu'il y a bel et bien des raisons, y compris scientifiques, qui les motivent, et qu'il est tout à fait vain de chercher à les niveler. Certains, dont je fais partie, les considèrent même comme la condition indispensable pour atteindre la meilleure adéquation aux objectifs de recherche…

Références citées, une sélection[23]

BAUDE Olivier (Dir.) (2006). *Corpus oraux. Guide des bonnes pratiques*, Presses Universitaires d'Orléans/CNRS Editions.

[23] Les quelques références qui suivent n'ont aucune prétention à quelque exhaustivité. Il s'agit seulement de documenter les thèmes abordés dans cette introduction (il n'y est donc pas question d'exploitation des corpus). Pour des références plus complètes, il est tout à fait impossible de renvoyer à un document unique. Voir cependant, parmi d'autres, Vincent 2009, Pusch & Raible 2002, Gadet 2011. Pour tout ce qui concerne les publications issues des différents grands projets cités ici, on se référera à leurs sites respectifs.

BLANCHE-BENVENISTE Claire (2010). *Le français. Usages de la langue parlée*, Peeters, Leuven & Paris.
BLONDEAU Hélène, Gillian SANKOFF & Ann CHARITY (2002). "Parcours individuels dans deux changements linguistiques en cours en français montréalais", *Revue québécoise de linguistique*, Vol 31 n° 1, 13-38.
BOISVERT Lionel & LAURENDEAU Paul (1988). « Répertoire des corpus québécois de langue orale ». *Revue québécoise de linguistique* 17-2. 241- 262.
BRASSEUR Patrice & FALKERT Anika (Dir.) (2005). Français d'Amérique : *approches morphosyntaxiques*, Paris, L'Harmattan.
CAPPEAU Paul & GADET Françoise (2007). « Où en sont les corpus de français parlés ? », *RFLA* XII-1, 129-33.
CAPPEAU Paul & SEIJEDO Magali (2005). *Inventaire des corpus oraux en langue française*, www.dglflf.culture.gouv.fr
CHAUDENSON Robert, MOUGEON Raymond & BENIAK Edouard (1993). *Vers une approche panlectale de la variation du français*. Paris: Didier-Erudition.
CHESHIRE Jenny (*Ed*.) (1991). *English around the world: sociolinguistic perspectives*, Cambridge, Cambridge University Press.
CFA, Contemporary French in Africa, http://www.hf.uio.no/ikos/english/research/projects/cfa/index.html
CFPQ, Corpus de français parlé québécois, http://pages.usherbrooke.ca/cfpq/index.php
CIEL_F, Corpus International Ecologique de la Langue Française, http://ciel-f.org/
COHEN Marcel (1956). *Français élémentaire ? Non*. Paris. Editions sociales.
DITTMAR Norbert (2002). *Transkription. Ein Leitfaden mit Aufgaben für Studenten, Forscher und Laien*. Opladen, Leske & Budrich.
DRESCHER Martina & NEUMANN-HOLZSCHUH Ingrid (2010). « Les variétés non-hexagonales du français et la syntaxe de l'oral. Première approche », *in* M. Drescher & I. Neumann-Holzschuh (Dir.), *La syntaxe de l'oral dans les variétés non-hexagonales du français*, Tübingen : Stauffenburg Verlag, 9-35.
English Today, Cambridge University Press.
English World-Wide, John Benjamins.
FONSECA-GREBER Bonnibeth (2007). "The Emergence of Emphatic 'ne' in Conversational Swiss French", *JFLS* Vol 17 n° 3, 249-75.
Le français en Afrique, http://www.unice.fr/ILF/ofcaf (27 numéros parus jusqu'en 2012).
FRAN, Le français en Amérique du Nord, http://continent.uottawa.ca/corpus-et-ressources-electroniques/
GADET Françoise (2007). « La variation de tous les français », *LINX* 57, 155-64.
GADET Françoise (2011). « What can be learned about the grammar of French from corpora of French outside France », in *Grammatik und Corpora 2009*, *Hgg* M. KONOPKA, J. KUBCZAK, C. MAIR, F. STICHA & U. WASSNER (dir.), Tübingen : Narr Verlag, 87-120.
GADET Françoise, LUDWIG Ralph & PFÄNDER Stefan (2008). « Francophonie et typologie des situations ». *Cahiers de Linguistique Revue de Sociolinguistique et de Sociologie de la langue française*, Varia, 143-162.
GADET Françoise, LUDWIG Ralph, MONDADA Lorenza, PFAENDER Stefan & SIMON Anne-Catherine (2012). « CIEL_F : choix épistémologiques et réalisations empiriques d'un grand corpus de français parlé ». *Revue Française de Linguistique Appliquée* XVII-1, 19-54.
GADET Françoise & MARTINEAU France (2012). « Le français panfrancophone saisi à travers un maillage de réseaux », *Cahiers de Linguistique* 38-2, 63-88.
HATTIGER Jean-Louis (1981). *Morpho-syntaxe du groupe nominal dans un corpus de français populaire d'Abidjan*, Thèse de 3e cycle, Université de Strasbourg.
KORTMANN Berndt *et al*. (*Eds*.) (2004). *A Handbook of varieties of English*. Berlin / New York: de Gruyter.
LAFAGE, Suzanne (2002 et 2003). *Le lexique français de Côte d'Ivoire. Appropriation et créativité*. Vol 1 et 2. *Le français en Afrique* n° 16 et 17.
Langages, (2008). « Construction des faits en linguistique. La place des corpus ». N° 171.
LIDIL, (2005). « Corpus oraux et diversité des approches ». N° 31.
LÜDI Georges (1992). « French as a pluricentric language », *in* M. Clyne (*Ed*), *Pluricentric Languages. Differing Norms in Different Nations*, Berlin & New York, Mouton de Gruyter, 149-78.
MARTINEAU France (2012). « Les voix silencieuses de la sociolinguistique historique ». *Cahiers de linguistique* 38-1. 111-135.
MARTINEAU France (à paraître). *Répertoire pour la préservation des corpus*. Site "Polyphonies du français" (www.polyphonies.uottawa.ca)

MONDADA Lorenza (1998). « Technologies et interactions dans la fabrication du terrain du linguiste », *in Le travail du chercheur sur le terrain*, *Cahiers de l'ILSL*, Université de Lausanne n° 10, 39-68.
MOUGEON Raymond & BENIAK Edouard (1991). *Linguistic Consequences of Language Contact and Restriction : The Case of French in Ontario, Canada*, Oxford : Oxford University Press.
PETRAS Cristina Anca (2008). *Les emprunts et la dynamique linguistique*. Thèse non publiée des universités de Iasu (Roumanie) et d'Avignon.
PFC, Phonologie du Français Contemporain, http://www.projet-pfc.net/?accueil:intro
POPLACK Shana (1989). « The care and handling of a mega-corpus », *in* R. Fasold & D. Schiffrin (*Eds*), *Language Change and Variation*, Amsterdam, Benjamins, 411-51.
PUSCH Claus & RAIBLE Wolfgang (*Eds*) (2002). *Romance Corpus Linguistics*. Tübingen. Gunter Narr.
Recherches sur le français parlé, Revue de l'Université d'Aix-Marseille2, 18 numéros parus, entre 1977 et 2004.
Revue française de linguistique appliquée (*RFLA*) (1996). *Corpus: de leur constitution à leur exploitation*. Vol. 1-2, http://www.rfla-journal.org
Revue française de linguistique appliquée (*RFLA*) (2007). *Corpus: état des lieux et perspectives*. Vol. XII-1, http://www.rfla-journal.org
Revue française de linguistique appliquée (*RFLA*) (2012) http://www.rfla-journal.org
RIVENC Paul (2006). « Les auteurs du *Français fondamental* face à un objet nouveau et insolite : l'interaction orale », *Documents pour l'histoire du français langue étrangère ou seconde* [En ligne], 36 | 2006, mis en ligne le 24 août 2011, consulté le 25 janvier 2013. URL : http://dhfles.revues.org/1185
SANKOFF David, SANKOFF Gillian, LABERGE Suzanne & TOPHAM Marjory (1976). « Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale ». *Cahiers de linguistique de l'Université du Québec*. 6 : 85-125.
SCHNEIDER Edgar (2007). *Postcolonial English. Varieties around the world*. Cambridge: Cambridge University Press.
SEUTIN Emile (1975). *Description grammaticale du parler de l'Ile-aux-Coudres*. Montréal: Presses de l'Université de Montréal.
SZMRECSANYI Benedikt & KORTMANN Bernd (2009). « Vernacular Universals and Angloversals in a Typological Perspective », *in* M. Filppula, J. Klemola & H. Paulasto, Filppula (eds.). *Vernacular universals and language contacts. Evidence from varieties of English and beyond*. New York / London: Routledge. 33-53.
THIBAULT Pierrette & VINCENT Diane (1990). *Un corpus de français parlé*, Québec, Recherches sociolinguistiques 1.
VaLiBel http://www.uclouvain.be/valibel
Verbum (2008). *Corpus oraux : recueil et analyse de données*, Tome XXX n° 4.
VINCENT Diane (2009). « Corpus, banques de données, collections d'exemples. Réflexions et expériences », *Cahiers de Linguistique*, 33/2, 81-96.
World Englishes, Wiley-Blackwell.

Françoise Gadet
Université de Paris Ouest Nanterre la Défense & MoDyCo
gadet@u-paris10.fr