Quels outils pour étudier la variation du français ? L'apport de la linguistique de corpus à l'exemple d'un diatopisme polysémique, *prime* (adj.)

La linguistique française dispose désormais de nombreux corpus préparés au cours des dernières décennies, établis avec des objectifs variables. Ces outils deviennent de plus en plus incontournables pour décrire le français à partir de réalisations effectives de la langue ; en même temps, ils imposent au chercheur de maîtriser les outils de la linguistique de corpus et de prendre un ensemble de précautions (cf. Cappeau/Gadet, 2007). Pour décrire la langue générale, des projets comme le TLF ont ainsi démontré la valeur ajoutée d'un dictionnaire qui s'appuie exclusivement sur un corpus (Frantext), plutôt que sur ses prédécesseurs. Pourtant, en linguistique variationnelle, et en lexicographie francophone en particulier (où *francophone* est utilisé au sens de « qui porte sur le français dans l'espace francophone »), la variation spatiale est encore largement identifiée selon une approche différentielle par rapport à un ensemble métalinguistique. Cette approche consiste pour l'essentiel à exploiter les ouvrages de référence qui font autorité, comme le TLF et FEW et les dictionnaires d'états anciens de la langue (*français de référence* selon Poirier, 2005 : 497). Ceux-ci sont complétés de sources spécialisées qui portent sur des domaines mal ou non représentés en lexicographie générale, en particulier sur des langues en contact et la variation diatopique du français même (du DSR au DLF ; BDLP). Ces dictionnaires s'appuient quant à eux sur les données lexicographiques antérieures, des enquêtes de terrain, ou encore des ensembles textuels, écrits et oraux. Cet ensemble reste métalinguistique, mais constitue un outil de travail objectivable, contrairement à l'idée abstraite que le lexicographe peut se faire du français général, ou même d'un 'bon usage', voire d'un français 'de Paris' ou 'de France' – qui eux aussi s'intègrent dans un diasystème variable (cf. Coseriu, 1981). Dans l'approche différentielle, les corpus sont donc bien utilisés pour analyser l'une ou l'autre variété de français, mais ont pour l'essentiel un rôle complémentaire, venant diversifier la nature de données majoritairement lexicographiques. Or, les outils de la linguistique de corpus, le nombre croissant de ressources textuelles, leur taille et leur présence dans différentes zones de l'espace francophone ne permettraient-ils pas d'identifier ce qui relève de la variation diatopique dans la langue par une comparaison *pan*francophone des corpus ? Dans une telle optique, quel est l'apport effectif des ressources textuelles ? Permettent-elles de se passer de la documentation lexicographique traditionnelle ?

Pour répondre à cette problématique, cette communication mobilise les concepts et outils de la lexicologie philologique historique et de l'approche comparative panfrancophone en particulier, complétée de la méthode fine de l'étymologie-reconstruction. Celle-ci vise à reconstruire progressivement les étapes d'évolution d'un élément de la langue. Elle s'appuie sur une comparaison des régularités formelles et sur les évolutions sémantiques parallèles des formes d'une famille lexicale, et donc sur les traces laissées par la documentation historique, mais aussi sur la documentation contemporaine – qui reste la plus vaste et la mieux établie (Chauveau, 2013 : 178-179). L'approche lexicographique traditionnelle sera confrontée à une exploitation de données textuelles dans une visée panfrancophone. Je constituerai pour cela un ensemble cohérent de corpus, en comprenant *corpus* comme un ensemble de réalisations discursives qui a été conçu pour des fins d'analyse linguistique dans le respect de critères de sélection particuliers et qui est exploitable par l'intermédiaire d'un moteur de recherche. J'ai sélectionné les grands corpus du français contemporain monolingues dans la francophonie qui dépassent une taille d'environ 5 millions de mots occurrences, qui permettent des requêtes lexicales, et qui sont rendus accessibles en ligne à la communauté scientifique. Ce corpus est susceptible de permettre une analyse efficace et rentable, même pour des unités lexicales marquées. Quoique sélectif, il a pour grand avantage de constituer un outil de travail vérifiable par d'autres membres de la communauté scientifique, pouvant servir de corpus 'de référence', comme pendant textuel des dictionnaires et grammaires de référence.

Si la taille des corpus n'est pas un critère satisfaisant, il est pertinent surtout pour l'analyse de noms, verbes, et adjectifs – donc d'unités qui appartiennent aux classes majeures, comme celles analysées ici – moins fréquents dans l'absolu que les unités appartenant aux classes mineures, comme les conjonctions ou les prépositions. Le seuil retenu ici est donc minimal. Même parmi les spécialistes de corpus qui s'intéressent non pas à des dimensions variationnelles de la langue, mais à la langue générale, la taille actuelle des corpus oraux (autour de quelques milliers de mots), ne permet pas « de faire des recherches lexicales ni d'établir des statistiques fiables sur les usages » (Baude, 2006 : 29). Selon les standards internationaux, un corpus d'étude conforme donnerait accès à 3 millions de mots à l'oral, 15 millions à l'écrit (ORFÉO). Dans une visée diatopique, le critère de la taille est d'autant plus critique au sens où il doit s'appliquer non pas à l'ensemble des données exploitées, mais à *chaque zone* de l'espace francophone. Il s'agit donc bien d'un des critères de valeur pour permettre des analyses lexicales qualitativement fructueuses sur la variation diatopique, à côté de paramètres comme la spécificité des sujets et des genres discursifs qu'ils accueillent, et les types de requêtes que permettent leurs moteurs de recherche (Wissner, 2012 : 252). Les limites d'exploitation de corpus oraux pour des recherches lexicales dans une visée diatopique avaient déjà été soulevées dans le cadre de la description des diatopismes du français en Belgique et en Afrique (Queffélec, 1997). La situation a-t-elle considérablement changé en plus de vingt ans ?

Pour répondre à cette question seront ici exploités des corpus de trois types. Il s'agit tout d'abord des corpus traditionnels qui donnent accès à des données surtout littéraires : Frantext (francophonie) et FLI (Canada) ; s'y joint Varitext, pour des données journalistiques et littéraires en Europe et en Afrique. Un deuxième type de corpus contient des données tirées du web francophone, notamment FrWac – devenu le plus grand corpus de la langue française – qui donnent accès à des données variées en termes thématiques, situationnels et discursifs, et apportent des données de types de discours propres (comme les blogs). Un troisième type est celui des corpus de transcriptions d'enregistrements oraux, où seule la base ESLO répond au seuil quantitatif fixé (France : Ouest). Approchant les sept millions de mots selon des approximations, c'est le corpus oral le plus grand du français qui soit entièrement rendu accessible à la communauté scientifique. S'y joignent les corpus d'enregistrements permettant des analyses lexicales à visée panfrancophone pour différentes zones de la francophonie : PFC et CIÉL-F (établis par une méthodologie unifiée), en attendant l'accès au Corpus d'Étude pour le Français Contemporain (CEFC) pour l'usage en Europe surtout. Pour l'Amérique du Nord, le corpus panaméricain FRAN est complété de la BDTS (dont un sous-ensemble est en accès libre), du corpus 'oral' CFPQ et du corpus MCVF pour les états antérieurs du français. Sont exclus de l'analyse les ensembles textuels qui n'ont pas été conçus pour l'analyse linguistique, comme les ressources journalistiques et les ressources web instables.

Dans le but de vérifier l'apport effectif des corpus dans une optique historico-comparative panfrancophone, cette communication propose une étude de cas plutôt qu'une approche exhaustive. Elle portera sur une forme polysémique du français moderne, *prime* (adj.) (français de référence : *précoce* ; *vif* ; *soupe au lait*…). Une étude détaillée permettra d'évaluer l'apport des corpus, y compris des nouveaux corpus panfrancophones, pour l'identification des caractéristiques, sémantiques, syntagmatiques, géolinguistiques et historiques des unités lexicales. L'analyse retracera non seulement les différents emplois de *prime* en Europe, mais aussi leur trajet historico-variétal à travers le temps et l'espace, d'une variété à l'autre, y compris outre-mer où le français fut durablement exporté à partir du XVII^{e} siècle. C'est dans un deuxième temps qu'il s'agira d'apporter des éléments de réponse à la problématique méthodologique soulevée, en évaluant l'apport des dictionnaires de français, d'un côté, et des grands corpus dans l'espace francophone, de l'autre.

Références bibliographiques

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

BDTS. *Banque de données textuelles de Sherbrooke*, < https://www.usherbrooke.ca/crifuq/>.

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

CEFC. Corpus d'Étude pour le Français Contemporain : corpus de français en Europe surtout en cours de construction, <http://www.projet-orfeo.fr/>.

CFPQ. *Corpus de français parlé au Québec*, <http://pages.usherbrooke.ca/cfpq/corpus.php>.

Chauveau, J.-P. (2013). Fr. *ébarouir* : étymologie-histoire et étymologie-reconstruction. *RLiR* 77, 167-182.

CIÉL-F. *Corpus International Écologique de la Langue Française* de français oral en interaction collecté de 2006 à 2012 dans quinze zones de l'espace francophone, <www.ciel-f.org>.

Coseriu, E. (1981 [1958]). Los conceptos de ‹ dialecto ›, ‹ nivel › y ‹ estilo de lengua › y el sentido propio de la dialectología. *Lingüística española actual* 3, 1-32.

DLF : Valdman, A. *et al*. (2010). *Dictionary of Louisiana French*, Jackson : University Press of Mississippi.

DSR : Thibault, A. (1997). *Dictionnaire suisse romand. Particularités lexicales du français contemporain. Une contribution au Trésor des Vocabulaires francophones*. Genève : Zoé.

ESLO. *Enquête Sociolinguistique à Orléans*, <http://eslo.tge-adonis.fr>.

FEW : Wartburg, W. von (1928-2003). *Französisches etymologisches Wörterbuch*. Bonn *et al.* : Klopp *et al*.

FLI. *Fichier lexical informatisé* du français québécois, <http://www.tlfq.ulaval.ca/fichier/>.

FRAN. *Corpus des français d'Amérique du Nord*, <http://continent.uottawa.ca/fr/corpus/corpus/corpus-interrogeable-fran/>.

Frantext. Base de données textuelles de la littérature française : corpus à dominante littéraire constitué de quelque 248 millions de mots du XVI^{e} au XXI^{e} siècles, <http://www.Frantext.fr/>.

FrWac. Corpus textuel du domaine « fr » du Web d'environ 1,6 milliard de mots, construit dans le cadre du projet WaCky Wide Web, <http://wacky.sslmit. unibo.it/doku.php?id=download>.

MCVF. *Modéliser le changement : les voies du français*, <http://www.voies.uottawa.ca/ corpus_pg_fr.html>.

PFC. La Phonologie du Français Contemporain : usages, variétés et structure (PFC) – 'Recherche'. Base de données recueillies en 48 points d'enquête dans la francophonie, <www.projet-pfc.net>.

Poirier, Cl. (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 Frey, C./Latin, D. (éds.), *Le Corpus lexicographique*. Louvain-la-Neuve : Duculot, 353-368.

TLF : Imbs, P./Quemada, B. (1971-1994). *Trésor de la langue française*. Paris : Gallimard.

Varitext. *Corpus des variétés nationales du français* <http://syrah.uni-koeln.de/varitext/>.

Wissner, I. (2012). Les grands corpus du français moderne : des outils pour étudier le lexique diatopiquement marqué ?, *SKY Journal of Linguistics* 25, 233-272.