Bruno COURBON

PRATIQUES SÉMANTIQUES ET DIFFÉRENCES INTERINDIVIDUELLES À L'ÈRE DES CORPUS INFORMATISÉS

> […] dès que les relations sont nombreuses, elles sont soumises à des interférences et aussitôt l'enquête discursive des approximations devient une nécessité épistémologique[1].
>
> […] la connaissance théorique n'est pas antérieure à la recherche empirique du point de vue temporel ("chronologique"), mais du point de vue logique ou rationnel : c'est-à-dire qu'elle est le fondement de toute recherche sur des faits (même dans les cas où ce fondement n'est qu'intuitif et reste sous-entendu)[2].

1. Corpus et sémantique : les nécessités de l'oxymore[3]

Le dualisme corps / esprit prend une tournure spécifique lorsqu'il s'agit de langage. On le retrouve tant au niveau général, dans l'appréhension des rapports entre langage et pensée, qu'au niveau particulier des rapports entre signe et concept, forme signifiante et forme signifiée. Cette relation entre formes matérielles et immatérialité, qui traverse les représentations collectives de la science occidentale, se rejoue dans l'association moderne entre corpus et sémantique. Longtemps sous la dépendance de la philosophie, la sémantique a eu tendance à aborder son objet de façon très conceptuelle, donc… incorporelle. Les analogies entre corps et esprit, langage et pensée,

1 G. Bachelard (2004 [1938]) : *La formation de l'esprit scientifique*, Paris, Vrin, p. 254.

2 Coseriu (1998b : 47).

3 Condamines (2005 : 15-16) : « [la] combinaison [sémantique et corpus] pourrait presque donner un effet d'oxymoron où se verrait opposer la notion de possibilité de stabiliser le système (ce qui était un des objectifs de la sémantique à ses débuts) et celle de variation des usages ».

corpus et sémantique sont manifestes, à la différence près (fondamentale) que ce dernier rapport se situe sur le plan de l'analyse. Il ne renvoie qu'indirectement à la réalité linguistique. Malgré tous les efforts pour la recontextualiser de façon appropriée, la signification observée n'est déjà plus qu'une trace de sens discursif, l'objet d'un regard qui, nécessairement, l'abstrait de ses conditions de manifestation premières, lui ôtant ainsi le fondement de sa valeur effective.

Toutefois, bien que le recours à des instruments de saisie et d'enregistrement de « données » linguistiques présente une vision artificiellement simplifiée de la réalité langagière, les corpus constituent aujourd'hui l'un des meilleurs moyens dont disposent les sémanticiens pour décrire des usages sémantiques en circulation. Tout le monde s'entend sur le fait que les techniques modernes utilisées dans le traitement de corpus permettent aux linguistes, en augmentant la réalité empirique examinée, d'ouvrir le champ des observables pertinents. La sémantique ne fait pas exception à cette règle (par ex., l'analyse de corpus aide à identifier des isotopies, à relever des cooccurrences régulières, à expliquer certains « segments répétés », etc.). Corollairement, le volume et la nature des « données » disponibles – *i. e*. de l'ensemble des prises linguistiques –, engendrent une complexification des phénomènes à analyser. L'analyse de corpus invite à reprendre à nouveaux frais le travail d'identification des faits sémantiques. Par-delà les significations toutes faites exposées dans les dictionnaires, les sémanticiens sont amenés à relever des phénomènes sémantisables de nature variable, dont l'identité, les limites et les relations avec l'« environnement » énonciatif peuvent être difficiles à dégager en dehors d'une situation référentielle concrète. L'hétérogénéité des objets linguistiques recueillis dans un corpus pose des défis théoriques qui étaient en partie insoupçonnés avant l'ère des corpus informatisés.

Obligeant à repenser la distinction entre parole/discours (acte) et langue (potentiel), la fréquentation de corpus peut porter à croire qu'il serait possible d'accéder à la diffusion et à la régularisation d'usages sémantiques (Courbon 2012a). Mais ce qui vaut pour les productions prises en corpus n'a pas nécessairement valeur générale[4]. Tandis que la démarche classique consiste à décrire des actualisations de signifiés largement décontextualisées, les vastes corpus permettent d'envisager d'étudier le processus inverse, de

4 Radicalisée en opposition, la distinction acte/potentiel tend à ne laisser voir l'individuel que comme de l'anecdotique (une simple réalisation), tandis que sont survalorisées l'hypothèse d'un système total ou celle de l'objectivité du texte. La conscience interdiscursive et, en général, les mises en relation morphosémantiques se rejouent pourtant chaque fois chez l'individu, lequel n'a jamais accès à l'ensemble des productions d'un genre ou d'un domaine, ni à un macro-système complet ou, *a fortiori*, à un diasystème, mais toujours à des parcelles de langue, qui prennent la forme d'éléments et de relations qu'il interprète et (re)systématise diversement selon ses besoins et sa sensibilité linguistiques.

« potentialisation » ou, plus précisément, de (re)structuration sémantique de formes signifiantes (constitution de faits sémantiques).

Aussi paradoxal que cela puisse paraître, ce ne sont ni la nature des « données », ni les formes de présence de la signification linguistique qui constituent le cœur de l'objet empirique d'une sémantique sur corpus, mais la référence. Le fondement empirique de la sémantique – au moyen de corpus ou d'autres types de « data » (Laks 2008) – révèle que référentialisme et différentialisme, loin de s'opposer, se complètent (les sujets, tout en référant, contrastent). Se pose alors la question de la relation entre expérience / référence, d'une part, et contraste / structuration, de l'autre. Les limites liées à un matériau fortement décontextualisé invitent à prendre en compte des « données empiriques » plus immédiates, constitutives de l'univers de référence des sujets. C'est dans cette perspective que s'inscrit le projet de créer un observatoire du comportement linguistique d'internautes dans/sur des forums électroniques, avec un intérêt particulier pour les différences diatopiques (section 3). Précisons que l'objectif de ce texte est non pas de mettre à l'épreuve de corpus une approche théorique de la signification, mais, d'abord, de réfléchir à une possibilité d'envisager la relation apparemment aporétique entre corpus et analyse sémantique. Quelques observations empiriques illustreront le propos ; elles ne correspondent en aucun cas à des résultats d'analyse.

2. Des corpus en diasémantique : d'une aporie à l'autre

2.1. *Corpus, diversité linguistique et considérations sémantiques*

Entre le travail pionnier de Mortureux (1981)[5] et les synthèses de Condamines (2005) et de Rastier (2011)[6], les problèmes de corpus n'ont presque pas été traités par les sémanticiens. La diversité sémantique, en outre, n'a que très peu retenu l'attention. Elle dérange. Ainsi, Gadet (1989 : 55) définissait la variation relativement aux trois domaines suivants : « phonologie, morphologie et syntaxe ». Sa position correspond à une conception courante. Bien qu'elle soulève le problème de la significativité qui se corrèle avec la variation (dia)linguistique (Gadet 2003 : 13), sa synthèse du « matériau variationnel qu'exploite le français » (*ibid.* : 43-44) ne comporte qu'un sous-point pour la dimension sémantique, intitulé « créativité sémantique » et inscrit dans le domaine « lexique et discours », lui-même assez limité (1/7 dudit matériau). La question du sens préoccupe cependant

5 Cet article présente une réflexion sur le rôle du corpus en sémantique ; l'analyse porte sur les différences dans le rapport aux « données » entre trois travaux académiques.

6 Ouvrage collectif introduit par A. Condamines dans le premier cas, regroupement de textes écrits ou co-écrits par l'auteur dans le second.

l'auteure, qui dénonce « le postulat de la synonymie » (*ibid.* : 109) : « Est-ce que le style (ou la variété) suppose la stabilité du sens ? D'ordre sémantique, ou pragmatique ? Autrement dit, est-ce qu'on peut dire la même chose en disant différemment, selon différents styles ? » (Gadet 2004 : 3).

Les segments (phoniques, graphiques, morphologiques, lexicaux…) ou les composés ([morpho]phonologiques, lexicaux, syntaxiques…) étant des objets plus directement identifiables que le sens, les linguistes se souciant de « variation » ont eu tendance à leur accorder la priorité[7]. La tendance à négliger la dimension sémantique ou à la subordonner à la dimension physique de la langue se retrouve jusqu'en sémantique même, où le sens est parfois conçu comme le résultat d'habitudes sémiotiques (on peut y voir un effet de l'orientation typiquement sémasiologique de la perspective interprétative). Pourtant, bien qu'elle se rapporte à des formes matérielles, la signification linguistique ne peut en aucun cas y être réduite : elle constitue un objet d'observation à part entière.

Il se peut par ailleurs que la vision accidentaliste du passage – nommé *actualisation* – d'un signifié abstrait à l'une de ses multiples formes de réalisation (le sens discursif) ait en quelque sorte masqué la question du caractère plus ou moins régulier de la variabilité dudit sens. Dans cette perspective, le signifié (ou sens abstrait) est posé comme un invariant à reconstruire (« signifié de puissance » chez Guillaume, « forme schématique » chez Culioli…), tandis que le sens discursif n'est que le résultat effectif produit par le contact du signifié avec un « réel extralinguistique » que sa contingence n'autorise pas à intégrer à l'analyse sémantique proprement dite (on parle alors d'effets de « contexte », de paramètres pragmatiques, etc.). Identifié à l'éphémère et au chaos, le sens discursif reste en périphérie. Or, le sens manifeste de profondes régularités à travers la diversité des « formes » qu'il prend en discours. Du fait de leur intangibilité et du nombre quasi illimité de leurs formes de réalisation, les usages sémantiques présentent peut-être une variabilité plus grande que les usages phonétiques ou syntaxiques.

On distinguera au moins trois niveaux de « variations » sémantiques : 1) la « variation » du sens effectif d'une unité, qui dépend en partie d'éléments

7 L'escamotage de la dimension sémantique est chose courante. On l'observe dans la définition par extension que donne Coseriu (1998b : 34) de la « grammaire structurale » : « morphosyntaxe, phonologie et lexicologie descriptives ». Dans l'ouvrage qu'elle dirige sur la linguistique de corpus, Bilger (2000 : 7) met de l'avant la syntaxe, la sociolinguistique et l'analyse du discours. Dans Habert *et al.* (1997 : 184), les objets touchés par la « variation » sont principalement d'ordre matériel : « graphèmes, formes, lemmes, lexies, système de catégories grammaticales, séquences, etc. ». Pour Vincent (2009 : 91), « tous les niveaux d'une variété de langue parlée » sont définis comme suit : « phonétique, morphologique, syntaxique, lexical, discursif ».

référentiels spécifiques (c'est à partir de cette variation que le signifié de l'unité peut être établi ou « reconstitué », tout comme le sont, plus près de la représentation concrète, les usages sémantiques [= 3)]) ; 2) la « variation » de la forme matérielle que prend un phénomène sémantique, éventuellement déjà établi en signifié (nous avons ainsi pu observer les nombreuses variantes de la combinaison *la partie émergée de l'iceberg* [Courbon 2012b] ; les relations de type [para] synonymie ou antonymie forment les extrêmes de cette « variation » morphosémantique) ; 3) la « variation » de l'usage sémantique d'un fait sémiotique établi, généralisée ou non dans la pratique individuelle des sujets. À chaque niveau de manifestation de la variabilité sémantique correspondent des degrés divers d'abstraction et de complexité. En outre, parce qu'elle est inévitablement formulaire, l'analyse du sens linguistique entraîne aussi une déformation, qui consiste à désincorporer celui-ci du faisceau de conditions empiriques dont il procède.

C'est au troisième niveau de « variation » du sens qu'est typiquement associé le marquage lexicographique. Le fait que les phénomènes sémantiques prennent des formes matérielles plus ou moins semblables pour l'usager est la raison qui justifie, temporairement, l'utilisation du terme *variation*. Du point de vue du sujet, il n'y aurait de variation que là où coexistent des formes « équivalentes » dans l'usage qu'il fait de la langue ou dans l'appréhension d'usages qu'il ne pratique pas, mais qu'il comprend. C'est aussi la prise en compte de ce troisième niveau qui permet d'expliquer des différences fréquentielles importantes (voir annexe I et tableau 2), puisque c'est à ce niveau que sont instituées les normes sémantiques, au sens « syn(/m)- » (-topique, -chronique, -phasique…) de la typologie générique proposée par Flydal (1952) et augmentée par Coseriu (1966). Sur le modèle de la distinction temporelle synchronie / diachronie, ces auteurs ont battu en brèche la vision macrosystémique proposée par les générations précédentes. Mais la conception diasystémique telle que la développent notamment Coseriu ou Weinreich reste très abstraite. À l'instar de la priorité qui a été accordée à la vision descendante du signifié vers le sens « contextuel », les pratiques effectives dans la conception traditionnelle du diasystème correspondent à autant d'actes ponctuels, qui ne sont que des symptômes « actualisant » (« activant », dans une terminologie plus récente) des propriétés systémiques établies dans une classe d'usagers déterminée[8].

8 Sans pour autant nier l'incidence des appartenances sociales sur les pratiques langagières des individus, la perspective adoptée ici consiste à « délivrer » dans une certaine mesure la pratique individuelle de la double détermination sociale (au sens large) et « contextuelle ». Examiner les pratiques individuelles convergentes ancrées dans des référentiels partagés (lesquels prévalent en général sur les « contextes » spécifiques) conduit à appréhender des trames communes ; celles-ci, en tant que faits collectifs, présentent le gros avantage d'être beaucoup plus concrètes que ne le sont les catégories génériques fondées sur l'intuition collective de différences dia-lectales (cf. la « méthode de l'abstraction » qu'évoque Bloomfield 1933 : 45). La définition de ces classes *a priori* d'usagers

Cette vision de la langue est empreinte d'un fort déterminisme social ; posé *a priori* et bien qu'il renvoie à des réalités empiriques, le marquage de classe est dans sa forme structurelle le reflet d'une époque où la notion de mobilité dialinguistique n'était que très peu considérée (Haugen, de qui nous reprenons le terme *dialinguistique*, était à cet égard avant-gardiste). Comme chez Weinreich *et al*. (1968), les « variétés » sont associées à des ensembles fixes définis en termes d'espaces physiques ou symboliques. Le rôle actif du sujet dans la transformation des classes auxquelles son lecte doit appartenir s'efface au profit de critères apparemment plus stables, du moins plus « généraux ». La conception classiste – qui est en partie (indirectement) fondée – a été un passage obligé dans l'histoire de la linguistique entre le postulat d'une langue commune uniforme et l'observation de pratiques plus ou moins partagées selon les individus, premiers artisans de la linguiversité.

Les études sémantiques classiques se situent principalement aux deux premiers niveaux, transindividuels, de « variation » du sens. La recherche dia-sémantique avec corpus favorise l'observation de la « variation » des usages sémantiques, en tenant compte des différences intra- et interindividuelles. Il ne s'agit pas d'exalter l'individu en réduisant la factualité linguistique à cette seule dimension. Néanmoins, la pratique des corpus incite à reconsidérer cette dimension qui fut longtemps minorée, voire négligée dans l'analyse des faits empiriques, pour des raisons tant pratiques que théoriques (exception faite des approches discursives, la sémantique du XX^e^ siècle a largement fait abstraction de la réalité linguistique des sujets de langue). Mettre en œuvre une telle démarche nécessitait la collecte d'une masse critique de productions verbales, afin de distinguer les conditions de « variation » dans les pratiques sémantiques des individus : outre les paramètres de l'origine sociale / géographique, de l'âge et du cadre énonciatif, la thématique, l'intérêt pour un objet particulier, la qualité de l'attention portée à la précision des termes, le mimétisme, le désir de se démarquer… font partie des facteurs à considérer. Négliger la variabilité des pratiques (inter)individuelles accroît le risque de reconduire, sous couvert d'une démarche empirique, l'image d'une langue-« système » homogène, constituée de « sous-systèmes » clairement définis, synthétisés à partir de productions désituées, et surtout détachées de la compétence effective de leurs principaux artisans, les sujets-agents de la langue.

---

est d'ailleurs largement escamotée, ou atomisée sous l'effet d'une conception paramétrique qui en éclate les contours – et pour cause : l'« étiquetage » socio-catégoriel donne la priorité aux conceptions, par nature abstraites, plutôt qu'aux réalisations ; or, les deux plans doivent être considérés. Outre que l'observation de la variabilité et de l'invariabilité sémantiques interindividuelles est une façon de mettre à l'épreuve des faits empiriques certaines catégories conceptionnelles, leur description sert aussi à mieux comprendre les raisons de leur maintien dans l'imaginaire linguistique (Houdebine). Quelques pistes de la démarche envisagée seront esquissées dans la section 3.

2.2. *« Ceci n'est pas un corpus » : la vision techniciste de la ressource*

Comme le souligne Vincent (2009 : 82), la notion de corpus est loin d'être uniforme. On remarque en particulier un éloignement de sa valeur première de recueil de productions discursives au profit d'une vision très technique. La conception classique de la notion de corpus persiste cependant dans la lexicographie, tant générale que spécialisée[9]. La définition qu'en donnait Mortureux (1981 : 49) précède manifestement la révolution technicienne : « l'ensemble des segments en langue naturelle […] fonctionnant comme autonymes […] dans les discours des études concernées, et supportant les métadiscours […] qu'elles constituent ». C'est encore ce genre de définition que l'on trouve dans le *Nouveau dictionnaire encyclopédique des sciences du langage* (Ducrot et Schaeffer 1995 : 60) : « Étudier une langue, c'est […] avant tout réunir un ensemble, aussi varié que possible, d'énoncés effectivement émis par des utilisateurs de cette langue à une époque donnée (cet ensemble = le corpus) ».

On conçoit aisément l'utilité pratique de distinguer entre bases ou banques de données, collections d'exemples et corpus, voire la pertinence méthodologique de définir des types et des sous-types de corpus. Mais de telles distinctions sont établies d'abord pour répondre à des objectifs de recherche particuliers. Sans mise en perspective, elles peuvent devenir des obstacles heuristiques à la reconstitution des faits. Prises dans une logique manichéenne (et prescriptive), ces catégories pratiques peuvent détourner le regard de l'objet principal – les phénomènes linguistiques –, en érigeant en théorie un purisme de corpus (faut-il y voir un « déplacement » symptomatique des prescriptions de l'« éternel grammairien » [Berrendonner] ? un rejet de la gratuité du donné ? ou bien encore le simple désir, compréhensible, de contrôler l'objet ?)[10]. La valeur effective des faits recueillis devrait primer, quelle que soit la forme sous laquelle ils sont présentés. À cet égard, il convient de ne pas confondre objectif technique et objet d'étude scientifique (cf. Rastier 2011 : 46). Échantillonnage, étiquetage, catégorisation, (en)codage, normalisation, équilibrage, représentativité correspondent à des possibilités techniques. Les subordonner aux normes de l'appareil statistique serait absurde. Dans les faits, la (méta)structuration de textes n'est pertinente qu'à partir du moment où

9 Le *Petit Robert* définit le corpus comme un « ensemble fini d'énoncés réels réunis en vue de l'étude d'un phénomène linguistique » (2013, sous CORPUS), définition très semblable à celle de Tournier et Tournier (2009).

10 Ce purisme, qui valorise l'homogénéité ou l'« hétérogénéité réglée » (Marchello-Nizia 1985 : 486), a précédé l'ère du tout-informatique, comme on le constate dans l'article CORPUS de Mounin (1974) : « La méthode du corpus a le désavantage d'inclure des matériaux hétérogènes, variations stylistiques ou dialectales, répétitions, phrases inachevées, que le sujet parlant a prononcées ». On préférerait sans doute une réalité plus « propre ». Disons à la décharge de Mounin qu'il n'est pas l'auteur de l'article en question.

elle représente l'activité des sujets, premiers responsables de la structuration linguistique effective. Si la méta-analyse structurelle des « données » arrive trop tôt, elle peut biaiser, voire bloquer l'analyse proprement dite[11].

La visée du traitement automatisé de discours (TAD) ne doit pas être confondue avec les objectifs d'une science du langage. Pour le TAD, le corpus est l'objet, son interprétation une fin en soi ; c'est la raison pour laquelle l'alimentation de la machine doit être optimisée à partir du donné. En linguistique, en revanche, le corpus est un outil heuristique (on parle ainsi de « linguistique outillée » [Habert] ou de « sémantique instrumentée » [Rastier]). La définition de la notion de corpus est par conséquent relative à l'usage qu'on en fait ; elle doit être reformulée selon les objectifs et les besoins de recherche. Qu'adviendrait-il en astrophysique si les spécialistes du télescope édictaient les normes de calibrage et d'utilisation de l'instrument sans se soucier des usages en partie imprévisibles qu'en feront les scientifiques ? Savoir si l'on a affaire ou non à un « véritable » corpus relève par conséquent d'un débat technocratique : le recours à des « tranches » de réel validables et, si possible, identifiables n'a d'autre but que de servir l'analyse linguistique.

L'objet est de toute façon bien imparfait, mais on peut s'entendre sur une définition de base. L'usage commun du terme *corpus* fait le lien entre la vision étroite (exclusive) d'un ensemble de données pré-calibrées à partir de paramètres externes et la vision plus large (inclusive) d'une ressource exploitable[12]. Cet usage se retrouve dans la définition de Kilgariff et Greffenstette (2008 : 90) : « *a corpus is a collection of texts when considered as an object of language or literary study* ». Vu sous cet angle, et bien qu'abstrait de ses conditions premières, le corpus est une représentation de discours qui se définit par sa fonction ustensile. Si l'on accepte cette définition, il convient de reconnaître – comme l'ont fait très tôt nombre de linguistes – la valeur empirique des productions inscrites sur la Toile, et en particulier l'image sans précédent de diversité qu'elles présentent[13]. L'écriture virtuelle

11 Habert *et al*. (1997 : 23) mentionnent la circularité des démarches typologiques habituelles.

12 La vision contemporaine du corpus est l'héritière de la conception traditionnelle du corpus canonique, qui devait correspondre à un ensemble uniforme, délimité et clos. Rey (2008 : 28) rappelle que le terme *corpus* « désign[e] par définition un ensemble fermé ». On retrouve cette propriété dans la plupart des définitions contemporaines : « [a corpus] may be considered under four main headings: sampling representativeness, finite size, machine-readable form, a standard reference » (McEnery and Wilson 1996 : 21) ; les corpus correspondent à des « compendiums de faits, clos, structurés, stables et publiquement partagés » (Laks 2008 : 5).

13 Keller et Lapata (2003) montraient ainsi que l'application d'instruments probabilistes aux ressources du Web donne des résultats que des corpus pré-structurés, du fait de leur moindre volume, ne permettent pas d'obtenir.

constituent une vaste mémoire discursive susceptible d'être exploitée sur le plan linguistique. Il est possible d'en rationaliser le contenu suivant des critères externes (lieu de publication, genre de textes, thématique, âge des intervenants, etc.). La Toile *dans son ensemble* comporte une part de bruit liée à la répétition mécanique de certaines informations : il est aussi nécessaire d'en penser le traitement. Les sondages réalisés dans ce vaste *terrain* d'observation indiquent toutefois des tendances collectives remarquables. Dans une perspective diatopique très grossière (considérant seulement la localisation des sites), les graphiques réalisés à partir du logiciel Diatopix (annexe I) confirment l'observation *in situ* de différences d'usage lexical. Les requêtes à partir de séquences dont la sémantisation n'est pas fortement dia-différenciée sur le plan géographique – par ex., *mon chien* / *ma chienne* – affichent des proportions relatives très semblables, ce qui confirme la validité approximative de l'outil.

2.3. *Le parti-pris des formes*

2.3.1. *La forme linguistique comme intermédiaire relatif*

« Le concordancier travaille […] sur une chaîne de caractères sans tenir compte de la nature grammaticale des éléments », écrit Debaisieux (2005 : 17). Nous ajouterions « ni de leur nature sémantique ». À l'ère des grands corpus électroniques, la contrainte technique commence en général dès la requête, par la sélection de formes physiques pertinentes qui puissent en faire l'objet. L'interrogation du corpus repose donc sur l'énorme biais de la forme choisie (interprétée comme suite de caractères par l'outil, éventuellement signifiante pour l'analyste). L'affichage des « résultats » qui en découlent ne révèle donc que les affinités propres aux formes qui ont alimenté la requête. Étant donné la nature mécanique de l'outil, il est évident que les combinaisons décrites ne sont jamais directement d'ordre sémantique, mais physique (elles devraient cependant révéler des affinités sémantiques).

Les phénomènes sémantiques – en particulier lorsqu'ils sont saillants – se manifestent aux sujets par l'intermédiaire d'une multiplicité de formes sémiotiques utilisées pour évoquer dans le discours un objet référentiel, quelle que soit la précision avec laquelle cet objet est représenté[14]. En général, un phénomène sémantique n'est appréhendé qu'à travers une multiplicité de formes que les usagers, bon gré mal gré, altèrent ou déclinent diversement. Pour reprendre l'exemple de la séquence *du point A au point B* (annexe I), soulignons que, telle quelle, la différence fréquentielle diatopique ne peut être analysée que comme l'indice d'une *possible* différence sémantique[15]. En

14 Les termes *objet* et *référent* renvoient ici à des réalités mentales (ou notionnelles).

15 Les 6 occurrences présentes dans Frantext sont toutes inscrites dans des textes scientifiques ou techniques français.

fait, puisque tel semble être le cas, il s'agit d'un « symptôme » lexical parmi d'autres, qui pourrait révéler l'existence, entre l'Amérique du Nord et l'Europe, d'un contraste dia-sémantique beaucoup plus général dans la représentation de l'espace et de l'activité – donc du déplacement et du changement –, qui se manifeste de façon remarquable par des formes de rationalisation – sémantisées diversement – de la distance et de la durée (on pourrait qualifier ces formes de *sémantico-culturelles*). Du point de vue sémantique, si elle est isolée d'autres « façons » d'exprimer le rapport à l'espace ou au temps, la combinaison régulière *du point A au point B* relève alors de l'épiphénomène lexical (noyées dans le bain des particularismes, les « expressions » suivent souvent dans l'imaginaire collectif la logique désémantisante de l'anecdote : elles identifient des différences spécifiques plus qu'elles ne révèlent des principes d'usage linguistique généraux). La variabilité dans l'usage des prépositions *à*, *dans*/*en* et *sur* pourrait refléter le même type de phénomène. L'examen de leur distribution privilégiée dans des syntagmes tels que *sur Internet* ou *dans Internet*, *en doctorat* ou *au doctorat*, *en vélo* ou *à vélo*, *sur la rue* ou *dans la rue*, *à l'hiver* ou *en hiver*, *sur Marseille* ou *à Marseille*, *sur la Belgique* ou *en Belgique*, *à ce temps-ci* ou *dans ce temps-là*, *sur l'étagère* ou *dans l'étagère*, etc. pourrait mettre en lumière les éventuelles régularités sémantiques qui en motivent la « variation[16] ». Fondée sur la mise en évidence de systématicités sémantiques effectives, l'observation de « variations » linguistiques intercorrélées permettra de vérifier la validité – ou d'affiner certains aspects – du modèle descriptif en termes de classes d'usagers regroupés selon leur origine sociale ou géographique, leur statut économique, etc.

Puisque l'association entre forme et sens (ou signifiant et signifié) n'est pas orientée *a priori*, l'analyse d'un phénomène sémantique doit toujours *passer* par l'identification des formes sémiotiques correspondantes (variations morpho-syntaxiques incluses)[17]. Celle-ci implique de la part de l'analyste la fréquentation, voire l'immersion dans des corpus, ainsi qu'un retour dans sa mémoire discursive, c'est-à-dire ni plus ni moins qu'une confrontation au réel linguistique – le fameux « principe de réalité [empirique] » (Condamines et Rastier, d'après Freud). Ce « bain empirique » facilite aussi l'identification de complexes

16 La variabilité dans l'usage de ces prépositions est panfrancophone (et historique), avec des préférences régionales pour certains usages. Par exemple, la tendance à la généralisation de *sur* en France, ou une préférence marquée pour *à* au Québec, dont les dictionnaires font peu état ; la dernière production de la lexicographie institutionnelle québécoise, *Usito* (2013), ne consigne pas l'usage caractéristique de *à* au Québec ; l'article SUR du *Petit Robert* (éd. 2013) ne comporte quant à lui qu'une mention, dans une citation, du remplacement fréquent de *à* par *sur* dans un sens locatif.

17 Le champ de manifestation d'un phénomène sémantique (c'est-à-dire le polymorphisme du sens) se révèle en grande partie dans le nombre, la diversité et la « productivité » des formes sémiotiques qui en constituent l'expression.

morphosémantiques qui entrent en relation d'équivalence ou de concurrence avec le phénomène étudié. En termes classiques, ce sont des potentialités de (re)paradigmatisation, sur une base sémantique, de « formes » lexicales (unités et combinaisons) qui n'ont pas toujours été reliées entre elles, ou ne le sont pas pour tous ni partout (pour faciliter la tâche, ces formes ne s'inscrivent pas toujours dans une continuité séquentielle). Les (re)structurations paradigmatiques – nous l'avions montré dans un mémoire (Courbon 2004) – passent par des configurations lexicales prototypiques (par ex., *fortune* + QUANTITÉ dans le cas du sens « matériel » dont nous examinions alors l'intégration diachronique).

D'abord basées sur des choix descriptifs (donc sur l'intuition), les requêtes effectuées dans de grands corpus informatisés présentent l'avantage de donner un accès rapide et facile à des « contextes » et à des « données » chiffrables. Cette démarche comporte cependant les limites de toute observation du réel médiée par un outil. Du point de vue linguistique, l'affirmation de Muller (1984 : VI) : « [la] perte d'information […] commence dès lors que l'on extrait la forme de son contexte » accorde certainement un poids trop grand à l'aspect matériel de la signifiance (qui serait identifiable *a priori*). La perte – ou le décrochage empirique – commence en fait en amont, à partir du moment où une séquence est choisie comme forme pertinente. Un problème majeur, somme toute récent, est celui de l'invisibilité de l'ensemble des « données » de corpus. Trop peu explorée, cette question n'est abordée ici que de façon superficielle, à travers le choix des formes d'« interface[18] ». Une conséquence de l'invisibilité globale des corpus est la nécessaire segmentation du donné en « unités » pertinentes, correspondant à des séquences graphiques de taille variable. Ce séquençage relève surtout de l'interface humain-machine (forme que prennent les requêtes et l'affichage des résultats) ; il peut avoir une incidence sur la représentation de réalités linguistiques. Bien qu'elle ne soit pas le propre de la linguistique outillée, le découpage en segments a prévalu dans le traitement assisté de corpus (calcul de la fréquence d'occurrence de formes prédécoupées, relevé de « segments répétés », « significativité » de la cooccurrence d'unités…). Cette prévalence du traitement segmental dessine une certaine vision des faits de langue. Elle contraint même l'analyse : la mise au jour d'observables effectivement significatifs est un objectif descriptif – voire un résultat d'analyse –, non une évidence de départ. L'un des principaux inconvénients de ce procédé réside dans le masquage relatif de la discontinuité et de la variabilité de l'étendue syntagmatique des phénomènes sémiotiques, si l'on admet le fait que leur grandeur varie et qu'elles ne se présentent pas nécessairement de façon séquentielle. Ces caractéristiques ne peuvent être analysées qu'à la condition qu'une signifiance globale soit perçue, ce qui exige l'identification

18 Mortureux (1981) envisageait les questions de la nature et de l'étendue d'un corpus, mais, pour des raisons compréhensibles, pas celle de ses conditions d'accès.

préalable des phénomènes correspondants (cf. les variantes lexicales dites « développées » d'expressions, et plus généralement toutes les transformations lexicales). La machine n'est pas armée pour ce type d'analyse, qui nécessite une saisie compréhensive. Ce genre de caractéristiques sémiotiques explique, le temps faisant son œuvre, que des « locutions figées », comme *au fur et à mesure*, ou de banales unités, comme *antidote*, ne soient plus que partiellement décomposables. Cela permet aussi de comprendre sur le plan sémantique l'étendue variable d'un champ de « collocations[19] », qui dépend en partie de la productivité des configurations sur lesquelles il repose.

La « variation » de la ou des formes matérielles associées à un phénomène sémantique est déterminée par la signifiance, et non l'inverse. En cela, l'analyse sémantique ne devrait pas recourir à la lemmatisation *a priori*, qui gomme, au profit d'une formalisation abstraite, la singularité des usages, et par conséquent occulte la significativité de certaines formes « grammaticales » ; sans compter que la notion même de contexte, déjà, est décontextualisante, le sens « contextuel » (= discursif) étant par définition transversal, *i. e.* non réductible à une unité lexico-grammaticale localisée en syntaxe, qui n'est qu'un épiphénomène. Ainsi, à un niveau d'analyse abstrait de toutes conditions énonciatives, il est possible d'identifier *se sucrer le bec*, *faire dur* ou (*c'est*) *du bonbon* comme des usages sémantisés au Québec, mais ce sont les discours qui donnent accès aux combinables effectifs de ces abstractions métasémiotiques présentées ici. Seule l'intuition acquise permet de produire les « bonnes » formes, adéquates sur le plan de leurs caractéristiques axiologiques et de leur valeur sémantique.

2.3.2. *De la forme au chiffre : la machine statistique*

Le rapport entre sens et fréquence reste à théoriser. Puisque le sens est d'ordre d'abord qualitatif, se pose la question des modes de quantifiabilité de phénomènes qualitatifs. L'utilisation singulière d'un mot, par exemple, peut marquer toute une communauté d'usagers, qui reprendra celui-ci dans sa nouvelle signification (augmentation fréquentielle). Si la rareté entraîne parfois des usages très variables sur le plan sémantique (l'instabilité du sens se corrélerait avec la faible exposition des sujets aux formes)[20], la fréquence élevée de

19 Nous reprenons ici, par commodité, la terminologie traditionnelle. Elle est cependant critiquable à plusieurs titres : parce qu'elle repose sur une conception séquentialiste des processus sémiosémantiques, parce qu'elle rend statique la représentation de ces derniers, et parce qu'elle donne à voir non des potentiels sémantico-combinatoires, mais des effets de contexte.

20 Nous avons entendu des emplois « comportementaux » dépréciatifs des unités lexicales *huppé* ou *érudit* (sens : « arrogant, condescendant »). Si ces usages ne sont pas répertoriés dans la lexicographie générale (comme c'est souvent le cas des usages « marginaux »), et s'il peut être très fastidieux d'en trouver des attestations dans les corpus traditionnels, doit-on pour autant les exclure du champ

certaines unités (simples ou complexes, référentiellement « arrimées » ou non) ne semble cependant pas garantir leur invariabilité sémantique. Concevoir la « richesse lexicale » en termes exclusivement formels et quantitatifs (en nombre de « mots ») est certainement inadéquat. La variabilité des combinaisons, leur couverture référentielle, ainsi que la diversité de leur sémantisation devraient entrer dans la définition de la « richesse » lexicale individuelle. Les différences remarquables entre univers de référence (section 3) rendraient par ailleurs sa mesure impossible dans l'absolu.

Par ailleurs, l'instauration d'un rapport quantitatif aux « données » (c'est-à-dire à du matériau présélectionné, rassemblé et typé) produit des effets de miroir grossissant sur certains phénomènes sémantiques abstraits de leurs conditions référentielles. L'aspect déréalisant de la (méta)analyse rend parfois difficile l'interprétation de la significativité des mesures, étant donné que les normes comptables sont établies d'abord par rapport à la composition *textuelle* des corpus (volume / types / identité de textes…), donc décentrées des habitudes linguistiques des usagers qui leur ont donné corps et voix. Cela devrait suffire à justifier la prise en compte de la variabilité individuelle lors de la collecte de « données ».

L'amélioration technique des ressources ne devrait pas déboucher sur l'exclusivité numérique. En tant qu'indicateurs, les chiffres sont des outils d'interprétation, non des catégories d'analyse.

2.4. *L'authenticité des faits inventoriés*

2.4.1. *Désubjectiver pour mieux authentifier ?*

Le gain technique ne devrait pas non plus conduire à une survalorisation de l'authenticité linguistique : l'un des premiers risques est de réduire l'image de la réalité linguistique aux seuls faits de discours collectés. Vincent (2009 : 83) parle à cet égard de « dictature des "pro-données-authentiques" ». La formule est un peu forte, mais on ne peut nier une tendance – qui entre en tension avec l'hégémonie des grands modèles – à juger tant l'introspection *per se* que l'écrit sous toutes ses formes comme des artefacts de moindre valeur. En ce qui concerne la dimension sémantique des « paroles » mises en corpus, on pourrait défendre le point de vue opposé : étant donné que tout fait linguistique détaché de ses conditions d'énonciation – et surtout : désubjectivé – perd de son authenticité dès lors qu'il est enregistré, les témoignages écrits, notamment lorsqu'ils ont été produits de façon spontanée, seraient plus authentiques et « fidèles » à l'original que ne l'est l'oral retranscrit[21]. La pratique de corpus s'est érigée contre le recours

---

de recherche sémantique, au prétexte qu'il s'agirait d'« erreurs » individuelles, que démontrerait du reste leur très faible fréquence ? Non.

21 Bien qu'hyper-significatif, « le matériel voco-prosodique et mimo-gestuel » qui accompagne le discours verbal (Kerbrat-Orecchioni et Constantin de

à l'exemple forgé, qui était parfois aussi peu convaincant que le jugement qui l'accompagnait. Étant donné que le *datum* et l'*exemplum* (Laks 2008) remplissent l'un et l'autre une fonction argumentative (Vincent 2009), il convient de déterminer les avantages et les inconvénients du passage d'une rhétorique de l'illustration par l'exemple à une rhétorique de l'illustration par le « donné ». Les lexicographes le savent, les énoncés choisis[22] ne sont pas nécessairement de moindre qualité par le seul fait qu'ils ont été formulés par des sujets-linguistes. À l'inverse, les énoncés dits « authentiques » ne possèdent pas une meilleure qualité intrinsèque du seul fait qu'ils ont été cueillis dans un corpus. D'une part, un énoncé « forgé » peut refléter un usage généralisé (c'est le cas de la plupart des exemples de la lexicographie moderne), tandis qu'un extrait de corpus peut n'avoir qu'une valeur incidente (c'était le cas de nombreux exemples-citations extraits de corpus littéraires dans la lexicographie traditionnelle). Plutôt que dans un donné brut, c'est davantage dans le rapport raisonné aux faits relevés dans le discours ou inscrits dans la mémoire des sujets que se situe la question de l'authenticité et de la portée générale d'une description (l'authenticité du fait *X* à un type d'usage Y ne permet *a priori* que de généraliser *X* relativement à l'habitude de Y).

L'analyse sémantique passe par l'intuition du sens. Les formes linguistiques regroupées en corpus peuvent servir à vérifier, confirmer ou infirmer des intuitions, ou à les préciser, voire à indiquer des possibilités qui n'avaient pas été envisagées[23]. Mais ces formes même n'existeraient pas ou seraient dépourvues de sens sans l'intuition de sujets, qui les *fait* telles. L'apparente immédiateté des composantes de corpus n'est que le résultat de processus sémantiques intersubjectifs (l'« immédiat » primitif), et c'est *entre autres* sur ces traces que les sémanticiens fondent leurs analyses. C'est pourquoi il importe de ne pas accorder un statut transcendant aux ressources de corpus. Leur exploitation reste instrumentale, et ne devrait pas détourner le regard de ce qui importe en premier lieu, à savoir l'identification, la description, voire les tentatives d'explication de faits linguistiques.

Chanay 2007 : 311) comporte une telle complexité que les analyses interactionnelles fines ne peuvent porter que sur de brefs passages. Même lorsqu'elle inclut des éléments to-textuels (Cosnier), la transcription de l'oral en quantité est nécessairement très en deçà de la réalité des productions, indiquant au mieux les rires, les gestes les plus saillants et les mimiques les plus signifiantes.

22 Qu'on les nomme « données idéalisées » (Legallois et Kwon 2006 : 145) ou « données auto-générées » (Vincent 2009 : 83) ; Gadet (1989 : 137) parlait aussi à ce sujet de « corpus théorique ».

23 En cela, les raretés sémantico-discursives peuvent éclairer le potentiel sémantique d'une forme. Nous rejoignons en partie Legallois et Kwon (2006 : 147) au sujet du caractère archétypal des emplois dits (= perçus comme) « figurés ».

2.4.2. *Magie du don et oubli du donneur : la désaffection du sens*

> [L]e corps devient un signifiant sans voix[24]

Dans la vision classique, le corpus est détaché de ses conditions d'énonciation et *surtout* des compétences de production et d'interprétation des sujets qui, malgré leur absence, en sont les principaux responsables. Il est par définition normatif et tend à désauthentifier l'image du réel qu'il présente. À la juste rigueur, le corpus le plus authentique serait composé, à l'image de certaines enquêtes, de l'ensemble des productions concrètes d'un individu *et* des interprétations qu'il en donne (corpus à échelle humaine)[25]. Les (très) grands corpus relèvent à cet égard de la réalité augmentée : la masse de « données » qu'ils comportent peut donner l'impression qu'ils offrent une image de la langue communément partagée, là où ne figurent en fait les productions que d'une très petite minorité de membres censés représenter « la » communauté linguistique. Ainsi, Frantext, qui fut longtemps le plus grand corpus de français, n'offre à entendre pour les derniers siècles la voix que de quelques centaines de francophones[26]. D'autres corpus, limités pour des raisons pratiques, ne présentent les productions, elles aussi limitées, que de quelques dizaines de donneurs. Si des généralisations sont possibles pour des phénomènes linguistiques récurrents sur les plans phonétique, morphosyntaxique ou discursif, celles-ci sont beaucoup plus difficiles sur le plan sémantique pour au moins deux raisons : le volume minime de prises et la faible diversité interindividuelle enregistrée. Le biais du donné en corpus ne consiste pas seulement dans l'impression d'immédiateté qu'il suscite, ni dans l'image de totalité qui dispose à la généralisation, mais dans le problème éthique qu'il soulève. Il est certes irréaliste de redonner voix aux centaines de millions de francophones qui se sont exprimés au cours des derniers siècles, mais peut-on objectivement faire abstraction du fait que le matériau rassemblé manifeste la domination verbale d'une minorité – en général lettrée, voire beaucoup plus lettrée que la moyenne de la population –, tandis que la majorité silencieuse… reste muette ? La plupart des différences interindividuelles sur le plan des pratiques sémantiques se trouvent *de facto* occultées.

Il importe, afin de restituer au fait sémantique sa part subjective, de réfléchir et de travailler à l'élaboration de ressources-miroirs du réel

---

24 C. Brousseau (1989) : « "De la tristesse"…», *Bulletin de la Société des amis de Montaigne*, 17-18, p. 40.

25 C'est dans cette perspective que Sauvageot (1957 : 9-11) présentait ses analyses linguistiques comme le résultat de son point de vue de sujet parlant singulier.

26 Produit hybride issu de la rencontre entre philologie et lexicométrie, le corpus Frantext est présenté comme une base de *textes* comportant un volume précis de *mots* pour une période donnée. Les unités de mesure, de nature typiquement texto-grammaticale, ne tiennent pas compte des sources premières, c'est-à-dire des auteurs : ils sont certes identifiés comme auteurs des *textes*, mais leur présence n'est pas quantifiée, même après la sélection d'un ensemble de *textes*.

linguistique *tel qu'il se présente et se trouve perçu par « les » usagers* dans la diversité de leurs pratiques de sémantisation, c'est-à-dire de compréhension, d'appropriation et de production de formes signifiantes. Le déplacement du point de vue oblige à définir les concepts méthodologiques d'authenticité, de validité et de représentativité des faits identifiés relativement aux phénomènes présents dans l'univers de référence des sujets, phénomènes qui dépendent des besoins épistémiques et praxiques des sujets en question.

En définitive, les principaux éléments (interdépendants) qui tendent à limiter l'exploitabilité d'un corpus sur le plan sémantique sont les suivants : son détachement des sources énonciatives, sa clôture, sa (trop) grande sélectivité et – surtout – sa matérialité. Or, c'est à la fois en amont et en aval des formes matérielles que le sens linguistique prend forme dans l'esprit des sujets, en dehors des limites propres aux textes et aux discours.

3. En deçà des corpus : référentiels et univers de référence

> [La fréquence] me semble déterminée […] par la probabilité d'actualisation du noème et par conséquent du signifié. Et cette probabilité est liée à des situations qui varient non en fonction de l'idiome, mais en fonction de la civilisation et du vécu dont chaque idiome est l'expression[27].

3.1. *Référentiel et univers de référence*

Afin d'appréhender les pratiques de sémantisation en incluant les dimensions attitudinale et perceptuelle, il semble nécessaire de distinguer entre, d'une part, l'univers de référence des sujets, et, d'autre part, les référentiels auxquels il leur est possible de se reporter. L'univers de référence des sujets correspond à l'ensemble des savoirs, gestes, affects et autres expériences – incluant celle de la langue – auxquels chacun est exposé, habitué et dont il est susceptible de parler, qu'il le fasse ou non. L'univers de référence évolue au gré des expériences. Il a partie liée avec l'activité discursive et interdiscursive du sujet ou des énonciateurs avec lesquels il entre en contact, de façon directe ou indirecte. Il peut être très personnel par certains aspects (forte singularité), ou au contraire très semblable à celui d'autres sujets (faible singularité). Bien qu'en partie intersubjectifs, puisqu'ils se recoupent ou se rejoignent, les univers de référence sont d'abord le résultat de parcours individuels. Les référentiels, en revanche, correspondent à des ensembles thématiques de référence, stabilisés dans des communautés plus ou moins larges (dimension collective). Ce ne sont à proprement parler ni des domaines, ni des genres, ni des secteurs définis, mais des espaces de discours

27 C. Muller (1971) : « Fréquence des signifiés ou fréquence des signifiants », *ELA*, 2, p. 87.

potentiels, regroupés autour de réalités identifiées (par ex., les remerciements se rattachent au référentiel des formules de politesse). Les référentiels jouent un rôle non négligeable dans la constitution de classes sémantiques, puisqu'ils conditionnent le regroupement de phénomènes linguistiques, c'est-à-dire de façons de parler de réalités saillantes (ainsi, la terminologie d'un domaine est d'autant plus étendue que les réalités auxquelles il est fait référence sont nombreuses). Les univers de référence se définissent par rapport aux parcours toujours singuliers des individus, tandis que les référentiels réfléchissent l'objectivité relative d'expériences partagées dans une communauté. Ces expériences peuvent être affectives, pratiques, cognitives, physiques…

Sauf à être tout à fait novatrices, les pratiques référentielles sont rarement isolées. Par conséquent, les univers de référence tendent à « s'alimenter » à la source de référentiels préexistants (démarche mimétique), ou bien à converger vers ceux-ci (démarche exploratoire). Le novice dans un domaine (la photographie, par ex.) peut référer à des réalités (par ex., les types de cadrage) sans connaître les pratiques terminologiques du domaine (lesquelles correspondent au référentiel du cadrage photographique) ; il peut ainsi retrouver par lui-même des distinctions établies dans le domaine en question (son univers de référence en matière de photographie s'approchera du référentiel de la photographie) ; il peut aussi apprendre le vocabulaire usuel et avoir une expérience pratique du domaine en question, auquel cas la zone de son univers de référence relative à ce domaine se conformera à peu près – selon son appropriation dudit domaine – au référentiel source. Peu de temps après s'être renseigné sur le sujet, notre amateur photographe aura déjà acquis un certain nombre de « compétences » sémantiques ; son univers de référence s'en trouve altéré d'autant. L'analyse de telles compétences devrait donc prendre en considération l'univers de référence sous forme de micro-analyses thématiques, par exemple. À cet égard, la comparaison de plusieurs lieux d'énonciation, sous le rapport très superficiel de la fréquence, pour quelques formes lexicales référentiellement typées autour de la maladie mentale, est révélatrice (tableau 1).

	corpus numérisés				forums électroniques	
	Frantext	**CFPQ**	**CLaPI**	***Devoir***	**Nutrition**	**Psychologie**
millions d'occ.	**8**	**0,5**	**≈ 0,8**	**≈ 144**	**≈ 409**	**≈ 808**
schizo(s)	2	0	0	15	73	16 687
schizophrène(s)	16	2	0	146	53	13 117
schizophrénie(s)	10	0	0	231	94	16 758

Tableau 1 : Fréquence d'occurrence des formes *schizo*(*s*), *schizophrène*(*s*) et *schizophrénie*(*s*) dans plusieurs types de corpus[28]

28 À l'exception de CLaPI, dont la partie la plus ancienne remonte aux années 1980, tous les corpus comportent des énoncés produits dans la période 2000-2013 (entre 2004 et 2012 pour *Le Devoir* et Frantext – quelques textes rassemblés

Deux observations méthodologiques seront faites. La première concerne l'importance du volume textuel : seuls les corpus de plus de cent millions de vocables donnent accès à un volume « satisfaisant » d'observables pour des formes de fréquence moyenne[29]. D'autre part, on remarque l'aspect déterminant de la thématique référentielle, qui prédomine sur le « genre » (on peut ainsi déduire du très faible nombre d'attestations des formes cherchées dans les corpus d'oral que le thème de la schizophrénie n'a que peu été abordé).

Dans le champ des possibles, tout n'est pas dit, tout non plus « ne se dit pas » ; mais, selon le cadre énonciatif, certaines « choses » se disent plus que d'autres – c'est de cette réalité primordiale, objet de discours, que se constitue le référentiel en tant qu'ensemble d'habitudes thématiques. Dans cette perspective, la référence n'est ni matérielle ni temporelle, autrement dit elle ne se situe ni dans le monde extra-mental, ni dans un avant ou dans un après de la sémiose[30]. Une part importante de la variabilité sémantique résulte du jeu complexe entre référentiels établis et extension des univers de référence des sujets parlant du monde. Aussi, les ressources empiriques d'une microsémantique variabiliste devraient représenter des aspects particuliers de l'univers de référence de sujets réels (champs d'intérêt, habitudes, comportements…).

Une distinction doit être faite ici entre les corpus, d'une part, qui recueillent des traces énonciatives (quelles qu'en soient les modalités de traitement), et, d'autre part, les univers de référence, qui sont « avivés » par les discours environnants, mais restent immatériels. Le recours à des ressources discursives, aussi nombreuses et diverses puissent-elles être, n'a jamais été aussi indispensable. Néanmoins, il est nécessaire de réfléchir aux façons d'accéder à l'en-deçà et à l'au-delà des « données » de corpus pour mieux les

---

dans ce dernier corpus ont cependant été écrits plus tôt, mais l'interface retient l'année de publication comme seul critère de sélection). Je remercie ici Samuel Dion-Girardeau et Hugo Mailhot pour l'aide substantielle qu'ils ont apportée dans la collecte de données numériques (projet de recherche DiffLex financé par le FRQ-SC, 2013-2016).

29 Il ne s'agit pas de célébrer le culte du chiffre, mais l'on est obligé de reconnaître qu'en deçà d'un certain seuil, assez élevé – qui dépend de la nature des composantes (en particulier du nombre d'énonciateurs) –, très peu de faits sémantiques pertinents sont présents en corpus. Il y a en cela une différence nette entre les faits linguistiques ancrés dans l'expression et ceux qui relèvent d'abord du « contenu » : ainsi, les objets phonétiques ou morphosyntaxiques sont limités à quelques dizaines d'objets (ce qui explique leur forte récurrence en corpus) ; en revanche, les objets sémantiques dépassent très largement l'ordre de la dizaine de milliers d'entités, ce qui explique leur moindre récurrence – rappelons néanmoins que la qualité primant sur la quantité, le *nombre* d'associations ou d'usages mémorisés n'est qu'un argument méthodologique en faveur d'un corpus de grande taille (les objets de la sémantique étant typiquement non segmentaux, la quantification s'avère plus délicate – du moins la significativité de son application).

30 *Atemporelle* ne signifie pas que la référence soit anhistorique.

recontextualiser – c'est-à-dire les (ré)inscrire dans des pratiques –, et redonner ainsi un peu de leur portée phénoménale aux discours prisonniers des corpus.

Si l'on compte avec les paramètres référentiels, les corpus d'exemples forgés et les corpus d'énoncés « authentiques » ne présentent guère de différence : ils proviennent dans un cas comme dans l'autre de l'intuition linguistique d'un sujet (référée à un usage virtuel dans le premier cas, à un usage effectif dans le second). Ce n'est pas tant la nature des productions que la mauvaise qualité des jugements afférents qui pose problème : jugements d'inacceptabilité débrayés de toute situation concrète dans un cas (décret de non-existence que les attestations de corpus révisent en déni de réalité linguistique) ; dans l'autre, généralisations hâtives à partir d'énoncés singuliers, avec peu de considération pour leurs conditions référentielles.

3.2. *Variabilité du sens et univers de référence : des corpus-passerelles ?*

3.2.1. *Observer les pratiques sémantiques : les corpus d'écrit interactionnel*

Comme le rappellent Kilgariff et Greffenstette (2008 : 99), un corpus n'est jamais représentatif que de ce qu'on y trouve. On peut en induire des normes d'usage, à condition de définir l'échelle de généralisation (une classe d'usagers, son poids démographique…) et les critères de généralisabilité (la fréquence, le gradient de similarité…). En dernière analyse, seuls les premiers concernés par les énoncés, c'est-à-dire ceux qui les ont formés, seraient en droit de statuer, en situation, sur la validité des jugements qui s'y rapportent. Même s'il va de soi que la carte ne sera jamais le territoire (tant qu'il y aura des sujets, celui-ci continuera d'évoluer), on peut s'entendre sur l'échelle de description (donc sur la portée de celle-ci) : cartographie à grande échelle, présentant une langue-monde (ou langue historique [Coseriu]) ; cartographie à échelle moyenne, présentant des langues-classes ou langues-communautés (langues fonctionnelles, avec leurs traditions respectives [*id*.]) ; cartographie à petite échelle, présentant des pratiques convergentes ou divergentes. Tout corpus offre une vision nécessairement tronquée du réel à petite échelle. Par conséquent, les changements d'échelle posent problème.

Le fait qu'ils soient externes à la pratique propre à l'analyste confère aux discours assemblés en corpus une certaine valeur d'« objectivité » (par leur volume, notamment). Mais le fait qu'ils soient bien souvent décrochés de la réalité subjective de leurs auteurs (et notamment, des affects, des connaissances préalables et des intérêts de ces derniers) ne les rend pas moins artificiels que des énoncés forgés (on parlera d'expérimentation *in vitro*). Il y a des cas limites : par exemple, les passages dialogués dans de l'oral conversationnel ou bien les interactions électroniques spontanées. Ce genre de témoignages rend possible une observation suivie, quasi *in vivo*. Ainsi, en changeant de niveau de granularité, peut-on envisager de décrire des parcours sémantiques individuels et interindividuels. Les productions enregistrées dans/sur des forums

électroniques, notamment, présentent les avantages suivants : les interventions se définissent par leur caractère interactionnel et par une grande spontanéité ; une collecte volumineuse peut être effectuée (certains forums dépassent par exemple le milliard de vocables) ; chaque commentaire est identifié, au minimum par l'indication de son auteur, de la date et de l'heure d'affichage, mais aussi, dans certains cas, au moyen d'informations plus précises sur les auteurs (âge, sexe, origine géographique, nombre de commentaires publiés, etc.). La partie personnelle de ces informations étant pseudonymisée ou facultative, on peut estimer qu'elles sont fiables la plupart du temps.

Un autre avantage d'examiner le contenu de commentaires enregistrés dans/sur des forums de discussion consiste dans leur profondeur historique[31] : la plupart des forums de discussion ont plus de dix ans d'âge, et il s'y trouve des fidèles qui, depuis l'ouverture des sites, n'ont cessé d'y intervenir régulièrement. Cet aspect n'est pas négligeable, dans la mesure où la dimension sémantique est certainement, de toutes les composantes de la langue, celle qui présente le plus de « mutabilité » intra-individuelle. C'est pourquoi, pour des raisons d'abord méthodologiques, nous situons la synchronie au plan de l'acte d'énonciation (« idiosynchronie » minimale [Saussure]) ; la diachronie commence donc potentiellement avec l'échange, ou du moins avec l'exposition à d'autres discours (influence interindividuelle)[32]. De ce point de vue, il peut suffire d'une seule exposition à une forme lexicale dans un sens déterminé (par ex., *historique* au sens de « mémorable ») pour que le nouvel usage en puissance soit associé à une valeur sociodiscursive (en l'occurrence, superlative) et intégré à un micro-système de différences et de référence.

Les forums électroniques constituent donc un *terrain* d'observation diachronique privilégié du sens linguistique en situation. Ce matériau empirique permet de suivre des profils individuels renseignés, ce que beaucoup de grands corpus ne permettent pas (Gadet 2003 : 13). Il offre aussi la possibilité de dégager des styles sémantiques interindividuels[33],

---

31 Kehoe (2006) indiquait ainsi les possibilités que les productions inscrites sur la Toile présentent pour l'étude diachronique de la langue.

32 Nous rejoignons ici Haugen (1972 : 303) : « Every performance alters one's competence by the increment of what one has learned or unlearned during the performance ». La question est ensuite de trouver un moyen d'observer les effets durables des dynamiques verbales sur les « compétences » partagées.

33 Homme de son temps par sa conception classiste de la société, Coseriu (1998b : 28-29) proposait de situer à un haut niveau de généralité la dimension qu'il avait d'abord nommée « sym-/diaphatique » : « Les types très généraux de styles apparentés correspondent à des aspects généraux de la vie et de la culture et à des types similaires de circonstances ». Cela revenait à poser l'existence d'un référentiel partagé de tous, indépendamment des horizons culturels ou sociaux ; nous situons ici la notion de style sémantique à un niveau de

mais également d'observer la résonance ou le « degré » de compréhensibilité (d'intercompréhensibilité) des formes utilisées, et d'appréhender ainsi des occurrences de variabilité sémantique *vive*. Il est alors possible d'identifier, sur le plan intra- et interindividuel, les différences par rapport à des pratiques routinisées, en cherchant à savoir si ces différences sont perçues, et comment elles sont reçues, comprises et reprises par différents coénonciateurs[34].

3.2. *L'hypothèse du milieu langagier partagé : du référentiel à la sémantisation*

3.2.2.1. Sous la surface des formes : le miroir de la fréquence

Les habitudes linguistiques – que Saussure mettait à raison au cœur de sa définition de la langue – ne sont ni celles de la langue, ni celles d'une « variété », mais d'abord celles des sujets socialisés. Or, ceux-ci appartiennent à une pluralité de groupes de taille variable, que les catégories sociales intuitives importées dans l'analyse ne permettent pas de saisir dans leur complexité. Les témoignages contrastés qu'offre la Toile (annexe II), ainsi que les fréquences présentées dans le tableau 2 l'illustrent.

	Devoir	***Monde***	**forums**
millions d'occ.	**≈ 144**	**≈ 141**	**≈ 141**
bon café	15	0	286
petit café	121	23	318
critique	1 120	1 085	2 720
plaisant	272	216	355
très sécuritaire	32	5	9
dépanneur	388	10	58
orignal	201	3	5
tomber amoureux	44	68	146
tomber en amour	21	3	1
argent américain	18	8	1
dollar(s)	26 246	17 634	529
dollar américain	658	54	3
dollars américains	672	77	4
dollar canadien	2 039	20	5
dollars canadiens	323	164	2
euro(s)	5 762	61 777	10 400
zone euro	1 721	4 955	0

Tableau 2 : Fréquence d'occurrence de diverses formes lexicales dans les quotidiens *Le Devoir* et *Le Monde* et dans un ensemble de forums du site Doctissimo[35]

généralité moindre, puisqu'elle se rapporte à la systématisation de convergences ou de divergences intersubjectives, au croisement des univers de référence avec des référentiels qui peuvent éventuellement jouer le rôle de repères sémantiques.

34 Le caractère interactionnel d'un discours est en général le lieu d'une plus grande spontanéité énonciative, ce qui permet, en corpus, le repérage de zones plus néophiles (Courbon 2007).

35 Les corpus ont été équilibrés à environ 140 millions de vocables. Pour le quotidien *Le Devoir*, cela correspond à la période de 2004 à 2012 et pour *Le Monde* à

Cet aspect de la démarche est très superficiel, puisque seuls des segments sémiotiques prédécoupés et présélectionnés sont pris en compte, et aucunement les significations[36], ni les valeurs discursives, ni les réactions métasémiotiques, ni l'environnement propice à l'apparition des phénomènes correspondants, en d'autres termes aucun des principaux aspects que devrait considérer une sémantique empirique. Ces aspects exigent de se placer à un niveau de granularité plus fin que celui auquel sont situés les segments lexico-grammaticaux. La plupart de ces différences fréquentielles, néanmoins, sont référentiellement significatives. Nous ne ferons ici que quelques observations (l'objectif, rappelons-le, n'étant pas de fournir des analyses dia-sémantiques, mais de s'interroger sur leurs conditions de possibilité empiriques). Tout d'abord, ce qui touche les sujets (affectivement) est (très) présent dans les forums (*bon café*, *petit café*, *tomber amoureux*, *critique*, *euro*(*s*) – dans une certaine mesure *dollar*(*s*) –, et, malgré le typage dia-sémantique, *plaisant* [relativement formel dans l'usage hexagonal]) ; en revanche, ce qui reste extérieur à l'univers de référence des sujets, dans ce cadre énonciatif particulier (vie quotidienne), est (très) peu présent dans les forums (*dollar*(*s*) *américain*(*s*) / *dollar*(*s*) *canadien*(*s*), *zone euro*). Les forums sont publiés sur Doctissimo, site français majoritairement fréquenté par des francophones européens, ce qui explique la moindre proportion de *très sécuritaire* (« qui ne comporte aucun danger »), *tomber en amour* (vs *tomber amoureux*), *dépanneur* (qui – pour des raisons alimentaires – se trouve ancré dans le référentiel des francophones nord-américains, ce qui n'est pas le cas en Europe – seules 6 des 58 occurrences du mot présentent le sens d'« épicerie »), *orignal* (nom d'un animal qui n'a rien d'anecdotique, puisqu'il peut causer la mort par accident de la route)[37], *dollar*(*s*) et *argent américain* (expression informelle pour parler de *l'autre* dollar), et, dans une certaine mesure, *plaisant*, peut-être moins courant(s) dans les usages européens que dans les usages nord-américains. La mise en regard des deux quotidiens stylistiquement équivalents *Le Devoir* (Montréal) et *Le Monde* (Paris) confirme l'ancrage référentiel des signes, même segmentés et extraits de leur « milieu » énonciatif (la disproportion, dans *Le Devoir*, de *dollar canadien* au singulier s'explique par la présence d'un référentiel spécifique

---

celle qui s'étend de 2008 à 2012. Les forums suivants du site Doctissimo ont été examinés (2002-2013) : Environnement, Forme et Sport, Loisirs, Médicaments.

36 Une signification très étroite *transparaît* toutefois à travers l'existence même des « signes » retenus. Il s'agit d'un effet de la mémoire sémiotique, ou, plus largement, de l'impression qu'une forme constitue un signe, donc qu'elle possède une signification (usage présupposé).

37 On relève en fait 26 occurrences de la forme *orignal*, mais 21 d'entre elles – soit la grande majorité – résultent de la transformation graphique (coquille) de la forme *original*.

[chronique boursière])[38]. Les styles sémantiques des individus sont rarement uniques. Ils s'inscrivent en général dans des communautés d'appartenance, qui débordent le cadre des « origines », sociale ou géographique, et peuvent être motivés par des intérêts cognitifs, pratiques, idéologiques, etc., ou tout simplement par l'expérience de la confrontation à la diversité linguistique.

3.2.2.2. Univers de référence et parcours de sémantisation

On peut faire l'hypothèse, presque triviale, que les dimensions les plus « présentes » de l'univers de référence des sujets orientent ceux-ci dans leurs pratiques sémantiques. Par conséquent, plus les univers de référence convergent vers un référentiel commun (qui les modèle), plus la sémantisation aura tendance à être partagée (cf. le « vocabulaire de base ») ; à l'inverse, lorsque les sujets ne partagent pas l'une des dimensions de leur univers de référence avec d'autres sujets et qu'ils n'ont pas accès à un référentiel-repère, il se peut que leur façon de parler des réalités correspondantes diverge des pratiques des autres.

L'apprentissage lexical, notamment à l'âge adulte, illustre la souplesse des parcours référentiels, qui sont la condition première de la sémantisation (cf. le concept d'« expertence » que propose Nyckees 2007). La rencontre et l'interpénétration des univers de référence prennent la forme, au niveau discursif, de demandes de clarification, d'adaptation par anticipation, de plaisanteries, etc., qui sont autant de « micro-dynamiques de diffusion » (Siouffi *et al*. 2012 : 221). Nous en présentons quelques cas de figure dans l'annexe II. Les forums de discussion offrent de nombreux exemples de ces cas typiques du « discours en interaction » (Kerbrat-Orecchioni). Interrogations épisémantiques, micro-évènements didactiques, voire « négociations » sur le matériau verbal (Kerbrat-Orecchioni 2000) ont une incidence à plus ou moins long terme sur la compétence linguistique des sujets. La recherche lexicale est parfois collective[39], mais on doit considérer également que les tiers muets apprennent autant que les plus volubiles : une étude longitudinale (diachronico-développementale) de l'affectation de sens à des formes sémiotiques jusque-là non sémantisées[40] devrait tenir compte, au niveau microlinguistique, des parcours de lecture et, plus généralement,

38 Concernant, dans la presse, les effets sémantiques de thématiques, voir Paquet-Gauthier (à par.).

39 Blanche-Benveniste (2005) donne quelques illustrations de ces formes de « collaboration » à la recherche lexicale.

40 Nous avons eu l'occasion de parler à cet effet d'« impression sémiotique » (Courbon 2012a). Il arrive de prendre conscience de ce phénomène lorsqu'une forme, qui était jusqu'alors inconnue ou limitée à un domaine d'usage spécifique, nous parvient dans une diversité de « contextes » nouveaux. La forme, en prenant sens dans un référentiel spécifique, fait signe. C'est pourquoi elle peut être thématisée, qu'elle qu'en soit la taille (il peut par exemple s'agir d'un énoncé en circulation).

de l'exposition *concrète* des sujets à la diversité des usages environnants (… tels qu'ils les comprennent)[41].

Quoique relatifs aux parcours de sujets-individus, les univers de référence sont nécessairement partagés. La confluence des univers de référence crée des zones de forte convergence sémantique interindividuelle (c'est du moins une façon de se représenter l'articulation entre langue et discours du point de vue sémantique). En outre, aussi paradoxal que cela puisse paraître, les expériences sémiotisées constitutives de l'univers de référence d'un sujet sont *a priori* dégagées d'une quelconque identité lectale. En d'autres termes, l'acte de référence prime sur l'identité sociocatégorielle et sur les contraintes situationnelles ou générationnelles : le conditionnement lectal existe, mais il est second et peut être rompu. On observe ainsi au quotidien une variabilité extra-{topique, stratique, phasique, technique, etc.}, qui peut conserver sa valeur périphérique, devenir doxale (Rastier), ou encore inviter à reposer la question de l'identité des unités « dia- ». Car, comme l'a bien montré Coseriu (1998a), c'est non seulement la délimitation des unités, mais également l'identité spécifique des faits qui posent problème. Le point de vue adopté accordant une place à la subjectivité des individus, le premier problème reste théorique, dans la mesure où le plus important pour l'usager n'est pas tant de délimiter (hors contexte didactique) que d'identifier des formes signifiantes (des contours) perçues et associées à des phénomènes sémantiques. Contrainte par le choix des « unités » (supposant leur disponibilité et leur « orthonymie » [Pottier]), la variabilité sémantique est ainsi orientée par la référence. La tournure que prend le second problème découle de l'articulation entre individuel et collectif. Se pose en effet la question du caractère plus ou moins généralisé des phénomènes produits par les sujets : ceux-ci reproduisent-ils des figures référentielles usuelles en répétant des combinaisons régulières ou bien innovent-ils, et dans ce cas, quels sont les motifs et les effets de leurs innovations (par exemple, s'agit-il pour eux d'un positionnement par rapport au référentiel ? la [nouvelle] façon de présenter une réalité se diffuse-t-elle ?).

Insister sur la singularité des parcours de sémantisation ne signifie pas rejeter la valeur sociale de la signification, au contraire. Les études sémantiques empiriques peuvent difficilement faire l'économie d'un questionnement sur la dimension individuelle, et, en particulier, sur l'articulation entre, d'une part, les convergences et les divergences interindividuelles et, d'autre part, la reconnaissance épilinguistique de normes massives.

---

41 L'acquisition sémantique se poursuit tout au long de la vie des sujets, ce que montrent notamment Dommès et Le Rouzo (2007), qui constatent, à niveau d'instruction équivalent, un accroissement qualitatif et quantitatif des « connaissances lexicales ». Selon le point de vue présenté ici, le développement et la fixation des ressources sémiosémantiques par les sujets résultent en particulier de l'expansion de leur univers de référence.

4. Éléments de conclusion

L'un des principaux problèmes de l'analyse linguistique tient à l'identification des faits pertinents, à leur constitution même en objets d'étude. Ce problème se pose de façon aiguë en sémantique. D'une part, parce que les faits analysés sont d'ordre immatériel (ou conceptuel), d'autre part, parce que les objets sémantiques classiques sont des produits dérivés de la tradition lexico-grammaticale, identifiés à des segments définis (signification d'unités lexicales, d'« expressions », etc.). Cette tradition repose sur la vision d'une réalité linguistique composée d'éléments discrets (selon un modèle typiquement lexicographique)[42] et de relations (selon le modèle syntaxique de la phrase canonique). Afin d'accéder au réel sémantique et d'en proposer une description, l'accès aux productions discursives nécessite de s'émanciper de ce cadre, en reposant, quitte à en définir le sens, la question de l'identité des phénomènes signifiants. Cela ne doit pas conduire à éliminer les unités signifiantes que l'approche compositionnaliste traditionnelle avait installées au premier rang, mais à en repenser la place et la nature relativement aux ensembles, continus ou discontinus, qu'elles intègrent ou dont elles procèdent.

Reconnaître la linguiversité au niveau phénoménologique implique de ne plus identifier ses formes de manifestation exclusivement à des ensembles homogènes (que l'on puisse caractériser au moyen d'identités « dia- »), mais d'y voir des phénomènes présents et sémantisés à des degrés divers à travers différents espaces d'expression. Il faut aussi accepter que ces phénomènes n'en soient pas pour tous et que, lorsque les usagers les perçoivent, ils soient diversement signifiants selon l'expérience que ceux-ci ont de la langue, l'attention qu'ils y portent ou leur sensibilité aux formes d'expression langagière. L'objectif n'est pas de chercher à décrire des « langues individuelles » en soi (ce serait faire se contredire les termes), mais, à partir des singularités constitutives de la diversité des pratiques (notamment des pratiques sémantiques), d'analyser les zones de convergence ou de divergence interindividuelles à plus grande échelle. Il est possible que les convergences sémantiques, qui dépendent en grande partie d'expériences collectives (communauté de pensée, centres d'intérêt partagés, habitudes communes, etc.), traversent les traditionnels groupes sociolinguistiques. C'est pourquoi, avant que certaines d'entre elles ne soient mémorisées comme signes (conception segmentale), les formes discursives sont d'abord des phénomènes sémiotiques diversement appropriés, qui donnent lieu à des reconfigurations et à des transformations (conception appréhensive et variabiliste). Passer d'un

42 C'est cette image traditionnelle de description du lexique que reproduisait Chomsky (1975 : 11) dans sa tentative de modéliser la sémantique ; c'est aussi ce même modèle que Véronis (2010) proposait de remplacer par une grammaire d'usages. Dans les faits, les sens sélectionnés par les lexicographes correspondent en règle générale à des usages sémantiques.

type de représentation de la signification linguistique relativement stable – comme composante d'un système de systèmes dans lequel se trouvent identifiés des occurrences à des types préexistants, des unités à des classes d'unités préétablies, des emplois ponctuels à des usages conventionnels – à un type de représentation des dynamiques sémantiques qui opèrent au sein d'ensembles de phénomènes diversement saisis, plus ou moins systématisés, mais (re)systématisables selon les besoins et les aspirations des sujets, équivaut à ajouter à la complexité des faits (structurelle/normative) un autre niveau de complexité (la dimension intersubjective des phénomènes).

Cet horizon conceptuel implique de revoir le rapport au « donné » linguistique. D'absolus dans le domaine du droit ou de la littérature, les corpus sont devenus relatifs en linguistique descriptive, où ils ne peuvent plus être exhaustifs que par rapport à des productions très particulières (textes inscrits dans un domaine bien défini, discours d'un personnage politique, etc.). Dans une approche désubjectivée de la langue et des discours, le corpus est, déjà, un pis-aller. La récurrence de certains faits relevés en corpus pose la question de leur niveau de généralisabilité : a-t-on affaire à des licences stylistiques ou bien à une tendance massive, largement sémiotisée ? Lorsque l'analyse est située au niveau de la langue, la généralisation tend à s'identifier à une « représentation » abstraite, qui néglige la dimension intersubjective à l'origine des faits relevés… ou absents du corpus. Dans une conception de la langue comme ensemble de phénomènes sémiotiques diversement systématisés, le corpus forme un point d'entrée dont les observations ne peuvent avoir qu'une valeur indicative. La nature des productions qui s'y trouvent rassemblées (notamment leur volume, leur « âge » relatif, l'identification de leur auteur ou de leur point d'insertion dans une éventuelle séquence d'interaction) est déterminante. L'analyse de corpus peut guider la compréhension des dynamiques verbales collectives. On doit reconnaître l'efficacité méthodologique de la conception techniciste du corpus. Toutefois, les « données » préstructurées, échantillonnées, délimitées et « enrichies » doivent être « exploitées » avec circonspection, en gardant à l'esprit la distance qui les sépare du réel linguistique des sujets de langue. Le matériau empirique extrait dudit réel devrait être associé à d'autres types de ressources linguistiques, qui incluent des renseignements sur les parcours linguistiques individuels, qu'ils soient directs (productions langagières antérieures) ou indirects. Ces renseignements ne peuvent être définis de façon exclusivement externe – c'est-à-dire qu'ils ne sauraient se limiter aux traditionnels paramètres sociolinguistiques, qui, pour respecter le principe covariationniste, sont pour l'essentiel d'ordre sociodémographique –; ils devraient comprendre aussi des facteurs pertinents à la saisie et à la constitution du sens linguistique par les sujets. Ces facteurs sont relatifs aux objets étudiés et à la finesse du grain de l'étude.

Les ressources empiriques envisagées ici, qui prolongent et dépassent les corpus exploités jusqu'à présent, correspondent à des ensembles ouverts,

beaucoup plus labiles, évolutifs, clairement associés à des foyers énonciatifs (sujets et conditions d'énonciation). Les faits constitutifs de ces ensembles sont par conséquent sémantisés à travers des pratiques et des rapports au langage identifiables. La significativité statistique de corpus convient mal à une sémantique qui accorde une place importante à la responsabilité des sujets. Car la significativité des phénomènes sémiotiques est d'abord subjective. Elle identifie le sujet à sa pratique de la langue, relativement à celle d'autres sujets. Elle est aussi plus proche des formes – constituées en « faits de langue » – auxquelles le sujet affecte du sens. La perspective dia-sémantique devrait contribuer au repérage des faits et des styles sémantiques relativement aux référentiels auxquels les sujets se rapportent. Les structurations sémantiques plus ou moins singulières dont ils sont responsables ne peuvent être représentées qu'à condition que soit pris en compte leur univers de référence.

Bruno COURBON
Université Laval

BIBLIOGRAPHIE

Études

BIBER Douglas (2008) [1993] : « Representativeness in corpus design », in T. Fontenelle (dir.), *Practical lexicography*, Oxford, Oxford University Press, p. 63-87.

BILGER Mireille (dir.) (2000) : *Linguistique sur corpus*, Perpignan, Presses Universitaires de Perpignan.

BLANCHE-BENVENISTE Claire (2005), « Les aspects dynamiques de la composition sémantique de l'oral », in A. Condamines (dir.), *Sémantique et corpus*, Paris, Hermès et Lavoisier, p. 39-73.

BLOOMFIELD Leonard (1933) : *Language*, New York, Holt.

CHOMSKY Noam (1975 [1972]) : *Questions de sémantique*, Paris, Seuil.

CONDAMINES Anne (dir.) (2005) : *Sémantique et corpus*, Paris, Hermès et Lavoisier.

COSERIU Eugenio (1966) : *Probleme der romanischen Semantik*, Tübingen, Narr.

— (1998a) : « Le double problème des unités "dia-s" », *Les Cahiers δια*, 1, p. 9-16.

— (1998b) [1981] : « Sens et tâches de la dialectologie », *Les Cahiers δια*, 1, p. 17-56.

COURBON Bruno (2004) : *Le sens « matériel » du mot* fortune *en français. Étude diachronique et historique d'une signification lexicale*, mémoire de DEA, Université Lyon 2.

— (2007) : « Usage(s) d'une approche pragmatique en sémantique diachronique : des lieux d'apparition et de stabilisation de l'innovation sémantique lexicale en français moderne », *Revue de sémantique et pragmatique*, 21-22, p. 149-173.

— (2012a) [2007] : « Une réutilisation possible du concept d'usage en sémantique diachronique ? », in J. Glikman *et al.* (dir.), *Le vocabulaire scientifique et technique en sciences du langage. Coldoc 2007*, p. 102-128.

— (2012b) : « Intégration syntagmatique du sens lexical et établissement de rapports synonymiques », in F. Berlan et G. Berthomieu (dir.), *La synonymie*, Paris, Presses de l'Université Paris-Sorbonne, p. 329-341.

DEBAISIEUX Jeanne-Marie (2005) : « Les corpus oraux : situation, exploitation linguistique, bilan et perspectives », *Scolia*, 19, p. 9-40 (la pagination suivie est celle du manuscrit de l'auteure, consulté sur le site http://halshs.archives-ouvertes.fr).

DOMMÈS Aurélie et LE ROUZO Marie-Louise (2007) : « Compréhension d'énoncés contenant une ambiguïté lexicale chez les adultes jeunes et âgés : effets de contexte, de familiarité et de fréquence », *Bulletin de psychologie*, 60, p. 59-69.

DUCROT Oswald et SCHAEFFER Jean-Marie (1995) : *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil.

FLYDAL Leiv (1952) : « Remarques sur certains rapports entre le style et l'état de langue », *Norsk tidsskrift for sprogvidenskap*, 16, p. 240-257.

FONTENELLE Thierry (2008) : *Practical lexicography*, Oxford, Oxford University Press.

GADET Françoise (1989) : *Le français ordinaire*, Paris, Armand Colin.

— (2003) : *La variation sociale en français*, Paris et Gap, Ophrys.

— (2004) : « Le style comme perspective sur la dynamique des langues. Introduction », *Langage et société*, 109, p. 1-8.

HABERT Benoît, NAZARENKO Adeline et SALEM André (1997) : *Les linguistiques de corpus*, Paris, Armand Colin.

HABERT Benoît (2000) : « Des corpus représentatifs : de quoi, pour quoi, comment ? », in M. Bilger (dir.), *Linguistique sur corpus*, Perpignan, Presses Universitaires de Perpignan, p. 11-58.

HAUGEN Einar (1972) : *The ecology of Language*, Stanford, Stanford University Press.

KEHOE Andrew (2006) : « Diachronic linguistic analysis on the web with WebCorp », in A. Renouf et A. Kehoe (dir.), *The changing face of corpus linguistics*, Amsterdam, Rodopi, p. 297-307.

KERBRAT-ORECCHIONI Catherine (2000) : « L'analyse des interactions verbales : la notion de "négociation conversationnelle" », *Lalies*, 20, p. 64-141.

KERBRAT-ORECCHIONI Catherine et CONSTANTIN DE CHANAY Hugues (2007) : « *100 minutes pour convaincre* : l'éthos en action de Nicolas Sarkozy », in M. Broth *et al*., *Le français parlé des médias*, Stockholm, Acta Universitatis Stokholmiensis, p. 309-329.

KILGARIFF Adam et GREFFENSTETTE Gregory (2008) [2003] : « Introduction to the special issue on the web as corpus », in T. Fontenelle (dir.), *Practical lexicography*, Oxford, Oxford University Press, p. 89-101.

LAKS Bernard (2008) : « Pour une phonologie de corpus », *Journal of French language studies*, 18, p. 3-32.

LEGALLOIS Dominique et KWON Song-Nim (2006) : « Sémantique lexicale et examen écologique de la co-occurrence », *Cahiers de lexicologie*, 89, p. 143-162.

MARCHELLO-NIZIA Christiane (1985) : « Question de méthode », *Romania*, 106, p. 481-492.

MCENERY Tony et WILSON Andrew (1996) : *Corpus linguistics*, Edinburgh, Edinburgh University Press.

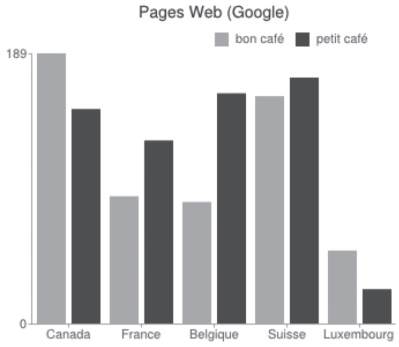
MOUNIN Georges (dir.) (2004) [1974] : *Dictionnaire de la linguistique*, Paris, Presses universitaires de France.

MORTUREUX Marie-Françoise (1981) : « Le “corpus” dans les études de lexique-sémantique », *LINX*, 4, p. 47-68.
MULLER Charles (1984) : « Préface », in P. Lafon, *Dépouillements et statistiques en lexicométrie*, Paris, Champion.
NYCKEES Vincent (2006) : « Rien n'est sans raison : les bases d'une théorie continuiste de l'évolution sémantique », in D. Candel et F. Gaudin (dir.), *Aspects diachroniques du vocabulaire*, Mont-Saint-Aignan, Éditions Universitaires de Rouen et du Havre, p. 15-88.
— (2007) : « La cognition humaine saisie par le langage : de la sémantique cognitive au médiationnisme », *Corela*, [En ligne], http://corela.revues.org/1538.
PAQUET-GAUTHIER Myriam (à paraître), « Ouvrages normatifs et “anglicismes sémantiques” à l'heure des grands corpus informatisés », *Actes des XXVIII^e Journées de linguistique*, Université Laval, Québec.
RASTIER François (2011) : *La mesure et le grain*, Paris, Champion.
REY Alain (2008) : *De l'artisanat du dictionnaire à une science du mot*, Paris, Armand Colin.
SAUVAGEOT Aurélien (1957) : *Les procédés expressifs du français contemporain*, Paris, Klincksieck.
SIOUFFI Gilles, STEUCKARDT Agnès et WIONET Chantal (2012) : « Comment enquêter sur des diachronies courtes et contemporaines ? », *Actes du 3^e Congrès Mondial de Linguistique Française*, Lyon, p. 215-226.
VÉRONIS Jean (2004) : « Quels dictionnaires pour l'étiquetage sémantique ? », *Le français moderne*, 72, p. 27-38.
VINCENT Diane (2009) : « Corpus, banques de données, collections d'exemples. Réflexions et expériences », *Cahiers de linguistique*, 33, p. 81-96.
WEINREICH Uriel, LABOV William et HERZOG Marvin (1968) : « Empirical foundations for a theory of language change », in W. P. Lehmann et Y. Malkiel (dir.), *Directions for historical linguistics*, Austin, University of Texas Press, 1968, p. 97-195.

Principales ressources exploitées

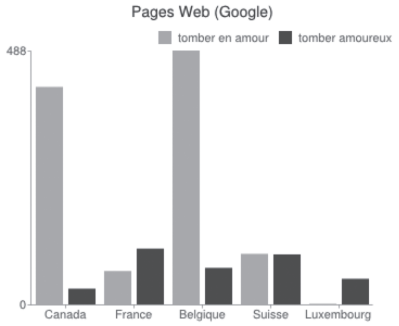
CFPQ, Corpus de français parlé au Québec, Université de Sherbrooke, http://recherche.flsh.usherbrooke.ca/cfpq, novembre 2013.
CLaPI, Corpus de langue parlée en interaction, Université Lyon 2, http://clapi.univ-lyon2.fr, novembre 2013.
DIATOPIX, logiciel d'affichage de requêtes Google par régions, Université de Montréal, http://olst.ling.umontreal.ca/~drouinp/diatopix, avril 2014.
DOCTISSIMO, site de forums électroniques (santé), http://www.doctissimo.fr, septembre 2013.
EUREKA, banque de textes journalistiques, http://www.eureka.cc, mars 2014.
FRANTEXT, banque de textes, http://www.frantext.fr, février 2014.
Le Petit Robert 2014 (2013) : Paris, Le Robert.
TOURNIER Jean et TOURNIER Nicole (2009) : *Dictionnaire de lexicologie française*, Paris, Ellipses.
USITO (2013) : *Dictionnaire*, Université de Sherbrooke, Delisme, http://www.usito.com, mars 2014.

ANNEXE I
SONDAGES RÉALISÉS DANS/SUR LA TOILE : LE REFLET DE TENDANCES DIATOPIQUES[43]



Valeurs relatives (par million de pages)

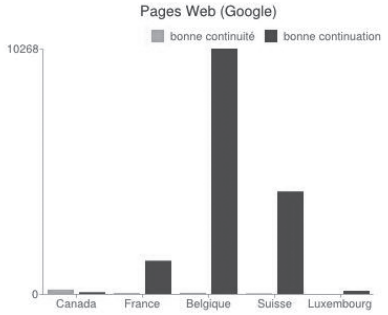
	bon café	*petit café*
Canada	189	150
France	89	128
Belgique	85	161
Suisse	159	172
Luxembourg	51	24



Valeurs relatives (par million de pages)

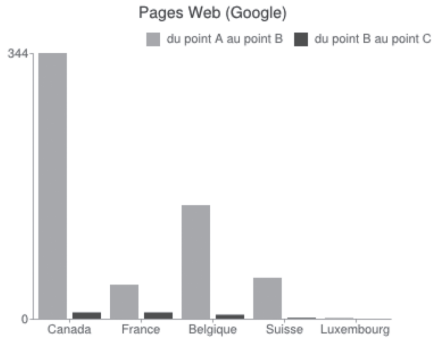
	tomber en amour	*tomber amoureux*
Canada	418	30
France	64	107
Belgique	488	70
Suisse	97	96
Luxembourg	1	49

43 Ces sondages ont été réalisés en avril 2014 au moyen du logiciel Diatopix, développé par Patrick Drouin à l'Université de Montréal.



Valeurs relatives (par million de pages)

	bonne continuité	*bonne continuation*
Canada	183	71
France	36	1388
Belgique	47	10268
Suisse	21	4289
Luxembourg	1	130



Valeurs relatives (par million de pages)

	du point A au point B	*du point B au point C*
Canada	344	8
France	44	8
Belgique	147	5
Suisse	53	1
Luxembourg	1	0

ANNEXE II
EXEMPLES D'INTERACTIONS SUR LE SENS

L'interrogation sur l'usage des signes / sur leur signification

1) « *Pourquoi tu utilise les dollars ?* Tu n'es pas française ? »
(italique rajouté ; maelyspupuce ; 9 octobre 2011 18:40:34 ; Doctissimo, forum Loisirs)

2) « par : LAMPADAIRE1
il me reste des recettes de cuisine canadiennes mais j'ai doute sur les convertions *1 tasse égale combien de gramme*?
posté le 5/12/03 à 12:53 »
(Italique rajouté ; http://forum.aufeminin.com/forum/cuisine1/__f42761_cuisine1-1-tasse-de-grammes.html)

Des réponses aux questions sur les signes / leur signification

2) « KYÔ SEN'NYU
[…]
Nombre de messages : 1211
Age : 22 Localisation : dans sa déprime habituelle :P (arf, que je suis méchaaante !!) !
Passion(s) : déprimer xD
Date d'inscription : 04/06/2006
Sujet : Re : +Horloge parlante+ tic tac tic tac.. Mar 1 Jan - 18:07
18h07 => grrr ! Je t'emm***, je fais du 95 C !
Bon, ça veut dire quoi, botché ?

AMILY TRYNAS
[…]
Nombre de messages : 1141
Age : 19
Localisation : Québec loin de vous les amiis =(
Date d'inscription : 06/03/2006
Sujet : Re : +Horloge parlante+ tic tac tic tac.. Mar 1 Jan - 18:14
Botché… Raté… Fais vite et mal… Bref… »

(Italique rajouté ; http://fee-ecole.forumactif.com/t5381p90-horloge-parlante-tic-tac-tic-tac)

3) « Re : Le coin des garçons
par Rebecca Buck le 03 Oct 2007, 23:39

Rituelle a écrit : […] il se leve debout quand ils font un but. est vraiment déçu quand l'autre équipe en score un. lors des batailles, il frappe dans les airs et me dis que sa aide le joueur
le lendemain quand il en parle avec *ses* «*chum*» au lieu de dire le canadien a bien jouer. il dit ON a bien jouer…
une chance que lui y'était la! il a fait toute la différence dans la game pour pouvoir dire ON a bien jouer!

Hahaha je reconnais beaucoup d'gens dans *le portrait d'ton chum*. Au fait, ça m'a toujours intriguée ce "ON"… et ils ne rateront pas une occasion d'la sortir celle-là !
Rebecca Buck

Messages : 326
Inscrit le : 17 Sep 2007, 12 :01
Âge : 27 ans
Sexe : Féminin
Localisation : QC/CH

Re : Le coin des garçons
par Parker Jones le 04 Oct 2007, 00:10
Euh… excusez-moi de paraître véritablement demeurée mais… *ça veut dire quoi "chum" ?*
Parker Jones

Propriétaire
Messages : 1858
Inscrit le : 23 Nov 2006, 13:39
Âge : 32 ans
Sexe : Féminin
Localisation : Un petit village paumé en Normandie avec l'ADSL

Re : Le coin des garçons
par Akeens le 04 Oct 2007, 00:16
Chum = mec ou pote.
[…]

Messages : 465
Inscrit le : 20 Sep 2005, 23:39
Sexe : Masculin

Re : Le coin des garçons
par Rituelle le 04 Oct 2007, 07:22
Parker Jones a écrit : Euh… excusez-moi de paraître véritablement demeurée mais… *ça veut dire quoi "chum" ?*
chum c'est soit un ou une ami(e)… mais plus souvent un amoureux… on dit du mec un chum et d'la fille une blonde
donc je suis la blonde de mon chum
Rituelle

Messages : 82
Inscrit le : 20 Sep 2005, 19:39
Âge : 34 ans
Sexe : Féminin
Localisation : Drummondville »

(italique rajouté ; http://www.simsagora.net/viewtopic.php?f=15&t=5358&start=40)

La reprise interdiscursive

4) « babybelle19
Profil : Doctinaute d'or
Posté le 14-07-2009 à 20:15:27
J'ai acheté la crème de jour et *la crème de "coucher" comme ils l'appellent*… »
(italique rajouté ; Doctissimo, forum Environnement)

5) « Moi je m'en carre pas des bots Mortal, parce que de 1 ils peuvent tjrs boter *"dans leur coin" comme tu dis si bien*, pour monter leur paragon easy, et de 2 lors de l'arrivé des ladder, ca risque de gueuler quand tu verras que tout les TOP c'est quasi que des boters » (italique rajouté ; HanGPIErr ; 3 avril 2014 ; http://eu.battle.net/d3/fr/forum/topic/10271827269)

6) PrincesseSara
Inscrit : 30 Juil. 2004
Messages : 7 971
[…]
Mon pere fait ca a la perfection …. lait, bananes, oranges, fraises, amandes en option, + une (voire 2) boule de glace a la vanille …
… ca dechire (comme ils disent les djeuns')
:D
PrincesseSara, 5 Mars 2006 »
(italique rajouté ; http://www.bladi.net/forum/threads/panach-boule-glace.61614)

L'illustration lexicale

7) « Une **gélatine**, on appelle ça familièrement un **jujube**, au Québec. Le *jujube* est en réalité le fruit du *jujubier*, un arbrisseau d'origine chinoise. Quel rapport avec les gélatines? Franchement, j'en ai aucune idée, mais des jujubes, c'est bon! »
(gras original ; Guy Labbé ; février 2009 ; http://www.candide.ca/blog/friandises/jujube-gelatine.html)

8) « *TRIHOREAU* dit :
30 juin 2011 à 18 h 55 min
Merci Chrystèle de nous faire à nouveau vivre ces moments et paysages très agréables *avec toute « la gang »* …comme on dit au Québec! »
(italique rajouté ; http://www.aventuresnouvellefrance.com/blog/panorama-voyage-quebec)

L'adaptation dia-lectale

9) « J'utilise de l'huile de canola (pour les Français, *le canola, c'est du colza canadien*)!!! »
(italique rajouté ; 14 octobre 2010 ; blogue http://poeleaunez.canalblog.com)

10) « seawitch
plus moche la mort
Profil : Doctinaute Hors Compétition Posté le 07-01-2011 à 08:25:46

> femmequebec a écrit :
> Donc *arrêter de capoter franchement.*

> *ça veut dire quoi capoter ?*

[…]

> femmequebec
> Quoi de neuf?
> Profil : Fidèle Posté le 08-01-2011 à 01:26:47
> […]
> Donc *arrêter d'avoir peur(de capoter)* »

(italique rajouté ; Doctissimo, forum Vie pratique)

11) « À l'époque, il y a 4 ans environ, j'étais allé à l'urgence en détresse respiratoire pour me faire dire qu'*il s'agissait d'hyperventilation. Je crois que les Français appellent ça la spasmophilie*. »
(italique rajouté ; grabuge – bisbille ; 04-10-2011 ; Doctissimo, Santé)

12) « 10 % de la chaleur corporelle s'échappe par la tête : si vous avez froid aux pieds *mettez un bonnet ou une tuque comme on dit au Québec*! Les moufles sont bien plus efficaces pour tenir au chaud mais un peu moins pratiques pour saisir (si vous le pouvez prenez une paire de moufles et une paire de gants de "travail"). » (gras rajouté ; non daté ; http://www.voyagekweikweiquebec.com/formules/conseils_ours.html)

Jouer sur le sens incertain d'un mot

> 13) « manon said…
> […] euh Amandine ça veut dire quoi capoter ? *je suppose que c'est du québecois*?)
> 10:00 AM

> Amandiine said…
> *Capoter veut dire un truc du genre "Je suis en train d'hallucinéééé"* lol
> […]
> 10:02 AM

> manon said…
> je me doutais bien mais tu me rassure un peu ! Parce que *un mec qui dis à une fille "je capote là" ça a de quoi être ambigue.* […]
> 5:42 PM »

(italique rajouté ; blogue ; http://4mand.blogspot.com/2006/04/new-york.html)

> 14) « Kanapeach
> Rédacteur invité (tmp)
> Hero Member
> ********
> Messages : 4613

Sexe: Homme
[…]
Re : Présentez-vous!
« Réponse #11071 le: 12 décembre 2011, 18:59:00 »
Bienvenue aux quelques nouveaux que j'ai loupé!

> Citation de: allbrice le 12 décembre 2011, 18:56:02
> *Ça veut dire quoi capoter ?*

Je vois qu'on a pensé à la même chose =D

[…]

Staff AK 3
Hero Member
****
Messages: 718
Sexe: Femme
Re : Re : Présentez-vous!
« Réponse #11073 le: 12 décembre 2011, 21:20:32 »

> Citation de: allbrice le 12 décembre 2011, 18:56:02
> *Ça veut dire quoi capoter ?*

Capoter (Québec) (Familier) S'exciter, s'emballer.
Au passage, bienvenue !

[…]

Hero Member
*****
Messages : 5184
[…]
Re : Présentez-vous!
« Réponse #11074 le : 12 décembre 2011, 21:22:45 »
T'aurais peut-être pas dû nous traduire ça comme ça, yarashii… :D

[…]

shadow8
Hero Member
*****
Messages : 5680
Sexe: Homme
[…]
Re : Présentez-vous!
« Réponse #11075 le : 12 décembre 2011, 22:12:42 »
Bien, *c'est dans le genre d'aimer ou adorer*
Vous êtes satisfais maintenant, bande de français. :P »

(italique rajouté ; http://www.anime-kun.net/forums/index.php?topic=9.11070)