Le corpus entre données, analyse et théorie

Dalbera Jean-Philippe

Numéro 1 Corpus et recherches linguistiques - novembre 2002

Résumé : L'usage de corpus n'est pas réservé aux linguistes. Néanmoins ceux-ci en sont des utilisateurs patentés, du fait, entre autres, que leurs analyses portent sur des productions linguistiques ou langagières non finies dont l'étude ne peut s'opérer que sur un échantillon. Mais pour que l'analyse prétende à quelque validité, cet échantillon doit être représentatif. Représentatif de quoi ? D'une réalité qui à la fois préexiste à l'analyse et qu'il contribue à cerner et à établir. D'où toute une palette de corpus dont les principaux types en usage dans la discipline, selon les matériaux utilisés, selon la clôture imaginée, selon la fonction assignée…, sont brièvement rappelés. La réflexion est ensuite centrée sur la délimitation de la place et de la fonction du corpus entre faits, analyses et théories ; il est montré, quelques exemples à l'appui, empruntés à la démarche du dialectologue et du lexicologue, que le corpus ne saurait être qu'un construit et que sa construction fait partie intégrante du prisme théorique à travers lequel le linguiste entend appréhender le réel.

Habeas corpus ! On est tenté de détourner le sens de cette formule et d'en faire une adresse au linguiste pour l'enjoindre à présenter ses données, tant il est vrai que le corpus joue un rôle de premier plan dans son travail quotidien. Pourtant la banalité et la fréquence d'emploi de ce terme dissimulent à peine des conceptions notablement diversifiées, en liaison avec les domaines d'étude et les approches spécifiques des uns et des autres. Nous nous proposons, pour notre part, d'illustrer, à partir d'expériences personnelles, quelques types de corpus en soulignant leur mode de constitution. Nous nous placerons successivement, en l'occurrence, dans la position du dialectologue et du lexicologue. Et nous essaierons, après quelques considérations d'ordre général, de mettre en évidence, à propos de cette catégorie de corpus, la dialectique du terrain et de la théorie ou du donné et du construit.

1. Quelques rappels

Au sens trivial – si sens trivial il y a –, tel, du moins, qu'il est recensé dans les dictionnaires d'usage courant, un corpus est un recueil de pièces ou de documents qui concernent une même matière, discipline ou doctrine. Mais cette acception large et floue se spécifie dans ses usages et se spécialise dans certains secteurs de la connaissance. En droit, le *corpus* renvoie par ellipse, au *corpus juris*, c'est-à-dire au corps de droit romain tandis qu'en physique *corpus* n'est guère utilisé et *corpuscule* qui prend le sens de « particule », constituant discret de la matière n'apparaît guère comme son diminutif.

Dans les sciences du langage – cette définition apparaît dans les dictionnaires les plus récents – un corpus est un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique. Le terme a pourtant conservé, en linguistique, un peu de son acception d'origine, d'où une certaine ambivalence. *Corpus* renvoie effectivement, en un premier sens, à une collection de textes présentant une certaine unité de genre ou bien d'époque ; ainsi furent élaborés au XIX^{ème} siècle le *Corpus inscriptionum graecarum* et le *Corpus inscriptionum latinarum*. *Corpus* devient même un mot français à part entière dès lors qu'il ne s'inscrit plus dans un syntagme latin, et l'on parle de *Corpus des Troubadours* ou de *Corpus des poètes de la Renaissance*.

Ce type de corpus n'est nullement l'apanage du linguiste : l'historien, le philologue, le juriste entre autres travaillent de leur côté sur des objets analogues même si leur perspective heuristique se révèle sensiblement différente ou si les matériaux eux-mêmes sont différents. Robert mentionne prudemment « recueil de pièces », et de fait, ce corpus peut être constitué de textes certes, mais aussi de mots, de témoignages oraux (enregistrés ou transcrits), etc.

Il est relativement facile de recenser les principaux critères de classification des corpus. On peut distinguer ainsi, selon la nature des matériaux constitutifs, les corpus de textes et les corpus d'unités de langue (entendons par là des phrases, des mots, des phonèmes…). On gagne à dissocier également parmi ces derniers, compte tenu notamment des méthodologies induites et de la pesanteur des traditions, ceux qui relèvent de l'écrit et ceux qui relèvent de l'oralité. Une autre division est opérée entre les corpus conçus comme des échantillons représentatifs des faits linguistiques et ceux qui se veulent exhaustifs dans un champ donné. Sont à prendre en compte également le caractère clos ou non-clos d'une part, brut ou traité de l'autre, des *data* constitutives du corpus. Les combinaisons multiples de ces choix binaires[1] engendrent évidemment, en définitive, une palette assez riche de corpus.

Reste une distinction fondamentale sur laquelle on reviendra dans ce qui suit : le corpus est-il de l'ordre du donné ou du construit ? Et, question partiellement

attachée à ce qui précède, la fonction assignée au corpus est-elle de démonstration ou de validation (d'une hypothèse) ?

2. Le corpus du linguiste

La question a été posée plus haut de savoir à quoi rime un corpus ; pourquoi le linguiste, en l'occurrence, use-t-il d'un corpus ? La réponse paraît évidente : quel que soit le domaine ou le champ linguistique à étudier, le volume de données est si considérable que l'on ne saurait tout prendre en compte dans le cours de l'analyse (que peut bien signifier *tout* d'ailleurs ?). De sorte que l'on est conduit à faire l'hypothèse (le pari) que les régularités susceptibles d'être découvertes par l'analyste sont potentiellement récursives et donc qu'une analyse limitée à un sous-ensemble de faits peut être de nature à rendre compte de l'ensemble[2].

Le travail sur échantillon se révèle donc être un impératif pratique. Mais pour que l'analyse prétende à quelque validité, on ne saurait se contenter d'un échantillon aléatoire. Tendre un micro et enregistrer quelques heures de discussions dans une salle de réunion ou de café permet sans doute de recueillir quelques éléments intéressants (selon les points de vue auxquels on se place, tours spécifiques de l'oralité, stratégies discursives, courbes intonatives, coloration régionale...), voire indiscrets, mais ne saurait tenir lieu d'échantillon représentatif du français parlé.

Par ailleurs, construire un corpus n'implique pas nécessairement une analyse spécifique en arrière plan ; mais le type de données sélectionnées n'est jamais innocent et traduit une préoccupation sous-jacente. Pour prendre un exemple simple, les corpus rassemblés par les dialectologues chargés d'enquêter, dans le cadre de la même entreprise, en obéissant aux mêmes consignes, afin de réaliser les Atlas Linguistiques de la France par régions ne sont pas complètement analogues. Certains dialectologues se sont soucié, lors des enquêtes, de noter les données négatives[3], tandis que les autres ne l'ont pas fait. Cela ne peut manquer d'avoir un impact sur les interprétations ultérieures[4].

On distingue donc schématiquement deux phases dans une étude linguistique : la phase d'analyse d'un ensemble fini de données et la phase de confrontation des résultats de cette analyse, c'est-à-dire des hypothèses avancées, à la réalité. Il s'ensuit que le recours explicite au corpus peut intervenir dans une phase liminaire de la recherche au moment où l'on tente de cerner les faits pertinents ou en fin de recherche au moment de valider les hypothèses émises.

Dans le cadre de la première attitude, lorsqu'on travaille avec un corpus-échantillon, on délimite les faits à étudier puis on procède à leur analyse. Cela

implique deux conséquences : la clôture du corpus relève de la responsabilité du chercheur, et la représentativité du corpus – dont dépend la validité de l'analyse – est exclusivement du ressort du chercheur. Le corpus apparaît dès lors clairement comme un objet construit. On discerne alors le double glissement de la notion « générique » de corpus :

1 – ensemble de faits présentant une certaine homogénéité

2 – ensemble de faits pertinents

3 – ensemble construit de faits.

De la notion de « collection d'objets » réunis parce qu'ayant en partage, au moins superficiellement, une ou plusieurs propriétés, on passe à un ensemble trié d'objets, c'est-à-dire à un ensemble de données filtrées, puis à un ensemble de données construit, c'est-à-dire complété ou remodelé par rapport à l'ensemble précédent de manière telle qu'il soit susceptible d'attester les possibles que l'analyse de l'ensemble précédent a suggérés.

3. Le corpus : échantillon représentatif

Revenons à présent, sur un exemple, à la phase de délimitation des données.

Soit un phonologue qui se penche sur la notion d'attaque de la syllabe en français. Quand il examine les possibilités qui s'offrent à l'initiale du mot, il se trouve en présence d'un inventaire assez complexe qu'il suspecte d'hétérogénéité. Il est notamment confronté à la question de la clôture des données ; quels sont les mots à prendre en compte ? On serait tenté de répondre que tous doivent l'être. Mais encore ?... Les mots de type *look, week-end, ciao, bye*... font-ils partie intégrante des données ? Et *slave, slalomer*..., et *psalmodier, ptérodactyle, xylophone, tmèse*, etc. ? Il a donc affaire à un paradigme composite comprenant aussi bien *pr-, tr- kr-, br-, dr-, gr-, pl-, kl-, bl-, kl-, fr-, fl-* que *spr-, str-, skr*..., *spl- skl-*, mais aussi *sl-*, mais aussi *ps-, pt-, ts-, ks-, kt-, tm-, kn*...

Il est donc relativement aisé de délimiter un échantillon représentatif de données, à condition bien sûr, d'assumer les exclusions. Mais, même dans un cas aussi élémentaire, la partition des données du corpus peut conduire plus loin.

Si le linguiste veut parvenir à un résultat et fournir une analyse cohérente, il sera peut-être tenté d'agir sur le corpus, non pas, certes, en modifiant les données, mais en découpant celles-ci de manière à rendre les structurations plus sensibles. Il pourra, par exemple, supposer une stratification des formes inventoriées,

renvoyant à la matrice lexicogénique (source grecque des mots du vocabulaire scientifique, marquage « savant » ...) ou à tout autre paramètre ; il pourra modeler son corpus en faisant référence aux niveaux ou variétés de langue[5]. Il lui restera ensuite à mettre en évidence les mécanismes à l'œuvre dans chacune de ces strates, à évaluer les interactions entre ceux-ci et, bien entendu, à justifier empiriquement son découpage. Mais ceci est une autre histoire.

Le point qui nous intéresse est que le corpus n'est pas un simple sous-ensemble des données de la réalité mais que cet échantillon est déjà travaillé. Il reste que l'analyse ne vaut que ce que vaut le corpus. On a trop souvent critiqué les corpus *ad hoc* ou les corpus introspectifs pour qu'il soit nécessaire d'y insister ici.

4. Le corpus : échantillon construit

Un deuxième exemple, extrait de la pratique du dialectologue, peut permettre d'illustrer la phase dynamique de construction du corpus. Le linguiste qui « fait du terrain », qui explore une aire dialectale peu ou mal connue, vit en permanence les métamorphoses de son corpus. Pour dire les choses très schématiquement, le dialectologue, en face d'un idiome nouveau, qu'il découvre dans le cadre d'une enquête, recueille dans un premier temps du « tout venant », ne sachant évidemment pas à l'avance quel type de traits caractérise, aréalement et diachroniquement, le parler dont il consigne pour la première fois les manifestations. Il procède ensuite, avant de revenir au terrain, à une analyse des faits engrangés et se trouve confronté à un certain nombre de difficultés (séquences phoniques inhabituelles pour lui et pour lesquelles il ne dispose que d'un nombre insuffisant d'exemples, mode de fonctionnement des enchaînements (liaison, élision) pour lequel les séquences enregistrées n'illustrent pas tous les cas de figure possibles, opposition phonématiques fugitives ou douteuses, corrélations morphologiques incomplètes, etc. Ce sont ces points qui vont aiguiller la suite de l'enquête : les bonnes questions à poser aux témoins, celles qui vont faire la lumière sur les spécificités du parler, celles qui vont faire surgir les réponses les plus riches d'enseignement, que ce soit dans une perspective de description synchronique, de reconstruction diachronique ou de comparaison aréale, ce sont celles qui vont amener des réponses aux originalités que le premier regard a cru déceler, celles qui vont livrer les clefs de tous ces comportements imprévus que la langue aura donné à voir lors de ses premières manifestations : les résultats (provisoires) de l'analyse conditionnent pour partie les questions et configurent le futur corpus représentatif. « On ne trouve que ce qu'on cherche » n'est nullement une lapalissade. Combien de fois nous a-t-il été donné de vérifier que l'enquêteur ne recevait et n'exploitait vraiment que ce à propos de quoi il s'était posé des questions.

Pour ne citer qu'une illustration très succincte (mais authentique)6, le fait d'enregistrer, dans le Haut-Taravo, en Corse, une forme comme *rinovime* pour le « levain » aurait pu demeurer un fait singulier, anecdotique. Cette notion était demandée partout, la réponse généralement fournie mais énigmatique au plan de sa source (étymologique ou motivationnelle) était : *nuirme, rnuvime, rnuime, rruime, rinuime, nuirmu.* Ces termes avaient fait l'objet d'hypothèses étymologiques diverses mais n'étant guère de nature à entraîner l'adhésion. L'enquêteur du NALC, sensibilisé au problème du motif de cette désignation, recueille effectivement au gré de ses enquêtes la plupart de ces formes déjà attestées par le passé, mais, le jour où un témoin lui donne pour réponse *rinuime* ou *rinuvime*, il ne se tient pas quitte de ce résultat et tente d'aller plus loin dans l'élucidation : le témoin explique alors à sa demande la technique de fabrication du pain. Ce qui est utilisé pour faire le pain, pour obtenir qu'il lève, c'est une boule de pâte conservée de la fabrication précédente, qui fait donc office de levain. Le levain, c'est donc le *ri-novime*, le « renouvellement ». Le motif devient limpide, et l'étymologie du terme une évidence : le levain c'est le machin à « renouveler » le pain ; la série de termes obscurs énumérée plus haut ne représente que des déformations de *rinovime* ayant conduit à l'opacification de son motif. Elémentaire, bien sûr, mais cette interprétation « naturelle » n'a été fournie par le témoin que sur sollicitation : encore fallait-il poser la bonne question ou du moins être sensible à la réponse. Et cette sensibilité à la réponse ou plus généralement au ***fait*** ne se trouve que chez celui qui, pour dire les choses lapidairement, a les questions en tête. C'est une des raisons pour lesquelles l'opposition que l'on fait parfois entre dialectologues de terrain (chargés d'arpenter le territoire et de recueillir le corpus de données) et linguistes de cabinet (chargés de faire l'analyse de celles-ci) est une fiction qui émane de gens qui n'ont jamais été confrontés aux données réelles.

C'est de proche en proche que le corpus s'élabore, d'hypothèses trop hâtives balayées par les faits en propositions plus subtiles qui cadrent mieux les données, de retouches en retouches et en vérification (indirecte et implicite, évidemment) auprès des témoins. La trame structurelle du parler se dessine ainsi progressivement en même temps que le corpus se construit. De l'exécution d'un questionnaire qui se voulait standard au départ, le travail d'enquête, au fur et à mesure que les faits se dévoilent, s'adapte et s'approfondit là où cela en vaut la peine et le produit résultant, ce que nous nommons le « responsaire », peut n'avoir plus qu'une relation lointaine avec le questionnaire initial qui a servi de base. Il en va d'un protocole d'enquête dialectologique comme du choix et de l'appropriation progressive d'une tenue vestimentaire. On se voit dans certains cas affublé d'un vêtement anonyme, passe-partout (uniforme, unisexe et taille unique) comme l'enquêteur se voit confier un questionnaire standard ; puis celui-ci devient à mesure de la progression de l'enquête et de l'analyse, l'équivalent d'un « prêt à porter » pour finir comme un « sur mesure » bien ajusté. Le

questionnement mécanique conçu a priori se corrige, se complète et se peaufine tout au long de l'enquête et, par suite le corpus de réponses auxquelles il donne lieu et, en définitive, l'échantillon même qui en est issu. ***Le corpus est indissociable de l'analyse***.

5. La clôture du corpus partie intégrante de la théorie

Les exemples que nous venons de citer, l'analyse phonématique, l'analyse étymologique et la conduite de l'enquête dialectologique de terrain répondent à des cas où le modèle d'analyse est connu et défini par avance. Mais il n'en est pas toujours ainsi. Dans certains domaines, même le découpage approximatif initial (celui qui est censé intervenir avant corrections ou affinements) fait défaut. Si l'on se propose d'étudier telles ou telles structures lexicales d'une langue à un moment donné, sait-on vraiment en quoi consiste l'étude ? A priori, cela semble évident : il faut procéder à l'analyse d'un corpus lexical, susceptible d'être fourni, par exemple, par la consultation des dictionnaires du moment. Mais en retenant quoi ? En privilégiant la forme phonique signifiante, le sens, le référent ? Selon quelle pertinence ?

Le signe lexical possède une double caractéristique : il se conçoit négativement à l'intérieur d'un paradigme, par rapport aux autres signes susceptibles d'apparaître dans le cadre de celui-ci : cela représente sa « valeur » ; mais il dépend par ailleurs (en tant qu'il constitue l'interface avec le monde) de sa relation au référent, c'est-à-dire de ce que l'on appelle le « motif ».

Or il s'avère que les relations fondamentales qui structurent le lexique sont tout à fait distinctes et indépendantes des relations que nous concevons entre les référents. Rendre compte des relations entre des unités lexicales n'a rien à voir avec donner une description différentielle des référents auxquels ces mots renvoient. *Gorges*, pour un locuteur français aujourd'hui, fait partie d'un ensemble de termes qui désignent un passage rétréci, dans la montagne et il se distingue d'autres termes tels que *défilé, canyon, vallon, col, pas*..., mais d'une part cette précision référentielle n'est pas obligatoire : l'hyperonyme est souvent suffisant et fixé par l'usage sans valeur oppositive par rapport aux autres possibles ; d'autre part, un autre terme pourrait être employé à la place de *gorge* sans problème communicationnel majeur ; en troisième lieu, cette comparaison des membres du paradigme ne peut rendre compte des emplois de *gorge* tels que *coupe-gorge, soutien-gorge, gorgé d'eau, régurgiter* ...

Bref, la confrontation sur base référentielle ne peut fournir de résultats probants ; la confrontation sur la base du motif, quand elle est possible, peut éclairer bien des choses, mais encore faut-il accéder au motif. Or le motif est

bien souvent masqué ; on le perçoit souvent intuitivement mais l'établissement de relations sûres suppose des régularités, des schémas qui se répètent, et un état de langue n'est pas à même de livrer des séries attestant de cette régularité. La quête du motif conduit impérativement à élargir le champ ; car le sémantisme créatif ne prend corps que si on le trouve répété, reproduit sous divers oripeaux indépendants les uns des autres et attestant par là-même que c'est bien cette vision que les locuteurs d'une langue ont dans la tête lorsqu'ils créent ce mot. C'est donc seulement si l'on regarde au delà d'un état de langue, en fouillant dans les stades antérieurs de celle-ci, dans les parlers populaires, les jargons techniques, les argots, dans les régionalismes, dans les langues et dialectes voisins ... que l'on se met en position de repérer quelques structures et quelques mécanismes lexicologiquement intéressants. La variation devient la clef, variation au niveau d'un macrosystème s'entend, dont chaque état de langue ne représente qu'une manifestation particulière. Mais alors, le corpus ? On voit bien que celui-ci ne saurait préexister à l'analyse ; il s'élabore, il se dévoile au fur et à mesure que l'investigation avance. De sorte que ***c'est finalement le corpus qui fait la théorie***.

Essayons d'illustrer ce point. En quoi consiste une étude lexicale qui s'attacherait à la notion de *toupie* en français ? Quel type de corpus serait à même de fournir les éléments susceptibles d'éclairer le propos ?

Une première approche qui vient à l'esprit consiste à inventorier les éléments du paradigme ou du champ auquel appartient *toupie* ; mais comment cerner ce champ ? *toupie* s'associe spontanément à toute une série de jouets d'enfants, frustes et connus depuis fort longtemps ; on verrait bien *toupie* dans une série comme *toupie, sabot, toton, dé, osselets* ... peut-être aussi *cerceau* ou encore *crécelle, pipeau* ...Mais le mot *toupie* peut-il se définir comparativement – négativement – dans un ensemble de ce genre ? Dans lequel au juste ? On ne procède là qu'à une comparaison d'objets : la dimension linguistique n'est nullement prise en compte.

Faut-il l'envisager par rapport à ses différents référents ? Qu'est-ce qu'une toupie ? Le dictionnaire répond : « (1) jouet d'enfant, formé d'une masse conique, sphéroïdale, munie d'une pointe sur laquelle elle peut se maintenir en équilibre en tournant ; (2) outil, sorte de tour pour évider (le bois, le métal..) ; (3) femme peu vertueuse ».

Faut-il envisager le mot en contexte ? Le dictionnaire propose lancer, fouetter une toupie, toupie à musique, tourner sur lui-même comme une toupie.

Faut-il prendre en compte le champ morphologique, la famille ? Il semble que le champ se réduise à *toupiller* « (1) tourner comme une toupie, (2) évider avec la

toupie », *toupilleur* « ouvrier du bois travaillant à la toupie » et *toupilleuse* « tour, machine-outil munie d'une toupie ».

Faut-il recourir à des considérations étymologiques ? Le linguiste accoutumé à traiter des problèmes de phonologie, de morphologie ou de syntaxe est réticent. Non que l'analyse diachronique soit exclue de sa sphère d'étude ; mais s'il est clair pour lui que l'étude diachronique et l'étude synchronique se composent pour donner une image en relief de la langue et de son évolution, les deux perspectives ne se confondent pas ; la comparaison saussurienne de la langue et du jeu d'échecs reste présente à sa mémoire : chaque état de l'échiquier a une existence indépendante des coups qui y ont abouti. Un nouvel arrivant peut reprendre, sans dommages, une partie en cours ; l'ignorance des phases qui ont précédé ne le pénalise pas.

La valeur d'un phonème ou d'un morphème grammatical ne dépend pas de sa valeur dans un stade antérieur du système. Il peut être intéressant de savoir d'où est issu tel phonème ou tel morphème mais cela n'intervient en rien dans l'établissement de la valeur de celui-ci à une étape donnée.

En est-il autrement pour une unité lexicale ? Ce qui est en cause, c'est le déroulement du cycle /motivation – convention – arbitraire/. Le motif de départ, celui qui joue lors de la création, représente le sémantisme fondamental ; c'est lui qui relie les unes aux autres, à travers diverses figures bien connues, métaphores, métonymies..., les différents emplois que nous connaissons et dont le dictionnaire fait état. Mais l'opacification du motif qui résulte de l'usage purement conventionnel du terme et des aléas de l'évolution dénoue les fils constitutifs du réseau d'emplois. De sorte que l'accès à la structure lexicale est tributaire de la reconstitution du motif. Cette reconstitution n'est pas nécessairement un problème diachronique ; la création lexicale est, en langue, un mécanisme permanent. Mais dans bien des cas, lorsque la création est ancienne, l'appréhension du motif suppose le recours à des états autres du système, qu'il s'agisse d'états diachroniquement antérieurs ou d'états parallèles (ceux, par exemple, que livre la variation diatopique ou la variation sociolinguistique).

L'analyse de la variation reste, encore une fois, la clef de voûte de l'édifice ; mais celle-ci ne peut s'appréhender qu'à l'intérieur d'un macro-système qui excède largement le cadre d'un état de langue.

Que révèle en l'occurrence la prise en compte d'un champ élargi, incluant temps et espace pour *toupie* ? Un premier élément digne d'intérêt est fourni par les attestations anciennes, en français, de notre mot ; celui-ci apparaît sous la forme *tourpie* et connaît des dérivés de type *toupiller* ou *toupier* « tournoyer comme une toupie ». Ces formes révèlent que les mots de cette famille ont connu deux

flottements au plan de l'expression : relativement à la présence d'un *-r* implosif en position intérieure et relativement à l'interprétation de la partie finale : *-ier* réalisé [je] ou [ije] avec diérèse, ou bien *-iller*, suffixe au demeurant banal [-i´e].

La prise de conscience de ces variations a pour conséquence que *toupie* se trouve rapproché d'un autre mot français que les dictionnaires traitent comme sans rapport aucun avec *toupie* : il s'agit du mot *torpille*. Si l'on veut bien relire la définition de *toupie* « (jouet) ***formé d'une masse conique*** » (c'est nous qui soulignons) ou « sorte de tour pour évider le bois », on se rend compte que non seulement les signifiants *tourpie* et *torpille* se ressemblent mais que les référents de ces deux mots ont également un air de famille très marqué : une torpille, ce n'est autre qu'un projectile formé d'une masse conique. A y réfléchir de plus près, le rapprochement *toupie* - *torpille* n'est pas si étrange que cela : une *toupie* se dit aussi en français *sabot* ; or *saboter* (un projet, par ex.) cela revient à le *torpiller*.

Si à présent on se penche sur les désignations dialectales de la toupie, avec l'idée de déceler à travers les variantes géolexicales le motif de désignation, une sorte de sémantisme fondamental, on trouve des termes comme *giravolta, girardola, virottola*... Nous avons montré ailleurs que le niç. *gavòudoula* appartenait également à cette série : composés tautologiques formés à partir de verbes de mouvement et exprimant l'idée d'un mouvement rotatif ; la toupie, fondamentalement, apparaît comme la petite rouleuse, coureuse, sauteuse, virevolteuse, tournevireuse et c'est cette vision que la langue traduit : le premier segment de tou(r)pie n'est sans doute autre que tourn-er. Là encore, ce sémantisme éclaire les emplois « secondaires » de *toupie*, aussi bien « machine-outil de type tour » (évidemment) que « femme peu vertueuse » ; la toupie tourne en effet mais sans but, à l'aventure, elle vagabonde, c'est une petite galvaudeuse, ou encore plus familièrement une girelle.

On n'ira pas plus loin, ici, dans l'analyse du mot *toupie* et des champs lexicaux dans lesquels celui-ci s'inscrit. Le point que nous entendons souligner est simplement que, dans le cas d'espèce, la construction du corpus servant d'assise à l'étude lexicale entreprise conduit non seulement à opérer des sélections dans les données à disposition mais surtout à élaborer un véritable modèle pour la description lexicale, modèle qui fait éclater les frontières ordinairement respectées dans le cadre d'une étude ponctuelle linguistique. ***Le corpus devient là indissociable de la théorie.***

Bref, le corpus du linguiste est a priori l'ensemble des faits sur la base desquels celui-ci entend conduire son analyse. Ce corpus est, au premier chef, de l'ordre des ***données brutes*** : il consiste en un certain nombre d'unités linguistiques recueillies selon divers modes et rassemblées. L'extrapolation qu'il convient de faire pour étendre les résultats de l'analyse de l'échantillon à la langue impose

que cet échantillon ait un caractère représentatif. La clôture du corpus ne peut plus être aléatoire ni seulement d'ordre quantitatif ; des contraintes qualitatives viennent s'ajouter, le corpus est alors de l'ordre des ***données pertinentes***. Par ailleurs la décision de garder le corpus ouvert a pour corollaire l'implication plus franche du linguiste dans le modelage de celui-ci ; le corpus est alors de l'ordre des ***données construites***.

On voit bien que les relations (corpus de) données – (faisceau d') hypothèses peuvent aisément s'inverser, en ce sens que, bien vite, ce peut être la délimitation du corpus qui « fait » l'objet et qui, pour partie du moins, configure la théorie.

Références bibliographiques

Bilger, M. (éd.) (2000). Corpus, méthodologies et applications linguistiques. Paris : Champion.

Dalbera, J.-Ph. (1988). « Dans le sillage de la toupie nissarde. Notes d'étymologie et de géographie linguistique ». Espaces Romans I. Grenoble : Ellug, pp. 193-204.

Dalbera, J.-Ph. (1996). « Aspects heuristiques : strates et représentations dans une base de données dialectales ». In G. Moracchini (éd.) Bases de données linguistiques : conceptions, réalisations, exploitations. Corte, pp. 103-116.

Dalbera-Stefanaggi, M.-J. (1988). « Evolution phonétique et démotivation : le « levain » corse ». Espaces Romans I. Grenoble : Ellug, pp. 205-212.

Godefroy, F. (1881). Dictionnaire de l'ancienne langue française et de tous ses dialectes, du XI^{ème} au XV^{ème} siècle.

Imbs, P. & B. Quemada (1971-1994). Trésor de la Langue Française. Dictionnaire de la langue du XIX^{ème} et du XX^{ème} siècle, Nancy.

Ravier, X. (1965). « Le traitement des données négatives dans l'Atlas Linguistique et ethnographique de la Gascogne ». Revue de Linguistique Romane, 115/116 : 262-274.

Réseau, P.(2001). Dictionnaire des régionalismes de France. Louvain-la-Neuve : De Boeck, Duculot.

Robert, P. (1985). Dictionnaire alphabétique et analogique de la langue française. Paris : Ed. Le Robert.

Collection des Nouveaux Atlas Linguistiques de la France par région. Paris : CNRS Editions.

1 On se gardera néanmoins d'accorder une importance excessive à des dichotomies qui ne sont qu'indicatives. Nous serions tenté, par exemple, d'étiqueter le corpus des régionalismes de France comme *unités de langue - grand corpus - non clos - brut - oralité - non construit*. Ce corpus est en effet un corpus de mots (et expressions) ; il est de grande taille, ouvert (susceptible d'être enrichi à la lumière d'études nouvelles), brut (c'est un recueil de formes attestées), fait de matériaux de l'oralité, non construit (il se donne à voir comme tel sans constituer l'argument d'une démonstration). Pourtant, à y regarder de plus près, bien des traits apparaissent en porte-à-faux : qu'est-ce qu'un régionalisme linguistique ? Y a-t-il notion plus fugitive et plus controversée ? Certes, P. Réseau a répondu à cette question; il a, du moins, donné *sa* réponse. Mais peut-on encore parler de corpus brut ou non construit ? Par ailleurs, dans sa méthode de travail, P. Réseau a, de fait, accordé une priorité aux régionalismes présents certes dans l'usage quotidien parlé mais dont on possède aussi une attestation dans des textes publiés. Peut-on encore parler de corpus de l'oralité ?
2 Notons que, le plus souvent, les questions que soulève le linguiste mettent en jeu non pas une grande quantité de données mais une infinité de données, dans la mesure où celui-ci doit rendre compte non seulement des faits attestés mais encore des faits possibles.
3 La notion de « données négatives » renvoie au refus, de la part du témoin, d'un terme que l'enquêteur lui propose ; ainsi, par exemple, si à la question « nèfle » le témoin ne répond pas, l'enquêteur fait une suggestion: ne dites-vous pas *nèspo* ? Le témoin peut alors accuser sa mémoire de trahison (« mais bien sûr !... »), ou refuser (« ça ne se dit pas ici ... »).
4 Cf. Ravier (1965).
5 Il a noté, par exemple, que dans le parler familier *pneu* se dit [pønø] et non [pnø].
6 Cf. Dalbera-Stefanaggi (1988).