Corpus : limites, représentativité et choix

Mortéza MAHMOUDIAN

1. HISTORIQUE

A l'origine, le corpus est conçu comme un moyen d'assurer l'objectivité de la recherche en linguistique. A l'époque — vers les années 30-50 du siècle dernier —, les débats entre linguistes portaient souvent sur deux types de problèmes: *i/* Le phénomène φ existe-t-il ? et *ii/* Le phénomène φ a-t-il telle ou telle structure? Autant *ii/* est normal dans le débat scientifique, autant *i/* est à éviter.

Ainsi, dans les sciences, on ne se demande pas si l'*esprit de sel* existe, mais bien quelle en est la composition, de quels corps simples il est composé. Tel n'était pas le cas en linguistique. Considérons des suites comme *On est égaux* ou *On est belle*. La réponse à la question «Ces énoncés existent-ils en français?» ne faisait pas unanimité. Par le recours au corpus, on espérait éviter ce genre de questions et débats. Dans l'ensemble, le concept *corpus* a largement contribué à éviter de faux problèmes. On entend par corpus un *ensemble de matériaux linguistiques concrets*. Le terme concret désigne ici un phénomène qui peut être situé dans le temps et l'espace. Si les matériaux réunis montrent qu'hier à 15 heures, au carrefour des Quatre-Routes, M.Duparc a dit à M. Dubanc *On est égaux*, l'existence d'un tel énoncé ne peut être remise en question.

Même si le corpus pouvait assurer l'objectivité totale des données — ce qui n'est pas le cas, comme nous verrons plus loin § 7 —, reste une autre condition à satisfaire pour que la linguistique puisse prétendre à l'objectivité : une méthode bien définie. Car après tout, l'objectivité signifie l'indépendance de la recherche par rapport à la subjectivité du chercheur. Ainsi, deux linguistes étudiant le même corpus en suivant la même méthode, pourront arriver aux mêmes

résultats. Pour l'objectivité, le corpus est donc une condition nécessaire mais non suffisante.

2. LIMITES

Le corpus présente un avantage certain, malgré les problèmes qu'il rencontre. Le corpus n'est intéressant que dans la mesure où il représente la langue ; car ce qu'on cherche à travers l'analyse du corpus, c'est la structure de la langue. Mais, le corpus peut-il représenter la langue ? Une réponse affirmative implique que1° on trouve dans le corpus tout ce qui est dans la langue 2° on ne trouve dans le corpus rien qui ne soit pas dans la langue. L'expérience montre qu'aucune des deux propositions ne va sans problèmes.

Pour résoudre les problèmes, ont été formulées de nombreuses propositions sur les conditions de la collecte du corpus ou sur son caractère intangible ou extensible du corpus.

3. CORPUS, INTANGIBLE OU EXTENSIBLE ?

Que faire s'il manque dans le corpus des éléments dont l'existence est certaine pour le chercheur (par exemple, des types de phrase qu'il a entendu en dehors de la situation de collecte)? Dans la sa stricte observance du principe «corpus», on doit fonder la description exclusivement sur les données contenues dans le corpus. Ce faisant, on se conformerait au principe de «corpus intangible» ; en même temps on renoncerait à la description totale de la langue.

Une solution alternative serait de chercher des compléments au corpus ; ce que permet le principe de «corpus extensible». Le complément faisant appel à la subjectivité du linguiste, le corpus ne risque-t-il pas de perdre sa raison d'être ?

4. CONDITIONS DE COLLECTE

Sous cet aspect, on a pu constater que le corpus était sensible aux circonstances de collecte ; et que les données recueillies variaient selon ces conditions. Constats qui ont conduit à chercher solution à des problèmes tels que : Comment s'y prendre pour éviter que l'informateur ne produise sa langue endimanchée ? Que faire pour que

l'informateur ne produise pas un mélange d'usages variés ? Comment choisir l'informateur ? Et ainsi de suite.

Ces constats font problèmes quand on part de l'hypothèse que la langue est dotée d'une structure homogène. Ce qui implique que la structure de la langue st une et invariable, et que tous les sujets d'une langue utilisent les mêmes règles et les mêmes unités. Les solutions proposées sont — me semble-t-il — autant d'astuces pour contourner ces problèmes, voire les escamoter.

Or il y a un problème inhérent à la conception et de la fonction de la langue et de la structure linguistique. Car, on se trouve dans l'impasse si l'on admet que a/ la langue est un instrument de communication, et que b/ la langue présente des variétés. Comment la communication serait possible si le sens était variable, donc non partagé ? On pense naturellement à Ionesco et le quiproquo entre deux sujets dont l'un attribuerait au mot *oreiller* le sens "fromage" et non sa signification courante.

Il est vain d'éluder les difficultés. La démarche constructive serait de regarder le problème en face, et d'y chercher une issue. Vue sous cet angle, la question pertinente serait : Comment la communication est possible malgré les variétés ?

Pour y répondre, je crois devoir faire un détour par le signe linguistique et la structure de la langue.

5. CARACTERES DU SIGNE LINGUISTIQUE

Le sens et la dimension psychique sont des propriétés indissociables du signe linguistique. Sans ces deux propriétés le concept de communication est vidé de sa substance. Or, les données contenues dans le corpus ne donnent directement accès ni à l'un ni l'autre.

Quand en compulsant le corpus, nous attribuons un sens aux monèmes ou une pertinence aux sons, nous le faisons par recours à l'introspection du descripteur que nous sommes ou à celle du locuteur qu'est l'informateur. C'est la subjectivité du locuteur que nous prenons comme critère quand nous lui demandons si [bẽ] est la même

chose que [vẽ].Autrement dit, les données du corpus — à elles seules — ne rendent pas possible une description linguistique ; si elles le font, c'est grâce à l'intuition. Ce qui revient à dire que le corpus ne dispense pas le linguiste du recours à la subjectivité.

Pour éviter le recours à l'intuition, Zelig S. Harris propose toute une gamme de procédures distributionnelles. Même avec cette panoplie hypertrophiée, il n'a pas réussi à la contourner. Ainsi, cherchant à trancher certaines questions délicates comme «Deux phonèmes ou deux variantes ?», Harris se trouve obligé de recourir au jugement des locuteurs[1].

6. SYSTEME, FERME OU OUVERT ?

A ses débuts, le structuralisme concevait le système comme fermé, c'est-à-dire constitué d'un nombre fini d'éléments (unités et règles) et doué de frontières précises. Face aux embûches rencontrées (Cf. *supra* §4), on a de plus en plus tendance à adopter le concept de système ouvert. Ouvert dans une double acception : *a/*le nombre des unités et règles qui constituent le système n'est pas déterminé[2], et *b/* tout système déborde sur les systèmes voisins, empiète sur leur champ.

Pour rester sur un plan plus proche du thème du colloque, prenons le point *b/*. Dans ce cadre, la représentativité du corpusn'est pas concevable, étant donné que les limites de l'idiome — qu'il s'agisse de langue, dialecte ou parler importe peu — ne sont pas fixées. Pour le chercheur qui étudie une aire linguistique, il n'est pas toujours facile de trancher des questions telles que

- A-t-on affaire à deux dialectes ou deux parlers du même dialecte?

ou bien

- A-t-on affaire à deux variétés du même parler ou deux parlers distincts ?

Une délimitation de l'aire linguistique est concevable dès qu'on fonde sur des critères précis. Des critères qui soient clairs et nets

même si un peu arbitraires. Ce sera alors une représentativité relative, valable dans des limites qu'on peut définir statistiquement, par approximation

Noter que recourir à l'approximation ne veut pas dire se contenter des *à peu près*. L'approximation peut être une opération bien précise comme dans une opération de division (par ex. 19 par 3), où l'on décide de ne retenir que deux, trois ou *n* décimales.

7. STRUCTURE RELATIVE, HIERARCHISEE

Le concept de système ouvert n'implique pas que les usages de la langue ne soient soumis à aucune règle, libre de toute contrainte ; mais que les unités et règles ne sont pas toutes communes à tous les usagers de la langue. Cette dissymétrie permet d'établir une hiérarchie. Ainsi, certains éléments sont partagés par tous les membres de la communauté linguistique (*i*). A l'opposé, il y des éléments qui sont peu partagés, voire individuels, idiosyncrasiques (ii). Entre les deux extrêmes, on trouve des éléments plus ou moins partagés. Appelons les deux extrêmes respectivement *centre* (*i*) et *marges* (*ii*) de la structure linguistique.

Quand nous voulons constituer un corpus, nous devons nous fixer un but. Disons d'emblée qu'il n'est ni possible ni souhaitable de chercher un corpus qui intègre toutes les habitudes individuelles. On peut alors chercher des matériaux qui ne représentent que la zone centrale du système ? Ou bien inclure aussi certaines parties de la marge, dont les éléments ont une extension importante dans la communauté?

C'est le choix délicat auquel on est confronté dès les premiers pas ; choix qui aura des implications non négligeables pour l'usage que nous voulons en faire.

Revenons au paradoxe du corpus dans la linguistique des années 30-50 (*Cf. supra* § 1), où l'on se posait des questions sur les conditions de la collecte du corpus : Comment s'y prendre pour éviter que l'informateur produise son usage réel, et non sa langue

endimanchée ? Que faire pour que l'informateur ne produise pas un mélange d'usages variés ? Comment choisir l'informateur ?

De telles questions n'auraient aucun sens si la langue avait une structure homogène à travers toutes les fractions de la communauté linguistique, et que tout le monde l'utilisât de manière identique, quelles que fussent les conditions de communication : situation, contexte, objet de l'échange, … Ceux qui posaient ces questions s'imaginaient — à raison — que la description de la langue serait tributaire des caractéristiques du corpus réuni. Et ces questions étaient pertinentes car on était en quête de l'usage spontané de la majorité de la population, et non l'usage normatif qui n'est en réalité que celui de la classe dominante.

8. IMPLICATIONS DE LA STRUCTURE RELATIVE

Ainsi conçue, la structure linguistique ne relève pas de la logique du «oui ou non» ; elle est de nature relative, statistique. Il en découle qu'une opposition phonologique —par ex. /ɛ/ ~ /e/ — ne vaut pas de manière absolue en français ; elle n'est valable qu'à un certain degré si l'on considère la communauté francophone dans son ensemble.

Autrement dit, la structure qui se dégage de la description dépend de la variété que nous considérons. Si la variété considérée est l'usage marseillais, l'opposition /ɛ/ ~ /e/ est très faible, et peut être négligée dans une approximation grossière. En revanche, si nous examinons l'usage vaudois, la même opposition est tellement générale que rien ne permet d'en faire abstraction.

Cela revient à admettre que la représentativité du corpus dépend de la population auprès de laquelle nous l'avons recueilli. Autrement dit, la représentativité du corpus est relative, sélective.

9. COMPLEMENTARITE CORPUS/ENQUETE

Il n'est pas assuré que le recours au corpus soit la seule voie d'accès aux données. L'expérience montre que dans certains cas,

l'enquête permet d'obtenir des matériaux linguistiques adéquats au but visé.

Soit un problème de linguistique appliquée : initiation des jeunes locuteurs à l'écriture, et prenons pour acquis deux principes :

a) les locuteurs —dès l'âge de scolarisation — ont conscience des phonèmes qu'ils utilisent.

b) ils ont capacité à représenter les phonèmes par des substituts (images, dessins, gommettes, …)

Il s'ensuit que est la meilleure façon d'initier à l'écriture est de partir des phonèmes, et de leur substituer un succédané. Le problème sera alors de mettre en évidence les phonèmes qu'ils utilisent et dont ils ont conscience.

Où chercher le système phonologique des enfants de 5-6 ans ? On peut certes penser à un corpus global de la langue. Mais imagine-t-on la taille qu'il devrait avoir s'il doit contenir — outre toutes les variétés sociogéographiques — les phases successives de l'acquisition de la phonologie ? Et les moyens que requerrait la constitution d'un tel corpus ?

Il semble que dans ce cas l'enquête soit un moyen à la fois économique et efficace, si l'on en croit les résultats des expériences menées sur le français[3] ou l'anglais[4]. Et la recherche entreprise sur tamazight par Ramdane Boukherouf — selon les résultats de la pré-enquête[5] — semble aller dans le même sens.

10. CRITERES DE SELECTION

Dans ce qui précède, j'ai essayé de montrer les embûches dont est semé le chemin du corpus. D'attirer aussi l'attention sur le fait que les critères du choix des matériaux ne sont pas neutres, et risquent de favoriser certains usages au détriment d'autres.

Quel critère retenir pour la collecte des matériaux. Corpus ou enquête ? Et pour le corpus, vers quel contenu orienter la collecte ? Le choix s'avère délicat. Cependant certaines pistes se présentent qui méritent examen et réflexion. Le premier critère qui vient à l'esprit est

sans doute l'extension sociale; ce qui est judicieux. Car, ce qu'on cherche, ce sont des matériaux qui permettent d'obtenir une image globale de l'idiome. Dans cette perspective, une variété — dialecte, parler, … — qui est pratiquée par une grande majorité de la communauté mériterait une place plus importante. De même qu'il serait normal de faire abstraction de l'usage qui ne représente qu'une infime fraction — disons 0,02% — de la communauté.

D'autres pistes aussi pourraient être envisagées qui ont leur valeur. Toutes les variétés d'un idiome n'étant pas également compréhensibles à travers la communauté, on pourrait poser comme critère le degré d'intercompréhension. Ce qui conduit à accorder une prépondérance à certaines variétés. On peut penser aussi à la dynamique et le sens de l'évolution qu'on peut mettre en évidence par l'observation des convergences et divergences résultant de l'évolution. Enfin, pourraient aussi intervenir les dispositions affectives des sujets parlants face à l'une ou l'autre des variétés dialectales. Bien que de nature subjective, elles devraient être prises en considération, car la subjectivité des sujets parlants fait partie de l'objet de la linguistique. Et cette liste n'est pas énumération exhaustive.

Les considérations qui précèdent ne définissent certes ni une procédure ni une méthode. Les facteurs qui entrent en jeu étant de tendances conflictuelles, il en résulte un équilibre difficile à établir. Cependant, si la fonction assignée au corpus est clairement définie, on a de fortes chances d'y parvenir. Ne serait-ce que empiriquement ; autrement dit par essai et erreur, pour traduire l'expression anglaise *trial and error.*

1- Zelig S. HARRIS, *Structural Linguistics*, The University of Chaicago Press, Chaicago, 1957 et Zelig S. HARRIS, Linguistique distributionnelle, in *Langages 20.*

2- André MARTINET, *Eléments de linguistique générale*, Paris, Colin, 1960.

3- Alfonic et Raphael.

4- ITA (Initial Teaching Alphabet).

5- *Cf.* ici même, L'acquisition phonologique de l'enfant Kabyle.