Cahiers de l'ILSL, n° 36, 2013, pp. 105–139

Corpus, enquête, système. Et la langue?

Réflexions sur l'objet de la linguistique.

Mortéza MAHMOUDIAN
Université de Lausanne

Où est l'objet langue dans les recherches qui mettent l'accent sur les conditions formelles? Pour illustrer mon interrogation je prendrai l'exemple du corpus, actuellement en vogue. L'idée de corpus remonte aux années 30-50 du siècle dernier. A cette époque, précurseurs et structuralistes cherchaient à définir pour la linguistique un objet concret, et à garantir ainsi son objectivité, donc sa scientificité. A la même époque, et par le même souci d'objectivité, est proposée une autre technique d'observation: enquête par questionnaire. Cet effort – méritoire – n'en soulève pas moins des questions sur les plans tant théorique que pratique. Ces concepts – corpus et enquête – s'inscrivent dans un cadre théorique où l'on conçoit les langues dotées d'une structure *sui generis*. Dans cette perspective, la structure de chaque langue a des spécificités que l'on ne peut circonscrire qu'à travers l'observation et l'étude de son usage. La présente étude vise à montrer la portée, mais aussi les limites du recours à ces techniques. On y soulève un paradoxe dans le rapport entre une structure censément finie et des données (recueillies par corpus et/ou enquête) extensibles *ad libitum*. Une issue possible de ce paradoxe serait d'opter pour

une conception relative et complexe de la structure et d'abandonner la structure finie, et les implications qui en découlent.[1]

Where is language as object *per se* when research focuses on formal conditions? To illustrate this question let us examine, for instance, corpus, which is currently very popular. The idea of corpus dates back to last century, to the thirties and fifties. At that time, precursors and structuralists sought to define a concrete object for linguistics, and thus to ensure its objectivity and hence its scientific status. At the same time, and for the sake of objectivity, another technique was proposed for observation: questionnaire surveys. This effort – however worthy it was – raises questions on both theoretical and practical levels. These concepts – corpora and surveys – proceed from a theoretical framework in which we conceive language as endowed with a sui generis structure. In this perspective, the structure of each language has specific features that can only be identified through the observation and study of its use. The present study aims to show the bearing but also the limits of such techniques. It raises a paradox in the relationship between a supposedly finite structure and data (collected through corpus and/or survey) expandable *ad libitum*. A possible outcome of this paradox would be to conceive structure as relative and complex, in itself, and give up finite structure, and its resulting implications.

[1] Mes vifs remerciements à Aris Xanthos et Alexei Prikhodkine pour leur lecture attentive d'une première version de cet article. Leurs pertinentes remarques et suggestions m'ont permis de préciser certains points et de combler certaines lacunes.

1. LE CORPUS

Le concept de corpus remonte aux années 30–50 du siècle dernier. A cette époque, les linguistes cherchaient à définir un objet concret pour leur discipline, et à en garantir ainsi l'objectivité donc la scientificité. Cet effort – méritoire, s'il en est – n'en soulève pas moins des questions sur les plans tant théorique que pratique. Dont 1° Le corpus peut-il être représentatif de la langue à l'étude? Si oui, totalement ou sous certaines réserves, dans des limites qui restent à déterminer? 2° Le recours au corpus garantit-il l'objectivité de la description? Si oui, absolument ou dans une certaine mesure?

Le concept de corpus s'inscrit dans un cadre théorique où l'on conçoit les langues dotées d'une structure *sui generis* (ou arbitraire, si l'on préfère). Dans cette perspective, la structure de chaque langue a des spécificités que l'on ne peut circonscrire qu'à travers l'observation et l'étude de son usage.

Avec l'arrivée de la vague générativiste, ce cadre de référence et les problèmes qu'il soulève ont été abandonnés – du moins pour un certain temps – par bon nombre de linguistes. Partant de la thèse d'innéité et d'universalité du langage, Chomsky et ses adeptes cherchaient les constantes biologiques du langage dans une structure profonde, censée être commune à toutes les langues. Les différences de structure entre les langues relèveraient, soutenaient-ils, de la structure de surface, et ne mériteraient guère l'attention des chercheurs.

Le retour triomphal du corpus semble dû à, au moins, deux causes. Avec la fin de la grammaire générative, est levée l'hypothèque sur les recherches qui prennent en compte diversités et variations linguistiques. A cette cause théorique s'en ajoute une autre, de nature technique: le développement spectaculaire

de l'outil informatique. Il est désormais possible de collecter des masses impressionnantes de données linguistiques, et de les dépouiller, compulser en vue d'une hypothèse déterminée ou d'un ensemble d'hypothèses.

Le regain d'intérêt pour le corpus s'inscrit dans un mouvement plus vaste: les recherches empiriques dans leur ensemble sont considérées avec plus de sérieux. Ce qui est – à mes yeux – un progrès. Cependant un examen sans complaisance de ces problèmes révèle – chez nombre de ceux qui proclament en théorie la nécessité du corpus ou y ont recours dans leur pratique – des questions laissées ouvertes, voire des paradoxes. Comme exemple de paradoxe, on citera l'existence, dans la même construction théorique, de la conception de la langue comme une structure finie composée d'éléments discrets d'une part, et de l'autre, du constat qu'il existe – dans une seule et même langue – des sous-systèmes variés dont le nombre et les caractéristiques ne sont pas déterminés (ni déterminables, ce me semble). Ce qui a pour conséquence que le linguiste ne dispose d'aucun moyen d'atteindre un système clos; il ne lui est donc pas possible de circonscrire précisément les unités constitutives d'une langue.

Ces problèmes ne sont pas de simples réminiscences historiques; on les retrouve dans les courants récents. Ils sont complexes en ce que les thèses énoncées permettent d'apporter réponse à certaines questions, mais rencontrent des difficultés dans d'autres domaines. La complexité vient de ce qu'on ne peut les accepter ni rejeter en bloc; et qu'il faut en déterminer la portée et les limites.

Dans ce qui suit, je partirai de la perspective structurale "classique". Les conclusions auxquelles on arrive valent cependant pour les recherches plus récentes, étant donné l'identité des hypothèses sous-jacentes.

2. IDÉE ET CONCEPT

Le corpus fait son apparition en linguistique quand on se pose le problème: *comment observer l'objet langue, ses manifestations, sa structure?* Bloomfield, suivant la psychologie behavioriste, soutient que l'intuition du sujet parlant n'est pas un moyen fiable, et que la donnée objective est le comportement du sujet. De là procède l'idée – et plus tard le terme – de corpus: la description linguistique doit prendre appui sur des données concrètes, réelles et non celles qu'imagine le linguiste, ni celles qui émanent de la subjectivité, tant du descripteur que de l'usager de la langue (cf. Bloomfield, 1933: §§ 2 et 10.1).

C'est dans les années 50 que le *corpus* reçoit une définition relativement précise, et est employé en tant que terme technique. Harris, par exemple, conçoit le corpus comme un ensemble "d'énoncés enregistrés" provenant d'un "seul dialecte" (Harris, 1951: § 2.33). Malgré cette définition restrictive, le concept contient des ambiguïtés et inadéquations qui ne sont pas spécifiques au structuralisme américain; et qui appellent examen et réflexion.

Ici, des constats de nature historique semblent bienvenus: 1° les données dont se servent Bloomfield ou Harris ne proviennent pas toutes de l'observation du comportement. Comme leurs prédécesseurs, ils utilisent essentiellement ce que l'on connaît – ou que l'on croit connaître – de telle ou telle langue, sans en vérifier l'objectivité ou l'origine[2]. Cependant, le concept a assez tôt montré son utilité, et des deux côtés de

[2] On trouve cependant dans leurs œuvres des études qui prennent effectivement appui sur des corpus. On peut citer comme exemple Harris (1955).

l'Atlantique, la majorité des linguistes y ont trouvé un outil de recherche intéressant; 2° jusqu'alors, la nature et la voie d'accès aux données linguistiques paraissaient évidentes aux précurseurs et aux premiers structuralistes. Ainsi, Saussure ou Troubetzkoy ne se posent pas de questions sur les formes et les normes des données ni sur les techniques de l'observation; 3° avec le développement de recherches structurales, apparaît la nécessité d'un contrôle empirique des données. D'où le recours, dès les années 40, à des enquêtes – comme l'enquête phonologique de Martinet (1945) – qui mettent en évidence les variations de la structure à l'intérieur d'une seule et même langue. D'où aussi la proposition de recourir au corpus; 4° On se trouve ainsi face à deux méthodes empiriques – corpus et enquête – qui sont proposées à partir de principes théoriques incompatibles: l'enquête est fondée sur le jugement intuitif du sujet parlant, alors que le recours au corpus – à l'origine – trouve sa justification dans la méfiance à l'égard de l'intuition du sujet parlant. J'y reviendrai, cf. *infra* § 7.3.

3. CORPUS EN STRUCTURALISME

Le corpus est un moyen d'observation des faits de langue; en tant que tel, il ne remplace pas la théorie. Il revient à la théorie de concevoir la structure linguistique, les unités qui en relèvent et les règles qui régissent le comportement de celles-ci. Au départ, la structure et ses éléments constitutifs sont de pures hypothèses qui peuvent être confortées, ébranlées, réfutées, … par leur confrontation avec des faits empiriques provenant du comportement et de l'intuition du sujet parlant. Le corpus est un moyen, une technique qui permet cette confrontation.

Cette position – clairement proclamée en linguistique structurale – assigne au corpus un rôle bien défini. Dans les travaux récents consacrés au corpus, cette position n'est pas toujours maintenue. Pour le moment, je me borne à dire que grâce au

corpus, se noue un lien étroit entre la théorie et ses applications.

Je prendrai un exemple classique: l'expérience du phonologue peut lui suggérer qu'en français, les phonèmes /e/ et /ɛ/ s'opposent en finale, et qu'à l'intérieur du mot, cette opposition est neutralisée. Autrement dit, les sons [e] et [ɛ] s'opposent en finale comme dans [pike] *piqué* et [pikɛ] *piquet*. Non à l'intérieur des mots, où ils sont en distribution complémentaire: [ɛ] en syllabe fermée comme dans [ɛm] *aime*, et [e] en syllabe ouverte: [eme] *aimer*. Hypothèse qu'on peut présenter par une formule ramassée:

> ***H*$_1$**. L'opposition /e~ɛ/ en français est neutralisée à l'intérieur du mot, [e] et [ɛ] étant en distribution complémentaire.

Cette hypothèse, si vraisemblable soit-elle, doit recevoir la confirmation de l'empirie. Plus précisément, l'hypothèse est confirmée par *expérimentation*, entendue comme l'observation provoquée de faits linguistiques, et reproductible dans des conditions identiques. Le corpus permet une telle observation: nous pouvons enregistrer la prononciation des francophones et vérifier si la distribution de [e] et [ɛ] est conforme à l'hypothèse. Nous pouvons aussi répéter cette observation plus d'une fois sur des populations différentes, mais répondant aux mêmes conditions: origines géographiques, classes sociales, tranches d'âge, …

Le recours au corpus peut confirmer ou infirmer l'hypothèse, certes. Mais en même temps il laisse ouverte une question: *L'hypothèse* H_1 *est-elle valable pour la langue française? Ou pour une variété du français?* Si elle vaut pour le français, alors la neutralisation s'appliquera au français de Nice comme à celui de Lille. On ne risque pas gros à parier que – appliquée à ces usages – la même méthode donnera des résultats différents de celui obtenu pour l'usage du français parisien.

Le problème est de taille: il touche à la conception même de la langue, de sa structure, de ses variétés et de leurs interrelations. Si la langue est une structure fermée, comme on l'affirme en linguistique structurale, elle est identique chez tous les locuteurs: elle comporte donc les mêmes unités, qui sont soumises aux mêmes règles. On s'attend dès lors à des résultats identiques, que les données soient collectées auprès d'un (groupe de) locuteur(s) ou d'un autre. Il n'y aurait aucune raison de prendre des précautions dans la collecte des matériaux quant à la situation sociogéographique des locuteurs. Encore moins justifiée serait l'exigence que les données proviennent d'un même individu parlant, dans le même style, etc.

4. CORPUS: SA RAISON D'ÊTRE

L'examen des résultats du corpus, dans le cadre de la linguistique structurale, conduit à d'autres interrogations, dont je relèverai quelques-unes. Au préalable, je rappelle que la phonologie est fondée sur le principe que les unités phoniques d'une langue ne peuvent être identifiées par l'étude de leurs attributs physiques seuls; et qu'elles dépendent du facteur humain. Ce qui fait de /s/ et /z/ deux phonèmes distincts, ce ne sont pas leurs caractéristiques physiques seules, mais ces caractéristiques physiques quand elles sont utilisées et reconnues comme différentes dans une langue. En d'autres mots, les unités sont identifiées par leurs fonctions.

Comment reconnaître la fonction (donc la pertinence) d'un fait phonique? La réponse à cette question a changé au fur et à mesure de l'évolution de la phonologie. Dans un premier temps, la pertinence phonologique est conçue comme une évidence: dans la mesure où, moi descripteur, je connais la langue que je décris, je sais que la différence entre sourde et sonore est pertinente dans le cas des sifflantes (*cinq* vs *zinc*), et qu'elle ne l'est pas dans le cas des latérales (*oncle* vs *ongle*):

d'où deux phonèmes /s/ et /z/, mais un seul phonème /l/ en français.

Avec le développement des études phonologiques, on s'est rendu compte que la structure phonologique d'une langue comprend des variétés, et que certaines d'entre elles peuvent échapper au descripteur; de là, la quête de l'objectivité dans la collecte des données. Par objectivité, il faut entendre l'indépendance par rapport à la subjectivité individuelle, en l'occurrence, par rapport à celle du descripteur. Pour atteindre ce but, deux voies ont été proposées, à quelques années d'écart: corpus en linguistique américaine et enquête dans la lignée praguoise.

Le problème est qu'en structuralisme américain, on présente le corpus comme seul moyen d'observation théoriquement justifié. Or, on conçoit en même temps la langue comme instrument de communication. C'est dans cette double assertion que se trouve un paradoxe. Car, dégager la structure à travers l'usage implique qu'on l'observe dans l'utilisation qu'en font le locuteur et son interlocuteur. Le recours exclusif au corpus présente une grande faille en ce que le processus de la communication est mutilé de sa moitié: on ne cherche pas à savoir si le récepteur perçoit deux éléments physiques – disons [s] et [z] – comme deux unités linguistiques différentes /s/ et /z/. Ce, par mesure d'hygiène méthodologique qui interdit tout recours à l'intuition[3].

Ainsi conçue, la mesure d'hygiène ou de rigueur méthodologique semble procéder de la confusion entre l'intuition

[3] Il est matériellement impossible de substituer à l'intuition le comportement; et à ma connaissance, personne ne s'est avisé à se lancer dans une recherche linguistique d'envergure tout en s'interdisant tout recours à l'intuition.

de l'usager de la langue et celle du linguiste descripteur. Il est certain que la linguistique, si elle se veut science, doit éviter toute immixtion de la subjectivité du chercheur dans la conduite et les résultats de la recherche. Mais l'intuition de l'usager est partie intégrante de l'objet de la linguistique. C'est cette subjectivité qui fait la spécificité des sciences humaines, et les distingue des sciences de la nature. On peut dire avec Jean Piaget que l'objectivité n'est rien d'autre que la subjectivité collective. L'intégration de la subjectivité du sujet parlant est aussi indispensable que source de complexité. Il est certes plus simple d'en faire abstraction. Mais la simplicité justifie-t-elle tout choix théorique? Je ne crois pas. Il semble plus judicieux d'élaborer un appareil théorique complexe quand la complexité de l'objet l'exige.

5. PROBLÈMES ET LIMITES

Un examen critique du concept et de l'usage du corpus montre une ambiguïté depuis l'origine.

D'une part, le recours au corpus est proposé comme un moyen pour conserver l'indépendance de la description par rapport à l'intuition du descripteur; pour éviter que le descripteur ne soit tenté de présenter ses propres habitudes pour une vérité générale, quand il se trouve en face de cas litigieux, là où les faits de langue n'ont pas un statut clair; par exemple, quand on ne sait si telle opposition phonologique et ou telle construction syntaxique est possible. C'est ce qu'on peut appeler l'acception faible (*a/*).

D'autre part, on s'attend que le corpus soit représentatif de la langue. C'est l'acception forte (*b/*).

On admettra volontiers que les faits invoqués, dans la mesure où ils sont attestés dans un corpus, ne sont pas l'invention du descripteur. Ce qui revient à reconnaître le bien fondé de *a/*. (Encore qu'il soit excessif de restreindre le rôle du corpus à l'attestation de faits.).

La conception maximaliste (*b/*) pose un problème: le corpus peut-il représenter la langue? C'est une question de taille, car elle touche à la conception même du système – ou de la structure, selon la terminologie[4] – linguistique. Par représentation, on peut entendre que tout ce que comporte une structure linguistique – unités, règles, … – est attesté dans le corpus. Et corrélativement, il n'y a rien dans le corpus qui ne relève pas de la langue. Cela suppose qu'on puisse déterminer ce qu'est ce *tout*. Dans cette acception, la représentativité d'un quelconque ensemble de données n'est concevable que si l'on peut énumérer exhaustivement les éléments constitutifs – unités et règles – de la langue à l'étude. Autrement dit, la question de représentativité du corpus suppose que réponse soit donnée à une autre question: la langue consiste-t-elle en un système fermé?

La question a été posée depuis longtemps, mais assez souvent[5], les réponses apportées manquent de clarté; voire renferment des paradoxes. Les pères fondateurs affirment d'une part le caractère clos du système de la langue, et admettent en même temps que toute langue comporte des variétés dont les spécificités peuvent entraver l'intercompréhension. Ainsi Bloomfield, qui affirme que les unités d'une langue sont soit identiques soit différentes sans moyen palier

[4] J'emploie les termes *structure* et *système* comme synonymes, désignant un ensemble constitué d'éléments interdépendants que sont *unités* et *règles*.

[5] Mais pas toujours. De nombreuses études récentes abordent les problèmes posés par le recours au corpus. Benoît Habert (2000) relève de multiples questions soulevées sous cet aspect. Anne Condamine (2005) montre clairement que le corpus limité ne peut être représentatif *stricto sensu* du système linguistique illimité.

(Bloomfield, 1926: § 2). Il constate en même temps qu'on n'a pas affaire aux mêmes phonèmes en anglais selon qu'il est parlé à Chicago ou à Londres (cf. *infra Appendice*).

L'ambiguïté n'est pas levée chez les premiers structuralistes qui recourent au corpus. Harris, par exemple, part du postulat que les unités d'une langue sont discrètes (cf. Harris, 1951: § 3.1). Il se pose en même temps des questions sur la représentativité du corpus (cf. Harris, 1951: §§ 2.33, 20.3 et 20.4). Entre autres, il prône que les matériaux constitutifs du corpus soient recueillis auprès d'un même locuteur et dans les mêmes conditions, afin d'éviter que des changements de style n'en altèrent l'homogénéité. De deux choses l'une: ou bien la langue – l'anglais, par exemple – est une; alors le respect de l'unité de style est une précaution superflue. Ou bien la langue comporte des variétés, irréductibles les unes aux autres, alors le corpus ne peut représenter que lui-même, ou au mieux, une variété donnée, un style déterminé. Ce qui reviendrait à ôter au corpus toute valeur de représentativité au sens qu'on vient de voir.

Le paradoxe n'est pas spécifique au structuralisme américain. On le retrouve aussi dans les courants européens de la linguistique de la deuxième moitié du XX^{e} siècle. Ainsi André Martinet affirme que les phonèmes d'une langue sont en nombre déterminé (cf. Martinet, 1960: § 1-14), et constate en même temps qu'il n'est pas possible d'énumérer les phonèmes d'une langue (Martinet, 1960: § 1-13).

Les propos critiques que je viens de tenir n'invalident pas le recours au corpus. Ils ne montrent qu'une chose: il y a une incompatibilité entre le corpus ainsi défini et la structure linguistique telle qu'elle est conçue. Qu'en conclure? Faut-il abandonner ou modifier l'un ou l'autre? Voire l'un et l'autre?

6. CORPUS: VALIDITÉ RELATIVE

Empiriquement, on a des raisons de prêter foi à l'efficacité du corpus. Des expériences montrent que le gros d'un système phonologique peut être obtenu par l'examen d'un corpus de taille relativement modeste; même au niveau syntaxique, un corpus – un peu plus conséquent, mais restreint encore – conduit à des résultats non négligeables.

Non seulement le corpus permet de montrer que les éléments dégagés ne sont pas pure invention du descripteur; les résultats obtenus par l'analyse d'un corpus se retrouvent – du moins pour l'essentiel – dans d'autres corpus, même s'ils n'ont pas tous été recueillis dans des conditions comparables. C'est ce que montre une étude menée par Remi Jolivet (1982) sur un échantillon – de 262 phrases ou 2'600 mots, extrait d'un corpus comprenant 17'000 mots – recueilli auprès d'écoliers de 12 ans. De la comparaison des résultats de son analyse avec ceux d'autres études quantitatives, l'auteur conclut "qu'il y [a] bien “quelque chose” qu'on [peut] appeler *le* français et que manifestent toutes les productions possibles dans cette langue" (p. 458). C'est dire que la validité du corpus dépasse le cadre strict de la conception minimaliste (*a/*). Mais elle se heurte à des limites qu'il serait intéressant de caractériser. On aurait ainsi des chances de mieux définir les deux concepts: le système et le corpus.

On peut étayer cette affirmation sur des données chiffrées, mais cela se conçoit aisément: si j'établis l'inventaire des phonèmes en partant de l'enregistrement d'un Vaudois, certains des résultats valent pour d'autres francophones: les oppositions /b~d/, /s~z/, /m~n/, /k~g/ à l'initiale du mot, sont très vraisemblablement attestées dans d'autres régions de Suisse Romande ou de France métropolitaine. En revanche, quand on se penche sur l'opposition de longueur (comme /i~i:/ dans *poli* et *polie*) ou d'aperture (/o~ɔ/, dans *maux* et *mot*), on constate

que les résultats varient selon les régions. Ceci illustre bien la complexité dont je parlais: la valeur d'un corpus n'est ni totale (ou absolue) ni nulle.

7. TECHNIQUES D'OBSERVATION, LEURS VICES ET VERTUS

Reprenons l'enquête par questionnaire, technique d'observation qui est – comme je l'ai dit – quasi contemporaine du corpus. Quels en sont les avantages et les inconvénients relativement au corpus? Trois aspects peuvent être retenus pour la comparaison: exhaustivité, authenticité et pertinence.

7.1. EXHAUSTIVITÉ

Du point de l'exhaustivité, les possibilités offertes par le corpus risquent d'être limitées: le chercheur n'est pas assuré d'y trouver – quand les questions posées sont très spécifiques – tous les éléments dont il se propose de traiter. Sous cet aspect, l'enquête présente un double avantage: elle peut couvrir un champ étendu de phénomènes, d'une part, et de l'autre, le chercheur peut orienter la collecte vers les phénomènes de son choix, en fonction des problèmes qu'il se pose. Soit la morphologie du verbe *asseoir*. Même dans un corpus de taille considérable, j'ai peu de chances de trouver les formes /asej/ (dans *je m'asseye*), /asje/ (*je m'assied*) et /asua/ (*je m'assois*). Dans une enquête, je peux concevoir à ma guise les questions posées aux informateurs.

7.2. PERTINENCE

Un autre problème que soulève le corpus est celui des éléments physiquement présents dans l'usage et leur fonction. Le chercheur qui s'intéresse au statut phonologique de [ɲ] et [nj], peut trouver de nombreuses occurrences de ces deux éléments. Mais le corpus ne fournit aucune indication quant à la pertinence de cette différence: sont-ce deux variantes phono-

logiques? Dans ce cas, *peignez* et *peiniez* se confondent. Sont-ce deux séquences phonologiques distinctes /ɲ/ et /nj/? Dans ce cas, peignez et peiniez restent distinctes. Le recours à l'enquête permet de déterminer le statut fonctionnel de ces éléments.

Le problème a son importance, puisqu'il touche au fondement de la phonologie.

7.3. AUTHENTICITÉ

Le reproche le plus souvent adressé à l'enquête est qu'elle risque d'influencer le comportement du sujet parlant; et de faire porter l'analyse linguistique non sur des données authentiques, mais sur des données déformées par les conditions d'observation. Le risque n'est pas à écarter d'un revers de main. Mais justifie-t-il l'abandon de tout recours à l'intuition du sujet parlant? Considérons les arguments évoqués pour cet abandon. Le sujet parlant n'a pas, dit-on, connaissance de sa langue. Les faits d'expérience prouvent le contraire. En témoignent les réponses convergentes des locuteurs face à des questions comme: "Prononcez-vous de façon identique *bain* et *pain, ton* et *don, camp* et *gant*?" Maintes enquêtes l'ont révélé – et il est facile d'en refaire d'autres. Pareils résultats montrent une chose: les sujets ont conscience[6] de l'identité ou différences des unités de leur langue, du moins dans certaines parties de la structure linguistique.

En linguistique structurale (au sens large du terme), la thèse "sujet parlant, ignorant de la structure linguistique" est issue

[6] On peut avoir conscience d'une différence, sans être en mesure de l'exprimer par des formules linguistiques.

du dogme behavioriste plutôt que de l'observation scientifique des faits. Les tenants de cette thèse rejoignent la posture de la grammaire traditionnelle selon laquelle la langue serait une et invariable; il relèverait de la juridiction du grammairien de délimiter la langue, donc de décider ce qui est correct – donc dans la langue – et ce qui est une faute – donc hors de la langue.

C'est un parti pris qui a une étonnante longévité: depuis Bloomfield et Chomsky jusqu'aujourd'hui. Présupposé d'autant moins défendable qu'il mène à l'impasse: dans ses postulats, Bloomfield affirme que dans une communauté certains sons sont identiques. Par cette proposition, on entend que l'identité linguistique des sons ne réside pas dans sa constitution physique seule, et qu'elle varie d'une langue à l'autre. Comment connaître l'identité des sons d'une langue? La réponse qu'on y apporte dans les propos théoriques est: par l'observation des réactions comportementales des sujets parlants. Certes, un francophone ne réagirait pas de la même façon à *prenez la rampe* et *prenez la lampe*. Il est aussi évident que l'apprentissage du langage chez l'enfant passe par l'observation du comportement de son entourage. Mais de là à faire du recours au comportement le principe unique d'identification des entités linguistiques, il y a un pas qui est allégrement franchi. Car, on ne voit pas quels indices comportementaux permettraient de saisir le sens de *casoar*, *électron* ou *mantisse*. Sans reprendre les longues discussions sur la validité d'une telle position théorique, et restant sur le plan pratique, on peut constater que Bloomfield lui-même n'a pas appliqué ce principe dans ses études sur le tagalog ou le menomini. L'impasse apparaît à l'évidence: on ne peut avoir accès à la connaissance de l'identité des éléments linguistiques par recours exclusif aux réactions comportementales. Le recours au jugement intuitif du sujet reste une voie incontournable.

En structuralisme classique, deux solutions sont adoptées dans la pratique descriptive. Pour les langues "exotiques", les

linguistes se fient à l'intuition du sujet parlant. Mais quand il s'agit des langues connues et décrites depuis longtemps, ils s'arrogent le droit de juger les sons que le sujet identifie et ceux qu'il distingue. La conséquence de cette attitude est de supplanter l'intuition du sujet parlant par celle du linguiste. A mes yeux, ceux qui optent pour une "linguistique pure" (cf. Lazard, 2009) suivent le même chemin. Le linguiste qui se refuse à tenir compte du statut des phénomènes linguistiques dans la collectivité et dans le psychisme des sujets, se réserve la compétence, le droit de décider ce que sont ces phénomènes.

C'est là le paradoxe: la quête de l'objectivité conduit à faire sortir par la porte l'intuition du sujet parlant; et faire entrer par la fenêtre l'intuition du descripteur. Est-ce là l'idéal de l'objectivité?

8. TECHNIQUES D'OBSERVATION, LEUR COMPLÉMENTARITÉ

La validité de l'enquête a aussi ses limites. Prenons des exemples sémantiques. Pour les mêmes éléments, les réactions font montre d'une grande variabilité, quand il s'agit de certains sens. Ainsi du sens "partie de l'arme à feu" ou du sens "talon de carte dans le jeu de tarot" pour *chien*. Ce, sans modification aucune des conditions d'enquête, le questionnaire et les informateurs étant les mêmes. C'est dire que les sujets ont une connaissance partagée de certains sens des unités significatives de leur langue, mais non de tous les sens de ces mêmes unités.

Bloomfield a sans doute raison d'affirmer qu'un anglophone – sans entraînement spécial – ne peut attribuer aucun sens précis à *cran-* dans *cranberry* "airelle" (Bloomfield, 1932: § 10.1); pas plus que le francophone moyen à *aub-* dans *aubépine*. Mais cela ne permet pas de conclure que le sujet n'a connaissance du sens d'aucune unité linguistique; encore

moins d'en arguer à l'inanité du recours à l'intuition du sujet. En toute conséquence, on doit s'interroger sur les conditions dans lesquelles l'intuition du sujet parlant peut être valablement observée. Font partie de ces conditions les caractéristiques des éléments soumis à l'examen, dont sa classe, son intégration dans le système, sa fréquence dans l'usage, …

Parmi les conditions d'observation, on peut penser à la manière de solliciter le jugement de l'informateur. Les oppositions phonologiques que fournit l'informateur varient selon qu'on lui demande de lire un texte ou de juger de l'identité ou de la différence de paires minimales (Labov, 1976: § 4)

De même, pour la signification, les réponses varient suivant la façon dont la question est formulée. Quand on demande à l'informateur d'énumérer les sens du mot *chien*, il y a de fortes chances que le sens "animal" soit donné. Mais pas nécessairement le sens "partie de l'arme à feu"; sens qui aurait plus de chance d'apparaître si on lui demande: "Le mot *chien* peut-il signifier partie de l'arme à feu?"

A ce point d'exposé, quelques remarques seraient bienvenues:

1° Diverses formulations de questions constituent différentes techniques d'observation, qui ne donnent pas toutes accès aux mêmes strates de la structure linguistique.

2° Demander à l'informateur d'énumérer les sens d'un mot fait apparaître les sens les plus immédiatement accessibles.

3° Pour les zones plus reculées de la structure sémantique, on aura plus de chance de recueillir des jugements intuitifs si l'on utilise des questions directes comme: "Le mot *m* peut-il avoir le sens *s*?"

4° La disparité des résultats d'enquête a sa contrepartie dans les résultats du corpus: les zones les plus reculées de la structure sémantique sont difficilement accessibles par un corpus de taille restreinte. Il faudrait sans doute un corpus de taille considérable pour trouver des occurrences du sens "talon de cartes dans le jeu de tarot" pour le mot *chien*.

J'en conclurai que le corpus et l'enquête sont des techniques complexes et complémentaires. Ils aboutissent à des résultats variables dont chacune donne une image partielle de l'objet. Il n'y a donc pas une technique idéale qui donne l'image fidèle et globale de la langue. Il incombe à la réflexion théorique de procéder à une synthèse permettant d'approcher davantage le réel de la langue. (Cf. *infra* § 11).

9. REPRÉSENTATIVITÉ, OBJECTIVITÉ, LEURS LIMITES

Le souci de cohérence exigerait qu'on recherche les conditions de la validité – ou de l'efficacité – du corpus, et les limites qu'elle rencontre. Cette quête conduit à relativiser la valeur des techniques d'observation et de collecte des données: celle du corpus autant que celle de l'enquête.

La valeur de ces techniques semble dépendre de deux types de facteurs:

i/ Facteurs internes: un système linguistique comporte de multiples strates dont toutes ne sont pas accessibles avec une égale facilité. Quelle que soit la technique utilisée, on aura plus de difficulté à faire sortir, par exemple, des constructions syntaxiques désuètes, que les constructions courantes.

ii/ Facteurs externes: toutes les catégories des locuteurs n'ont pas le même usage ni la même intuition linguistiques. On aura vraisemblablement plus de chances de trouver pour le mot *chien* le sens "partie de l'arme à feu" si notre informateur est armurier de son état.

On peut revenir maintenant aux questions posées ci-dessus (cf. *supra* § 1) sur représentativité et objectivité.

Qu'entend-on par représentativité? Rappelons-en l'acception forte: on qualifie de représentatifs les matériaux où l'on trouve tout ce qui est dans la langue. Ce qui implique que rien ne se

trouve dans la langue qu'on ne trouve déjà dans les matériaux collectés.

Rien ne permet d'affirmer qu'un quelconque ensemble de matériaux linguistiques puisse satisfaire à cette acception de représentativité (cf. *b/*, § 5), Cependant les données réunies sont sous-estimées suivant l'acception faible (cf. *a/*, § 5), qui tend à en limiter la valeur à ceci: ces faits linguistiques ne sont pas pures inventions du descripteur.

Quant à la question d'objectivité, nous avons remarqué que les techniques d'observation influencent, chacune à sa façon, les résultats. Le nombre des techniques étant passablement élevé, il n'est matériellement pas possible de les appliquer toutes dans une recherche. A supposer que – en mobilisant d'énormes moyens – on en applique un bon nombre, on se trouvera confronté à un nouveau problème: comment faire la synthèse de toutes les données réunies? Il paraît évident que toutes les données ne peuvent pas être comptabilisées par l'adjonction pure et simple des unes aux autres. Grand est l'écart entre la bonne vieille méthode troubetzkoyenne – épreuve de la commutation – qui fait ressortir ce dont le sujet parlant a une connaissance immédiate, d'une part, et d'autre part, la technique du locuteur masqué (angl. *matched guise*) de Labov qui ouvre l'accès à ce qui gît dans zones les plus éloignées de la mémoire du locuteur (cf. Labov, 1976).

L'objectivité ne peut donc être totale, dans la mesure où le descripteur choisit l'une et/ou l'autre des techniques d'observation.

10. APPROXIMATION

Revenons au paradoxe: "Comment concilier la large gamme des potentialités descriptives et le postulat de système clos?" En effet, les possibilités offertes par les techniques de description sont pratiquement illimitées. On peut étudier les habitudes linguistiques d'une vaste communauté, d'un groupe restreint,

d'un individu, voire de l'individu dans une circonstance donnée. Et à chaque niveau, on trouve de nouveaux éléments, des variations jusqu'alors inconnues. Autant dire que l'analyse n'a pas de limite. Et pourtant l'hypothèse "langue–structure" laisse attendre une structure contenue dans des limites précises.

Dans la pratique, le descripteur arrête d'affiner la technique d'observation quand celle-ci donne des résultats nouveaux qu'il estime négligeables, insignifiants. Il y a là une approximation qui – si justifiée soit-elle – n'en mérite pas moins interrogation et réflexion. Qu'est-ce qui justifie pareille approximation?

Noter que l'approximation n'est pas en soi un vice dans une démarche scientifique. En mathématique, on y a recours, mais en adoptant une mesure précise. Dans des opérations arithmétiques, par exemple, on peut décider de ne retenir que deux décimales. Mesure qu'on appliquera systématiquement.

Or, en "structuralisme classique" cette mesure n'est pas proposée. Pareille mesure n'est d'ailleurs pas compatible avec une théorie linguistique où l'on conçoit la structure de la langue comme formelle, c'est-à-dire constituée d'un nombre finie d'éléments. Les tenants de ces positions font comme s'il n'y avait là aucun problème; et que leur bon sens suffisait pour décider où arrêter l'analyse. Qu'advient alors de l'objectivité tant recherchée si l'on se contente des à peu près?

Face à ce problème, deux solutions conséquentes s'offrent au linguiste. Soit maintenir le cadre formel et exclure tout recours à la mesure et l'approximation; c'est la solution adoptée par la grammaire générative, entre autres. Soit reconnaître la pertinence de la mesure et de l'approximation; ce qui implique qu'on remette en cause le cadre formel, et qu'on le remplace par une structure où mesure et nombre trouve leur place.

Il convient de remarquer que les linguistes dans leur majorité se sont accommodés, des décennies durant, de cette

inconséquence. Et le corpus et l'enquête ont continué à rendre de bons et loyaux services. Malgré ces à peu près, les données obtenues apportent réponse à des questions, et suggèrent des solutions à certains problèmes. Comment l'expliquer? Quelles en sont les raisons?

L'une des raisons est sans doute le fait que, passé certaines limites, les données – qu'elles proviennent du corpus ou de l'enquête – ne paraissent pas convaincantes. Prenons un exemple syntaxique: le syntagme nominal comportant le déterminant "partitif" comme *du vin* qu'on rencontre fréquemment dans certaines fonctions. Ainsi *il n'achète que du beau* ou *y a de la flicaille partout dans la ville*. Ce type de syntagme pose des problèmes en fonction sujet. Des constructions comme *du beau se vend comme du petit pain* ou *du vin ne me convient pas* ou encore *de la flicaille grouille sur la place* sont rares dans le corpus. Et quand on l'interroge, le sujet parlant est plutôt incertain dans ses jugements. Martinet propose de qualifier ces parties comme des *marges* de la structure linguistique.

Penchons-nous un instant sur les particularités des marges. On constate d'abord que les habitudes linguistiques ne se forment pas en vase clos, mais dans l'échange et l'interaction entre sujets. Dans les marges, ces échanges peuvent mettre en contact des variantes, offrir à l'individu parlant une gamme de choix. Comment fait-il son choix? Vraisemblablement, en raison des facteurs extérieurs à la langue, en fonction des circonstances dans lesquelles il a pris connaissance d'une forme. Il rejetterait la forme /il s asua/ *il s'assoit* parce qu'elle lui rappelle un réceptionniste qui ne lui a pas été sympathique. Il aurait des réticences à l'égard de /il s asej/ *il s'asseye* car il a trouvé vulgaire la personne qui l'a proféré. Ainsi de suite. Malgré le choix d'une forme – /il s asie/ *il s'assied* par exemple – il n'exclut pas les autres formes de son usage dit "passif". Pour assurer l'intercompréhension, il est amené à s'accommoder de ces variations et différences. Les sujets sont peu ou

prou conscients de ces zones d'incertitudes, d'insécurité linguistique. A plus forte raison, l'observateur qu'est le descripteur en est conscient aussi. S'il estime certains faits comme quantité négligeable, c'est, entre autres, en se référant aux incertitudes dans sa propre intuition.

Un constat s'impose: reconnaître l'existence de marges dans un système linguistique, c'est reconnaître qu'il y des phénomènes qui ne sont pas tout à fait dans la langue ni tout à fait hors de la langue. C'est déroger au principe de "langue, système fermé". A première vue, il y a là une contradiction dans les termes.

Ces prises de positions – paradoxales s'il en est – sont-elles vraiment indéfendables? Cela n'est pas si sûr, car, certains paradoxes proviennent du fait que la proposition théorique est trop simple comparée à la grande complexité de l'objet. Mettre en évidence le décalage entre théorie et empirie pourrait être le premier pas vers la recherche d'une solution plus adéquate.

11. HIÉRARCHIE ET STRATES

L'étape suivante consiste en la quête d'une théorie plus adéquate. Admettre l'existence de marges dans une structure, c'est reconnaître que ses éléments constitutifs ne se valent pas; qu'ils peuvent être hiérarchisés selon la place qu'ils tiennent dans le système ou – ce qui revient au même – le rôle qu'ils jouent dans la communication. Mais alors pourquoi se contenter de la distinction centre/marges? Rien *a priori* ne justifie une dichotomie en la matière. Il semble plus adéquat d'introduire une hiérarchie à échelles multiples qui permette de décrire finement les strates de la structure, le degré de finesse pouvant être choisi en fonction du but poursuivi.

Dans plusieurs directions de recherche, on trouve cette conception de la structure, appliquée à des domaines restreints

dans une langue ou une autre. Il en existe plusieurs modèles. J'en citerai deux.

Le premier est un ensemble d'études[7] réalisées par une équipe de chercheurs de l'Université de Lausanne dans laquelle Remi Jolivet a pris une part très active. Ces recherches partent de l'idée que la structure linguistique est formée de strates multiples et hiérarchisées; et que cette hiérarchie peut être établie par l'observation des dimensions sociale et psychique. Plus précisément, les strates les plus haut placées sont connues par de larges fractions de la communauté parlante par opposition aux strates du bas de l'échelle qui ne sont connues que de groupes limités de la communauté. Sur le plan psychique, les strates au sommet de la hiérarchie correspondent à un haut degré de certitude, alors que les strates situées au bas de l'échelle suscitent de fortes hésitations. Des recherches empiriques vérifient dans une large mesure cette hiérarchie. Sans entrer dans les détails, j'illustrerai ce propos par des exemples lexicaux. *Rose*, *maison*, *murger* et *rudéral* sont des mots de la langue française, si je me fie au dictionnaire, *Le petit Larousse*, par exemple. Si la langue a son siège dans "l'âme collective" pour employer la formule de Saussure, il s'ensuit que tous les francophones les connaissent, et qu'ils sont tous sûrs de la signification de ces mots. Il y a des raisons d'en douter. On peut soumettre ce doute au verdict de l'enquête; mais j'en réfère au jugement intuitif du lecteur. Pareilles observations montrent qu'une autre conception de la structure de la langue est possible: structure conçue comme plus complexe et à la fois plus adéquate à l'objet. Plus complexe, en ce qu'une langue n'est pas un système homogène, mais bien un système

[7] Pour une vue d'ensemble, on peut consulter *La Linguistique,16(1)* (1980), où sont publiées cinq contributions (pp. 5–117) consacrées à ce thème.

représenté par de multiples variantes, ou – pour employer la formule consacrée – un système de systèmes. Plus adéquate, parce que les sous-systèmes peuvent faire l'objet de contrôle empirique si l'on les rapporte à leurs corrélats psychiques et sociaux. En l'occurrence, on peut vérifier la signification de ces mots si l'on prend en compte qui les emploie et dans quelles conditions.

Pour le second modèle, je prendrai celui de John Goldsmith (2001) consacré à l'apprentissage non supervisé de la morphologie (entendue comme la combinaison des morphèmes dans le cadre du mot). Bien que son but soit de nature pratique (l'apprentissage des langues), le projet est fondé sur des principes théoriques qui sont en rapport avec l'objet de notre discussion.

L'objectif est de construire des procédures heuristiques permettant de segmenter les mots en morphèmes. La procédure adoptée consiste à traiter par des programmes informatiques un grand corpus – de 5'000 à 500'000 mots. L'expérience – portant sur plusieurs langues européennes – aboutit à des résultats satisfaisants dans une large mesure. Différents degrés de finesse peuvent être atteints par l'analyse. Les résultats montrent en outre que la taille du corpus nécessaire varie selon les échelles: plus l'analyse recherchée est fine et plus la taille du corpus nécessaire est grande. Et inversement, plus l'approximation descriptive est grossière, plus la taille du corpus adéquat est restreinte. Mais comment juger de la validité des règles morphologiques ainsi obtenues? Pour ce faire, on compare celles-ci avec les règles issues de l'analyse faite par un morphologue humain (*human morphologist*).

La confrontation de ces deux directions de recherches est riche en enseignements. Je me bornerai à trois constats: *i)* le corpus et l'enquête apparaissent comme deux techniques d'observation complémentaires (cf. *supra* § 8). L'analyse du corpus

peut certes être effectuée par des instruments (physiques comme ordinateurs ou théoriques comme logiciels). Mais elle ne peut se passer du jugement intuitif du sujet parlant. Et c'est à raison que Goldsmith a recours à l'analyse faite par des humains; *ii)* une enquête approfondie – étant donné les ressources humaines qu'elle exige – est nécessairement limitée à une petite portion de l'espace social. Elle ne peut donc mettre en évidence l'extension sociale des phénomènes étudiées; ce qui est accessible par l'analyse d'un grand corpus; *iii)* dans un premier temps, le recours à l'intuition est limité à celle du ou des linguistes. Il est fort probable que les problèmes rencontrés conduisent à ouvrir la perspective, et à faire entrer en ligne de compte l'intuition des usagers non linguistes. A l'instar de ce qui s'est passé dans l'évolution de la phonologie. Ce sera alors un nouvel aspect de la complémentarité corpus/enquête.

12. A LA RECHERCHE DE L'OBJET PERDU

On vient de voir que la linguistique dispose désormais de techniques efficaces pour l'analyse et la description fines. Grâce à elles, on est en mesure d'observer dans le détail, le comportement d'un phénomène microscopique dans des circonstances données; d'en évaluer aussi les éventuelles variations et de rapporter ces variations aux facteurs qui les déterminent: groupes sociaux, classes d'âge, conditions d'observation, contexte linguistique, origine géographique…. Un peu comme Higgins dans *My fair lady*.

Au cours de leur évolution, les techniques descriptives ont dû emprunter des chemins tortueux. Dans ce cheminement, la linguistique aurait-elle perdu sa destination? Le but de la linguistique reste-t-il la langue et sa structure? La fragmentation des études sur le langage ne semble pas y mener; elle semble même y dresser des barrières.

Pareil émiettement condamne chacune des disciplines – ou "sciences du langage" – à l'isolement, en coupe les liens

avec la théorie du langage et à la fois avec les disciplines voisines.

Il y a un siècle, Ferdinand de Saussure assignait à la linguistique la tâche de circonscrire la langue, objet à la fois concret et intégral de la linguistique. Il situait l'objet langue dans l'âme collective. Cette proposition n'a rien perdu de son actualité dans la linguistique saussurienne, du moins dans l'une de ses interprétations. En ce sens que, le système qu'est la langue, et qui sert à la communication, doit nécessairement avoir son siège dans "l'esprit" du sujet parlant, correspondre à une réalité psychique, d'une part et de l'autre, cette réalité psychique doit avoir une extension sociale, être partagée par les membres de la communauté linguistique.

Ces corrélats psychosociaux peuvent-ils faire l'objet d'une étude scientifique? Non, répondaient les behavioristes. Et de proposer le corpus comme un dispositif permettant d'éviter les dangers mentalistes. A l'origine, le corpus est un alternatif choisi et fondé sur la conviction que l'étude de la réalité psychique du langage est impossible. C'est une prise de positions métaphysique, dont l'opposé serait aussi recevable que discutable.

Les positions prises par Emile Durkheim dans *Les règles de la méthode sociologique* (Durkheim, 1894/1967) sont à cet égard très intéressantes:

> Car les phénomènes sociaux ne se distinguent des précédents [phénomènes physico-chimiques et biologiques] que par une complexité plus grande. Cette différence peut bien impliquer que l'emploi du raisonnement expérimental en sociologie offre plus de difficultés encore que dans les autres sciences; mais on ne voit pas pourquoi il y serait radicalement impossible.

Ces positions sont diamétralement opposées à celles de Bloomfield.

Les propos de Durkheim par leur teneur paraissent applicables à toutes les sciences de l'homme, y compris la linguistique. Les enquêtes constituent une manière de soumettre les hypothèses du linguiste à l'épreuve expérimentale. Sous cet aspect, les progrès des sciences connexes – neurosciences, entre autres – semblent prometteurs, et pourraient ouvrir la perspective d'autres types de recherches expérimentales.

13. LA LANGUE DANS L'IMAGINAIRE

Si la langue a son siège dans l'âme collective, tout sujet a de sa langue une image. Il est permis de penser que c'est à l'aune de cette image, de ses composants, de ses attributs qu'il reconnaît telle parole proférée comme relevant – ou ne relevant pas – de sa langue.

Quelle est cette image? Et quels, ses attributs, ses composants? Comment les atteindre?

Les résultats des recherches empiriques nous livrent des traces ou fragments de cette image. Soit le mot. Comment le sujet parlant le reconnaît-il? Quelle est l'image qu'il en a? Les enquêtes sémantiques montrent que l'écheveau s'effiloche d'autant plus que l'enquête use de techniques plus fines. Qu'en affinant la technique, on fait apparaître des couches de plus en plus enfouies de la signification du mot: des sens oubliés, voire redécouverts par l'informateur même.

Est-ce ainsi l'image que garde le sujet de la langue et de ses constituants: une image sommes toutes vague, aux frontières indéterminées, aux contours flous?

Une autre issue semble possible. Il n'est pas interdit d'envisager que le sujet a de sa langue une image globale, relativement claire. Et que le halo dont se trouve entouré l'objet langue est une conséquence des techniques employées.

On peut penser aux premières décennies du XXe siècle où les progrès techniques – en phonétique instrumentale, physiologique, acoustique, … – rendent possible une connaissance

très poussée des caractéristiques physiques du son. Par ces moyens, on parvient à caractériser les sons bien au-delà de ce dont a conscience le sujet parlant. Mais le matériel phonique des langues est perdu dans ces études savantes, susceptibles de mesurer la longueur en millisecondes, les degrés de nasalité, etc. On fonde l'identité du phonème sur les affinités physiques: le phonème est conçu comme une famille de sons; les doctes phonéticiens sont scandalisés quand on propose de fonder l'identité du phonème sur sa fonction distinctive: quelle hérésie que d'assimiler à la même unité des sons aussi différents que [r] et [ʀ]! L'apport, la nouveauté de la phonologie consiste en l'abandon du point de vue purement physique pour le remplacer par le point de vue fonctionnel. Les phonèmes sont dès lors identifiés par leur fonction. Le succès de la phonologie provient du choix d'un point de vue nouveau. Choix qui permet de ramener le matériel phonique d'une langue à quelques dizaines d'unités. De ce choix découlent deux conséquences: la simplicité, comme je viens de le dire, mais aussi et surtout la conformité des résultats à l'intuition et au comportement du sujet parlant.

Admettons pour un instant que le sujet a de sa langue une image globale, relativement claire. Le statut de l'objet langue dans l'imaginaire du sujet n'est pas tellement différent de celui d'autres objets. Prenons le citron en guise d'exemple. Le sujet connaît ce qu'est un citron, même si celui-ci possède des attributs changeants. Le citron est jaune, mais un citron pas mûr, donc vert, est quand-même un citron; le citron est ovoïde, mais un citron quasi sphérique reste tout de même citron, etc. C'est ce genre de clarté relative que je crois devoir chercher dans l'image que se donne le sujet de sa langue. Quels sont les moyens qui manquent à la linguistique pour accéder à cette image?

Je suis tenté de penser que ce ne sont pas des moyens techniques. Nous en avons en suffisance, comme la phonétique des années 30 du siècle dernier. Ce qui manque ce sont – me semble-t-il – des moyens conceptuels, c'est une vision nouvelle.

Faut-il chercher à connaître l'image de la langue dans l'intuition du sujet parlant? Je crois que oui. Parce qu'en dernière analyse, l'une des innovations – la plus importante, à mes yeux – de la phonologie praguoise est de reconnaître l'imaginaire du locuteur comme partie intégrante des unités phoniques.

APPENDICE: LE CORPUS VS DISCRÉTION, FINITUDE ET VARIATIONS

Bloomfield considère que les unités d'une langue sont discrètes et en nombre fini. Il écrit:

> Assumption 1. Within certain communities, successive utterances are alike or partly alike … Outside of our science these similarities are only relative; within it they are absolute. (Bloomfield, 1926: § 2)

Cette formule correspond au concept de "tiers exclu" ou "éléments discrets". A ma connaissance, le terme *discret* est utilisé pour la première fois dans Harris (1951), cf. § 2.6 *discrete elements* et § 2.1 *discrete parts*. Bloomfield écrit encore: "Assumption 3. The forms of a language are finite in number" (1926: §14)

Curieusement, en guise d'illustration pour le public anglophone auquel il s'adresse, il ne donne pas le système phonologique de l'anglais en général, mais celui de "l'anglais américain (Chicago)" (cf. Bloomfield, 1933/1970: § 8.2). Noter en passant que son ouvrage *Language* – dont la première édition date de 1933 – suit d'assez près les "Postulats" (1926); les divergences entre les deux textes ne sont vraisemblablement pas le fait d'une évolution de la pensée de

Bloomfield. On est dès lors amené à se poser des questions sur la langue, son identité et ses limites. L'anglais est-il *une* langue? Si oui, alors les phonèmes de cette langue doivent – en vertu de l'hypothèse 1 – être absolument identiques partout. Qu'a-t-on alors besoin de spécifier que l'inventaire des phonèmes est celui de l'anglais américain, et qui plus est de Chicago. Tout porte à croire que Bloomfield se trouve dans l'embarras quand il tente d'énumérer les phonèmes de l'anglais; et que pour finir, il reconnaît implicitement qu'il y a des différences non seulement entre l'anglais britannique et l'anglais américain, mais aussi entre l'anglais de Chicago et celui d'une autre métropole. Ce constat est-il conciliable avec le caractère absolu des identités (=similarities)? Je ne crois pas; surtout si l'on tient compte d'une anecdote que rapporte Bloomfield de son voyage en Angleterre. A Londres, il demande à un chauffeur de taxi de le conduire à *Comedy Theatre*. Et le chauffeur de lui répondre qu'il n'y a pas de *Carmedy Theatre* à Londres. Ce qui revient à reconnaître que les voyelles de l'anglais ne sont pas tout à fait identiques de part et d'autre de l'Atlantique.

Zellig Harris n'échappe pas à ce genre de paradoxe. Ses positions sont différentes de celles de Bloomfield. D'abord, il ne donne aucune définition de la langue. Et quand il parle de la langue, il l'oppose au sens qu'il désigne par le terme "situation sociale" (Harris, 1951: § 2.0, *passim*); Harris identifie ainsi la langue à son expression phonique. C'est en fait la radicalisation des positions bloomfieldiennes: si Bloomfield considère la signification et tout autre phénomène mental avec méfiance, Harris les exclut totalement – en théorie, du moins – du champ linguistique. Ce faisant, il vise une procédure

descriptive mécanique, libre de tout recours aux réactions intuitives autant que comportementales du sujet parlant.

La procédure descriptive qu'il préconise consiste en le découpage – *a priori* arbitraire[8] – de la chaîne parlée en des éléments successifs; ces éléments seront ensuite caractérisés par leurs propriétés distributionnelles, et réparties en classes d'éléments. Ce qui aboutit à une structure comportant des classes d'unités et des règles. Plusieurs structures peuvent être obtenues selon différents découpages, tous également arbitraires. Sera retenue comme *la* structure de la langue, celle qui sera la plus simple, c'est-à-dire celle qui comportera le nombre le moins élevé de classes, les classes les plus extensives et les règles les plus générales.

L'examen sans complaisance des thèses de Harris aboutit à la conclusion qu'il prend des positions paradoxales. D'une part il prône de fonder l'analyse du corpus sur la distribution des éléments conçue comme seul critère d'identification (tant des unités que des classes): "The only relation which will be accepted as relevant in the present survey is the distribution …" et "The present survey is explicitly limited to questions of distribution" (Harris, 1951: § 2.1). D'autre part, il postule que les unités ont un caractère discret, et les classes un contour exact. Or, les occurrences des éléments ne peuvent constituer un critère permettant de tracer une frontière nette et sans bavure entre deux classes d'unités; et Harris le reconnaît quand il parle de "d'une probabilité fondée sur la fréquence" d'occurrences. (Harris, 1954/1970: 15).

On peut prendre acte de la contradiction inhérente du distributionnalisme comme théorie du langage, et en rester là.

---

[8] "La première opération consiste simplement en une segmentation, arbitraire si c'est nécessaire" (Harris, 1970: 15). La segmentation peut-elle être non arbitraire quand on récuse les critères du sens et de l'intuition du sujet?

On peut aussi débarrasser le distributionnalisme de ses dogmes – unités discrètes, classes exactement délimitées, non recours au sens et plus généralement à l'introspection –, incompatibles avec un modèle probabiliste. Il y aura alors lieu de s'interroger sur ce que peut apporter le distributionnalisme revu et corrigé à l'analyse du corpus. C'est ainsi que Goldsmith utilise – ce me semble – les idées de Harris. Les conséquences n'en sont pas négligeables: une unité, /ɛ/ par exemple, n'est pas unité à 100%, mais à un certain degré. Selon les spécificités du corpus (sa taille, origines des informateurs, etc.) on aura une unité /E/ ou deux /ɛ/ et /e/. Il en va de même pour les classes. D'une manière générale, le "néodistributionnalisme" sera tout à fait susceptible d'offrir des descriptions à échelles multiples. Et les problèmes que pose Harris sur la représentativité du corpus prennent tout leur sens.

En fait, les germes de cette révision existent déjà dans les textes de Harris. Je me contenterai d'un seul exemple. Au § 20.3. de *Structural Linguistics*, Harris pose le problème de la description de la structure de la langue que je paraphrase ainsi: en admettant que le corpus soit un échantillon adéquat, nous sommes en mesure de dire, sur la foi de la description, que certaines séquences apparaissent dans les énoncés de cette langue. Mais cela ne veut pas dire que d'autres séquences ou d'autres éléments sont exclus. Nous sommes cependant à même de dire que certaines séquences n'apparaissent presque jamais. Cette affirmation peut prendre appui sur des tests directs; ou bien sur le fait que ces séquences vont à l'encontre des règles les plus générales du corpus.

Je crois pouvoir relever deux principes: 1° recours à l'introspection. Introspection du descripteur d'abord, puisqu'il lui revient de décider si un corpus est "un échantillon adéquat" de la langue. Introspection du sujet parlant ensuite, étant donné qu'il est permis de recourir au "test direct" où le locuteur est

appelé à se prononcer sur la possibilité de certaines séquences d'unités; 2° recours à hiérarchie et échelles. Il est permis d'exclure les séquences qui enfreindraient les règles les plus générales. Pareille exclusion repose sur le principe implicite que toutes les règles ne se valent pas; et qu'elles peuvent être hiérarchisées selon leur généralité ou restriction.

Cette prise de position n'est pas tout à fait isolée. Ailleurs aussi, Harris propose de recourir à l'enquête pour trancher la question "les sons *A* et *A′* sont-ils identiques ou différents?" pour les cas où les procédures descriptives ne peuvent y apporter réponse (Harris, 1951: § 4.22, cf. aussi pp. 38, 173, 361 et 264).

L'ambiguïté du concept corpus n'est pas une particularité du structuralisme américain. On la rencontre dans d'autres courants structuralistes. Chez Martinet, par exemple, on trouve: *a/* un décalage entre les principes énoncés – finitude et discrétion – et la pratique effective. Martinet affirme que les phonèmes sont des "unités discrètes" (Martinet, 1960: § 1-17) et en "nombre déterminé" (Martinet, 1960: § 1-14). Il constate cependant que "la réponse à la question “Combien telle langue a-t-elle de phonèmes?” [est] souvent délicate" (Martinet, 1960: § 1-13); *b/* ambiguïté du concept corpus. Le corpus, une fois constitué, est-il à considérer comme intangible? Le descripteur peut-il le compléter quand il en sent le besoin? La réponse n'est pas claire (cf. Martinet, 1960: § 2-4). Dans l'ensemble, on a l'impression que le rôle du corpus est limité à celui d'un recueil de données, garant de l'authenticité des matériaux analysés et décrits. Cette conception du corpus est parfois clairement énoncée. Ainsi dans Martinet (1979).

RÉFÉRENCES

Bloomfield L. (1926). A Set of Postulates for the Science of Language, *Language, 2*: 153–164.

Bloomfield L. (1970). *Le langage*. Paris: Payot. (Parution en anglais 1933).

Condamine A. (éd.). (2005). *Sémantique et corpus*. Paris: Hermès Sciences.

Durkheim E. (1967). *Les règles de la méthode sociologique (16e édition)*. Paris: Presses Universitaires de France. (Première édition 1894).

Goldsmith J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics, 27(2):* 153–198.

Habert B. (2000). Des corpus représentatifs: de quoi, pour quoi, comment? In M. Bilger (dir.), *Linguistique sur corpus: études et réflexions*. Perpignan: Presses universitaires de Perpignan, pp. 11–58.

Harris Z. S. (1951). *Methods in Structural Linguistics*. Chicago: The Chicago University Press.

Harris Z. S. (1955). From Phoneme to Morpheme. *Language, 31*: 190–222.

Harris Z. S. (1970). La structure distributionnelle. *Langages, 20*: 14–34. (Parution en anglais 1954).

Jolivet R. (1982). *Descriptions quantifiées en syntaxe du français. Approche fonctionnelle*. Genève-Paris: Slatkine.

Labov W. (1976). *Sociolinguistique*. Paris: Minuit.

Lazard G. (2009). Pour une linguistique pure. *Bulletin de la Société de Linguistique de Paris, 104(1)*: 1–16.

Martinet A. (1960). *Eléments de linguistique générale*. Paris: Colin.

Martinet A. (1979). *Grammaire fonctionnelle du français*. Paris: Didier.