CORPUS ET ETUDES SUR LA LANGUE PARLÉE

Claire Blanche-Benveniste
Université de Provence

Nous utilisons aujourd'hui largement le terme de *corpus*, souvent sans bien le définir, comme un équivalent de "sources attestées". Dire que l'on travaille sur des corpus, parlés ou écrits, ou, plus brièvement "sur corpus", c'est habituellement une façon de dire qu'on utilise des exemples attestés et qu'on ne fabrique pas les données linguistiques à partir de sa propre intuition. Mais il y a souvent une tendance – en France, en tout cas – à réduire l'extension du mot *corpus*, pour l'appliquer seulement à des enregistrements de langue parlée collectés par des enquêtes, et, du coup, à restreindre le sens même de ce qu'on entend lorsqu'on traite des données linguistiques sur lesquelles il convient de travailler. En ce cas, "travailler sur corpus" implique souvent qu'on travaille exclusivement sur l'oral. L'histoire récente du terme, telle qu'on peut la suivre en France, montre qu'il y a là des malentendus, qui viennent en grande partie des limites mal définies entre *corpus* et "autres sources de documentation" sur la langue.

L'idée d'étudier la langue parlée par le moyen de corpus a beaucoup évolué dans les années récentes. Pendant tout un temps, un travail de ce genre se plaçait presque nécessairement dans une perspective sociolinguistique. Puis, l'intérêt s'est porté sur les recherches typologiques. Le français parlé jouait, parmi les langues romanes, un rôle très particulier. Plus récemment, les interactions ont paru attirer l'attention. Enfin, la perspective de description linguistique a été renouvelée par le développement assez nouveau des comparaisons entre langues romanes parlées.

1. Quelques avatars de la notion de corpus

1.1. Le corpus comme source majeure de documentation

Pour la linguistique structurale américaine, telle que la présentait Z. Harris en 1951, lorsqu'il s'agissait de décrire des langues non écrites, que l'on connaissait par les témoignages d'informateurs de la langue, la notion de corpus avait un sens particulier. Le corpus est une collection d'énoncés fournie par l'informateur. Cette collection, prise comme un "échantillon" de la langue, peut avoir des dimensions variables. On se contente d'un corpus plus petit pour une description phonologique que pour une enquête sur la morphologie (Z. Harris, 1951, p. 13). Mais des centaines d'heures d'entretien avec un informateur peuvent se révéler insuffisantes pour établir une description syntaxique exhaustive. Il y faudrait une sorte de corpus idéal, qui reste toujours inacccessible (ibid., p. 253).

Le descripteur est donc amené à compléter sa collection d'énoncés par d'autres procédures bien contrôlées, que Z. Harris résumait sous le terme de "eliciting" (ibid. p. 12 et p. 368): en questionnant son informateur, en faisant répéter, en créant des situations qui peuvent faire apparaître une forme ou une tournure, etc. Ces procédures sont très délicates; l'informateur n'a pas toujours la possibilité de dire s'il utilise réellement telle ou telle forme qu'on lui soumet; des obstacles surgissent, qui tiennent aussi bien aux relations interpersonnelles qu'à différents éléments culturels difficiles à maîtriser:

> "The eliciting of forms from an informant has to be planned with care because of suggestibility in certain interpersonnal and intercultural relations and because it may not always be possible for the informant to say whether a form which is proposed by the linguist occurs in his language" (p. 12).

Ce complément d'information, difficile à récolter, est jugé nécessaire pour constituer un corpus représentatif de la langue.

Le terme de *corpus*, qui avait été lancé en France vers 1923 (d'après A. Rey)[1], paraissait encore bizarre dans les années 1970, lorsque R. L. Wagner s'y est intéressé. Il s'agissait de susciter des études sur le français parlé, en se défiant des intuitions que tout un chacun croit avoir sur sa langue maternelle, et qui se révèlent souvent être

> "des intuitions tâtonnantes, contradictoires, que l'on couvre du beau nom de compétence" (R. L. Wagner, 1973, p. 112).

L'idée de corpus à l'américaine était présente; le mot était encore neuf. "Le mot est pédant (autant que celui de *campus)*", écrivait R. L. Wagner.

> "Il s'est insinué dans la nomenclature des linguistes français et y atteint un degré de fréquence excessif" (1973, p. 100).

Pour en définir la portée, R. L. Wagner proposait de dire que la collection d'enregistrements faits par enquêtes sur le français parlé ne constituait qu'une partie de la documentation que l'on devait accumuler. Ces enquêtes devraient être complétées par des questionnaires, des énoncés notés au vol, des vérifications directes ou indirectes. Et, de cet ensemble, qui formerait une documentation générale, on pourrait extraire une "documentation de base élaborée et triée", qui formerait alors l'outil de travail qu'il nommait "un corpus critique" (p. 102).

> "L'expérience prouve que l'enregistrement de conversations libres ou de narrations d'histoires ne couvre pas, de loin, tout le champ du dicible. Il doit être complété par une enquête de contrôle: réponses à des questions soigneusement préparées en vue de tester les sujets sur des points précis de morphologie et de syntaxe" (1980, p. 9).

Dans cette acception, le corpus n'était pas une simple accumulation de documents. C'était le résultat d'un travail d'échantillonnage, à recommencer pour chaque nouvelle question examinée. L'enregistrement mécanique des productions de français parlé n'en était qu'une sous-partie.

Pour la description du français parlé contemporain, il semble qu'on n'ait pas exploré systématiquement ce recours à différentes sources de données.

Les dialectologues faisaient des questionnaires; dans le modèle de Edmont, des mots isolés et une centaine de phrases (cf. G. Bergounioux, 1994, p. 313). A leurs débuts, les recherches sur le français parlé (*Français Fondamental, Enquête d'Orléans, Corpus Sankoff-Cedergren*) ont enregistré essentiellement des entretiens, en cherchant, pour des raisons sociolinguistiques, à les avoir le plus "spontanés" possible. Cependant, on peut prévoir que certains phénomènes linguistiques sont liés aux situations d'enregistrement et qu'une enquête un peu

large demande, de ce fait, différentes situations d'enregistrement. En français, les "passés simples", les emplois du relatif *dont*, (et tous les phénomènes que nous avons nommés de "langue du dimanche, Cf; Blanche-Benveniste 1982), les parenthèses ou les constructions à enchâssements, tout cela se rencontre plus dans certaines sources que dans d'autres. La distribution est, dans ces cas, liée à la différences des genres.

Les différences de genres ont amené à corriger certaines conclusions, un peu hâtives, sur l'évolution de la langue. Lorsque des enfants de 10-11 ans racontent un événement dramatique, et qu'ils le font debout devant le micro (Pazéry 1989), ils utilisent des passés simples dans leurs narrations, alors que ce temps est classé dans presque toutes les grammaires comme un "temps du langage écrit".

1. 2. Faible représentation des autres sources

1.2.1. Les notations "au vol"

Elles sont indispensables pour récolter des données qui apparaissent rarement dans les corpus, bien que les locuteurs puissent les produire très aisément. L'équipe d'Aix en a fait l'expérience à propos d'une enquête sur les accords du participe passé

Dans les productions orales, remarquait M. Audiber- Gibier (1992), seul un petit nombre de participes passés ont une finale féminine audible: on l'entend dans "je l'ai faite, dite, inscrite, assise", mais pas dans "je l'ai vue, rencontrée, suivie". Un calcul montre que, sur l'ensemble des participes passés qu'utilisent les locuteurs, seuls 6% peuvent fournir des accords audibles. Si on limitait l'enquête aux attestations que fournissent les transcriptions d'enregistrements, il y aurait une énorme déperdition. M. Gibier a obtenu l'essentiel de ses exemples en notant au vol des séquences verbales, en dehors des situations d'enregistrement.

Une grande partie des études sur langue parlée est complétée, de façon explicite ou non, par des exemples notés "à la volée". Il est évidemment impossible d'utilisr cette méthode dans les cas où l'on a besoin d'un large contexte.

1. 2. 2. Les questionnaires

Les questionnaires, qui étaient l'outil principal des recherches en dialectologie, ont été utilisés assez systématiquement pour les recueils de régionalismes (G. Tuaillon, 1983; J. Pohl 1990), et surtout pour tous les cas où il fallait travailler sur des "écarts" et où il était possible de solliciter les locuteurs.

Parmi les questions portant sur le français commun, pour lesquelles il semblait utile de recourir à des questionnaires, R. L. Wagner citait l'emploi de "lui" et "elle" appliqués à des référents non-humains (1973, p. 147):

– "Elle a acheté une mini-jupe et elle ne porte plus qu'elle".

Les 30 témoins qu'il avait sollicités lui montraient que "lui" et "elle" référant à des objets était majoritairement acceptés. L'équipe d'Aix a utilisé cette technique pour quelques études isolées, comme le choix fait par les locuteurs français actuels entre "mille cinq cents" et "quinze cents", selon les objets comptés et les situations de comptage (L. Veis, 1986). Mais il semble qu'on ait au total produit peu d'enquêtes grande envergure sous la forme du questionnaire. Cela tient sans doute en grande partie à l'orientation même des études prévues pour la langue parlée.

2. Quelques grandes orientations dans les recueils de français parlé

2.1. La sociolinguistique

Dans les années 1970-80, sous l'influence de William Labov, il s'est développé en France et dans les pays de langue française un fort intérêt pour l'étude sociolinguistique. Le principal corpus présenté à cette époque a été celui qu'on nomme "Sankoff-Cedergren", sur lequel de très nombreuses études ont été conduites.

La perspective adoptée visait même à obtenir des entretiens tous du même type, afin de pouvoir mieux comparer entre eux les différents locuteurs et leurs particularités socioculturelles. Le procédé le plus commode consistait à faire parler différents locuteurs sur un même type de sujet, et dans les mêmes circonstances (par exemple, les "changements survenus dans la ville"). Dans l'enquête d'Orléans, menée par une

équipe britannique, on demandait, en complément aux entretiens, de raconter "comment vous faites une omelette" (ce qui est curieusement la même consigne que l'on donne dans les hôpitaux pour étudier les maladies de langages !).

Cette perspective sociolinguistique était dynamique. Elle s'intéressait au devenir de la langue, vu à travers quelques projections idélogiques. Les chercheurs tendaient à poser l'hypothèse que les classes sociales défavorisées conduisaient l'évolution. Par exemple, c'est dans les quartiers "populaires" qu'on aurait noté des amorces d'évolution qui réduisaient l'emploi de l'auxiliaire "être" en français, supprimaient presque totalement le subjonctif, et transformaient la forme des interrogations. Un appareil statistique très raffiné, élaboré par David Sankoff, apportait à ces hypothèses une garantie technique solide.

2.2. La typologie des langues

Dans les années 1980-1990, les recherches en typologie des langues, principalement sous l'influence de Bernard Comrie, ont favorisé les études sur les langues parlées, étudiées d'après différentes sortes de corpus. Le français y a occupé une place de choix, en raison du grand décalage que l'on peut y constater entre la norme scolaire de langage et les différents usages. Une partie de ces décalages a été analysée comme une nouvelle orientation typologique du français, qui aurait été masquée par les descriptions classiques.

C'est ainsi que Martin Harris (1978) a lancé l'idée du *Drift of French Syntax*. Les dislocations du sujet, très fréquentes dans le français familier (*elle est venue, Marie)*, étaient interprétées comme des indices d'un changement de l'ordre des constituants, passant de SVO à VOS. De la même façon, l'absence du *ne* de négation, que l'on constate facilement dans le français des conversations, a été replacée dans un grand cycle d'évolution typologique, où le français, encore une fois, révélait des orientations caractéristiques (O. Dahl, 1979). Sur certains autres points, comme les constructions relatives, on peut montrer au contraire une grande convergence entre les langues romanes, à condition de prendre les exemples non dans les normes décrites par les grammaires, selon lesquelles ces langues apparaissent assez différentes, mais dans les usages des langues parlées, où elle semblent typologiquement beaucoup plus proches (Blanche-Benveniste, 1990).

2.3. L'étude des interactions

C'est sans doute E. Goffman (1981) qui a donné l'élan, en s'intéressant à des usages de la langue parlée peu étudiés jusque là. Le développement de la pragmatique a provoqué quantité d'analyses faites sur du langage "en situation". Les *Cahiers de Linguistique Française* de l'Université de Genève en ont montré beaucoup d'applications, en s'intéressant par exemple, aux formes de l'argumentation, ou aux formes de la référence. L'étude des conversations est devenu un domaine à part entière (cf. Kerbrat-Orrecchioni 1990).

M.A.K. Halliday a montré, dans sa perspective "fonctionaliste", comment l'information était formulée de façon différente par écrit et par oral, ce qui entraînait nécessairement des interactions différentes. Il a, à ce propos, introduit des analyses devenues classiques sur la condension sémantique de l'écrit, face à la complexité syntaxique de l'oral.

Les corpus de langue parlée révélaient davantage que des faits de grammaire: ils semblaient pouvoir permettre un accès aux relations pratiquées entre les locuteurs. Du coup la grammaire elle-même a été parfois délaissée.

2.4. Les descriptions grammaticales

Elles ont reçu une impulsion importante à partir du moment où les principales langues romanes, portugais, espagnol, italien, français, se sont dotées des moyens de travailler sur leurs langues parlées. Cela a permis de mettre en place un ensemble de ressemblances et de différences grammaticales qui étaient beaucoup moins sensibles auparavant.

La plupart des phénomènes liés à la production du langage parlé, comme les hésitations, les reprises, les bribes de discours, peuvent être traitées de façon identique dans les quatre langues. Par exemple, l'hésitation sur le choix des prépositions et sur les séquences de constructions verbales paraît très semblable dans les quatre cas (Cf. F Bacelar Nascimento, pour le portugais et le français).

Une grande partie de ce qu'on a pu appeler "la macro-syntaxe" (cf. Blanche-Benveniste 1990, Berrendonner et Reichler-Béguelin, 1989) semble être en commun. Ainsi, on peut montrer que l'utilisation des "compléments de phrases", ou la forme des grandes périodes peuvent se décrire selon des procédés très semblables (Cf. Renzi, N. Vincente).

En revanche, les corpus de langue parlée permettent de cerner, mieux semble-t-il que par des exemples de langue écrite, les particularités syntaxiques et morphologiques de ces langues: usage des temps, syntaxe des infinitifs et participes, syntaxe des sujets, usage des pronoms, etc.

La notion de grammaticalité, cruciale pour une partie des grammaires formelles, a été affinée et compliquée par les données orales. Par exemple, les recueils de liaisons faits par P. Encrevé ont permis de présenter une tout autre notion de grammaticalité que celle qui prévalait dans ce domaine.

3. Conclusion

Il semble que les corpus de langue parlée, pris au sens étroit ou au sens large du terme, ont permis de compléter la description grammaticale dans deux domaines importants.

Le domaine de la description syntaxique s'est étendu. Il est devenu banal d'y inclure les dislocations, les clivées, les pseudo-clivées ou les corrélations, qu'on classait, il y a peu, dans les faits de stylistique, de mis en relief et d'expressivité. Le fait que ces tournures apparaissent avec régularité dans les corpus de différentes langues parlées, selon des fonctionnements réglés, et en dehors des effets stylistiques classiques.

L'interprétation sémantique des faits de syntaxe a été renforcée. Là où on pouvait voir, en termes de sociolinguistique, des variations de forme pour un même contenu pragmatique, par exemple dans différentes formes d'interrogation ou dans la concurrence entre temps verbaux différents, on peut voir actuellement autant d'options sémantiques légèrement différentes. La notion de "cohérence textuelle", qui avait été pendant longtemps calculée d'après les impératifs de textes écrits, a été complètement revue et réinterprétée (cf. Reichler-Béguelin, 1994).

Evolution révélatrice: les plus récentes grammaires du français (Cf. Riegel et alii,1994), tiennent compte de la problématique des corpus de langue parlée. C'est là un signe certain de l'évolution des opinions sur ce sujet.

Notas

[1] Le modèle venait, précise A. Rey, du droit, où l'on employait depuis longtemps "corpus juris" pour désigner le recueil des lois du droit romain.

Bibliographie

AUDIBERT-GIBIER, Monique, 1992, "Etude de l'accord du participe passé sur des corpus de français parlé", *Langage et Société*, n° 617-30.

BACELAR DO NASCIMENTO, Maria-Fernanda, M.L.GARCIA MARQUES et M.L. SEGURA DA CRUZ, 1987, *Português fundamental. Metodos e Documentos. Tomo 1°, Inquérito de frequência.* Lisboa: INIC.

BERGOUGNIOUX, Gabriel, 1994, *Aux origines de la linguistique française.* Paris: Pocket.

BERRENDONNER, Alain et REICHLER-BÉGUELIN, Marie-José, 1989, "Décalages. Les niveaux de l'analyse linguistique", *Langue Franaçaise* n° 81, 110-135.

BLANCHE-BENVENISTE, Claire, 1982,"La escritura del lenguje Dominguero", in E.FERREIRO y M. GOMEZ-PALACIO (eds.) *Nuevas perspctias sobre los procesos de lectura y escritura.* Mexico: Editorial Siglo XXI.

BLANCHE-BENVENISTE, Claire, 1987, "Les études sur les langues parlées viennent-elles compliquer l'établissement d'une typologie?", *Cercle Linguistique d'Aix-en-Provence. Travaux 5, Typologie des langues*, 49-57.

BLANCHE-BENVENISTE, Claire, 1990, "Usages normatifs et non normatifs dans les relatives en français, en espagnol et en portugais", in Johannes BERCHERT, Giuliano BERNINI et Claude BURIDANT (eds.), *Towards a Typology of European Languages. Empirical Approachess to Language Typology 8.* Berlin/New York: Walter de Gruyter, 317-335.

BLANCHE-BENVENISTE, Claire et JEANJEAN, Colette, 1986, *Le français parlé. Edition et transcription.* Paris: Didier-Erudition.

BLANCHE-BENVENISTE, Claire, BILGER, Mireille, ROUGET Chritine et EYNDE, van de, Karel, 1990, *Le français parlé. Etudes grammaticales.* Paris: Editions du CNRS.

COMRIE, Bernard, 1981, *Language Universals and Linguistic Typology.* Oxford: Basil Blackwell.

COSTE, Elisabeth, 1989, "Etude syntaxique des problèmes posés par les parenthèses dans les corpus de français parlé", Mémoire de Maîtrise, Université de Provence.

DAHL, Osten, 1979), "Une typologie des négations de phrase", *Linguistics*, 17, 79-106.

ENCREVÉ, Pierre, *La liaison sans enchaînement*. Paris: Ed. du Seuil.

GOFFMAN, Ervin, 1981, *Forms of Talk*. Oxford: Basil Blackwell.

GOUGENHEIM,G., R. MICHEA, P. RIVENC et A. SAUVAGEOT, 1956, *L'élaboration du français élémentaire*. Paris: Didier.

HALLIDAY, M.A.K, 1985, *Written and Spoken Language*. Cambridge: Cambridge U.P.

HARRIS, Martin, 1978, *The Evolution of French Syntax. A comparative Approach*. London/ New York: Longman

HARRIS, Zellig S., *Structural Linguistics*. Chicago/London: Phoenix Books. The University of Chicago Press.

JEANJEAN, Colette, 1988, "Le futur simple et le futur périphrastique en français parlé. Etude distributionnelle", in Cl. Blanche-Benveniste, A. Chervel et M. Gross (eds.) *Grammaire et histoire de la grammaire. Hommage à la mémoire de Jean Stéfanini*. Aix-en-Provence: Publication de l'Université de Provence, pp. 235-258.

KERBRAT-ORECCHIONI, Catherine, 1990-1992, *Les interactions verbales*. Paris: A. Colin, 2 volumes.

LAROCHE-BOUVY A.D., 1984, *La conversation quotidienne*. Paris: Didier/Crédif.

PAZERY, Nelly, 1990, *Corpus de l'incendie de la Sainte Victoire*. Video et transcription. Département de Linguistique française, Université de Provence.

POHL, Jacques, 1979, *Les variétés régionales du français. Etudes belges (1945-1977)*. Bruxelles: Editions de l'Université de Bruxelles.

REICHLER-BÉGUELIN, Marie-José, 1994, "Anaphores pronominales en contexte d'héérogénéité énonciative: effets d'(in)cohérence", *Colloque "Relations anaphoriques et (in)cohérence"*, Anvers, 1-3 /12/ 1994.

RENZI, Lorenzo, 1985, *Nuova Introduzione alla Filologia Romanza*. Bologna: Il Mulino.

REY, Alain, 1992, *Dictionnaire Historique de la Langue Française*. Paris: éd. Le Robert, 2 volumes.

RIEGEL, Martin, 1994, *Grammaire méthodique du français*. Paris: P.U.F.

SANKOFF, David, SANKOFF Gilian, LABERGE, Suzanne et TOPHAM, M., 1976, "Méthode d'échantillonnnage et utilisation de l'ordinateur dans l'étude de la variation grammticale", *Cahiers de Linguistique de l'Université du Québec*, n°6, 85-125.

SANKOFF, Gilian and CEDERGREN, Henrietta J., 1971, Some results of a sociolinguistic study of Montreal French", in R. DARNELL (ed.) *Linguistic Diversity in Canadian Society*, 61-87.

TUAILLON, Gaston, 1983, *Matériaux pour l'étude des régionalisme du français. Les régoinalismes du français parlé à Vourey, village dauphinois*. Paris: Klincksieck.

VEIS, Luc, 1986, *Les pièges du comptage en français. Etude sur corpus. Systèmes et sous-systèmes.* Mémoire de maîtrise, Département de Linguistique française, Université de Provence.

VOGHERA, Miriam,1992, *Sintassi e Intonazione nell'italiano parlato.*Bologna: Il Mulino.

WAGNER, R. L., 1973, *La Grammaire française, volume 2. La grammaire moderne. Voies d'approche. Attitude des grammairiens.* Paris: SEDES.