Pour une phonologie de corpus*

BERNARD LAKS

Université de Paris X, laboratoire MoDyCo

(Received August 2007, revised September 2007)

'It has always seemed to me, ever since I first tried to become a grammarian, that grammar was a subject with too much theory and too little data' (Halliday, 1992:61)

ABSTRACT

Pour les sciences du langage, comme pour toutes les sciences, la question du rapport aux données est l'une des plus fondamentales qui soit. Les avancées technologiques récentes en matière d'analyse et de fouille de bases et de banques de données linguistiques ont donné à la question de la relation entre corpus et modèles une acuité nouvelle. Face aux constructions génératives qui voient la théorie comme fondamentalement sous déterminée par les données factuelles, de nouvelles approches ont été proposées qui mettent au premier plan les faits de langue observables et construisent l'analyse linguistique comme une modélisation des usages. Voir dans ce renversement de la relation entre données et modèles un simple effet des technologies nouvellement disponibles et défendre que les linguistiques de corpus ne sont rien d'autre que de simples dispositifs techniques au service d'une linguistique descriptive, empirique ou herméneutique, occulte les questions épistémologiques centrales qui commandent la relation de la science linguistique à ses observables. En réanalysant l'opposition classique entre sciences de l'*exemplum* et sciences du *datum*, je montre ici que la linguistique, et singulièrement la phonologie, se sont construites, contre la grammaire, comme des sciences empiriques ayant pour objet la modélisation des observables linguistiques. La notion de corpus apparaît ainsi l'une des plus anciennes qui soient. Comprendre son ancienneté et construire son historicité permettent de saisir ce qui est en jeu dans son resourcement technologique récent et permet de voir le moment génératif comme une parenthèse, certes un temps productive, dans la longue confrontation du linguiste

modélisateur à ses corpus d'observables. L'enjeu de la réanalyse de la notion de corpus en linguistique et en phonologie est donc considérable, contre la vulgate saussurienne qui doit si peu à la pensée du Maître genevois, il ne s'agit de rien de moins que frayer le chemin d'une linguistique de parole, condition *sine qua non*, comme il l'a souvent dit d'une linguistique de la langue.

Préambule : Une histoire nécessaire

Il y a quelques années déjà, avec Jacques Durand, puis avec John Goldsmith, j'ai plaidé pour le caractère fondamentalement cumulatif de la phonologie, pour autant qu'elle se pense et se définisse comme une science galiléenne (Durand et Laks, 1996 ; Goldsmith et Laks, 2005). Cette position a connu un certain succès dans les cercles phonologiques (*cf.* Scheer, 2004b). Elle implique de porter une attention soutenue à l'histoire et à l'épistémologie de la phonologie, tant il est vrai que, comme l'a souvent rappelé Bourdieu, l'histoire et l'épistémologie d'une science sont partie intégrante de cette science et constituent, par le travail d'anamnèse qu'elles supportent, l'une des conditions *sine qua non* de sa scientificité propre (Bourdieu, 2001). C'est pourquoi, il est toujours nécessaire, lorsqu'on se propose d'articuler une analyse des courants nouveaux qui émergent dans une discipline, d'en commencer par une mise en perspective historique et épistémologique. Dans ses divers linéaments, la ***linguistique de corpus*** est à la mode, et la ***phonologie de corpus*** s'impose dans le paysage académique contemporain. Eclairer l'une et travailler à promouvoir l'autre supposent donc quelque préambule historico-épistémologique.

1. Faits de langue : exemples et données linguistiques

Dans une perspective poppérienne, la qualité cumulative d'une science est étroitement liée au formatage et au statut de ses data ainsi qu'à l'explicitation et à la transparence des méthodologies qui permettent de les recueillir, et surtout de les reproduire. Or les questions du statut des données, des méthodes de leur collecte, du format dans lequel elles sont couchées, commentées et transmises, pour fondamentales qu'elles soient, sont parmi les moins abordées de la linguistique et de la phonologie contemporaines post-SPE. Elles ont pourtant fait l'objet de grandes attentions dans les périodes qui ont précédé. La philologie classique, la grammaire

historique et comparée, le structuralisme européen et surtout américain ont tous inscrit leur démarche scientifique dans une problématique plus ou moins explicite de collecte puis d'analyse de données accumulées en collections stabilisées que l'on peut appeler des corpora[1]. Cette pratique se distingue nettement d'une méthodologie plus classiquement grammairienne fondée sur la référence à des ***exemples*** très rarement constitués en collections stables et fermées, systématiques, classées, publiques et partagées[2]. En histoire et épistémologie des sciences, l'opposition entre sciences de ***l'exemplum*** et sciences du ***datum*** est fondamentale.

Cette opposition est pour une part constitutive de la rupture introduite par la grammaire générative dans la deuxième moitié du 20ème siècle. L'option cartésienne, réaffirmée lors du tournant cognitiviste des années 1965[3], fonde la grammaire générative sur une analyse des intuitions d'un locuteur-auditeur idéal et abstrait, appartenant à une communauté linguistique homogène qui apprend la langue instantanément et n'est affecté par aucune des limitations de la performance (Chomsky, 1965: 4). Outre leur labilité individuelle et inter individuelle, les jugements de grammaticalité ne sont jamais constitués en corpus publics et opposables permettant d'évaluer relativement ***consistance et complétude*** de telle analyse ou de telle argumentation générative.

En linguistique, il n'en a pas toujours été ainsi. Aux 19ème et 20ème siècles, les sciences du langage et leurs précurseurs se sont constituées comme des sciences du *datum*, inscrivant leur démarche dans la dynamique épistémologique qui depuis la Renaissance faisait émerger la science moderne comme une systématique adossée à de larges compendiums de faits. Du point de vue historique, la notion de corpus apparait en effet comme très ancienne, mais elle joue un rôle de première importance dans le développement de la pensée scientifique moderne. Il faut donc y regarder de plus près.

* Ce travail a été nourri de nombreuses discussions et échanges. Je remercie Chantal Lyche, Jacques Durand, John Goldsmith et Ernesto d'Andrade. J'ai également bénéficié de la relecture experte et anonyme de trois collègues que je remercie pour la pertinence de leurs remarques et suggestions. Tous blâmes sont miens, toutes laudes sont leurs.

[1] Comme Blanche-Benveniste (2000 : 2), je note que 'plus exacts latinistes que les Français, les linguistes des autres pays européens disent généralement un *corpus* et des *corpora*'. Comme elle, néanmoins, je me conforme dorénavant, à l'usage français.

[2] Pour une analyse historique du rôle et de la disposition des exemples ainsi que de leur relation à la norme *Cf.* Chevalier (2007).

[3] Pour une analyse *Cf.* Goldsmith et Huck (1995).

1.1 Généalogie du corpus

Sans entreprendre ici une véritable généalogie du concept de corpus, je note cependant qu'il remonte au moins à Justinien (527-565) qui fit compiler le *Corpus Juris Civilis* -recueil à vocation exhaustive qui contenait les constitutions impériales, un manuel de droit et l'ensemble de la jurisprudence commentés. En rappelant que le corpus de Justinien faisait pendant au *Corpus Juris Canonici*, on se souvient de ce que la notion de corpus doit à la pensée théologique, au moins dans les religions du Livre. L'empilement des commentaires consacrés, des exégèses canoniques et des références croisées forme dans les religions monothéistes de très vastes banques de données textuelles que l'on peut souvent qualifier anachroniquement d'hypertextuelles. Que l'on songe seulement, par exemple, à la Torah entourée des strates successives de ses commentaires reçus, aux Evangiles dits Synoptiques et aux travaux qu'a suscité le Problème Synoptique durant des siècles, ou encore aux Hadiths et à leur appareil exégétique. Il s'agit toujours, de vastes corpus structurés, clos, stables et publiquement acceptés. Ces corpus sont certes à vocation herméneutique ou religieuse, mais si l'on se tourne à présent du côté des sciences naissantes de la période moderne, on soulignera à nouveau le rôle heuristique des vastes compendiums de faits, clos, structurés, stables et publiquement partagés[4], rassemblés à cette époque.

De l'histoire naturelle de Buffon aux grands classements de Linné, l'accumulation des faits, des données et des descriptions est constitutive d'un classement raisonné qui fonde une première théorisation et une première modélisation. Adossée à d'énormes compendiums, la Science se dégage alors comme un raisonnement sur l'organisation des données, comme une contemplation, une θεωρία (*théoria)* conduite par la structuration interne des données. Dans cette émergence, l'importance de Carl von Linné fut considérable. Avec la notion de ***taxon*** (catégorie abstraite, construite et super ordonnée) il engage l'esprit scientifique moderne. Avec lui, le compendium s'analyse désormais en taxinomies raisonnées lesquelles constituent le socle même de toute théorisation et modélisation scientifique. Le lien entre corpus et modèle est enfin explicitement établi dans la Systématique dont Mendeleïev livrera à la fin du 19ème siècle l'un des monuments les plus accomplis. Il peut sembler que l'on est bien loin de la

[4] Faute de place, je ne développe pas ici la notion de représentation publique que j'applique implicitement aux corpus. Je renvoie à la Théorie de la Pertinence (Sperber et Wilson, 1989). L'existence de corpus publiquement acceptés me semble constituer un critère minimal de cumulativité scientifique.

linguistique[5], mais pour se convaincre du contraire il suffit de rappeler le parallélisme étroit qui existe entre le Tableau Périodique des Eléments, avec ses cases vides et leur poids atomique prédits par le modèle, et le *Mémoire sur le Système Primitif des Voyelles en Indo-Européen* (Saussure, 1878) lui-même également construit comme une théorisation d'un vaste compendium de données sanskrites.

Pour en revenir au 19ème siècle et aux sciences de la nature, après Cuvier et Lamarck, le lien entre corpus et modèle, théorie du corpus et théorie explicative, trouve son achèvement dans la théorie darwinienne, dont on a pu dire à juste titre, qu'elle constituait la théorie scientifique la plus considérable de l'époque moderne[6]. Dire que Darwin fût un homme de corpus est un pléonasme : le voyage sur le Beagle ***est une enquête***. Son analyse ***est une taxinomie*** qui fonde le monument théorique et conceptuel qu'est la théorie de l'évolution. On voit ainsi se dégager une ligne méthodologique très forte dans l'histoire des sciences : constitution de l'observable, taxinomie et systématique, théorisation et modélisation. Dans cette méthodologie *bottom-up*, la construction d'un corpus factuel joue, on le voit, un rôle considérable.

1.2 Corpus linguistiques actuels

On objectera que cette approche du corpus, pour fondée qu'elle soit dans l'histoire des sciences, est très différente de celle qui apparaît dans l'histoire toute récente de la linguistique et des traitements automatiques de la langue. Alors que Mellet (2002 : 6) propose une définition encore relativement ouverte du concept de corpus[7], Rastier (2005 : 33) propose ainsi une définition beaucoup plus restrictive de la ***Linguistique de Corpus***. Pour lui, la définition de corpus est intimement liée aux applications computationnelles qu'elle nourrit et le corpus est étroitement dépendant, pour ce qui concerne son nettoyage, son étiquetage, son

[5] Denis Lepesant me rappelle que Maurice Gross tenait Carl von Linné pour l'un des plus grands scientifiques de l'histoire. Revendiquant son héritage taxinomique, il le cite explicitement comme l'une de ses sources dans sa préface à *Méthodes en syntaxe* (Gross, 1975).

[6] A propos de l'influence de Darwin sur la philologie et les débuts du comparatisme *Cf*. Laks (2002); Laks *et al (*2006); Sériot (1999).

[7] 'Dans le champ linguistique, la notion s'est complexifiée au cours des dernières décennies en fonction de la diversité des pratiques et des objectifs assignés à la constitution et à l'exploitation des corpus. […] Un corpus ne peut être ***clos et exhaustif*** que dans le cadre d'une monographie, auquel cas il sera étudié en tant que tel, sans prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation. […] A l'opposé des corpus homogènes et exhaustifs se trouvent les ***corpus échantillonnés*** ; là, le problème se déplace : l'enjeu n'est plus celui de l'exhaustivité, mais celui de la ***représentativité***. Il s'agit alors de constituer des échantillons représentatifs d'une réalité plus large – en statistique on dirait : d'une population. […] Ces corpus se veulent généralement ***des corpus de référence***, exploitables pour des recherches variées par plusieurs générations de linguistes'. (Mellet, 2002: 3-6).

balisage et sa structuration interne, de ces mêmes applications computationnelles[8]. Cette position est partagée par Habert, Nazarenko et Salem (1997), Biber, Conrad et Reppen (1998), Habert (2000) et Mayaffre (2005). Elle apparaît néanmoins trop étroitement liée à une linguistique textuelle, ou linguistique des genres, particulière, pour pouvoir s'appliquer au vaste mouvement international qui se range aujourd'hui sous la bannière des linguistiques de corpus. Cette définition, par trop computationnelle, est également trop spécifiquement liée à une théorie linguistique de type herméneutique, et au type de corpus qu'elle instrumentalise, pour être adéquate. Elle occulte en particulier le lien qui existe entre la linguistique de corpus contemporaine et le vaste mouvement de réflexion sur les données et les corpus dont est tissée l'histoire de la linguistique des deux derniers siècles, et plus généralement encore, nous venons de le rappeler, l'histoire des sciences modernes.

Reconnaître et assumer cette filiation épistémologique est particulièrement important pour comprendre les enjeux théoriques de la linguistique de corpus actuelle. Comme le souligne Nelson (1992 : 17) citant les corpus linguistiques anglais des 18ème et 19ème siècles, la notion de corpus correspond à une très longue tradition en linguistique. Nelson, père du premier, et longtemps plus important corpus informatisé de l'anglais (le *Brown Corpus*), parlant dans la section 'Language corpora, a Historical Conspectus' de la fameuse conférence Nobel, qui en 1991 devait porter sur le devant de la scène la ***linguistique de corpus*** (Svartvik, 1992), précise ainsi que la notion de corpus n'est tributaire d'aucune technologie particulière : 'I will confine myself to corpora accumulated B.C., *i.e*. before the use of computers […] Some seem to believe that there were no corpora before that. The truth is that many important corpora of English were assembled long before the computer was invented'.

C'est cette approche que je retiens ici : la linguistique de corpus, loin de constituer un courant nouveau, apparaît en science du langage comme une orientation ancienne voire très ancienne, même s'il est vrai qu'elle ne fait retour en pleine lumière que nimbée de l'aura de technologies et d'outils très sophistiqués. Il faut donc, pour en comprendre les enjeux, faire à présent retour sur son histoire en grammaire et philologie.

[8]'Convenons d'une définition positive : Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de ' nettoyage ', leur codage, leur étiquetage ; enfin, la structuration même du corpus. […] Notre définition suppose qu'un corpus n'est pas un corpus de mots (cf. à l'inverse le projet européen Paroles) ; ni un corpus d'attestations ou d'exemples (comme Frantext, dès lors qu'on n'a pas accès aux textes-sources) ; ni un corpus de fragments (comme le British National Corpus, qui ne contient aucun texte complet, mais un échantillonnage).' Rastier (2005).

2. Linguistique et corpus

De façon générale, on peut dire que depuis l'origine, grammaire, philologie et linguistique sont fondées sur ***l'observation de faits de langue***. Ces faits sont organisés comme des compendiums, ensembles plus ou moins vastes de données et d'exemples formant ce que Auroux (1998) appelle ***un observatoire***. Les étiqueter tous corpus, comme fait Mellet (2002) est sans doute un peu rapide et il faudrait distinguer plus précisément ce qui ressortit à la base de données, voire à la banque de données quand l'objet est très vaste et particulièrement hétérogène, et ce qui ressortit au corpus[9].

2.1 Faits de langue, données linguistiques, corpus

Bases et banques de données sont des ensembles de données, souvent composites, agrégées sans autre objectif explicité que quantitatif et de documentation empirique. Les bases de données sonores modernes comprennent ainsi des éléments, de types et de statuts très divers, voire hétéroclites (émissions radiodiffusées ou télévisées, interviews, discours publics, enregistrement ethnographiques, enquêtes phonologiques etc.). Comme le note Blanche-Benveniste (2004), elles apparaissent dès le début du 20ème siècle avec les premières techniques d'enregistrement phonographique[10] et se multiplient très récemment avec le développement des archives sonores[11]. Elles sont aujourd'hui potentiellement surabondantes, ne nécessitant qu'un travail de collecte et d'établissement des métadonnées (*Cf*. la mise en ligne des archives de l'INA). Enfin, banques et bases de données sonores sont peu, ou pas, tributaires d'hypothèses linguistiques et phonologiques précises et ne peuvent servir qu'à des fins documentaires externes à l'analyse phonologique proprement dite. Lorsque l'on dispose d'une base de données sonores, le corpus correspondant reste à construire[12].

---

[9] Comme me le rappelle un relecteur anonyme, les bases de données constituent ce que plus classiquement on appelle des dictionnaires. La convergence récente entre dictionnaires et corpus (*Cf*. par exemple le COBUILD) appelle des commentaires dont je n'ai pas la place ici.

[10] Les enquêtes de Ferdinand Brunot et les Archives de la Parole qu'il fonde à l'Université de Paris en 1911 en constituent un bon exemple : discours radiodiffusés d'hommes politiques, interviews informels d'ouvriers parisiens, enquêtes ethnographiques et dialectologiques, chansons et littérature orale etc. *Cf.* Cordereix (2001) et Veken (1984).

[11] On trouve de nombreuses bases de données sonores sur le site de European Language Ressource Association (ELRA). Voir également le site du programme européen CLARIN. Un exemple de base de données orales, très composite, est fourni pour le français, l'italien, le portugais et l'espagnol par C-ORAL-ROM (coord. E. Cresti et M. Moneglia). On en trouvera une recension critique dans Romano (à paraître).

[12] Un exemple peut être donné avec mon corpus ‘'Hommes Politiques’ (HPOL, *Cf*. http://www.projet-pfc.net/?accueil:hpol). La base de données sonores m'a été fournie par le Département audio-visuel de la

Le corpus au contraire, se présente comme un objet explicitement construit, tributaire d'une méthodologie et d'un cadre théorique, fut-il implicite, dont la construction répond à un objectif documentaire très précis et affiché[13]. Halliday (1992) date leur apparition des années cinquante, au moment où les magnétophones plus ou moins portables permettent aux linguistes d'effectuer facilement leurs propres enregistrements. Cette datation est sans doute beaucoup trop restrictive, car on ne peut nier que l'enquête sur la prononciation du français de Martinet (1945), ou même que les grandes cacologies des 18ème et 19ème siècles, qui toutes incorporent des descriptions très précises et très nombreuses d'usages réels, ne constituent des corpus. Si le corpus phonologique contient des éléments permettant de documenter de façon fiable des usages, quelle que soit la forme physique de l'enregistrement (graphique, phonique, API, numérique etc.), et si le corpus phonologique est bien construit, au sein d'un cadre théorique précis, en fonction d'hypothèses à éprouver quant à la phonologie de la langue considérée, alors ces derniers exemples sont bien des corpus phonologiques, de même d'ailleurs que les grammaires de Baïf (1574), de Meigret (1542), ou la recension de Thurot (1881-1883). La preuve, tout à fait directe, en est fournie par la capacité qu'ont toujours eu les phonologues, travaillant ces compendiums en prenant en compte les cadres théoriques (parfois très idiosyncratiques comme chez Baïf) qui les ont vus naître, de construire des analyses phonologiques très élaborées des états de langue visés[14].

Le corpus, spécialement phonologique, est donc un objet défini par le linguiste et construit par lui en fonction d'hypothèses et d'objectifs précis, ne serait-ce que la recollection

---

Bibliothèque Nationale de France. Le corpus que j'ai construit avec la collaboration d'Antonio Balvet et d'Atanas Tchobanov incorpore des hypothèses très précises sur la synchronie et la diachronie de la liaison. Ces hypothèses permettent de tester rétrospectivement les analyses classiques de la liaison de Fouché et Grammont à Delattre et Encrevé. Le choix des segments intégralement transcrits, les catégories de transcription (pauses, hésitations, consonne tenue ou non), le codage contextuel en 21 catégories du contexte droit et autant du contexte gauche transforme la base de données en corpus instrumental. Pour plus de précision, *Cf.* Laks (2007).

[13] Je suis parfaitement d'accord avec Mellet (2002: 8) lorsqu'elle écrit : 'Nous sommes déjà loin de la conception naïve qui prévalait encore il y a une dizaine d'années, selon laquelle la constitution d'un corpus de données attestées devait permettre d'éviter toute manipulation artificielle de la réalité : le travail sur corpus faisait partie des *behavioral, natural methods.* Et encore : *The data of a corpus, more thoroughly than we have grown to expect in linguistics, are independent of the tenets of the theory they are required to test.* Bien au contraire, le constat s'est partout imposé que le corpus est un objet construit, que ce soit à travers l'effacement symbolique de tout ce qu'il ne contient pas […] ou que ce soit à travers la structuration, l'organisation des données retenues, voire leur enrichissement au moyen des procédures du balisage et de l'étiquetage.'.

[14] On pense aux travaux de Morin qui, travaillant ce que nous appelons l'archive phonologique, a produit de très nombreuses analyses synchroniques et diachroniques de la phonologie du français. *Cf.* par exemple Morin (2005a; b; c), ou encore Laks (2005a).

des usages attestées. A la différence des bases de données il est donc directement exploitable pour l'analyse, mais son utilisation est contrainte par la méthodologie et les actes de la recherche qui l'ont construit. Le corpus PFC (http://www.projet-pfc.net/?accueil:intro), construit sur la base d'une enquête systématique des usages contemporains, qui incorpore dans ses méthodologies de recollection, de transcription, d'étiquetage et de balisage, les principales hypothèses classiques de la phonologie de E muet et de la liaison en français en constitue un bon exemple. Sa profondeur historique au plan conceptuel permet d'éprouver les hypothèses anciennes, l'explicitation et la précision de son codage, les outils qu'il incorpore, permettent de construire des analyses innovantes (*Cf*. Durand, Laks et Lyche 2002, 2005; Durand et Lyche, ce volume).

Je reviendrai sur cette définition instrumentale des corpus sonores ci-dessous, pour poursuivre mon analyse historique de la notion de corpus en phonologie, je me conforme à l'usage, en quelque sorte pré-théorique du terme corpus, tel que le définit par exemple Francis (1992: 17) 'a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis'. De ce point de vue, grammaire et philologie des siècles passés regorgent de corpus.

2.2 Corpus grammaticaux et corpus philologiques

Certes, on ne saurait dire de la philologie classique ou de la grammaire historique et comparée des deux siècles précédents qu'elles apparaissent immédiatement comme des ***sciences du corpus***, même au sens faible que je viens de rappeler. Y font le plus souvent défaut cet exposé préliminaire des modes de collecte et d'établissement des données, cette présentation quantitative et qualitative des data et en définitive, cette mise au premier plan de la base de donnée sur laquelle s'établira l'analyse, auxquels se reconnaît aujourd'hui une analyse dite de corpus. Mais, si l'on y réfléchit un peu plus, par décision scientifique ou par nécessité empirique, ***toutes ces dimensions sont déjà implicitement présentes*** dans la linguistique du 19ème siècle. J'en prendrai quelques exemples.

Pour les langues anciennes, par exemple, le corpus est pensé clos, stable, parfaitement connu et établi: c'est une archive publique partagée[15]. Il s'agit certes d'un corpus de textes et documents écrits, même si l'on peut, comme pour le latin vulgaire, voire tardif, chercher à

[15] Comme me le rappelle un relecteur, Panini avait sans aucun doute appris par cœur l'ensemble du Véda qui constituait son corpus de référence.

l'enrichir en le complétant d'attestations d'usages réels de la langue vivante (Väänänen, 1962). Il reste que le corpus latin, dans ses diverses circonscriptions stylistiques (théâtre, histoire, épopée, patristique etc.), temporelles (archaïque, classique, tardif), voir géographiques, est très largement établi et stabilisé [16]. Comme pour toutes les langues anciennes, hébreu biblique (dit de Tibériade), arabe classique (littéraire ou coranique), grec (hellénistique ou biblique), sanskrit (védique) etc., l'établissement des données, peut être critiqué, amélioré, contesté à la marge, mais il fait l'objet d'un très large consensus.

La moindre conséquence de ce consensus n'est pas que ces corpus puissent faire l'objet de publications, éventuellement critiques, mais autonomes, sur tous types de supports : forts volumes papier hier, CD-ROM aujourd'hui. La modernité en la matière ne tient donc à rien d'autre que la technologie qui permet que se constituent aujourd'hui sur internet de vastes banques de données philologiques pour les études classiques[17]. Ces banques de données et les différents outils informatiques qui leur sont associés, constituent donc, mais explicitement cette fois, la philologie actuelle comme une ***science des corpus***[18].

Il n'est pas nécessaire, pour se convaincre à nouveau de l'ancienneté des travaux linguistiques sur corpus, de faire la même analyse concernant le comparatisme. Il suffit de renvoyer à la forme même du compendium de Schleicher (1861), aux travaux monumentaux de Bopp (1889) ou à ceux des néogrammairiens qui font émerger telle loi phonétique d'une ***observation systématique des correspondances***, et adossent sa formulation précise à une vérification minutieuse de son application dans les corpus parfaitement reçus. Ceci leur permet de traquer et réduire les éventuelles exceptions, dont on sait qu'elles constituent à la fois un problème épineux et le moteur même de la recherche comparatiste[19] (*Cf*. Osthoff et Brugmann, 1874).

---

[16] S'agissant du latin, même la grammaire générative de stricte obédience travaille aujourd'hui ces corpus. Pour une analyse *Cf.* Bortolussi (2008)

[17]Le XIVème Congrès international de l'Association Guillaume Budé (Limoges, 1998) a été consacré en grande partie à ces questions, *Cf.* Levet (2000) ou Beguin (1998). Ce dernier fournit sur son site une recension très complète des corpus et des ressources électroniques disponibles pour les études classiques : http://www.antiquite.ens.fr/pdf/IntroAntiquite2002.pdf.

[18] On peut remarquer au passage que ceci est également vrai des travaux d'histoire des langues. Ainsi, par exemple, les travaux d'ancien et de moyen français s'appuient fortement sur l'ensemble des ressources numériques (banques de textes, corpus, dictionnaires, lexiques, concordanciers) développés par le laboratoire ATILF de Nancy (http://www.atilf.fr/) et aujourd'hui maintenues par le Centre National de Ressources Textuelles et Lexicales (http://www.cnrtl.fr/).On peut également consulter, pour ce qui concerne les ressources linguistiques multilingues, le projet européen CLARIN et son site (http://www.mpi.nl/clarin/).

[19] On pense bien entendu aux exceptions et divers problèmes posés par la première mutation consonantique telle qu'elle est traitée par la loi de Grimm. La loi de Verner en réduit le plus grand nombre. De

La méthode comparatiste induit des effets méthodologiques sur l'ensemble de la philologie classique, effets clairement soulignés par Saint-Gérand (2001) 'C'est très probablement de cette époque [1860] que date en France l'avènement de la seconde philologie française. Gaston Meyer, Paul Paris en modèlent la forme scientifique en proclamant la double nécessité de manipuler désormais ***des corpus de documentation aussi complets et fiables que possible***, et d'employer une méthode historico-comparative qui respecte la régulation du développement des langues. Ainsi sera désormais délivré l'enseignement de l'*École Pratique des Hautes Études* et de l'*École des Chartes*. C'est la mort de la philologie impressionniste.'

S'agissant enfin des travaux anciens et modernes d'inspiration plus typologique, Hagège (2002) note justement que dès avant les grandes classifications modernes issues de la comparaison systématique de nombreuses descriptions de langues (Greenberg, 1963), au milieu et à la fin du 19ème siècle, la typologie linguistique travaille déjà des données descriptives formatées sous forme de corpus de référence. Dans la mouvance comparatiste et typologique, les travaux récents de Comrie (1981), Comrie, Matthews et Polinsky (2003), Comrie *et al*. (2005) attestent du même tournant que permet l'utilisation de technologies informatiques modernes, tournant marqué par l'importance croissante des bases de données linguistiques qui définissent à nouveau la typologie comme une ***science de corpus***.

2.3 Corpus anciens, corpus modernes

A jeter un regard rétrospectif sur nos disciplines se découvre ainsi un vaste panorama dans lequel l'utilisation systématique de descriptions ordonnées sous la forme de base de données et de corpus de référence constitue une pratique ancienne et récurrente, bien établie méthodologiquement et pratiquement en linguistique et en philologie. Il n'y a donc d'anachronisme que de façade à annexer ces pratiques à ***une linguistique de corpus*** apparue en tant que courant de recherche en sciences du langage de façon bien plus récente. Nous nous séparons ainsi à nouveau du courant herméneutique de la linguistique textuelle (Habert,

---

façon générale, le débat néogrammairien se nourrit pour une part de l'exhibition d'exceptions qu'il s'agit d'expliquer et de motiver afin de préserver les caractères aveugles, automatiques, de diffusion immédiate et sans exceptions des lois phonétiques *cf.* Labov(1981). Ceci suppose une clôture, une stabilité et accord public sur les corpus de référence.

Nazarenko et Salem, 1997 ; Rastier, 2005), qui fait du lien avec l'ingénierie informatique et le traitement automatique des langues la pierre de touche des linguistiques de corpus au sens contemporain. Il importe de reconnaître que le rapport aux données descriptives regroupées en vastes ensembles stabilisées, publiques et partagées ne date ni de l'apparition de l'ordinateur, ni de celle des outils sophistiqués de traitement automatique de ces bases.

3. Le 20ème siècle, la linguistique et les corpus

Cependant, si l'on tourne à présent son regard vers la linguistique contemporaine, celle du 20ème siècle, la situation est assez paradoxale. Pendant la large période de temps qui a vu la domination du paradigme générativiste sur les sciences du langage, la linguistique de corpus est restée peu ou pas visible[20]. Méprisée pour son empirisme et son agnosticisme théorique, vilipendée pour son manque d'intérêt et son incapacité explicative[21], rejetée pour son rapport quasi incestueux à un structuralisme et un behaviorisme caricaturé comme antédiluviens (*cf.* infra), elle est restée confinée à quelques secteurs assez marginaux. Les continuateurs de Harris[22] d'abord qui, avec Gross[23] ont continué à défendre une approche lexique-grammaire fondée sur une recension précise des constructions possibles et ont développé de nombreuses bases de données et dictionnaires électroniques syntactico-lexicologiques. Les sociolinguistiques (Encrevé, 1976; Labov, 1976 ; 1979; Laks, 1992), les dialectologies modernes et toutes les linguistiques de l'oral ensuite (Blanche-Benveniste, 1997; 2000; Gadet, 1989; 2003) qui ont défendu la primauté des usages réels et leur documentation sur une

[20] Aarts (2000: 5) ouvre son article par le rapport de son entrevue avec Chomsky à MIT : 'What is your view of modern corpus linguistics?' Noam Chomsky: 'It doesn't exist.'

[21] On ne peut ici s'empêcher de penser à la célèbre saillie de Fillmore (1992: 35) 'Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like (There isn't anybody exactly like this, but there are some approximations.). Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations.). These two don't speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, 'Why should I think that what you tell me is true?', and the armchair linguist says to the corpus linguist, 'Why should I think that what you tell me is interesting?).

[22] On sait l'importance théorique première que jouent les données et l'analyse quantitative de leur structure (dépendances, colocations, implications syntagmatiques, relation paradigmatiques privilégiées etc.) dans l'œuvre de Harris (*Cf.* par exemple Nevin (2002)). Concernant le rapport aux données et la conception du corpus chez Harris *Cf.* également Goldsmith (2005).

[23] *Cf.* Gross (1976) et spécialement Gross et Perrin (1989).

linguistique de l'exemple construit par et pour le linguiste. La linguistique descriptive enfin dont le statut a radicalement changé dès qu'elle s'est trouvée intimement liée au traitement automatique et aux industries de la langue nouvellement apparus[24].

Il y aurait ainsi une histoire détaillée à faire des trente dernières années du 20ème siècle, où, dans les marges et les friches abandonnées par le formalisme chomskyen et ses hypothèses linguistico-cognitives, ont silencieusement prospéré la linguistique de corpus et nombre d'alternatives qui apparaissent à présent en pleine lumière.

Cette période où le corpus est resté un instrument marginal a pourtant été très fructueuse. De très grosses enquêtes résultant en corpus de référence ont été conduites, de très grosses bases de données et de très gros corpus textuels ont été construits. Pour le français et les premières je renvoie aux inventaires compilés par Cappeau[25], pour les seconds à celui de Habert, Nazarenko et Salem (1997). On dispose ainsi aujourd'hui de très gros ensembles de données linguistiques. Par exemple pour les corpus textuels de l'anglais on peut citer les corpus suivants : Brown Corpus (1 million de mots), Lancaster-Oslo-Bergen (LOB, 1 million de mots), London-Lund (435 000 mots), Helsinki (diachronie de l'anglais, 1,5 million de mots), Archer (diachronie de l'anglais, 1,7 million de mots), British National Corpus (100 millions de mots) et plus récemment COBUILD Corpus (56 millions de mots). Pour le français, outre Frantext (4000 textes, 210 millions d'occurrences) et TLFI (100 000 mots) on peut citer les corpus oraux d'ESLO 1 et 2, du GARS (Corpus de Référence du Français Parlé, 440 000 mots) ou de PFC (900 000 mots). Pour une présentation plus détaillée des corpus français, on se reportera à Cappeau et Gadet (2007a).

3.1 La linguistique structurale et les corpus

Sans jamais s'éteindre au cours des derniers siècles et en développant même singulièrement sa couverture empirique, la linguistique de corpus fait donc un retour remarqué à la fin des années quatre-vingt dans les wagons de l'ingénierie linguistique. D'aucuns ont vu dans son maintien au temps de l'impérialisme générativiste un simple effet de la permanence épistémologique de l'empirisme anglo-saxon, et dans son retour sur le devant

[24] Pour l'anglais, et pour la linguistique descriptive dans son lien au traitement automatique des langues, il faut souligner le rôle central qu'a joué Jan Svartvik créateur du London-Lund corpus. *Cf*. par exemple Svartvik, Aijmer et Altenberg (1991).

[25] Les inventaires compilés par Paul Cappeau pour l'Observatoire des Pratiques linguistiques de la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF) sont consultables sur le site : http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire.pdf. *Cf*. également Cappeau et Gadet (2007)

de la scène linguistique la simple résurgence d'un structuralisme par ailleurs dépassé[26]. Tant pour ce qui concerne l'empirisme, que le structuralisme ou le behaviorisme, une telle appréciation méconnait très profondément ce qu'a été l'apport théorique fondamental de ces courants à la construction des sciences du langage et, reprenant la doxa chomskyenne la plus usée, empêche de comprendre ce qui constitue encore aujourd'hui leur actualité scientifique ainsi que la pertinence épistémologique des corpus[27]. Un retour rapide sur l'histoire de la linguistique moderne et les corpus est donc à nouveau nécessaire.

La linguistique moderne se dégage progressivement de la Grammaire, qu'elle soit normative ou non, en opposant à une rhétorique de l'*exemplum*[28], une science du *datum*, c'est à dire en définitive en proposant une nouvelle ***approche empirique des faits de langue***. C'est en effet sur la base d'un empirisme affirmé et d'une attention particulière aux faits empiriques de langue saisis comme des usages, c'est-à-dire comme des produits sociaux et culturels que Saussure et Whitney arrachent la linguistique à la grammaire et constituent la science du langage. La linguistique écrit Saussure dans ses notes (Saussure, 2001 : 273) est d'abord l'étude de la parole : '[la linguistique] comporte deux parties: l'une qui est plus près de la langue, dépôt passif, l'autre qui est plus près de la parole, force active et véritable origine des phénomènes qui s'aperçoivent ensuite peu à peu dans l'autre moitié du langage. […] En résumé 1° non ce qui est individuel mais ce qui est consacré par l'usage social, remplissant ainsi les conditions qui font qu'une chose est linguistique. 2° non nécessairement ce qui est écrit mais de préférence ce qui est parlé; 3° non dans un but normatif et pour donner des

[26] 'Faut-il voir dans cet engouement actuel pour les corpus le retour aux débuts de la linguistique structurale américaine des années cinquante? Après l'accent chomskyen sur la formalisation et l'intuition du locuteur natif, la revanche de l'empirisme? Le découragement serait de mise s'il y avait effectivement piétinement et ressassement. Or, l'étude des origines de ces travaux le montre, ce qui domine, ce sont les discontinuités qui l'emportent, ainsi que la diversité, voire l'éclatement des horizons théoriques et des réalisations pratiques' (Habert, Nazarenko et Salem, 1997: 8).

[27] J'ai montré ailleurs (Laks, 1996) quelle était l'actualité scientifique du néo structuralisme et d'un certain néo behaviorisme illustrés par les modèles connexionnistes, les grammaires d'usage ou les grammaires cognitives. Je n'y reviens pas ici. *Cf*. également (Maniglier, 2007).

[28] Parce qu'il y a toujours, qu'on le veuille ou non, une relation très forte entre *exemplum* et *imitatio*, toute rhétorique de l'exemple se résout nécessairement dans une approche normative. Chevalier (2007: 153) est assez clair sur ce point : ' Toute grammaire étant une régularisation de l'écrit tend à ramener toutes les formes de langage à un modèle qui a varié avec les époques et qui a ses finalités propres […] [la place et le rôle des exemples sont] la transposition en grammaire d'un dispositif de l'art oratoire qu'Aristote a décrit dans sa Poétique'

règles de bonne expression mais 4° enfin, avec le but de généraliser les observations, d'arriver à une théorie applicable aux langues.'[29]

Cette nouvelle approche empirique des faits de langue, si proche en définitive d'une linguistique des corpus, est partagée par les fondateurs de la linguistique américaine. Franz Boas comme Edward Sapir la construisent dans une relation étroite à l'enquête anthropologique et culturelle et il est difficile de contester que leurs travaux sur le Dakota, l'Esquimau central, le Paiute du Sud, ou encore le Navaho[30], ne soient fondés sur des corpus linguistiques méthodologiquement exemplaires. Au-delà même du corpus linguistique *per se*, l'ethnolinguistique américaine invente également la base de données linguistique, ethnologique et textuelle, vaste ensemble de documents textuels, iconographiques ou linguistiques de diverses formes et statuts permettant de documenter une culture et une langue[31].

De ce côté-ci de l'atlantique, il sera difficile aussi contester que la géolinguistique et ses atlas (Gilliéron et Edmont, 1902-1912) ne soient pas fondés sur des enquêtes et des techniques de recollection permettant de construire de vastes corpus lexicaux.[32] S'agissant du structuralisme européen, on ne saurait non plus contester que l'imposant travail de Troubetzkoy (1939) ne soit construit sur une sorte de corpus regroupant la description du système phonologique d'un très grand nombre de langues[33] ou que Martinet ne définisse et n'applique ***l'enquête phonologique*** qui permet de construire de véritables corpus linguistiques (Martinet, 1945; 1956)[34]. Pour la même période structuraliste on pourra encore citer la grande enquête de Gougenheim (Gougenheim *et al.*, 1956) visant à dégager un corpus de français dit fondamental, ou encore de façon plus récente, les travaux quantitatifs de Catach (Catach,

[29] On sait que la fameuse phrase 'Il faut considérer la langue en elle-même et pour elle-même' qui a autorisé tant de dérives grammaticales abstraites est totalement apocryphe. Sur la primauté de la linguistique de la parole *Cf*. Bouquet (1997).

[30] *Cf*. Boas (1964); Boas et Delora (1941); Sapir et Hoijer (1967); Sapir et Bright (1992); Sapir et Hoijer (1942).

[31] Par exemple Sapir (1939). On sait le rôle central qu'a joué à cet égard la Smithsonian Institution et les rapports d'enquêtes ethnographiques et linguistiques qu'elle délivrait chaque année au Président des Etats-Unis.

[32] Que ces corpus, augmentés de données nouvelles ou non, puissent se voir aisément traduits dans les formats des corpus et bases de données électroniques renforce notre analyse. *Cf*. par exemple la base de données THESOC maintenue à l'université de Nice : http://thesaurus.unice.fr/.

[33] Comme l'atteste l'édition récente de sa correspondance, Troubetzkoy entretenait des contacts épistolaires avec de très nombreux phonologues de par le monde afin de compléter et d'affiner sa base de données sur les systèmes phonologiques attestés. *Cf.* Troubezkoy (2006).

[34] Les travaux de Martinet et Walter (1973) ont été poursuivi par Walter et les membres de l'école fonctionnaliste, dans de nombreux ouvrages. Entre autres : Walter (1977; 1982).

1984; Catach et Jejcic, 1984) construits eux-aussi sur de véritables bases et banques de données[35].

Bref, sans qu'il soit nécessaire d'argumenter plus avant, on admettra que la linguistique structurale dès l'origine, s'est toujours vue et pensée comme ***une linguistique de corpus*** au plein sens du terme, et qu'elle a trouvé dans ***la construction du datum*** linguistique, sa taxinomie raisonnée et sa modélisation systémique, le terrain même de son accomplissement. Il n'est donc pas historiquement paradoxal de défendre que ***la linguistique de corpus commence bien avant la linguistique de corpus***, comme le note avec ironie Mair (1992: 98) dans son commentaire sur Chafe (1992). 'Wallace Chafe's paper conveys a comforting message to us: corpus linguistics is normal linguistics, and any other kind of linguistics is something odd or special that needs to be justified. For like Molière's Monsieur Jourdain, who late discovered that he had the ability to talk in prose, Wallace Chafe was not aware that he was a corpus linguist until this fact was pointed out to him after a distinguished career in the field'.

3.2 Corpus et intuitions

Dans cette continuité, la théorie chomskyenne fait rupture. Ce qui apparaît comme spécifique de la grammaire générative depuis le tournant cognitiviste de 1965, ce n'est pas tant le recours à l'intuition du sujet parlant et aux jugements de grammaticalité, pratiques partagées par les linguistes et les grammairiens depuis la nuit des temps, que le recours ***unique*** à ces données, le refus de baser l'analyse sur l'observation de l'usage comme source première, et, en définitive, ***l'instrumentalisation de ce type particulier de données*** dans des modes de raisonnement et des argumentaires très spécifiques. Comme l'a très bien dit Milner (1989) le recours à l'intuition ne vise pas tant à définir le possible de la langue, ce qui après tout consisterait à construire un corpus d'un type spécifique, qu'à cerner son ***impossible***, car c'est précisément cet impossible qui constitue le principe argumentatif premier du raisonnement chomskyen. Intuitions et jugements de grammaticalité permettent, pour les générativistes, de faire le départ entre ce que les langues font, et ce qu'elles ne font pas. C'est cet argument négatif qui permet de projeter la grammaticalité attestée par les jugements (dans différentes langues éventuellement) sur un universel présupposé. Comme l'a souvent défendu

[35] Je rappelle que le premier dictionnaire inverse et le premier dictionnaire statistique du français, produits qui ressortissent sans conteste pour Habert, Nazarenko et Salem (1997) à la linguistique de corpus, datent de la même époque et sont dus à Alphonse Juilland (Juilland, 1965; Juilland, Brodin et Davidovich, 1970).

Chomsky lui-même, il s'agit d'un raisonnement parfaitement abductif conduit dans le cadre d'une théorie où la modélisation est fondamentalement ***sous-déterminée par les données observables*** (Chomsky, 1966).

Ainsi ce qui est en jeu dans l'utilisation des intuitions du locuteur, c'est beaucoup moins l'adhésion ou le rejet d'un certain behaviorisme qu'***une conception de la démonstration et de l'argumentation en linguistique***. Chomsky lui-même, et toute la grammaire générative avec lui, sont très clairs sur ce point. Ce qui justifie le recours unique aux intuitions et aux jugements de grammaticalité c'est la conception même de la grammaire qui est la sienne : dès que l'on sépare définitivement le langage externe manifesté (*E-language*) du langage interne actualisant la faculté innée de langage (*I-langage*), le recours aux données externes n'a plus d'intérêt[36]. Ce qui intéresse le linguiste, ce ne sont pas les manifestations avérées de la grammaire, c'est la grammaire elle-même en tant qu'elle constitue un dispositif mental abstrait permettant d'éclairer le fonctionnement cognitif humain en général ainsi que ses préconditions innées[37]. Ce dispositif interne, seule l'intuition du sujet parlant permet de l'atteindre, définissant ainsi la grammaire générative comme une approche réaliste et cartésienne du fonctionnement mental (Laks, 2005a).

Dans un cadre structuraliste, il en va tout autrement. Si les intuitions peuvent être utiles comme jalons dans la recherche de régularités, ***seul le corpus de faits de langue permet d'éprouver la fiabilité d'un modèle*** et de projeter la construction théorique sur le système linguistique lui-même[38]. L'objet du linguiste reste ***le système de la langue*** et son objectif la modélisation de sa structure. Or, en toute rigueur saussurienne, cet objet n'est complet et stable dans le cerveau d'aucun individu, car contrairement à la ***parole***, la ***langue*** n'est pas un objet individuel. ***Trésor social commun partagé***, c'est un produit ***culturel***, ***contractuel*** et

[36] 'We want to stress the significance of seeing language as an internal fact about speakers, a form of knowledge, or I-language as Chomsky has called it, as opposed to its external manifestation as utterances, texts, sets of sentences, or social convention – E-language in Chomsky's terms. […] Actual practice often devotes more significance to these external phenomena than might be scientifically justified' (Anderson et Lightfoot, 2002: XIV).

[37] 'The language faculty has an initial state, genetically determined; in the normal course of development it passes through a series of states in early childhood, reaching a relatively stable steady state that undergoes little subsequent change apart from the lexicon. To a good first approximation, the initial state appears to be uniform for the species. (…) we call the theory of the state attained its *grammar* and the theory of the initial state Universal Grammar (UG). (…) The initial state is in crucial respects a special characteristic of humans, with properties that appear to be unusual in the biological world. (…) When we say that Jones has the language L, we now mean that Jones's language faculty is in the state L (…)To distinguish this concept of language from others, let us refer to it as *I-language*, where I is to suggest 'internal', individual', and intensional'. (Chomsky, 1995: 18-19).

[38] 'The aim of the corpus is not to limit the data to an allegedly representative sample but to provide a framework to find out what questions should be asked about language in general' (Mair, 1992: 99).

consensuel qui n'est complet et systémique que dans la ***masse parlante***, car il fait corps avec elle, et change avec elle[39]. L'attention toute particulière que portent les structuralistes aux faits de langue, à leur collecte, à leur accumulation sous forme de corpus organisés et à leur analyse comme faits d'usage, ne découle donc pas d'un penchant particulier pour l'entomologie, comme raille Chomsky à propos de la sociolinguistique labovienne[40], mais constitue ***une méthode scientifique*** qui découle directement, tout comme en grammaire générative, d'une théorie de la langue précise et articulée. Il en est de même du distributionnalisme et ses procédures de découvertes (tests paradigmatiques, permutations et commutations), si souvent moquées, qui renvoient à ***un mode argumentatif*** parfaitement cohérent avec cette théorie[41].

Ainsi, s'agissant du recours à l'intuition, Harris par exemple en conteste, non la pertinence ou l'intérêt de principe, mais ***le contenu informationnel trop pauvre***. Rappelant, contre les caricatures du distributionnalisme (*supra*), que l'objectif du linguiste reste une sorte de projection explicative sur l'ensemble de la langue et ne se limite donc pas à la description du corpus qu'il a sous les yeux, il écrit : 'The distributional investigations sketched above are carried out by recording utterances (as stretches of changing sound) and comparing them for partial similarities. We do not ask a speaker whether his language contains certain elements or whether they have certain dependences or substitutabilities. Even though his 'speaking habits' yield regular utterances, they are not sufficiently close to all the distributional details, nor is the speaker sufficiently aware of them. Hence we cannot directly investigate the rules of the 'language' via some system of habits or some neurological machine that generates all the utterances of the language. We have to investigate some actual corpus of utterances and derive there from such regularities as would have generated these utterances - and would presumably generate other utterances of the language than the ones in our corpus. Statements about distribution are always made on the basis of a corpus of occurring utterances; one hopes

[39] Il n'est pas nécessaire de commenter encore une fois de façon détaillée les grandes articulations de la construction saussurienne. Néanmoins *Cf.* Maniglier (2007) pour d'utiles mises au point et une mise en perspective, anthropologique, philosophique et culturelle.

[40] [Chomsky] 'Sociolinguistics is, I suppose, a discipline that seeks to apply principles of sociology to the study of language; but I suspect that it can draw little from sociology, and I wonder whether it is likely to contribute much to it. [Ronat]: In general one links a social class to a set of linguistic forms in a manner that is almost bi-unique. [Chomsky]:You can collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so' (Chomsky et Ronat, 1979: 56-58).

[41] De ces deux aspects, Harris a donné un exposé détaillé dans *Methods in structural linguistics* (Harris, 1951).

that these statements will also apply to other utterances which may occur.' (Harris, 1954: 146).

On voit donc que le structuralisme n'a jamais prohibé le recours à l'intuition[42]. Il a seulement défendu qu'il ne saurait constituer le seul et unique *datum* de la linguistique, car la reconstruction du système de la langue exige une observation des faits linguistiques beaucoup plus détaillée que ce qu'il permet. Certains structuralistes ont ajouté des données d'intuition (positive) à leur data, d'autres s'y sont refusés[43], mais pour des raisons tout autres que l'adhésion ou le refus de la compétence chomskyenne idéalisée.

Loin donc de constituer un débat secondaire marqué par les apories d'un structuralisme borné et d'un behaviorisme obsolète le débat qui oppose intuitions et corpus de faits linguistiques reste ainsi d'une actualité brûlante en ce qu'il porte *in fine* sur ce qui constitue un argument, ou une preuve, dans un raisonnement linguistique et sur ce qui permet, au-delà de l'observation conjoncturelle, d'en projeter les conclusions sur l'ensemble de la langue. C'est ce que souligne à juste titre Sampson (1975: 73) ‘Someone wishing to defend the use of intuition may object that all linguistics, including the pre-Chomskyan descriptivists as well as those, like the present author, who aims to reconcile Chomskyan linguistics with behaviourism, have in fact relied heavily on intuition. This objection again misses its mark through failure to appreciate how science works. We do not care where a scientist gets his theory from, only how he defends it against criticism. Any scientific theory is sure to incorporate many untested assumptions, guesses, and intuitive hunches of its creator. All that matters is that any feature of the theory which is doubted can be confirmed or refuted on empirical grounds'.

3.3 L'objet de la phonologie, les objets phonologiques

S’il est un domaine où le paradigme chomskyen et son approche des données a eu des conséquences profondes, c'est bien celui de la phonologie. Sans entreprendre ici une histoire de la phonologie pré et post chomskyenne (*Cf.* Goldsmith et Laks, 2005; Laks, 1997b; 2001) je voudrais seulement insister sur la rupture profonde introduite par la phonologie générative

[42] 'Harris’s view, from his earliest work through his final statements in the early 1990s, was that the best foundational chances for linguistics were to be found in establishing a science of EXTERNAL LINGUISTIC FACTS (such as corpora, though they would typically be augmented by other external facts, like speaker judgments), rather than a science of internalized speaker knowledge.' Goldsmith (2005: 720).

[43] Maurice Gross dont j'ai dit ci-dessus qu'il devait être considéré indubitablement comme un linguiste de corpus, travaillait ainsi sur des exemples forgés par le linguiste.

classique quant à la relation du phonologue à ses données telle qu'elle était pensée par le structuralisme.

La phonologie est en effet marquée par un singulier paradoxe, il n'existe pas d'intuitions phonologiques, ou si elles existent, elles ne sauraient avoir le même ***statut décisionnel et argumentatif*** que les intuitions syntaxiques : les intuitions phonologiques ne sont pas des jugements de grammaticalité, ***ce sont des appréciations des différents usages***[44]. Au phonologue, même génératif, il faut donc des données externes, et des observations précises d'usages pour construire une analyse. A contrario, l'un des rares phonologues génératifs qui avec beaucoup d'honnêteté scientifique a déclaré décrire sa propre compétence phonologique (Dell, 1973) n'a fait, on peut en juger aujourd'hui[45], que s'approcher d'une norme orthoépique moyenne, celle que Morin (2000) à baptisé Français de Référence[46].

De plus, la phonologie est toujours contrastive, d'une certaine manière. Il faut pour éprouver un modèle ou une proposition étudier un grand nombre de langues et de systèmes différents. Le rejet de l'enquête phonologique et de la méthodologie de recueil d'un *datum* précis a donc eu des conséquences très particulières en phonologie générative : les descriptions et les corpus anciens, ou encore les collections d'exemples des analyses classiques, mêmes orthoépiques, ont servi de données de départ.

Toute la première période de la phonologie générative est ainsi marquée, sous couvert d'avancée scientifique et d'innovation formelle, par la simple reprise d'analyses anciennes et de descriptions classiques reformulées dans un cadre dérivationnel. Les langues amérindiennes et singulièrement le corpus des publications de *International Journal of American Linguistics* constituent la source presque inépuisable de données, de descriptions et d'analyses, peu, mal ou pas du tout citées, qu'il suffit alors de réécrire dans un format

[44] Je prends ici 'intuition' au sens étroit, technique et conceptuel qu'il a en grammaire générative. Nul ne doute que la forme sonore puisse faire l'objet d'intuitions au sens large, de jugements sur les usages etc., mais il ne s'agit pas alors d'intuition sur le possible ou l'impossible de la structure.

[45] Pour une analyse comparative des différents corpus, d'observation ou d'intuition, concernant le comportement de schwa en français, *Cf*. Geerts (2007).

[46] Ce qui après tout est bien normal puisque les éléments autobiographiques qu'il fournit montrent qu'il a eu un cursus scolaire complet jusqu'aux études universitaires incluses, et qu'il a donc baigné pendant une vingtaine d'années au moins, dans un milieu où s'acquièrent et s'intériorisent, avec la littéracie la plus accomplie, les normes linguistiques les plus précises, celles en gros que Fouché (1959) s'est attaché à codifier.

dérivationnel génératif à la *Sound Pattern of English* (SPE) pour faire œuvre de phonologie contemporaine[47].

La phonologie du français, singulièrement celle de la liaison en offre une illustration encore plus détaillée (*Cf.* Durand et Lyche, ce volume). Avec Schane (1968) la phonologie générative proposait en effet une reformulation dans un cadre dérivationnel abstrait des exemples codifiés par Fouché (1959) et de l'approche ancienne centrée sur le concept de consonnes latentes initialement due à Pichon (1938). Dans un modèle théorique fondamentalement sous-déterminé par les données comme nous l'avons rappelé ci-dessus, le parallélisme ***purement formel*** entre la suppression d'une consonne devant consonne et la chute d'une voyelle devant voyelle conduisait ainsi Schane à la création d'une curieuse chimère, vite baptisée troncation, que rien, ni dans l'histoire de la langue ni dans sa phonotactique synchronique, ne permettait de fonder. Il faudra quelques années, et quelques dizaines de travaux, pour la déconstruire.

La simple réécriture formelle que la phonologie générative faisait passer pour une innovation théorique et conceptuelle radicale[48] est en définitive parfaitement illustrée par SPE lui-même dont il n'a pas été assez souligné qu'il analysait et formalisait des données toujours produites ***ailleurs*** et par ***d'autres***. Cette conception de la phonologie introduit un rapport totalement déstructuré aux données dont la production n'est plus ni contrainte ni contrôlée par aucune méthodologie. Il n'y a plus de corpus construit et analysé en tant que tel (*e.g.* avec ses apories et ses contre-exemples) et n'importe quelle notation éparse fait exemple et justifie n'importe quelle hypothèse. Scheer (2004b) en fournit deux illustrations saisissantes, sinon dramatiques. La première est erronée. Pour des raisons sans doute internes à son cadre conceptuel, il considère comme non attesté et inexistant le *Canadian Raising* qui apparaît pourtant très bien documenté dans les travaux récents de Chambers et de nombreux dialectologues de la région concernée[49]. La seconde, tout à fait juste et plus récente, renvoie à

[47] *Cf.* par exemple sur le Yawelmani et les dialectes du Yokuts, Archangeli (1984), Kisseberth (1969), Kuroda (1967) qui reprennent Newman (1944) ou encore, dans l'esprit de Kenstowicz et Kisseberth (1979), la réanalyse du Zoque (Wonderly 1951) par Dell (1973).

[48] Pour plus de détails, cf. Goldsmith et Laks (2005); Laks (2005; 2006).

[49] 'Fonder des raisonnements qui donnent lieu à des conclusions théoriques majeures sur quelques mots d'une langue dont on ne connaît rien et qu'on a repris de seconde ou troisième main rappelle une pratique de la phonologie générative naissante. Un cas d'école ici est le Canadian Raising : l'ensemble de données connu sous ce nom a servi, dans un article important (Bromberger et Halle 1989), à prouver que les règles phonologiques doivent être ordonnées. […] Or en cherchant à [le] localiser, s'aperçoit que la seule et unique source en est un article de trois pages écrit par Martin Joos (1942). Celui-ci dit avoir relevé les données décisives dans une classe en enseignant. Or les dialectologues de l'anglais canadien n'ont jamais pu en retrouver trace […], si bien que la question capitale des règles ordonnées et du dérivationnalisme tout entier a été débattue pendant des années sur

l'usage naïf d'internet comme source d'exemples : tel phonologue justifiant son analyse formelle par l'existence d'un bon nombre d'exemples congruents puisés sur le web sans s'apercevoir qu'il s'agit en fait d'occurrences tirés par Google des propres publications antérieures du dit phonologue (Scheer, 2004b: 137). Ces deux exemples illustrent à nouveau la nécessité d'établir des corpus empiriques, clos, stables et publiquement acceptés, si tant est que, comme Scheer d'ailleurs, on plaide pour le caractère scientifique et cumulatif de la phonologie.

On sait que la phonologie générative standard disparait corps et biens autour des années 1975, ne subsistant que de façon marginale, mais reconnue orthodoxe par Chomsky lui-même, chez Halle (Bromberger et Halle, 1989; 2001). J'ai montré ailleurs que la rupture introduite par la phonologie autosegmentale puis les morphophonologies non concaténatives, les phonologiques métriques puis harmoniques, pouvait se lire comme un ressourcement empirique de ces nouvelles phonologies (Laks, 1997a). Ressourcement empirique par la prise en compte d'un type nouveau de données (accentuelles, tonales, syllabiques, rythmiques etc.) et par la mise au travail de descriptions phonologiques de langues jusqu'alors peu considérées par la phonologie générale (domaine sémitique ou africain, langues à tons etc.). Dans tous les cas, il semble bien que le rapport méthodologique aux données se soit à nouveau inversé et qu'une attention nouvelle soit portée aux corpus et aux données (Scheer, 2004a; b).

Je n'en prendrai qu'un exemple récent. J'ai rappelé ci-dessus combien la sociolinguistique labovienne avait été méprisée et marginalisée, considérée comme une simple description de la performance, et combien ses concepts de variation (inhérente, sociale, stylistique) avaient été ignorés de même que les appareillages formels qui cherchaient à en rendre compte (règles variables, *Cf*. Sankoff et Labov, 1979). La phonologie mondiale est aujourd'hui dominée par le cadre théorique de l'optimalité (Prince et Smolensky, 2004). Il s'agit en fait d'une vaste nébuleuse de tendances et de courants, partageant un même cadre notationnel et quelques concepts centraux (*Cf*. les contraintes, hiérarchies, évaluations

---

la base de quelques mots de provenance franchement douteuse que personne n'a jamais pu vérifier. […] Canadian Raising est un fantôme qui encore aujourd'hui refait régulièrement surface lorsque la question dérivationnelle est traitée (par exemple à l'occasion du débat sur l'opacité au sein de la théorie de l'optimalité). (Scheer, 2004: 137). Scheer ne semble pas connaître les travaux de Chambers et d'autres sur l'anglais dans la région des Grands Lacs et au Canada.

parallèles des candidats etc.). Même si quelques analystes avisés ont pu montrer qu'il s'agissait en fait de la recomposition d'un cadre dérivationnel classique (Hulst et Ritter, 2000), il reste que les phonologies optimalistes constituent aujourd'hui le paradigme international de référence. Or dans ce cadre, la sociolinguistique variationniste a pris une grande place. Deux courants au moins, la phonologie fonctionnelle (Boersma, 1998) et la phonologie des variantes (Antilla, 1997; Antilla et Cho, 1998) ont totalement intégré les apports laboviens au cadre optimaliste. Antilla reprend ainsi non seulement les concepts laboviens de variation, mais intègre aussi les dispositifs formels de leur traitement (*i.e.* les règles variables) au nouveau cadre optimaliste (*i.e.* les contraintes variables et familles de tableaux). Plus récemment, Hayes et Cziráky Londe (2006) ont proposé une approche directement stochastique pour rendre compte des phénomènes de gradience et de variation. Cette intégration du variationnisme aux courants dominants de la phonologie contemporaine a des conséquences épistémologiques et méthodologiques sérieuses pour ces phonologies. Le variationnisme labovien est en effet explicitement ***une phonologie de corpus***. Prenant au sérieux la variation et le changement, ces nouvelles approches doivent ainsi nécessairement en revenir à l'enquête phonologique et à la construction raisonnée du corpus, seuls susceptibles d'exhiber la variation et le changement. En ces matières en effet, le recours à l'intuition et aux jugements n'est possible ni au plan méthodologique ni au plan théorique.

4. Conclusion : 'If usage is usage, then what might grammar be?', une réponse à Newmeyer (2003)

Comme j'ai tenté de le montrer ci-dessus, la question du corpus, en linguistique et en phonologie, n'est certainement pas une question secondaire, pratique ou technique, liée à l'apparition de tel type d'appareillage ou à la sophistication grandissante de tel type de moyens. C'est une question épistémologique et théorique de première importance qui parcourt toute l'histoire de la linguistique : qu'est-ce qu'un fait linguistique?, pour reprendre le titre du célèbre article de Labov (1975). Hormis la période de l'impérium génératif, la linguistique s'est toujours définie et construite, face à la rhétorique grammaticale de *l'exemplum* comme science du *datum*. La notion de linguistique ou de phonologie de corpus n'a donc en ce sens aucun caractère de nouveauté. Linguistique et phonologie ont toujours adossé leurs analyses et leurs modélisations à de vastes compendiums de faits collectés, colligés, classés et organisés en fonction d'hypothèses linguistiques et phonologiques précises. Ce n'est que dans

la période récente, générative et post générative que cette évidence s'est trouvée occultée par une construction théorique qui construisait son empirie de façon radicalement différente. D'une certaine façon, le débat entre l'approche chomskyenne du langage et celles qui l'ont précédée, et suivie, est un débat sur les faits et les données linguistiques dont il convient de partir. Envisageant la langue comme fondamentalement sous-déterminée par ses modalités d'acquisition et son usage en contexte[50], la grammaire générative a construit ses modélisations comme sur-déterminant les objets auxquels le linguiste devait prêter attention. Or ce rapport entre théorie et faits linguistiques a profondément évolué récemment.

Bien que souvent inaperçus par les linguistes eux-mêmes, le champ conceptuel de la linguistique contemporaine internationale a connu des changements récents très substantiels : ***la relation entre théorie et données, modèles abstraits et observables, s'est totalement inversée***. Des modèles sous-déterminés par les données, on est ainsi passé à des approches qui placent au premier plan les observables et les phénoménologies construites explicitement. Les modélisations ***quantitatives*** (Biber, Conrad et Reppen, 1998), ***statistiques*** (Manning et Schütze, 1999) et ***probabilistes*** (Chater et Manning, 2006) se sont développées. Les ***grammaires de construction*** (Goldberg, 1995) qui modifient profondément la conception que l'on peut avoir du stockage lexical via le concept langackerien de ***list/rule fallacy*** ((Langacker, 1987 ; 1991; 2000), les modèles ***occurrrentialistes*** (Bybee, 2006) qui accordent une importance centrale à la structure statistique des données, les modèles ***exemplaristes*** (Bybee, 2001) et leurs effets sur la conception de l'apprentissage, ont tous conjointement contribué à redonner aux données d'observation la place première qui était la leur dans les approches pré-génératives. Une autre façon, un peu plus tonitruante, de le dire consiste à remarquer que la grammaire générative *per se* est internationalement devenue minoritaire, sinon marginale[51]. Ce qui domine, ce sont, pris dans leur diversité, ce que Langacker (1987: 494) à baptisé les modèles basés sur l'usage (*Cf*. aussi Barlow et Kemmer, 2000).

[50] Dans une approche 'Principes et Paramètres' où GU est donné au départ, le LAD réduit l'acquisition à un simple réglage, c'est-à-dire à presque rien. Il en est de même du contexte, au sens le plus large, qui n'intervient éventuellement à titre interprétatif qu'après que tout ce qui est proprement interne ait agi. Sur ces questions *Cf.* par exemple Langacker (2000: 2,3).

[51] Ce constat est partagé par les générativistes eux-mêmes : 'Nor does generativist-oriented research predominate on the world scene today. I might speculate that a majority of North American linguists do one form or another of generative grammar, but in how many other countries is that the case? Perhaps in Britain and the Netherlands, but nowhere else. Germany, for example, has produced a number of outstanding generative syntacticians, but they represent a minority within the subfield. As Henk van Riemsdijk, the doyen of Dutch generative linguistics, has recently observed: 'Take a country like Germany. It is not a poor country, they've got a lot of universities. But if you want to pinpoint the centers of generative grammar, depending on your

Avec une grande honnêteté intellectuelle, Newmeyer, dans son adresse présidentielle à la *Linguistic Society of America*, dresse en 2003 le même constat de la marginalisation relative du paradigme grammatical chomskyen (Newmeyer, 2003). Parcourant le champ des théories linguistiques actuelles, il les range toutes (p. 682) sous le terme 'current anti-Saussurean USAGE-BASED MODELS' et réserve le label de linguistique saussurienne à la seule grammaire générative parce qu'elle prend pour objet la grammaire la plus abstraite telle qu'elle serait instanciée dans le cerveau d'un locuteur générique. Cette appropriation de Saussure ne manquera pas de faire sursauter un lecteur, même distrait, du Maître genevois. Saussure, qui tâche à 'étudier la vie des signes au sein de la vie sociale', défend en effet explicitement le contraire (*Cf. supra* et Maniglier (2007) pour une analyse récente).

Dans son parcours du champ linguistique contemporain qu'il voit structuré autour de l'opposition grammaire/usage, Newmeyer adopte une curieuse politique de la terre brûlée donnant en quelque sorte *quitus* aux linguistiques de l'usage quant à leur capacité à décrire, analyser et modéliser les pratiques linguistiques réelles des locuteurs en situation d'interlocution[52], ce que la grammaire générative ne parvient pas à faire. 'I argue that the mental grammar is only one of many systems that drive usage, since grammars are not actually well designed to meet language users' needs'. Abandonnant donc le terrain entier de la linguistique de la parole, et contestant (contre Saussure dont il se réclame pourtant!) qu'elle constitue la condition de la langue, Newmeyer défend que la grammaire mentale reste organisée autour de structures argumentales complètes dont il admet qu'elles n'apparaissent pratiquement jamais dans l'usage. Par une sorte de ***reductio ad absurdum*** il définit la grammaire mentale comme la condition synchroniquement et diachroniquement ***archaïque*** de la langue. Synchroniquement, la grammaire (universelle) n'est que le préliminaire abstrait de toute langue. Elle ne s'actualise jamais. Diachroniquement elle correspond à la rupture

---

generosity, you end up with maybe half of a dozen' (Van Riemsdijk, 1998, 18). One would not find even that many in, say, France or Italy'. (Newmeyer, 2005, 232).

[52] Je note au passage que Newmeyer, en forme d'argumentation defensive, réactive la rupture inaugurale du paradigme chomskyen : 'The last decade has seen the resurgence of many of the same ideas that were the hallmark of generative semantics. In particular, most of the ways of looking at form and meaning that fall under the rubric of COGNITIVE LINGUISTICS have reasserted—albeit in different form—the bulk of the ideas that characterized generative semantics. Langacker (1987:494) coined the term USAGE-BASED MODEL to refer to those approaches that reject a sharp distinction between language knowledge and language use […] My impression is that more linguists around the world do cognitive linguistics than do generative grammar.'' (Newmeyer, 2003: 683).

miraculeuse[53] qui a permis au genre *homo* de devenir *vocis*. Elle n'est même pas la langue originelle, mais correspond aux préconditions représentationnelles et mentales qui l'ont rendue possible. De cette grammaire mentale abstraite et universelle ***d'avant l'origine*** il n'y a bien entendu, par définition même aucun *datum* possible. Réservant ce terrain spéculatif à la grammaire générative chomskyenne, Newmeyer concède tous les autres aux linguistiques de l'usage.

Réduire la grammaire générative à une spéculation ***ab originem*** n'est pas propre à Newmeyer. Face aux désillusions engendrées par la relation entre grammaire générative et psycholinguistique[54], Chomsky dans ses derniers travaux argumente pour la même réduction et construit sous le nom de ***biolinguistique*** la nouvelle alliance susceptible de prendre en charge ce programme néo-cartésien qui, recomposant le problème de l'esprit et du corps à la manière du 18ème siècle, cherche dans la biologie les arguments d'une rupture avec le darwinisme permettant de comprendre le ***surgissement pur et préliminaire de la cognition spécifiquement humaine*** qu'il appelle grammaire mentale[55]. Outre la spéculation *ab originem*, Chomsky définit l'objet sa biolinguistique de façon aussi restrictive que Newmeyer (*Cf.* Chomsky, 2004; Hauser, Chomsky et Fitch, 2002; Jenkins, 2000). Il distingue très précisément ce qu'il appelle 'Faculty of Language in the Broad sense' (FLB) et 'Faculty of Language in the Narrow sense'(FLN). Seule, on s'en doute la seconde relève spécifiquement de la (bio)linguistique et est prise en charge par le programme chomskyen, tout le reste est plus ou moins externe, cognitivement mixte et relève peu ou prou de la mise en paroles (l'usage). Le schéma qui résume sa pensée, publié dans *Science* (Hauser, Chomsky et Fitch,

[53] Si [comme cela est nécessaire dans le système de Chomsky et de Fodor] le premier utilisateur du langage possédait déjà un modèle inné complet, c'est que serait intervenue une rupture miraculeuse dans la séquence de l'évolution.' (Putnam, 1978: 428, souligné par l'auteur).

[54] Si le programme chomskyen a pour un temps fourni un programme de travail à la psycholinguistique, force est de constater qu'aucune des avancées de ce domaine ni aucun de ses résultats n'a jamais fait évolué le programme chomskyen d'un pouce (*Cf.* Tomasello (2000; 2003). Newmeyer tire ainsi les conclusion du divorce: 'I believe that the great majority of psycholinguists around the world consider the competence-performance dichotomy to be fundamentally wrongheaded.'(Newmeyer, 2003: 683).

[55] 'The biolinguistic perspective views a person's language in all of its aspects – sound, meaning, structure -- as a state of some component of the mind, understanding "mind" in the sense of 18th century scientists who recognized that after Newton's demolition of the "mechanical philosophy," based on the intuitive concept of a material world, no coherent mind-body problem remains, and we can only regard aspects of the world "termed mental," as the result of "such an organical structure as that of the brain," as chemist-philosopher Joseph Priestley observed. Thought is a "little agitation of the brain," David Hume remarked; and as Darwin commented a century later, there is no reason why "thought, being a secretion of the brain," should be considered "more wonderful than gravity, a property of matter." By then, the more tempered view of the goals of science that Newton introduced had become scientific common sense: Newton's reluctant conclusion that we must be satisfied with the fact that universal gravity exists, even if we cannot explain it in terms of the self-evident 'mechanical philosophy.' As many commentators have observed, this intellectual move 'set forth a new view of science' in which the goal is 'not to seek ultimate explanations' but to find the best theoretical account we can of the phenomena of experience and experiment (I. Bernard Cohen)' (Chomsky, 2004).

2002, figure 2: 1571), est saisissant. Tout ce qui ne correspond pas aux principes récursifs et compositionnels de la grammaire au sens le plus abstrait et le plus restrictif de FLN est rejeté à la périphérie et relève au mieux de FLB, qui n'est pas en propre l'objet de la linguistique. C'est le cas de toute la phonologie et pas uniquement de la phonétique. La phonologie analyse donc des usages, et repose de ce fait même sur une recollection de faits attestés.

Voici donc une bonne nouvelle pour conclure: pour Chomsky lui-même, ***la phonologie est une science des usages et il ne saurait donc, ni en droit ni en fait, y avoir d'autre phonologie que de corpus!***

Bibliographie

Aarts, B. (2000). Corpus linguistics, Chomsky and fuzzy tree fragments. dans C. Mair et M. Hundt (dir.), *Corpus Linguistics and Linguistic Theory.*, Amsterdam/Atlanta: Rodopi, pp. 5-13.

Anderson, S. et Lightfoot, D. W. (2002). *The Language Organ : Linguistics as Cognitive Psychology*. Cambridge: Cambridge University Press.

Antilla, A. (1997). Deriving variation from grammar : A study of Finnish genetives. dans F. Hinkens, R. van Hout et L. Wetzels (dir.), *Variation, Change and Phonological Theory*, New-York: John Benjamins, pp. 35-68.

Antilla, A. et Cho, Y.-m. Y. (1998). Variation and change in Optimality Theory. *Lingua*, 104: 31-56.

Auroux, S. (1998). *La raison, le langage et les normes*. Paris: PUF.

Baïf, J. A. de. (1574). *Étrénes de poézie fransoêze an vers mesurés*. Paris.

Barlow, M. et Kemmer, S. (dir.) (2000). *Usage based models of language*. Stanford Cal.: CSLI.

Beguin, D. (1998). *Les CD-ROM de corpus littéraires et de données bibliographiques à l'usage des chercheurs antiquisants* XIVeme congrès international de l'Association Guillaume Budé. Limoges, 25-28 août 1998: Les Belles Lettres

Biber, D., Conrad, S. et Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure* Cambridge: Cambridge University Press.

Blanche-Benveniste, C. (1997). La notion de variation syntaxique dans la langue parlée. *Langue Française*, 115: 19-29.

Blanche-Benveniste, C. (2000). *Approches de la langue parlée en français.* Paris: Ophrys.

Blanche-Benveniste, C. (2004). Le singulier et le pluriel en français parlé contemporain. *Bulletin de la Société de Linguistique de Paris*, XCIX.1: 129-54.

Boersma, P. (1998). *Functional Phonology : Formalizing the interactions between articulatory and perceptual drives*. Amsterdam: LOT.

Bopp, F. (1889). *Grammaire comparée des langues Indo-Européennes- comprenant le Sanscrit, le Zend, l'Arménien,..* Introduction de Michel Bréal. Paris: Bibliothèque Nationale.

Bortolussi, B. (2008). La grammaire générative et le latin : exemples construits et utilisation des corpus. *Bulletin de la Société de Linguistique de Paris*: A paraître.

Bouquet, S. (1997). *Introduction à la lecture de Saussure*. Paris: Payot.

Bourdieu, P. (2001). *Science de la science et réflexivité*. Paris: Raisons d'agir.

Bromberger, S. et Halle, M. (1989). Why phonology is different. *Linguistic Inquiry*, 20.1: 51-70.

Bromberger, S. et Halle, M. (2001). The Onthology of Phonology (Revised). dans N. Burton-Roberts, P. Carr et G. Docherty (dir.), *Phonological Knowledge : Conceptual and Empirical Issues*, Oxford: Oxford Universtity Press, pp. 19-39.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2006). *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Cappeau, P. et Gadet, F. (2007a). L'exploitation sociolinguistique des grands corpus. Maitre-mot et pierre philosophale *Revue française de linguistique appliquée*: 99-110.

Cappeau, P. et Gadet, F. (2007b). Où en sont les corpus sur les français parlés? *Revue française de linguistique appliquée*: 129-33.

Catach, N. (1984). *La phonétisation automatique du français : les ambiguïtés de la langue française*. Paris: Editions du CNRS.

Catach, N. et Jejcic, F. (1984). *Les listes orthographiques de base*. Paris: Nathan.

Chafe, W. L. (1992). The importance of corpus linguistics to understanding the nature of language dans J. Svartvik (dir.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin - New York: Mouton de Gruyter, pp. 79-97.

Chater, N. et Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences.*, 10.7: 335.

Chevalier, J.-C. (2007). Les exemples et la norme dans les grammaires : étude historique. dans G. Siouffi et A. Steuckardt (dir.), *Les linguistes et la norme*, Berlin: Peter Lang, pp. 151-63.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: M.I.T. Press.

Chomsky, N. (1966). *Topics in the theory of generative grammar*: Janua linguarum. Series minor ; nr. 56. The Hague,: Mouton.

Chomsky, N. (1995). *The minimalist program*. Cambridge, Mass.: The MIT Press.

Chomsky, N. (2004). *Biolinguistics and the Human Capacity (Talk delivered at MTA, Budapest, May 17, 2004).* http://www.chomsky.info/talks/20040517.htm

Chomsky, N. et Ronat, M. (1979). *Language and responsibility : based on conversations with Mitsou Ronat*. New York: Pantheon Books.

Comrie, B. (1981). *Language universals and linguistic typology : syntax and morphology*. Chicago: University of Chicago Press.

Comrie, B., Haspelmath, M., Dryer, M. S. et Gil, D. (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Comrie, B., Matthews, S. et Polinsky, M. (2003). *The atlas of languages : the origin and development of languages throughout the world*. New York: Facts On File.

Cordereix, P. (2001). Ferdinand Brunot, le phonographe et les 'patois'. *Le Monde alpin et rhodanien*, 1-3: 39-54.

Dell, F. (1973). *Les règles et les sons : introduction à la phonologie générative*. Paris: Hermann.

Durand, J. et Laks, B. (1996). Why phonology is one? dans J. Durand et B. Laks (dir.), *Curent Trends in Phonology : models and methods*, Salford: Salford University Publications, pp. 1-15.

Durand, J., Laks, B. et Lyche, C. (2002). La phonologie du français contemporain : usages, variétés et structures. dans C. D. Pusch et W. Raible (dir.), *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache / Romance Corpus Linguistics: Corpora and Spoken Language*, Tübingen: Narr (Coll.ScriptOralia 16), pp. 93-106.

Durand, J., Laks, B. et Lyche, C. (2005). Un corpus numérisé pour la phonologie du français. dans G. Williams (dir.), *La linguistique de corpus*, Rennes: Presses Universitaires de Rennes, pp. 205-17.

Durand, J. et Lyche, C. (ce volume) French liaison in the light of corpus data.

Encrevé, P. (1976). *Présentation In Labov William 1976 : Sociolinguistique*. Paris: Editions de Minuit.

Fillmore, C. J. (1992). Corpus linguistics or computer-aided armchair linguistics. dans J. Svartvik (dir.), *Directions in Corpus Linguistics Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, New York Berlin: Mouton de Gruyter, pp. 35-61.

Fouché, P. (1959). *Traité de prononciation française*. Paris: Klincksieck.

Francis, N. W. (1992). Language corpora B.C. dans J. Svartvik (dir.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin - New York Mouton de Gruyter, pp. 17-35.

Gadet, F. (1989). *Le français ordinaire*. Paris: Armand Colin.

Gadet, F. (2003). *La variation sociale en français*. Paris: Ophrys.

Geerts, T. (2007). *More about less : fast speech phonology, the cases of French and Dutch.*, Thèse Université de Paris X et University Rabdoud de Nimègue.

Gilliéron, J. et Edmont, E. (1902-1912). *Atlas linguistique de la France*. Paris: Champion.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldsmith, J. et Huck, G. (1995). *Ideology and linguistic theory : Noam Chomsky and the deep structure debates.* New York: Routledge.

Goldsmith, J. et Laks, B. (2005). Generative Phonology and its successors. dans L. R. Waugh et J. E. Joseph (dir.), *The Cambridge History of Linguistics*, Cambridge: Cambridge University Press.

Goldsmith, J. A. (2005). Review : The Legacy of Zellig Harris: Language and information into the 21st century. *Language*, 81.3: 719-36.

Goldsmith, J. A. et Laks, B. (en cours). *Battle in the mind fields*.

Gougenheim, G., Michéa, P., Rivenc, P. et Sauvageot, A. (1956). *L'élaboration du français élémentaire*. Paris: Didier.

Greenberg, J. H. (dir.) (1963). *Universals of Language*. Cambridge Mass.: MIT Press.

Gross, M. (1975). *Méthodes en syntaxe : régime des constructions complétives*. Paris: Hermann.

Gross, M. (1976). *Grammaire transformationnelle du français : syntaxe du verbe*: Langue et langage. Paris: Larousse.

Gross, M. et Perrin, D. (1989). *Electronic dictionaries and automata in computational linguistics : LITP Spring School on theoretical computer science, Saint-Pierre d'Oléron, France, May 25-29, 1987 : proceedings*: Lecture notes in computer science, 377. Berlin ;; New York: Springer-Verlag.

Habert, B. (2000). Review of G. Kennedy 'An introduction to corpus linguistics'. *Journal of quantitative linguistics*, 7.3: 245-50.

Habert, B., Nazarenko, A. et Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

Hagège, C. (2002). Sous les ailes de Greenberg et au-delà. Pour un élargissement des perspectives de la typologie linguistique *Bulletin de la Société Linguistique de Paris*, 97.1: 5-36.

Halliday, M. A. K. (1992). Language as system and language as instance: The corpus as a theoretical construct. dans J. Svartvik (dir.), *Directions in Corpus Linguistics Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin, New-York: Mouton de Gruyter, pp. 61-79.

Harris, Z. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10.2-3: 146-62.

Hauser, M. D., Chomsky, N. et Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 29822, 1569-80.

Hayes, B. et Cziráky Londe, Z. (2006). Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony. *Phonology*, 23: 59-104.

Hulst Van der, H. et Ritter, N. A. (2000). The SPE-heritage of Optimality Theory. *The Linguistic Review*, 17: 259-89.

Jackendoff, R. (2002). *Foundation of Language : Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jenkins, L. (2000). *Biolinguistics, Exploring the Biology of Language*. Cambridge: Cambridge University Press.

Juilland, A. (1965). *Dictionnaire inverse du français*. La Haye: Mouton.

Juilland, A., Brodin, D. et Davidovich, C. (1970). *Frequency dictionary of French words*. La Haye: Mouton.

Labov, W. (1975). *What is a linguistic fact?* . Lisse: Peter de Ridder Press.

Labov, W. (1976). *Sociolinguistique*. Paris: Editions de Minuit.

Labov, W. (1979). *Le parler ordinaire*. Paris: Editions de Minuit.

Labov, W. (1981). Resolving the neogrammarian controversy. *Language*, 57.2: 267-309.

Laks, B. (1992). La linguistique variationniste comme méthode. *Langages*, 108: 34-51.

Laks, B. (1996). *Langage et cognition : l'approche connexionniste*. Paris: Hermès.

Laks, B. (1997). Nouvelles phonologies. *Langages*, 125: 3-14.

Laks, B. (2001). Un siècle de phonologie. *Modèles Linguistiques*, XXII-1: 75-103.

Laks, B. (2002). Le comparatisme de la généalogie à la génétique. *Langages*, 146: 19-46.

Laks, B. (2005a). Approches de la phonologie cognitive . dans J. Durand, N. Nguyen et V. Rey (dir.), *Nouvelles approches en phonétique et en phonologie.*, Paris: Hermès.

Laks, B. (2005b). La liaison et l'illusion. *Langages*, 158: 101-26.

Laks, B. (2007). Les hommes politiques français et la liaison (1908-1999). dans L. Baronian et F. Martineau (dir.), *Modéliser le changement : Les voies du français*, Montréal: Presses de l'Université de Montréal

Laks, B. (dir.) (1997). *Nouvelles phonologies*. Langages 125.

Laks, B., Cleuziou, S., Demoule, J.-P. et Encrevé, P. (dir.) (2007). *Origins and evolutions of languages : approaches, models , paradigms*. Londres: Equinox.

Langacker, R. (1987). *Foundations of cognitive grammar I: Theoretical perequisite*. Stanford: Stanford University Press.

Langacker, R. (1987, 1991). *Foundations of cognitive Grammars*. Stanford: Stanford University Press (2 vol.).

Langacker, R. (2000). A Dynamic Usage-Based Model. dans M. Barlow et S. Kemmer (dir.), *Usage Based Models of Language*, Standford: CSLI, Standford University, pp. 1-65.

Levet, J.-P. (2000). Utilisation des NTIC et démarche humaniste : observation et suggestions pour la formation des maîtres. *Revue de l'EPIC*, 4: 91-98.

Mair, C. (1992). 'Comments' dans J. Svartvik (dir.), *Directions in Corpus Linguistics Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin New-York: Mouton de Gruyter, pp. 79-98.

Maniglier, P. (2007). Processing culture : new trends in artificial intelligence and linguistics in the light of structuralism. dans S. Franchi et F. Bianchini (dir.), *Toward an archeology of Artificial Intelligence*, Berlin: Springer.

Manning, C. D. et Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Martinet, A. (1945). *La prononciation du français contemporain : témoignages recueillis en 1941 dans un camp d'officiers prisonniers*. Genève: Droz.

Martinet, A. (1956). *La description phonologique : avec application au parler franco-provençal d'Hauteville (Savoie)*. Genève: Librairie Droz.

Martinet, A. et Walter, H. (1973). *Dictionnaire de la prononciation française dans son usage réel*. Paris: France-Expansion.

Mayaffre, D. (2005). Statut des corpus en linguistique des corpus. Un observatoire ou un observé? dans Rôle et place des corpus en linguistique (JETOU 2005), Toulouse, Université de Toulouse.

Meigret, L. (1542). *Traité touchant le commun usage de l'écriture françoise*. Lyon: Republications Slatkine Genève, 1972.

Mellet, S. (2002). Corpus et recherches linguistiques *Corpus*, 1: 1-17.

Milner, J.-C. (1989). *Introduction à une science du langage*. Paris: Editions du seuil.

Morin, Y.-C. (2000). Le français de référence et les normes de prononciation. *Cahiers de l'Institut de linguistique de Louvain*, 26-1: 91-135.

Morin, Y.-C. (2005a). La liaison relève-t-elle d'une tendance à éviter les hiatus? Réflexions sur son évolution historique. *Langages*, 158: 8-23.

Morin, Y.-C. (2005b). La naissance de la rime normande. dans M. Murat (dir.), *Poétique de la rime*, Paris: Champion, pp. 220-52.

Morin, Y.-C. (2005c). Liaison et enchaînement dans les vers aux XVIe et XVIIe siècles. dans J.-M. Gouvard (dir.), *De la langue au style*, Lyon: Presses Universitaires de Lyon, pp. 299-318.

Nelson, F. (1992). Language corpora B.C. Directions in Corpus Linguistics, dans J. Svartvik (dir.), *Directions in Corpus Linguistics Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin New-York: Mouton de Gruyter, pp. 79-98.

Nevin, B. (dir.) (2002). *The legacy of Zellig Harris: Language and information into the 21st century : Philosophy of science, syntax and semantics*. vol. 1. Philadelphia: John Benjamins.

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language,* 79.4: 682-797.

Newmeyer, F. J. (2005). A reply to the critiques of 'Grammar is grammar and usage is usage'. *Language*, 81.1: 229-36.

Niedermann, M. (1953). *Phonétique historique du latin*. Paris: Klincksieck.

Osthoff, H. et Brugmann, K. (1874). *Morphologische Untersuchungen auf dem Gebiete des Indogermanischen Sprachen.I*. Leipzig.

Pichon, E. (1938). Genre et questions connexes (sur les pas de Mlle Durand). *Le français moderne*, 6: 107-26.

Prince, A. et Smolensky, P. (2004 [1993]). *Optimality Theory : constraints interaction in generative grammar*. Londres: Blackwell.

Putnam, H. (1978). *Meaning and the Moral Sciences* Boston-London: Routledge & Kegan Paul.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. dans G. Williams (dir.), *La linguistique de corpus*, Rennes: Presses Universitaires de Rennes, pp. 31-45.

Romano, A. (à paraître). À propos de deux bases de données de parole publiées récemment : compte-rendu de 'API – Archivio del Parlato Italiano' (coord. F. Albano Leoni) et de 'C-ORAL-ROM' (coord. E. Cresti - M. Moneglia). *International Journal of Italian Linguistics* A paraître.

Saint-Gérand, J.-P. (2001). *L'onomastique avant l'onomastique: naissance d'une discipline dans la France du XIX° siècle.* . L'onomastique au carrefour des sciences humaines. Lyon: http://www.chass.utoronto.ca/epc/langueXIX/

Sampson, G. (1975). *The Form of Language*. Londres: Weidenfeld and Nicolson.

Sankoff, D. et Labov, W. (1979). On the uses of variable rules. *Language in Society*, 8: 189-222.

Sapir, E. (1939). *Nootka Texts: Tales and Ethnological Narratives with Grammatical Notes and Lexical Materials*: William Dwight Whitney Linguistic Series. Philadelphia: Linguistic Society of America.

Saussure, F. de (1878). *Mémoire sur le système primitif des voyelles dans les langues indo-européennes.* Leipzig: B.G. Teubner (repris dans Saussure 1922).

Saussure, F. de (2002). *Ecrits de linguistique générale*. Paris: Gallimard.

Schane, S. A. (1968). *French phonology and morphology*. Cambridge, Mass: MIT Press.

Scheer, T. (2004a). Le corpus heuristique : un outil qui montre mais ne démontre pas. *Corpus*, 3: 153-92.

Scheer, T. (2004b). Usage des corpus en phonologie. *Corpus* 3: 5-84.

Schleicher, A. (1861). *Kompendium des vergleichenden Grammatik des indogermanischen Sprachen*. Weimar: Böhlau.

Sériot, P. (1999). *Structure et totalité*. Paris: PUF.

Sperber, D. et Wilson, D. (1989). *La pertinence : communication et cognition* Paris: Editions de Minuit.

Svartvik, J. (dir.) (1992). *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*. Berlin , New-York: Mouton de Gruyter.

Svartvik, J., Aijmer, K. et Altenberg, B. (1991). *English corpus linguistics : studies in honour of Jan Svartvik*. Londres; New-York: Longman.

Thurot, C. (1881-1883). *De la prononciation française depuis le commencement du XVI^{ème} siècle, d'après les témoignages des grammairiens*. Paris: Bibliothèque Nationale.

Tomasello, M. (2000). The item-based nature of children's early syntactic development *Trends in cognitive science*, 4.4: 156-63.

Tomasello, M. (2003). *Constructing a language : A usage based theory of language acquisition*. Cambridge Mass.: Harvard University Press.

Troubetzkoy, N. S. (1939). *Grundzüge der Phonologie*.vol. Traduction française (J. Cantineau) : *Principe de Phonologie,* Paris, Klincksieck, 1947.

Troubezkoy, N. S. (2006). *Correspondance avec Roman Jakobson et autres écrits* Paris: Payot.

Väänänen, V. (1962). *Introduction au latin vulgaire*. Paris: Klincksieck.

Vallée, N., Rousset, I. et Boë, L.-J. (2001). Des lexiques aux syllabes des langues du monde : Typologies, tendances et organisations structurelles. *Linx*, 45: 37-50.

Van Riemsdjk, H. (1998). GLOW 1978–1998. *Glot International*, 3: 18-19.

Veken, C. (1984). Le phonographe et le terrain : la mission Brunot-Bruneau dans les Ardennes en 1912. *Recherches sur le français parlé*, 6: 45-71.

Walter, H. (1977). *La phonologie du français contemporain*. Paris: Presses Universitaires de France.

Walter, H. (1982). *Enquêtes phonologiques et variétés régionales du français*. Paris: Presses Universitaires de France.