CHAPITRE V

LE LANGAGE AU FIL DU TEMPS : CORPUS ET DIACHRONIE

1. DEFINITIONS ET ENJEUX

L'écoulement du temps structure de nombreux corpus, sans qu'ils permettent pour autant la saisie de l'évolution du langage. La volonté de créer des dictionnaires reposant sur l'usage effectif et son changement a par exemple contribué à la création de corpus électroniques intégrant des données de différentes périodes. C'est le cas du *Trésor de la Langue Française* (INaLF, CNRS) qui s'appuie sur une base de textes de plus de 160 millions de mots, s'étalant du XVI^{e} au XX^{e} siècle. Toutefois, de tels corpus ne constituent pas forcément des corpus adaptés aux études diachroniques. Le registre littéraire y domine, au détriment d'autres registres. La dimension temporelle structure également d'autres corpus, encore plus spécialisés. Corpus mono-émetteur : c'est le cas de ***Mitterrand1***, dont les textes s'égrènent sur le premier septennat. Corpus pluri-locuteurs : c'est le cas des résolutions générales des quatre grandes confédérations syndicales ouvrières françaises étudiées entre 1971 et 1976 (Bergounioux *et al.*, 1982)[1]. Ces corpus sont de la même manière restreints à un registre (ou à des variations sur un même registre) : entretien, interview et discours de circonstance pour ***Mitterrand1***, résolutions de congrès pour (Bergounioux *et al.*, 1982). Le temps intervient, mais on ne peut saisir son rôle que sous un angle limité : une thématique, un domaine, ou un genre bien défini.

A côté de ces corpus de fait spécialisés, se constituent des corpus « historiques ». Ils sont destinés explicitement à l'étude de l'évolution de la

[1] Le chapitre IX aborde la mesure de l'évolution lexicale de tels corpus.

langue. Nous présentons en détail un corpus de ce type : ***Archer***, en section 2, ainsi que les problèmes de représentativité et de constitution de tels corpus. L'évolution de la langue peut être examinée sur la courte durée, sur le moyen terme, ou sur le long terme. Nous rendons compte d'études relevant de ces différentes temporalités en section 3. Nous abordons enfin en section 4 les problèmes méthodologiques propres aux corpus historiques.

2. UN CORPUS POUR L'ETUDE DE LA DIACHRONIE : *ARCHER*

Les analyses diachroniques de l'anglais disposent du corpus d'***Helsinki*** d'1,5 millions de mots (Kytö, 1993b). La période couverte va de 750 à 1700[2]. Le corpus ***Archer***[3] (Biber *et al.*, 1994) complète la tranche chronologique couverte[4]. D. Biber, E. Finegan et D. Atkinson (1994, p. 7-13) montrent les usages possibles d'un tel corpus historique. Ils utilisent par exemple la distinction établie par Biber (cf. chapitre I) entre production informationnelle (qui favorise noms, prépositions, adjectifs attributs etc.) et production « impliquée » (qui privilégie le présent, l'omission de *that*, les contractions, les démonstratifs, la première personne, le pronom *it*, *BE* comme verbe principal, les pronoms indéfinis, etc.). Si l'on compare les registres, théâtre, lettres et journaux intimes se font plus « impliqués » depuis le XVII^{e} siècle, tandis que la médecine devient plus « informationnelle ». La comparaison entre anglais et américain sur la même durée montre que les registres américains sont généralement plus informationnels que leurs équivalents anglais.

2.1 L'anglais et l'américain de 1650 à aujourd'hui

Archer[5] a été constitué pour permettre l'étude diachronique de l'anglais et de l'américain entre 1650 et aujourd'hui par le biais de dix « registres », qui mêlent thématiques et genres[6]. Les registres sont les suivants pour l'écrit : journaux intimes, lettres, fiction, écrits journalistiques, médecine (anglais seulement), science (anglais seulement), décisions de justice (américain seulement), et pour l'écrit lié à l'oral (c'est-à-dire imitant l'oral ou servant de base à une production orale) : les conversations fictives, le théâtre, les sermons et homélies.

[2] Des documents écossais (1450-1700) et américains (1600-1700) constituent deux corpus complémentaires (Kytö, 1993).
[3] A Representative Corpus of Historical English Registers.
[4] Il y a donc recouvrement pour la période 1650-1700, ce qui autorise des comparaisons fructueuses sur les choix faits pour représenter ce laps de temps (cf. *infra*).
[5] Le corpus rassemblé à Cambridge (English Faculty) pour la période 1600-1800 s'inscrit dans la même perspective (Wright, 1993).
[6] Dans la même acception qu'au chapitre I.

Archer est organisé par périodes de cinquante ans pour que l'on puisse examiner l'évolution, les flux et les stabilités sur des périodes relativement courtes. L'américain n'est dans l'immédiat représenté que par trois périodes : deuxième moitié des XVIII^{e}, XIX^{e} et XX^{e} siècles. L'anglais l'est pour les neuf périodes. Pour chaque période de cinquante ans et chaque registre, un échantillon de 20 000 mots[7] est constitué. ***Archer*** totalise 1,7 million de mots.

2.2 Echantillonnage des registres

Le choix de textes relevant des registres visés se heurte à plusieurs obstacles. En premier lieu, les ressources bibliographiques sont organisées thématiquement et non par registres. Ainsi, une des sources bibliographiques consultées, à l'entrée *lettres*, renvoie en fait aux manuels d'écriture de lettres, ce qui ne correspond pas à l'objectif visé : la correspondance privée authentique. En second lieu, « les distinctions de registre d'une période peuvent ne pas correspondre exactement avec celles d'une autre période[8]. Les registres ne restent pas nécessairement distincts l'un de l'autre au fil du temps. Bien sûr, les registres émergent à un moment donné de l'histoire, pas nécessairement tous au début d'une période d'investigation ni au début d'une période de cinquante ans retenue » (Biber *et al.*, 1994, p. 5).

M. Kytö (1993) témoigne de la complexité des paramètres à prendre en compte pour rassembler des données représentatives de l'américain entre 1600 et 1700 , dans le cadre d'un autre corpus historique. Seuls sont retenus les documents écrits (et éventuellement imprimés) aux Etats-Unis, et pour la période commençant en 1670, date qui sépare la première génération d'immigrants de ses descendants, provenant d'auteurs nés dans ces colonies (ou établis depuis suffisamment longtemps). Les dates d'installation différentes des colonies du Sud (Virginie, premières arrivées en 1607) et du Nord (Plymouth, 1620, baie de Massachusetts 1630, etc.) amènent à constituer des échantillons distincts pour rendre compte de leurs histoires langagières propres. Certains registres caractéristiques des colonies ont été intégrés : récits de captivité, témoignages, etc. L'appréhension de l'oral ne peut s'effectuer que par des biais : « Le langage de tous les jours trouvé dans la correspondance privée, certains journaux intimes ou des textes faits pour être dits fournissent un moyen d'approcher le langage parlé du passé, le vrai cœur du changement linguistique. De la même manière, les écrits des immigrants les moins éduqués, qui n'auraient peut-être pas pris la plume dans leur pays d'origine mais qui étaient forcés de le faire dans les colonies, peuvent aussi nous donner des aperçus [*glimpses*] de la langue parlée » (*ibid.*

[7] 10 fragments de 2 000 mots, pour diminuer le poids des idiolectes.
[8] Par exemple la correspondance peut relever de la littérature, voire de la philosophie, comme de l'échange purement privé aux XVI^{e} et XVIII^{e} siècles (Wright, 1993, p. 26). Finegan et Biber (1995, p. 249) expliquent l'incohérence relative de leurs résultats concernant les lettres par l'hétérogénéité de ce registre.

p. 5)[9].

Pour ***Archer***, au sein d'un registre, le choix des ouvrages repose sur une procédure aléatoire[10] (au sens probabiliste)[11]. Un protocole bien défini permet également, pour chaque registre, d'extraire des fragments (pas forcément continus) de 2 000 mots[12].

2.3 Structuration temporelle

L'échelonnement des documents retenus peut avoir comme logique une périodisation. C'est le choix d'***Archer***, qui distingue donc des périodes de cinquante ans : ce sont les blocs qui sont soumis ensuite à l'analyse linguistique et statistique.

Le parti pris du corpus couvrant l'anglais de 1600 à 1800, à Cambridge (Wright, 1993), est tout autre : un étalement continu des documents, avec une ossature formée de textes sélectionnés à dix ans d'intervalle. L'objectif est ici de permettre au chercheur de choisir les intervalles qui lui paraissent pertinents et de ne pas l'enfermer dans une périodisation qui peut s'avérer non valide pour sa recherche.

2.4 Représenter les états de langue ou des idiolectes ?

De quels usages les corpus historiques constitués sont-ils représentatifs ? Une des réponses possibles est celle qui sous-tend la création d'***Archer*** : les variations observées relèvent des genres ou des types textuels sous-jacents. Si l'on veut étudier l'évolution d'une langue, il faut articuler l'échelonnement des textes dans le temps avec leur stratification en genres qui ont une cohérence et un mouvement propres. D'où une démarche d'échantillonnage aléatoire, utilisant des extraits courts, mais nombreux. Finegan et Biber (1995, p. 252) soulignent ainsi que la représentation du genre *sermons* est probablement plus satisfaisante dans ***Archer*** que dans ***Helsinki***, même si ce dernier corpus comprend des textes entiers qui totalisent un nombre de mots plus important. ***Helsinki*** en effet utilise les sermons de deux prêcheurs seulement, tandis

[9] Cet article fournit des extraits significatifs de tels documents (*ibid*., p. 5-8).

[10] Ainsi, pour la fiction anglaise, le répertoire *Oxford Companion to English Literature* (*OCEL*) a été utilisé. Les 1 099 pages de l'*OCEL* ont été divisées par le produit du nombre de périodes et de textes requis pour chaque période, ce qui a fourni un intervalle de 13 pages. Le numéro de la première page considérée a été tiré au hasard, puis on a examiné la page suivante à 13 pages d'intervalles et ainsi de suite. Pour les textes de fiction par exemple, sur chaque page examinée, on a pris le premier auteur anglais ayant écrit un roman dans une des périodes retenues et on a choisi son 3e roman s'il y en avait 3 ou plus (ou son 2e ou son unique roman). On a continué jusqu'obtenir le nombre de textes nécessaires pour toutes les périodes (ce qui a nécessité plusieurs passes sur l'*OCEL*, en tirant à chaque fois un nouveau numéro au hasard pour la première page considérée).

[11] On reviendra au chapitre IX sur les raisons de ce choix.

[12] Par exemple, pour les textes journalistiques ou scientifiques anciens, les documents ont souvent une taille inférieure à 2 000 mots. Il faut alors regrouper. Inversement, dans les périodes récentes, la longueur des textes oblige à prélever les 500 premiers et derniers mots, ainsi qu'un empan de 1 000 mots au milieu, pour ne pas sur-représenter certains « sites » particuliers des textes (introduction, conclusion, etc.).

qu'un échantillon plus élevé de prêcheurs figure dans ***Archer***.

D'autres travaux (Wright, 1993, p. 27-29) insistent au contraire sur la dimension idiolectale des observations. S. Wright (*ibid.* p. 28) cite par exemple les recherches sur l'emploi de certains marqueurs relatifs : « [...] au début du XVII^{e} siècle, le système des relatives différait du système actuel en ce que le pronom *which* pouvait optionnellement servir à renvoyer à un antécédent humain aussi bien qu'à un antécédent non humain. Cependant, progressivement, c'est le pronom *who* (à la place de *which*) qui a été choisi pour renvoyer à des antécédents humains. Hope (1990) a montré que le choix des marqueurs relatifs dans les œuvres de Shakespeare et Fletcher était basés sur deux systèmes en compétition. Alors que celles de Fletcher sont typiques de l'association moderne entre le relatif *who* et des antécédents humains, l'usage suivi par Shakespeare suggère que ce trait n'est pas un facteur aussi significatif pour son choix. Pour ces deux écrivains donc, la sémantique du système de marqueurs relatifs a des valeurs différentes. » Le rassemblement de données textuelles plus importantes pour un groupe d'auteurs contemporains a pour objectif alors de caractériser l'usage commun de ce groupe par rapport aux idiolectes de chacun des auteurs[13]. Se pose aussi la question de la part de la manipulation stylistique de la langue, de l'idiolecte et de l'usage du moment.

3. ÉTUDES DE LA DIACHRONIE

Les corpus électroniques permettent d'examiner l'évolution de certains phénomènes langagiers sur de très courtes durées (d'une année sur l'autre, par exemple), sur le moyen terme (quelques décennies) et sur le long terme : on peut alors comparer des états de langue reconnus comme distincts dans la tradition linguistique (ancien français / moyen français / français classique / français moderne) ou examiner les changements au fil des siècles.

3.1 La courte durée

J. Sinclair a forgé le terme de « corpus de suivi » (*monitor corpus*) pour désigner des flux continus de textes permettant l'analyse chronologique, année par année par exemple, de données langagières. Cette notion était au départ une vue de l'esprit. De plus en plus de textes sont désormais directement sous forme électronique. C'est le cas de quotidiens employant une langue « tenue » comme *Le Monde*, *The Guardian*, édités sous forme de CD-ROM. C'est le cas aussi des bandes de

[13] Voir (Wright, 1993, p. 30-34) pour une discussion du statut à donner aux emplois par Joseph Addison des différentes formes de relatives. S. Wright prend nettement le contrepied de l'interprétation que fournissent Biber et Finegan des mêmes faits.

photocomposition de journaux mises à disposition des créateurs de corpus. On peut donc comparer les ensembles constitués pour chaque année, ou examiner les apports d'une année donnée[14].

A. Renouf (1993) détaille l'utilisation en ce sens du *Times*, de novembre 1990 à septembre 1991. Un premier filtrage isole les mots nouveaux, en les répartissant en noms propres, acronymes et mots « ordinaires ». Le classement de ces derniers renseigne sur les mécanismes à l'œuvre et leur productivité relative : formations à base d'onomatopées, jeux de mots, mots-valises, « composés », doublons dérivatifs (*indifferentness*), suffixations (*eco-terrorism*, *executivedom*), préfixations (*euroconvertible*) et conversions, etc. Par exemple, *gate*, par analogie avec *Watergate*, n'est guère productif en mars 1991 (seul ce mot est utilisé) mais donne naissance en fin 1992 à *iraq(-)gate*, *dianagate*, *camillagate*, *threshergate*. A. Renouf (*ibid.*, p. 286-287) donne aussi les 50 préfixes (*non-*, *re-*, *over-*, etc.) et suffixes (*-like*, *-based*, *-style*, etc.) les plus fréquents dans les composés de mars 1991.

3.2 Le moyen terme

La constitution des premiers corpus de référence pour l'anglais remonte aux années soixante, avec ***Brown*** et ***LOB***. Ces deux corpus fournissent un échantillon voulu représentatif de l'usage, américain d'un côté, anglais de l'autre, en 1961 précisément, au sein d'un certain nombre de registres. Plus de trente ans nous séparent de ces « instantanés » du début des années soixante. Aussi peut-on s'en servir pour examiner les écarts avec l'usage actuel.

C'est l'objectif de C. Mair (1995). Il compare l'emploi de *help* dans ***Brown*** et ***LOB*** avec l'usage en 1991. C. Mair a constitué pour ce faire un corpus selon les mêmes critères que ***LOB***, à ceci près que les textes retenus sont de 1991. Il appuie également son analyse sur le CD-ROM du journal *The Guardian* pour la même année. Il examine l'évolution des constructions suivantes de *help* :

> + to infinitif (*Maybe he will help to turn our fair city into a 'ghost town'*)
> + infinitif seul, éventuellement précédé d'un SN sujet « logique » de cet infinitif (*I helped him mend his bicycle*)

La deuxième construction est généralement présentée comme un américanisme dans les grammaires anglaises. Une étude détaillée indique que la première est effectivement la variante dominante en anglais dans les années soixante. Le corpus de 1991 montre (*ibid.*, p. 264) d'une part que la fréquence de *help* avec un complément infinitif s'accroît sensiblement par rapport à 1961 et d'autre part que la construction avec

[14] Pour les corpus de suivi, le problème n'est pas de réaliser une édition électronique « propre », exempte de coquilles, faisant autorité, mais de pouvoir utiliser au plus vite des données vastes qui vont se trouver rapidement remplacées par d'autres (Blackwell, 1993, p. 101). Le nettoyage ne vise pas la perfection. Il doit simplement permettre le fonctionnement des outils logiciels d'exploration des données. Vu la taille des données traitées, il doit être entièrement automatique ou limiter au maximum l'intervention humaine.

infinitif seul domine désormais (en particulier sans SN sujet « logique » de l'infinitif). La construction avec infinitif seul domine également dans le CD-ROM de 1991 du journal *The Guardian*. Comme il s'agit d'un journal dont la langue est « tenue », cette prédominance montre que la construction en cause a perdu la connotation de « relâchement » qui était la sienne trente ans auparavant. C. Mair voit dans cette évolution l'indice d'une « grammaticalisation », définie comme la transformation au fil du temps de certaines formes lexicales en simples marques grammaticales. *Help* se viderait progressivement de son sens et deviendrait un simple « étai » pour l'infinitif associé[15]. Pour C. Mair (*ibid.*, p. 267), en outre, l'opposition faite par les grammaires entre les deux constructions n'est pas tout à fait exacte. L'anglais et l'américain suivraient un mouvement parallèle, quoique décalé, dans l'évolution de l'utilisation de *help*.

3.3 La longue durée

3.3.1 La position des adjectifs en moyen anglais tardif

H. Raumolin-Brunberg (1994) étudie la position des adjectifs en moyen anglais tardif (1350-1500). Elle s'appuie sur les données d'***Helsinki***. Elle examine particulièrement l'hypothèse avancée par plusieurs chercheurs selon laquelle la position de base serait post-nominale : on trouverait globalement plus d'adjectifs après qu'avant le nom ; pour les adjectifs pouvant se présenter dans les deux positions, la post-position serait plus fréquente ; enfin, la position après le nom serait non marquée. H. Raumolin-Brunberg limite son étude à la prose pour que n'interviennent pas les contraintes sur l'ordre des mots propres à la poésie. Le sous-corpus examiné comprend 200 000 mots.

Les constats effectués dans ***Helsinki*** contrecarrent nettement l'hypothèse formulée ci-dessus. La comparaison de deux sous-périodes (1350-1420 et 1420-1500) ne montre pas d'évolution sur la position de l'adjectif, là encore contrairement à certaines propositions. En outre, l'écart entre les proportions pour les occurrences et les lemmes indique que beaucoup des adjectifs précédant habituellement le nom sont très fréquents (*great*, *good*, *holy*, etc.). Les post-posés sont au contraire peu fréquents[16], comme le montre le tableau suivant :

place de l'adjectif	occurrences	%	formes	%
pré-nominal	5 197	92,3	531	73,1
post-nominal	432	7,7	195	26,9
total	5 629	100	726	100

Les résultats obtenus sont également très proches d'études faites pour

[15] Un peu comme dans les constructions à verbe support du type *prendre peur* où le nom véhicule l'essentiel du sémantisme, le verbe apportant des indications temporelles et aspectuelles.
[16] Les adjectifs qui apparaissent uniquement post-posés sont à 90 % d'origine latine ou française.

l'anglais contemporain. Enfin, l'examen des divers registres représentés dans le sous-corpus ne manifeste pas d'écarts significatifs dans le placement des adjectifs par rapport aux constats globaux qui viennent d'être donnés.

Au regard de ces résultats, H. Raumolin-Brunberg conclut à la primauté de la position antéposée de l'adjectif en anglais, tout au long de son histoire.

3.3.2 L'alternance *that* / zéro

En anglais, après certains verbes comme *hear*, *hope*, *know*, *think*, *say* et *tell*, certaines propositions objet peuvent être introduites par *that* (*I hope that becoming a catholic will give you peace of mind*) ou rester non marquées *(I told him I had a letter from you*). Cette alternance et ses conditions ont largement été étudiées. Les données d'***Helsinki*** ont permis de montrer une tendance générale à la progression de la construction zéro entre 1350 et 1710.

Finegan et Biber (1995) reprennent l'étude de cette alternance en utilisant ***Archer***, sur la période allant de 1650 à 1990. Mais ils se restreignent à trois genres : les lettres, les sermons et les articles médicaux. Toutes périodes confondues, la répartition par construction et par registre est la suivante :

	that	zéro
sermons	89 %	11 %
médecine	83 %	17 %
lettres	53 %	47 %

Paradoxalement, les résultats pour les articles médicaux et les sermons vont à contrecourant de la tendance mise en évidence pour ***Helsinki***[17]. Au contraire, ces deux registres favorisent continûment et de plus en plus nettement la construction avec *that* par rapport à la construction zéro. Finegan et Biber interprètent ce décalage par une progression plus générale de ces registres vers une forme plus cultivée (*literate*) et moins orale. Les lettres témoignent d'une évolution comparable, mais plus atténuée (avec un étonnant renversement de tendance pour la période 1900-1949, où la construction zéro domine).

Ces évolutions décalées poussent à multiplier les points de vue dans l'analyse globale de changements linguistiques. Finegan et Biber examinent d'ailleurs les attirances de certains des verbes majeurs pour chacune de ces deux constructions, toutes périodes confondues : « [...] les verbes *say*, *tell* et *know* montrent une forte préférence pour *that* dans

[17] Finegan et Biber (*ibid.*, p. 251-253) montrent dans le détail les difficultés d'une comparaison des résultats sur ***Helsinki*** et sur ***Archer*** pour la période approximativement partagée par ces deux corpus (1640-1710 et 1650-1699 respectivement). Les principes d'échantillonnage diffèrent, on l'a vu. La taille réduite des parties correspondant à cette période pour les deux corpus fait aussi obstacle.

les trois registres, tandis que *think* montre une préférence nette pour la construction zéro, du moins en médecine et dans les lettres » (*ibid.*, p. 250).

3.3.3 L'évolution des démonstratifs en français

En français, les démonstratifs ont connu un changement morphologique radical. Aux XI^e^ et XII^e^ siècles, s'opposent sémantiquement deux paradigmes de démonstratifs. Le premier (désormais CIST) est issu du latin vulgaire *ecce iste*, le second (désormais CIL), d'*ecce ille*. Le premier exprime la proximité, le second l'éloignement, temporel ou spatial, soit par rapport à l'auteur, soit par rapport à l'un des personnages[18]. Chacune des formes peut être aussi bien pronom (*Cil vient*) que déterminant (*Cil chevaliers vient*) et il existe en outre des formes longues préfixées par *i-* : *icelui,* etc. Rappelons que l'ancien français possède une déclinaison opposant deux cas : le cas-sujet (issu du nominatif latin) et le cas-régime (issu de l'accusatif latin). S'ajoute parfois, c'est le cas pour les démonstratifs, un second cas-régime singulier (issu du datif latin). Au total, 14 formes différentes (28 si l'on inclut celles préfixées en *i-*). À partir du XVII^e^ siècle, le paradigme des pronoms (*Celui-ci vient*) est totalement séparé de celui des déterminants (*Cet homme vient*). Une étape marque le passage d'un système à l'autre. Au XII^e^ siècle, apparaît au nord de la France une nouvelle forme de cas-régime masculin pluriel : *ces*, toujours déterminant, va ensuite être employé également au féminin pluriel. Fin XII^e^-début XIII^e^ siècle, apparaît *ce*, déterminant masculin singulier au cas-régime, employé uniquement devant un mot commençant par une consonne (*ce chevalier*). C'est en fait un nouveau paradigme qui émerge, le troisième : *ce* / *ces*, uniquement déterminant et toujours atone, sans opposition de genre au pluriel, et sémantiquement indifférencié (pas d'opposition proximité / éloignement).

Ce changement profond n'a pas d'équivalent dans la plupart des autres langues romanes, où les formes de démonstratifs continuent à être employées à la fois comme déterminants et comme pronoms. Il reste énigmatique : les changements phonétiques ne suffisent à expliquer ni la spécialisation globale des paradigmes ni la sélection des formes survivantes au sein de chaque paradigme.

L'objectif de C. Marchello-Nizia (1995, p. 115-181) est d'expliquer dans le détail la répartition et l'évolution des différentes formes. Les hypothèses qu'elle propose s'appuient sur des constats que seul permet le traitement de très gros corpus[19]. Elle souligne en effet (*ibid.*, p. 138-139) : « Par généralisation ou simplification abusive, on gomme le fait que ce n'est pas tout le paradigme de CIL qui est devenu pur pronom, mais seulement

[18] L'opposition sémantique entre les deux séries, indéniable, est plus complexe. Elle a suscité de nombreuses analyses (Marchello-Nizia, p. 129-130). L'hypothèse actuellement la plus satisfaisante, selon C. Marchello-Nizia, est celle de G. Kleiber (*ibid*., p. 129-137). Pour ce dernier, les formes en CIST indiquent au destinataire qu'il faut opérer l'appariement référentiel à partir du contexte d'énonciation immédiat de l'occurrence (contexte spatio-temporel représenté ou contexte énonciatif ou discursif), ce qui n'est pas le cas pour les formes en CIL.

[19] Cf. section 4.1 sur la taille des corpus historiques.

quatre formes sur sept : *celui*, *celle*, *ceux*, *celles* ; *cil*, *cel* et *celi* ont disparu. Pour *cil*, on peut dire qu'il s'agissait d'une forme de cas-sujet (singulier ou pluriel), et dès lors que la déclinaison disparaissait, les formes qui instanciaient les différents cas devaient disparaître. Mais pourquoi est-ce *celui* qui s'est conservé et non *cel*, et pourquoi à l'inverse pour le féminin est-ce *celle* et non pas *celi* qui s'est conservé ? De même, ce n'est pas tout le paradigme de CIST qui s'est conservé en devenant pur déterminant. Sur six, seules deux formes, la forme du féminin singulier *cette*, et celle du masculin singulier devant voyelle *cet*, viennent directement du paradigme CIST. *Ce* n'en provient pas, non plus que proprement le pluriel épicène[20] *ces*[21]. Les autres formes, au nombre de quatre (*cist*, *cestui*, *cez*, *cestes*), ont disparu. *»*

C. Marchello-Nizia s'appuie sur un important corpus d'ancien et de moyen français. Pour l'ancien français, ont été utilisés seize textes en vers ou en prose (*ibid.* p. 147-148), soit près de 685 000 mots, s'échelonnant de 1100 environ à 1300 environ. Ces textes se situent dans le domaine littéraire, central dans les recherches des médiévistes, et une concordance est disponible pour chacun d'eux. Ils comprennent 8 237 démonstratifs. Pour le moyen français (XIV^{e} et XV^{e} siècles), le corpus utilisé pour la constitution du *Dictionnaire du Moyen Français* (INaLF, Nancy), qui compte environ 4 millions de mots et qui est d'origine plus variée[22], a fourni près de 36 000 occurrences de démonstratifs.

L'examen détaillé des concordances des formes longues (préfixées en *i-*, suffixées en *-ui* / *-i*, ou portant les deux affixes) dans le corpus d'ancien français[23] permet de mieux cerner les notions de *«* soulignement *»*, d'*«* expressivité *»*, de *«* renforcement *»*, utilisées jusqu'alors. Ces formes sont en effet employées en début de phrase ou de vers. Elles sont pronoms dans 3 cas sur 4 pour les formes suffixées en *-ui* / *i* et déterminants dans deux tiers des cas pour les formes préfixées en *-i* . Elles déterminent alors le plus souvent un substantif complément d'objet placé en tête de phrase. Elles mettent en évidence cette construction, marquée à cette époque.

A partir de ces observations, C. Marchello-Nizia (*ibid.*, p. 144) formule l'hypothèse d'une répartition des démonstratifs en trois groupes : les formes toujours atones (*ces* et *ce*), les formes toujours toniques (les formes longues) et les formes pouvant être atones ou toniques (*cil*, *cel*, *cele*, *ceus* et *cist*, *cest*, *ceste*). C'est dépasser l'opposition déterminant / pronom et prendre en compte la dimension accentuelle.

Les cas-sujets masculins singuliers *cil* et *cist* suivent bizarrement une évolution décalée : *cist* s'efface à partir de 1250, en lien avec la chute de la déclinaison, tandis que *cil* reste employé jusqu'à la moitié du XV^{e} siècle, où il connaît une disparition brutale. C'est un parallélisme avec le pronom personnel *il* qui expliquerait cette évolution de *cil* : on constate en effet une évolution parallèle de *il* et de *cil* (*ibid.* p. 164). En outre, les

[20] Utilisable au féminin et au masculin.
[21] Cette forme provient à la fois de *cez* (de la série CIST), par évolution phonétique de l'occluso-constrictive finale [ts] en [s] et de *cels* (de la série CIL), employé de façon inaccentuée et proclitique comme déterminant. *Ce* est fait par analogie sur *ces*.
[22] 182 œuvres différentes, de longueur inégale et de divers genres (chroniques, romans, chansons de geste, poésie lyrique ou didactique, chartes, traités philosophiques, etc.).
[23] 1 027 occurrences sur 8 237 démonstratifs.

comptages opérés montrent qu'en moyen français, les deux paradigmes CIST et CIL ne sont pas encore spécialisés, l'un pour les déterminants, l'autre pour les pronoms. Les emplois pronominaux sont occupés essentiellement par trois formes : *celui*, *celle*, et *cestui*. Ce serait là encore l'influence du système pronominal qui aurait joué. Ont en effet été conservées comme pronoms démonstratifs les formes (*celui*, *ceux*, *celle*, *celles*) ressemblant aux pronoms personnels employés de manière autonome (*lui*, *eux*, *elle*, *elles*), celles sans correspondant pronominal disparaissant (comme *celi*, *cesti*, *cestui*). Par ailleurs, les formes longues se spécialisent en moyen français dans la fonction de pronom, alors que dans la période précédente, la détermination focalisante les caractérisait. Ce serait aussi le contrecoup du remplacement progressif de l'accent tonique de mot à valeur distinctive, encore présent en ancien voire en moyen français, par l'accent en fin de groupe syntaxique, la détermination marquée trouvant dans *-ci* et *-là* post-fixés le moyen de souligner cet accent de groupe. Cette évolution est une deuxième étape dans le mouvement de distinction entre la catégorie du pronom et celle du déterminant, mouvement amorcé avec l'apparition du déterminant *ce* / *ces*, et achevé à la fin du moyen-âge par l'institution de formes purement pronoms.

4. PROBLEMES METHODOLOGIQUES

La constitution et l'annotation de corpus diachronique rencontrent des obstacles spécifiques. Les ressources résultantes permettent néanmoins de vérifier, de préciser les évolutions et de renouveler les explications qui en sont fournies.

4.1 Des corpus « petits » et peu annotés

La constitution même des corpus pose des problèmes spécifiques pour les états anciens d'une langue où les sources sont des manuscrits (l'ancien français par exemple). Les variantes graphiques d'une même forme peuvent être nombreuses[24]. Mais il est désormais possible de mémoriser et de relier différents types de documents. C'est le cas du projet *Charrette* dirigé par K. Uitti (Université de Princeton) : les transcriptions diplomatiques des huit manuscrits du XIIIe siècle du *Chevalier de la Charrette* de Chrétien de Troyes, soit près de 36 000 lignes pour un poème d'environ 7 100 lignes, sont reliées à une version électronique de l'édition Foulet-Uitti et aux images de ces manuscrits. La philologie voit ainsi s'ouvrir de nouvelles perspectives.

Nous l'avons vu, le développement des corpus électroniques a très

[24] Les 28 formes de démonstratifs repertoriées par C. Marchello-Nizia (1995) se réalisent en plus de 80 graphies.

largement bénéficié cette dernière décennie des apports, techniques et financiers, de la communauté du TALN qui voit là une étape indispensable pour la mise au point de systèmes de traitement du langage robustes. L'accent est bien sûr mis sur la langue contemporaine. Autrement dit, il n'y a pas vraiment de raisons que beaucoup de temps et d'énergie soit consacré à la recherche sur les états de langue anciens. On peut donc escompter un retard sensible dans les techniques et les moyens mis en œuvre pour l'annotation des corpus historiques. Les corpus historiques actuels sont d'ailleurs très sensiblement plus petits que les corpus synchroniques (Finegan et Biber, 1995). Que l'on compare le million et demi de mots d'***Helsinki*** ou d'***Archer*** avec les 100 millions de mots (étiquetés, au surplus) de ***BNC***.

En dehors de ces projets de corpus conçus pour étudier la diachronie, parce qu'il est coûteux de constituer des corpus bien répartis sur les genres et les périodes, les constats sont souvent établis sur les ensembles de textes qui sont effectivement disponibles sous forme électronique mais qui ne forment pas vraiment un corpus historique au sens d'***Archer*** par exemple. Cette situation biaise évidemment les observations et leur interprétation, sans que les chercheurs qui ont recours à ces rassemblements de circonstance en soient toujours conscients.

L'annotation de ces corpus se heurte en outre à des obstacles spécifiques. Une langue à cas comme l'ancien français connait une variation importante dans l'ordre des mots, alors que les étiqueteurs et parseurs disponibles ont été conçus pour des langues où l'ordre des mots est notablement plus contraint. La connaissance du lexique et de la syntaxe de ces états de langue n'offre pas non plus le même appui à une automatisation. À l'inverse, ces corpus historiques étant destinés, pour leur très grande majorité, à rester « nus », ils ne permettent pas facilement de valider ou d'invalider des hypothèses linguistiques. Ils supposent une analyse très souvent manuelle des données[25] pour trier les faits et proposer des hypothèses, mais aussi pour comparer la représentation formelle postulée avec le corpus. Ainsi, T. Nevalainen (1994), pour étudier l'évolution de l'opposition en anglais entre les formes des adverbes en *-ly* et sans suffixe (*slowly* / *slow*) en contrastant la période 1350-1420 avec la période 1640-1710, commence par extraire d'***Helsinki*** les formes se terminant en *-ly* (elle répertorie 14 variantes graphiques du suffixe), élimine celles qui ne sont pas des adverbes ainsi que les adverbes faits sur une base nominale (*namely*), et cherche les adjectifs ayant servi de base aux adverbes ainsi isolés. Ce sont encore de simples concordances qui sont employées par Finegan et Biber (1995, p. 245) dans leur étude de l'alternance *that* / zéro après certains verbes.

4.2 Vérifier et préciser les évolutions

C. Mair (1995, p. 260) résume assez bien ce que la linguistique

[25] Même si des environnements informatiques adéquats allègent parfois la charge.

diachronique va gagner dans ces nouvelles études : « L'approche du changement linguistique basée sur les corpus corrigera des distorsions évidentes dans la littérature actuelle sur le sujet. Il sera possible de séparer l'usuel et le normal de l'exceptionnel. À la différence de l'observateur qui enregistre l'exemple unique d'une nouvelle construction tout en omettant de noter les preuves massives de la persistance de l'ancienne construction, l'analyste de corpus sera en position de décrire les tendances statistiques avec précision. »

Ce constat se vérifie déjà pour l'exemple des démonstratifs en français. Les textes de la période effectivement disponibles sous forme électronique ne couvrent pas, loin s'en faut, tout ce qui est répertorié. Les conclusions et décomptes actuels seront donc sans doute infléchis[26]. Le recours au corpus permet néanmoins une finesse d'analyse de l'évolution, forme par forme, du système des démonstratifs, qui n'était pas envisageable auparavant. Il entraîne surprises, réévaluations et découvertes : « [...] le grand nombre des données qui nous sont désormais accessibles montre une situation fort inattendue en moyen français » (Marchello-Nizia, 1995, p. 165). Mais il en va de même pour l'opposition *that* / zéro, et pour la position des adjectifs en moyen anglais tardif.

C. Mair ajoute (1995, p. 260) : « [...] les innovations grammaticales généralement ne bouleversent pas le langage mais s'établissent d'abord dans des genres textuels spécifiques, des registres ou des niches fonctionnelles. Les corpus, comme témoignages de performance réelle, rendront plus faciles l'étude de ces types de contraintes. » Cette démarche est exemplifiée par l'étude de l'alternance *that* / zéro. Elle reste à entreprendre pour la position des adjectifs (seule la prose a été étudiée) et pour les démonstratifs. Il n'est pas exclu en effet que la distinction poésie / prose influence l'emploi des démonstratifs, en particulier pour la répartition entre déterminants et pronoms.

4.3 Acceptabilité et fréquence

Par définition, il n'existe pas, pour les états disparus d'une langue, de compétence du locuteur actuel. L'érudit contemporain ne saurait affirmer : cet énoncé n'est pas acceptable. En effet, sa connaissance de ce qui lui paraît possible ou non dans la période qu'il étudie provient uniquement de sa connaissance intime de textes en nombre fini, dont il a fini par abstraire les mécanismes lexicaux et syntaxiques dominants. Elle n'équivaut pas, loin s'en faut, à une capacité à produire des énoncés relevant de cet état de langue. La perception des régularités à l'œuvre est probablement distordue, dans les deux sens : certains faits de très faible fréquence peuvent avoir échappé à l'attention et, à l'inverse, certaines caractéristiques dominantes peuvent être sous-estimées. L'oral est par ailleurs insaisissable, sinon par les biais qu'offrent certaines types d'écrits,

[26] D'où des précautions légitimes comme : « [...] après 1340, au moins en l'état actuel de notre documentation, on ne trouve plus aucune trace de ce morphème *cist* en français » (Marchello-Nizia, 1995, p. 159).

avec le risque que rappelle C. Blanche-Benveniste (1997, p. 36) à propos de la *Grammaire des fautes* d'H. Frei de « confondre fautif et parlé », et de prendre « les fautes typiques de scripteurs inexpérimentés » pour des reflets de l'oral. La découverte de nouveaux documents, de nouvelles éditions critiques peuvent en plus amener à réévaluer la place de certains phénomènes[27].

Les corpus permettent par contre d'approcher les régularités centrales d'un état de langue oublié. Pour cerner les *«* impossibles de langue *»*, C. Marchello-Nizia (*ibid.*, p. 22) propose de recourir au raisonnement suivant : *«* On accordera [...] une importance privilégiée à l'absence de formes ou de constructions attendues, et corrélativement aux paraphrases. En effet, si un tour attendu n'est jamais attesté, et qu'on rencontre régulièrement sa paraphrase en lieu et place où on l'attendait, alors on a le droit de formuler l'hypothèse que le tour qu'on attendait là est, dans ce cas, agrammatical. *»*

La quantification occupe par conséquent une place centrale. Mais elle rencontre des difficultés sur des corpus d'états anciens de la langue. Lorsqu'il s'agit d'étudier des propriétés linguistiques *«* fines *»*, le nombre d'occurrences d'un phénomène donné dans une partie du corpus est souvent faible (inférieur à la dizaine). Il n'est d'ailleurs pas toujours possible, soit pour des raisons de coût soit plus fondamentalement parce que les sources sont lacunaires, de compléter les inventaires du phénomène visé. Ces petites quantités ne rendent cependant pas pour autant illégitime le recours à des modèles probabilistes appropriés pour évaluer leur significativité. Certains de ces modèles sont présentés au chapitre IX.

4.4 Affiner les explications

Le recours à des corpus diachroniques favorise pour l'analyse du système des démonstratifs en français un renouvellement de l'explication du changement morphologique. Traditionnellement, la causalité retenue était la suivante : un changement phonétique déclenche un changement morphologique qui peut lui-même entraîner un changement syntaxique. Les études récentes sur lesquelles s'appuie C. Marchello-Nizia poussent à relativiser dans ce cas le poids des changements proprement phonétiques (pour *ces*, par exemple). Parallèlement, les concordances facilitent l'étude détaillée des comportements syntaxiques (par exemple pour les formes préfixées en *i-*) et l'existence de textes enregistrés en nombre suffisant, une périodisation précise pour chaque forme (*cil* et *cist* par exemple). Ces données et ces outils permettent de donner consistance aux facteurs qui sont invoqués : l'évolution de l'accent, qui passe du mot au groupe syntaxique, et l'influence de parentés de plus haut niveau, de systèmes méta-morphologiques et sémantiques généraux

[27] « [...] les textes nous parviennent par copistes, et parfois générations de copistes interposées, auxquels s'ajoute inévitablement l'intervention de l'éditeur moderne ; jamais un texte n'est le pur reflet de l'usage de l'auteur ; il s'agit nécessairement d'une langue hybride [...] » (Marchello-Nizia, p. 22).

(avec la restructuration du système pronominal).

Nous avons vu l'usage de la notion d'analogie pour expliquer l'« invention » de *ce* : il viendrait compléter *ces* et faire pendant avec lui au couple *le* / *les*. C. Marchello-Nizia rappelle (*ibid.*, p. 176-178) les critiques qu'appelle l'usage de cette notion pour rendre compte, en dernière instance, de certaines évolutions[28]. L'analogie est le plus souvent utilisée au coup par coup. Elle fonctionne alors comme *«* explication » de la dernière chance. Elle est utilisée de manière *«* superficielle *»,* par opposition à des règles dûment formalisées.

Au delà des explications parfois hasardeuses par l'analogie, l'annotation linguistique de corpus étalés dans le temps fournit désormais la possibilité d'étudier des corrélations extrêmement complexes – et pratiquement non perceptibles sans appui informatique – entre des phénomènes situés aux différents niveaux de l'analyse linguistique ainsi que leur évolution au fil du temps. C'est le cas d'une des hypothèses majeures de C. Marchello-Nizia : la corrélation de l'évolution des démonstratifs avec celle des pronoms personnels. On souhaiterait alors tout naturellement dépasser le recours à des concordances et des comptages sur les seuls démonstratifs pour disposer de données chiffrées sur les deux systèmes et pouvoir examiner les corrélations, si elles existent, entre eux, par le recours, par exemple, à l'analyse multi-dimensionnelle (cf. chapitre IX). On progresserait vers le test effectif de l'hypothèse plus générale qui est posée (*ibid.*, p. 168) : *«* les systèmes morphologiques des langues s'organisent à un niveau supérieur en macro-systèmes sémantiques et formels plus abstraits, et ce sont ces méta-structures qui sont cause de certains des changements qui affectent les systèmes du niveau inférieur, immédiatement perceptibles, eux. *»* Dans une optique proche, les contraintes pesant sur l'omission du sujet pronominal en moyen français sont soumises dans (Dupuis *et al.*, 1992) à une analyse multivariable. À partir de l'examen de la distribution du sujet dans 10 textes s'échelonnant du premier tiers du XIVe siècle jusqu'à la fin du XV^{e} siècle, cette analyse montre que, parmi les facteurs examinés : la période du texte, l'opposition prose / poésie, le type de proposition et la personne du sujet, c'est le type de proposition dont l'influence ressort nettement : l'omission est plus souvent le fait des principales et des indépendantes que des enchâssées.

Les analogies réelles devraient être désormais plus facilement objectivables. La vision des causalités à l'œuvre dans le changement linguistique en sera probablement renouvelée. Ces causalités sont peut-être à chercher à des niveaux de structuration beaucoup plus abstraits (Kroch, 1990, p. 239) que ceux qui sont envisagés généralement.

[28] Cf. aussi (Kroch, 1990, p. 238)