Quand il ne restera que l'induction : corpus, hypothèses diachroniques et la nature de la description grammaticale

Mario Barra-Jover

Mario Barra Jover
Paris 8 / CNRS (UMR 7023)

S'il ne restait que l'induction : corpus, hypothèses diachroniques et la nature de la description grammaticale

> C'est donc peut-être un sujet digne d'éveiller la curiosité que de rechercher quelle est la nature de cette évidence qui nous assure de la réalité d'une existence et d'un fait au-delà du témoignage actuel des sens ou des rapports de notre mémoire.
> Hume, *Enquête sur l'entendement humain* (Hume 1758 / 1983 : 86)

1. Occurrences, représentations et hypothèses

Cet article[1] propose une réflexion épistémologique qui, partant de la méthodologie utilisée (et utilisable) en linguistique diachronique, touche

[1] Je remercie Brenda Laca pour ses remarques et ses conseils qui m'ont permis de rendre plus clairs quelques passages et de supprimer des erreurs et des maladresses.

finalement à notre conception de ce qu'est une description grammaticale synchronique. Il semble pertinent, d'emblée, de rendre explicite le sens dans lequel des termes comme « constatation », « corpus », « exemple », « hypothèse », « induction », « déduction » et autres seront utilisés. Pour ce faire, nous pouvons commencer par une présentation naïve de cinq opérations pouvant sous-tendre les descriptions grammaticales :

(1) Constatation des occurrences. Ainsi, on peut repérer et isoler l'énoncé *le chien a eu très peur*, prononcé dans une situation donnée, par Hélène Thibault dans la rue Magenta de Tours, le 5 février 2004 à 15h03. Il va de soi que toute occurrence est unique et non reproductible car il s'agit d'un fait historique.

(2) Promotion des occurrences au rang de représentations[2]. On suppose que la séquence en question peut se reproduire dans un nombre potentiellement infini de cas. *Le chien a eu très peur* devient ainsi un membre de l'ensemble {le chien a eu très peur}.

(3) Description de la représentation en termes de variables et établissement de catégories. On divisé la séquence en entités et on attribue à chaque entité l'appartenance à tel ou tel ensemble et sous-ensemble. Ainsi, {le chien a eu très peur} devient, par exemple, [Article + Nom + Verbe + Adverbe + Nom]

(4) Délimitation des ensembles et sous-ensembles pouvant spécifier les variables et établissement de règles combinatoires. On va repérer des fonctions, des relations, des hiérarchies et autres. Ainsi, [Article + Nom+ Verbe + Adverbe + Nom] devient, par exemple, $[_{Ph}\,[_{GN}$ *le chien* $[_{GV}$ *a eu* $[_{GN}$ *très peur*]]]]. A partir de là, il sera possible de le manipuler comme une structure et de proposer des contraintes plus ou moins fines aux possibilités combinatoires potentielles comme, par exemple, « cette construction n'est possible qu'avec des verbes comme *avoir* et *faire* lorsque leur objet n'apparaît pas introduit par un déterminant », ce qui permet de faire la prédiction que, étant donné que la représentation [le chien a eu très folie] est agrammaticale, l'occurrence **le chien a eu très folie* est impossible.

(5) Fournir une explication cohérente avec la théorie utilisée, c'est-à-dire avec l'ensemble d'hypothèses qui a été déployé pour réussir les opérations 3 et 4.

Ces cinq « opérations », ne sont pas nécessairement autant d'étapes ordonnées et certaines d'entre elles peuvent être superposées ou négligées. Elles sont, tout simplement, utiles pour rendre clair et stable le sens de certains termes ou formulations.
Ainsi, lorsqu'on donne un « exemple », on peut proposer une occurrence ou une représentation. *Le chien a eu très peur* est un exemple-occurrence s'il procède d'un corpus, parce qu'un corpus doit, avant tout, être un répertoire

[2] Le contraste entre « occurrence » et « représentation » peut, dans certaines circonstances, correspondre à l'opposition *token* / *type*, mais je l'utilise dans un sens plus large. C'est pourquoi j'ai préféré éviter les termes anglais.

organisé d'occurrences. Mais, une fois qu'il est manipulé ou, simplement, proposé comme objet reproductible, il devient un exemple-représentation. Les linguistes travaillant systématiquement sur corpus défendent l'idée que tout exemple-représentation doit être fondé sur des occurrences, ce qui n'est pas toujours le cas dans certains travaux où les exemples-représentations peuvent être proposés sans qu'une occurrence les légitime. Je peux, en effet, proposer l'exemple « Ma grand-mère a été élue Miss France à 80 ans » sans qu'il soit nécessaire de constater s'il existe une occurrence isomorphique. On parle, dans ce cas, d'exemples artificiels (opposés aux authentiques) mais le terme est bien malheureux car ils sont « fabriqués » avec le même mécanisme qui produit les occurrences[3].

D'autre part, on aimerait dire que le raisonnement inductif va de (1) à (5) tandis que le raisonnement déductif va de (5) à (1). Mais c'est loin d'être le cas, car il ne semble pas possible de procéder de façon strictement unidirectionnelle[4]. Qui plus est, il est souhaitable d'établir une différence entre les deux dimensions philosophiques de l'induction.

La première concerne le passage de l'opération (1), constatation des occurrences, à l'opération (2), promotion des occurrences au rang de représentations. Nous avons, dans ce cas, affaire au problème de l'induction tel que le pose Hume (1758/1983) dans le texte qui sert d'exergue à cet article. En termes linguistiques, il pourrait être formulé de la façon suivante :

(6) Existe-t-il un argument tiré de l'observation des données qui me fasse avoir la certitude logique qu'un énoncé quelconque, entendu ou lu même une seule fois, a pu apparaître et pourrait apparaître un nombre potentiellement infini de fois ?

Hume (1758/1983 : 96-97) reconnaît seulement la régularité comme source de certitudes (7a) mais il nie toute fondation logique aux prédictions (7b et 7c) :

(7) a. « C'est seulement après un long cours d'expériences uniformes d'un genre donné que nous atteignons une ferme confiance et de la sécurité à l'égard d'un événement particulier ».

b. « Quand on dit : j'ai trouvé, dans tous les cas passés, telles qualités sensibles conjointes à tels pouvoirs cachés, et quand on dit : des qualités sensibles semblables seront toujours conjointes à de semblables pouvoirs cachés, on ne se rend

[3]Un exemple artificiel est celui produit, par exemple, par une machine de traduction, dans la mesure où le mécanisme utilisé pour la production n'a (j'en ai la conviction) rien à voir avec le nôtre. Cf. Barra Jover (2000).

[4] Il me semble qu'un chercheur peut afficher un « discours inductif » lorsqu'il présente ses résultats. Il montre d'abord ses données de façon strictement observationnelle, puis il repère les questions qu'elles soulèvent, il avance ensuite des hypothèses qu'il teste et il propose une hypothèse finale comme solution. Mais je suis loin de croire que cette organisation du discours reproduise vraiment le parcours suivi dans le processus d'élaboration des hypothèses.

pas coupable d'une tautologie (…). Vous dites que l'une des propositions est une inférence tirée de l'autre. Mais il vous faut avouer que l'inférence n'est pas intuitive, et qu'elle n'est pas démonstrative ».

c. « Il est donc impossible qu'aucun argument tiré de l'expérience puisse prouver cette ressemblance du passé au futur, car tous les arguments se fondent sur la supposition de cette ressemblance ».

Mais, comme le signale Popper (1991 : 43-45), si nous posons le problème de Hume en termes psychologiques et non logiques, il peut exister une telle certitude. La question se pose alors de savoir quelle en est la source, s'il ne s'agit pas de la simple constatation d'une régularité. En termes à nouveau strictement linguistiques, ce problème peut sembler banal en synchronie mais il se pose de façon évidente en diachronie comme nous allons le voir lors de notre réflexion sur la représentativité des corpus.

La deuxième dimension de l'induction concerne le passage des opérations (1-2) aux opérations (3-5), c'est-à-dire, comment aller de façon empiriquement fondée du particulier au général. On peut appeler cela la solution de Bacon dont la position est la suivante :

(8) Bacon (1620 / 2004) :

> §21 : En effet, l'esprit brûle de sauter au plus général pour s'y reposer ; et, au moindre délai, il se dégoûte de l'expérience.
> §45 : L'entendement humain, en vertu de son caractère propre, est porté à supposer dans les choses plus d'ordre et d'égalité qu'il n'en découvre.
> §50 : Mais toute interprétation plus vraie de la nature s'obtient à l'aide d'instances et d'expériences convenables et appropriées. Là, les sens jugent de l'expérience seule, l'expérience, de la nature et de la chose même.

Nous avons ici affaire, bien entendu, à un problème d'ordre méthodologique concernant la démonstration empirique qui doit contrôler les hypothèses et non à un problème psychologique concernant la genèse même des hypothèses. Ce dernier aspect ne fera pas partie du développement qui suit mais je voudrais simplement insister sur l'idée qu'il est difficile d'accepter que les hypothèses scientifiques soient uniquement le résultat de l'observation de régularités dans les données. Ceci impliquerait aussi d'accepter qu'il y a des énoncés scientifiques strictement observationnels, alors qu'il est improbable qu'une constatation quelconque ne soit pas imprégnée de théorie. C'est pourquoi le problème de la genèse des hypothèses a été souvent négligé[5]. Popper (1990 : 56) expédie vite la discussion, non sans humour, d'ailleurs :

[5] Il est négligé dans les approches « réalistes » de la science, c'est-à-dire les approches qui partent du fait que les théories scientifiques sont des découvertes qui convergent progressivement avec la structure de la réalité. Il est moins négligé par les « anti-réalistes » (parmi lesquels l'auteur de ces lignes) qui partent du présupposé que les

(9) Car la question factuelle, d'ordre psychologique et historique : « comment en arrivons-nous à nos théories ? », si passionnante qu'elle puisse être en elle-même, n'a rien à voir avec la question logique, méthodologique et épistémologique de la validité (…).
Certains scientifiques, paraît-il, estiment qu'ils trouvent leurs meilleures idées en fumant, d'autres en buvant du café ou du whisky . Aussi n'y a-t-il aucune raison pour moi de ne pas admettre que certains puissent les trouver en observant ou grâce à des observations répétées.

Un peu plus loin, on essaiera de relier le problème de Hume et la solution de Bacon[6] à propos des hypothèses diachroniques. Pour l'instant, il ne reste qu'à expliciter le sens accordé au terme « déduction » lorsqu'on parle de linguistique, car nous ne saurions nous servir du terme dans son sens strict (une proposition déduite dans un cadre théorique est une proposition dont la vérité se démontre grâce à une chaîne d'implications logiques partant des axiomes de ce cadre). Je vais adopter l'interprétation informelle selon laquelle un chercheur opère de façon déductive lorsqu'il construit, à des fins théoriques, un raisonnement conceptuellement acceptable et qu'il cherche uniquement les données servant à le prouver. Autrement dit, lorsqu'il n'y a pas de démonstration empirique. Pour ce qui est des hypothèses diachroniques, cette méthode est adoptée lorsqu'un programme de recherche développé indépendamment de la diachronie aspire à conforter ces hypothèses avec des données historiques[7] ; mais elle est aussi adoptée dans

---

théories scientifiques sont des constructions proposant des entités et des états de choses à pouvoir prédictif et explicatif et pouvant être confrontées aux données mais sans qu'il existe d'isomorphisme entre leurs entités et leurs états de choses, et ceux de la réalité. Cette position n'entre pas, bien entendu, en contradiction avec le réalisme ontologique et c'est peut-être Wittgenstein qui l'a le mieux exprimé:
(i) « 4.2211- Quand même le monde serait infiniment complexe de telle sorte que chaque fait consistât en une infinité d'états de choses et que chaque état de choses se composât d'une infinité d'objets, il faudrait encore qu'il y ait des objets et des états de choses ». (Wittgenstein 1918/1961 : 58).
Il est évident que l'exemple le plus remarquable d'une théorie anti-réaliste des hypothèses scientifiques est Kant (bien explicitement dans Kant 1783/1993). Mais il y a eu plus récemment des contributions venant de la philosophie analytique, comme celle de Van Fraassen (1994), fondée sur le symétrie, ou celle de Corradi-Fiumara (1995), fondée sur la métaphore. L'idée de changement de métaphore dans le domaine linguistique a été suggérée dans Barra Jover (2000b) à propos du programme minimaliste.

[6] La solution de Bacon annonce l'« expérience cruciale » de Popper (1991). Toutefois, Popper, chose surprenante, ne fait pas la moindre allusion à Bacon.

[7] Comme nous allons le voir, le meilleur exemple de cette méthode est offert par les générativistes qui, comme Lightfoot, cherchent à prouver une théorie du changement avec des cas précis mais sans un vrai travail « de terrain » préalable. Dans la

des grammaires historiques romanes où le point de départ est un latin plus ou moins idéalisé et le point d'arrivée est l'état moderne de la langue ou des langues en question. Dans cette situation, les étapes intermédiaires sont conçues de façon déductive et sont confortées (si jamais elles le sont) à l'aide d'exemples disparates sélectionnés en fonction de leur correspondance avec les étapes intermédiaires postulées (Tekavcic 1972 en est un très bon exemple). Il va sans dire que la phonologie diachronique est souvent obligée d'embrasser, faute de données graphiques interprétables, ce procédé comme étant le seul accessible.

Cette présentation du cadre conceptuel dans lequel les termes seront utilisés permet d'annoncer certains des problèmes ponctuels qui seront abordés par la suite. Dans §2, nous parlerons de corpus en visant particulièrement ce qui touche à leur représentativité (c'est-à-dire le problème de Hume). Ceci nous permettra de présenter les obstacles qu'une discipline comme la linguistique diachronique, qui n'a d'autre source d'information que les corpus, semble ne pas pouvoir surmonter. Dans §3, sont exposées quelques unes des positions adoptées face à ces limitations et dans §4, est proposée ma propre solution : une méthodologie permettant d'introduire des expériences et, partant, d'ouvrir la voie à des démonstrations empiriques (c'est-à-dire la solution de Bacon). L'acceptation de cette méthodologie a des conséquences sur notre idée de ce qu'est une description grammaticale. C'est pourquoi, en §5, nous sortirons de l'espace diachronique pour nous poser quelques questions sur le type d'objet que nous créons lorsque nous procédons à une description grammaticale.

2. Corpus et représentativité

Retenons la définition de corpus avancée par Rastier (2005 :32) :

(10) « Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique et réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications ».

Il n'est pas rare que ce type de définition soit assorti d'une mention à l'authenticité. Par exemple :

(11) « Not everybody, of course, goes along with these assumptions but in general there is consensus that a corpus deals with natural, authentic language » (Tognini Bonelli 2001 : 2)

---

bibliographie française, l'école inspirée par Guillaume peut procéder aussi de cette façon.

2.1. La représentativité en général

Parmi les divers aspects qui se prêtent à commentaire, je vais me concentrer sur la question de la représentativité. Rastier (2005 : 32) ajoute à sa définition un commentaire qui en vaut d'autres :

(12) « Aucun corpus ne représente la langue : ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique qui comprend l'ensemble des documents disponibles dans une langue ».

L'idée essentielle de Rastier est que tout corpus est conçu à des fins spécifiques dont il est difficilement extrapolable. Il est donc question ici de représentativité qualitative.
Dans ce cas, les questions qui se posent touchent à la sélection des échantillons permettant d'assurer que toutes les variétés de ce complexe langagier que nous appelons, pour des raisons pas toujours raisonnables, « une langue » sont convenablement représentées (cf. Biber 1993). Un bon échantillonnage suppose ainsi que la dispersion d'une séquence à travers les différentes variétés permet de juger de son degré d'implantation. La représentativité quantitative est un paramètre plus intuitif mais elle est nécessaire pour que tout calcul de fréquence relative soit interprétable.
L'affirmation de Rastier me semble raisonnable dans la mesure où la représentativité d'un corpus ne peut pas être appréciée en dehors du type de questions que l'on va lui poser. Il est évident qu'un corpus de 100.000.000 mots comme le *British National Corpus* est représentatif du point de vue lexical, si nous tenons compte du fait que ce corpus est composé de quelques 30.000 items qui se répètent[8]. Il est moins évident qu'il le soit du point de vue syntaxique, dans la mesure où le système étudié est constitué d'une série non exhaustivement parcourable d'occurrences[9]. Les unités lexicales ou les systèmes morphologiques flexionnels ne produisent pas de représentations à combinaisons de variables potentiellement infinies, tandis que les séquences syntaxiques (et peut-être celles de la morphologie dérivationnelle) le font. Nous avons affaire ici à une sorte de piège conceptuel où il est plus facile d'entrer que de sortir. C'est pourquoi je propose de l'éviter en formulant le problème de la représentativité autrement.

---

[8] Je remercie Achim Stein pour m'avoir donné des arguments statistiques à ce propos.
[9] Admettons, de façon entièrement spéculative et comme simple expérience de pensée, que chaque locuteur d'une langue produit une phrase de cinq mots par minute et que chaque phrase peut contenir une séquence différente à partir des 30.000 unités composant le lexique d'une langue. En faisant abstraction, pour compenser le fait de négliger l'ordre, de l'existence de flexion et de la possibilité de répétition d'un mot dans la même phrase, nous obtenons en trente ans de production langagière dans un pays comme la France de milliards de milliards de phrases différentes qu'aucun corpus ne peut représenter statistiquement.

Lorsque l'on pose à un corpus la question « la séquence S est-elle attestée dans la langue L ? » et que la réponse est affirmative, qu'est-ce qui permet de l'interpréter comme « la séquence S appartient à la langue L » ? Et, surtout, qu'est-ce que veut dire ici « appartenir » ? En effet, dans le lexique, il est possible de faire et d'accepter des constatations restreintes, mais elles ne sont pas facilement acceptables en syntaxe dans la mesure où, toujours en synchronie, toute occurrence semble automatiquement promue au rang de représentation (dans le sens précis de (2)). Un exemple peut nous aider à mieux comprendre le problème.

Quelqu'un trouve, en tant que locuteur natif de l'espagnol, un terme comme *bejuquillo* dans un dictionnaire (cf. Casares 1982) et il se demande s'il est possible d'en trouver des occurrences. Le *CREA* (*Corpus de Referencia del Español Actual*, www.rae.es) en fournit trois dont deux apparaissent dans le même texte vénézuélien et une dans un texte espagnol[10]. On peut donc affirmer que la séquence est attestable, même si l'on ne comprend pas très bien le sens du terme. Si nous disons qu'elle appartient à l'espagnol, nous interprétons par là que, dans l'ensemble des locuteurs de l'espagnol, il y en a au moins deux qui « possèdent » le terme parce qu'ils l'ont appris. Il est donc pertinent de formuler la proposition existentielle : « il existe un terme x et il existe au moins un locuteur y tel que y possède x» et il n'est pas possible d'aller plus loin. Comme locuteur natif, je sais que je ne l'ai jamais utilisé et que, pour le faire, je devrais suivre un « apprentissage ». Il s'agit donc de la constatation, restreinte à un sous-ensemble, de trois occurrences, et l'interprétation des données est strictement quantitative. Il est question ici de l'espagnol en tant que langue « sociale » ou « externe ».

Si, par contre, nous nous demandons si toutes les prépositions espagnoles sont attestées dans des contextes où elles gouvernent une proposition introduite par *que* non relatif (proposition-*que*, dorénavant) et que nous consultons le même corpus, nous obtenons pour la préposition *hacia* ('vers') trois occurrences[11] interprétables comme *hacia* + proposition-*que*. La

[10] D'après le dictionnaire, un *bejuquillo* est une sorte de collier pour la taille.

[11] Les voici. Dans (iii), le sens de la séquence est le même qu'en (ii) :
(i) *se camina hacia que* la mitad de la demanda energética se concentre en estos espacios. (*Física y Sociedad*, nº 13, 2002)
'on s'achemine vers le fait que la moitié de la demande énergétique se concentre dans ces espaces'
(ii) Parece todo *apuntar hacia que* (con excepciones importantes) lo que ha fallado ha sido el empresario industrial. (*ABC*, 17/04/1982)
'Tout semble indiquer que ce qui a failli c'est l'entrepreneur industriel'
(iii) La última versión, difundida estos días desde Miami, *apunta hacia que* el propio multilaureado ex general Ochoa, ya fusilado como principal cabeza del "cartel cubano", estaba a punto de desertar… (*ABC*, 14/07/1989)

séquence est donc attestable et nous comprenons les phrases sans aucun problème. Si nous disons alors que cette séquence appartient à l'espagnol, nous entendons par là davantage que le fait que trois locuteurs la possèdent, car nous en faisons automatiquement une représentation qui permet d'avancer l'hypothèse qu'elle peut être produite, dans un contexte grammatical donné, par un nombre potentiellement infini de locuteurs sans qu'ils doivent apprendre quoi que ce soit. D'autant que il y a lieu de supposer que nous avons pu nous-mêmes la produire un nombre indéterminé de fois. On peut donc formuler une proposition universelle du type : « pour tout x (*hacia que*) et pour tout y (locuteur de l'espagnol), y peut produire ou avoir produit la séquence x dans le contexte z » (z étant une description grammaticale précise). Il va donc de soi que nous ne saurions rester dans la constatation restreinte et que la phrase « *hacia que* est possible en espagnol » a une interprétation qualitative indépendante des données quantitatives. Il est donc question ici de l'espagnol en tant que langue interne.
Qui plus est, si nous continuons nos recherches et que nous interrogeons le corpus sur la séquence *bajo* ('sous) + proposition-*que*, nous constatons que, parmi les 360 occurrences fournies, *que* n'est pas une conjonction et qu'il ne semble pas y avoir de cas de *bajo* + proposition-*que*. On en trouve cependant un qui semble aller à l'encontre de cette conclusion :

(13) Con su atenta de ayer recibí las copias de los testamentos *bajo que* falleció su buen esposo de usted y mi apreciable amigo (Victor Chamorro, *El muerto resucitado*, Madrid, Albia, 1984)

Ici, *bajo que* pourrait être interprété comme une sorte de '*sous le fait que' car le résultat peut faire sens ('Avec sa lettre d'hier, je reçus les copies des testaments sous le fait que votre mari décédât …'). Mais nous ne le faisons pas et nous nous rendons compte qu'il s'agit d'un registre juridique qui utilise l'expression « morir bajo un testamento » ('mourir sous un testament') et que la séquence veut dire ('les copies des testaments sous lesquels son mari décédât'). Si la lecture *bajo* + proposition-*que* est exclue c'est bien parce que nos connaissances de locuteur natif rejettent cette possibilité, ce qui n'arrive pas avec *hacia que* et ce qui n'arrive pas, non plus, avec la lecture relative acceptée, si étrange soit-elle.
Ceci revient à dire que l'induction humienne par observation et par régularité ne semble pas à l'œuvre en syntaxe, car aucun travail sur corpus synchronique ne se passe de l'introspection (propre ou d'autrui) comme déclencheur d'exemples-représentations et comme instrument de validation d'hypothèses. Il s'agit de la connaissance d'arrière-plan qui, selon Popper (1990), permet l'induction et que nous appelons, dans le cas des

connaissances langagières, compétence[12]. Tout cela peut aller de soi pour certains, mais il est nécessaire d'en faire constat afin de mieux percevoir ce qui arrive en diachronie.

2.2. Le problème de la représentativité en diachronie

Ce qui a été dit jusqu'à présent implique qu'en l'absence de locuteurs compétents un corpus diachronique ne peut pas promouvoir automatiquement les occurrences au rang des exemples-représentation. C'est dans ce cas que nous sommes vraiment dans une situation d'induction humienne et les conséquences sur la nature de la description grammaticale sont lourdes. Un exemple semblable aux précédents nous le montre facilement. Je cherche à savoir si la séquence *sobre* ('sur') + proposition-*que* existe en espagnol médiéval. Le dépouillement des textes semble indiquer, en général, que la forme n'existe pas, mais deux textes du XIIIe siècle offrent trois exemples auxquels on peut attribuer une interprétation sémantique plausible[13]. D'un point de vue quantitatif, nous sommes face à une situation qui nous rappelle celle présentée par *hacia que* en 2.1, mais du point de vue qualitatif la situation est tout autre, car on ne saurait inclure *sobre que* dans un répertoire de séquences productives. La situation est, de ce point de vue, la même que

---

[12] C'est pourquoi des propos comme les suivants me semblent pour le moins naïfs, dans la mesure où toute personne travaillant sur un corpus synchronique, qu'il le veuille ou non, qu'il le dise ou non, fait des hypothèses introspectives sur ce qui existe et ce qui n'existe pas :
(i) « A corpus and an introspection-based approach to linguistics are not mutually exclusive. In a very real sense they can be gainfully viewed as being complementary » (McEnery & Wilson 2001 : 19).

[13] Les données sont extraites de Barra Jover (2002 : 269). La séquence semble ici avoir un sens causal, d'autant plus que (i) traduit un passage de la *Vulgata* qui utilise *quoniam* et *quia* respectivement (tous les deux interprétables comme 'parce que').
(i) E fablaran e dizran: *sobre que* fizo el Criador assi [a] esta tierra e esta casa? (*sic*) *Sobre que* dexaron a so Sennor Dios que saco a sos parientes de tierra de Egipto e fiaronse en otros, aoraronlos e sirvieronlos; e aduxo el Criador sobrellos toda esta cosa. (Almerich, *La fazienda de Ultramar*, éd. M. Lazar, Salamanca, Acta Salmanticensia, 1965. 149).
[*Et ils parleront et ils diront : à cause du fait que le Créateur fit ainsi cette terre et cette maison. A cause du fait qu'ils abandonnèrent leur Seigneur Dieu…*]
(ii) Conocida cosa seya a todos los qui esta carta uieren, que sobre pleyto que querien mouer contra la ecclesia de Siguenza Garci Gilez e Pero Sanchez, *sobre que* dizien que demandando prouecho por ala ecclesia de Siguenza auien sacado sub contractu una summa. (Ramón Menéndez Pidal (éd.), *Documentos lingüísticos de España*. Madrid : Centro de Estudios Históricos, 1919, document 255).
[*Soit connu de tous (…) qu'à propos de la querelle contre l'église de (…) par rapport (?) au fait qu'ils disaient qu'en demandant un profit pour l'église de S. ils avaient gagné sous contrat une quantité*]

pour l'exemple lexical de *bejuquillo*, car nous pouvons en faire une proposition existentielle (« il existe au moins deux locuteurs qui produisent *sobre que* ») mais pas une universelle (« tout locuteur de l'époque peut produire *sobre que* ») qui demanderait l'observation d'une régularité quantitativement consistante comme celle observée, à un moment donné, pour *porque* ou *para que*, sur lesquels on peut risquer un énoncé universel. Bref, on se borne à constater l'occurrence et on s'abstient d'en faire une représentation.

Ceci n'est qu'un premier exemple des nombreux obstacles à la représentativité des corpus diachroniques. Ils peuvent être organisés dans la casuistique suivante :

a) Représentativité restreinte dans l'échantillonnage :

Il s'agit d'un premier cas d'anomalie dans la dispersion par genres dont un bon exemple nous est offert par les termes à valeur minimale pendant le Moyen-Âge. Il est facile de supposer que des expressions comme la suivante :

(14) N'en ai eü vaillant un oef pelé (*Charroi de Nîmes,* éd. J.-L. Perrier. Paris : Champion, v. 427)

étaient fréquentes pendant la période médiévale jusqu'au point d'être à l'origine de la négation renforcée avec *pas* qui a produit la négation discontinue du français standard ou la négation postverbale en occitan (cf. par exemple, Ramat & Bernini 1990). Il est également vrai que l'on trouve, en ancien français ou un ancien espagnol un répertoire important d'exemples. Mais il est aussi vrai que la plupart parmi eux sont des *hapax* (cf. Möhren 1980, pour le français et Llorens 1929, pour l'espagnol). Pis, ils apparaissent notamment en poésie et l'on peut les soupçonner de n'être, dans la plupart des cas, qu'un expédient commode permettant de résoudre un problème de rime (Möhren 1980 et Barra Jover 1992)[14]. Il nous est donc impossible d'affirmer que ces expressions à valeur minimale faisaient partie de la grammaire de n'importe quel locuteur du XIVe siècle.

Malgré tout, la particularité distributionnelle précédente est facile à comprendre à côté d'autres situations où le rapport entre un genre et une forme ne va pas de soi. C'est le cas, par exemple, des démonstratifs *este* et *ese* en espagnol médiéval. Un travail fondé sur des textes littéraires comme celui de Saéz Durán à propos de *ese* (1996) laisse l'impression que, dans la langue médiévale, la distribution des démonstratifs était très proche de celle de nos jours. Or, le dépouillement des textes juridiques donne un résultat bien différent. Ici dominent les démonstratifs *este* et *aquel*, tandis que *ese* est systématiquement relégué à l'expression *ese mismo* (cf. Barra Jover 2007a)

[14] En espagnol, la plupart des occurrences se produisent dans le dernier vers des quatrains mono-rimiques à rime riche.

qui renvoie aux origines étymologiques du terme. Nous ne pouvons donc pas savoir ce qu'un locuteur quelconque avait pu intérioriser comme système d'oppositions en dehors de l'élaboration d'un texte dans tel ou tel genre.
Ce problème d'échantillonnage par genres a la particularité de ne pas empêcher la formulation de représentations à partir des occurrences. Tout simplement, ces représentations doivent être restreintes à un sous-ensemble culturellement délimité de la production langagière, ce qui amène au paradoxe de ne pas pouvoir les intégrer, sans plus, dans la grammaire d'un locuteur lambda. Comme nous le verrons en §3, certains auteurs trouvent que cette solution est acceptable dans le cadre des « traditions discursives ».

b) Les occurrences ne peuvent pas être promues représentations :
Ce problème ne se pose pas uniquement avec les *hapax* ou les exemples vraiment anecdotiques comme ceux de l'espagnol *sobre que*. Il peut aussi se poser dans des circonstances plus complexes lorsque la dispersion des occurrences d'une séquence à travers le temps présente des lacunes non négligeables. Un exemple assez frappant de cette anomalie nous est fourni par les quantifieurs nominaux personnels *nadie* ('personne') et *alguien* ('quelqu'un') en espagnol médiéval. Il s'agit d'innovations médiévales, dans la mesure où les formes héritées du latin tardif étaient, comme dans les autres langues romanes, des dérivés des adjectifs-pronoms *nullus* ou *ne(c) unus* pour la négation et *aliqu-unus* pour l'affirmation (ancien espagnol *nul, ningún, algún*, ancien français *nul, nesun, aucun*). Nous pouvons nous demander, sans entrer ici dans les détails de leur motivation morphologique, à quel moment ils s'incorporent au dialecte castillan. Et la réponse n'est pas du tout facile. Dans un corpus composé de 60 textes (cf. Barra Jover 1992), la forme *alguien* apparaît trois fois au XIIIe siècle, une fois au XIV et deux au XVe[15]. La situation est encore plus surprenante pour *nadie*. Dans le premier texte littéraire castillan, la *Chanson du Cid*, il y a onze occurrences de la forme *nadi*. Trois autres textes de la première moitié du XIIIe siècle donnent uniquement quatre occurrences. Ensuite, il faut attendre la deuxième moitié du XVe siècle pour en trouver à nouveau des exemples, cette fois plus dispersés et réguliers[16].
Ces données sont difficiles, voire impossibles, à interpréter, notamment en ce qui concerne *nadi(e)*, car nous avons une forme qui apparaît régulièrement dans le premier texte littéraire et qui semble ensuite disparaître. La question se pose alors de savoir si l'on peut dire que *nadi(e)* « appartenait » à la grammaire de n'importe quel locuteur de l'ancien espagnol ou s'il est

[15] Les données venant d'autres sources ne changent pas grand-chose. Montgomery (1965) ajoute un cas unique de *algui* du XIIIe et Malkiel (1948) trois cas de *alguien* dans un même texte du XVe.
[16] Les données de Malkiel (1945) ajoutent deux cas de la première moitié du XIIIe et un cas du XIVe.

possible qu'un quantifieur de cette classe soit un « dormant » (on a souvent parlé d'un « état latent ») pendant des siècles tout en étant accessible à quelques locuteurs. Comme le lecteur a pu le constater, j'ai fait référence à des travaux réalisés avant l'arrivée des corpus informatisés. Ce n'est pas fortuit, car dans la section suivante, lorsque nous envisagerons des solutions à ces problèmes, nous nous demanderons si l'élargissement de la base empirique fournie par les corpus informatisés permet de sortir de ce type d'impasses.

c) Le problème du n+1 texte :
Dans Barra Jover (2000 : 12) ce problème apparaît formulé de la façon suivante :

(15) Problème du n+1 texte : une conclusion obtenue à partir d'un ensemble de n textes ne garantit jamais de prédictions sur ce qui peut arriver dans un n+1 texte.

Autrement dit, rien ne permet d'avoir la certitude que quelque chose soit impossible à une époque donnée, comme rien ne permet d'avoir la certitude que quelque chose va se reproduire régulièrement dans tous les textes d'une même époque.
Partons de quelque chose qui se présente actuellement comme une certitude. Dans des langues romanes comme l'italien, l'occitan, le catalan, l'espagnol et le français, il est impossible, à la différence du portugais, d'introduire un pronom personnel objet entre l'infinitif et la marque de futur et de conditionnel issue de l'auxiliaire HABERE (cf. français **je donner-vous-ai* / espagnol **dar-os-é* / portugais *dar-vos-ei*). Les choses sont bien différentes du point de vue historique. La construction avec interposition de pronom est attestée en portugais, espagnol, occitan et catalan médiévaux[17]. Il semblerait, pourtant, que l'italien et le français ne l'ont pas connue. Les grammaires ou les traités de morphologie verbale du français (cf., par exemple, Fouché 1967) n'en font pas mention, pas plus que Squartini (2007) et Cardinaletti (2007) qui s'occupent respectivement du verbe et du pronom oblique en ancien italien. Jensen (1990 : 351), pour sa part, signale explicitement que

[17] Par exemple :
(i) [CAT] e dix que ella tornaria a son marit e *acusar s'hia* de sos falliments (Ramon Llull. *Llibre d'Ave Maria*, dans H. Guiter. *Grammaire de la langue du « Llibre d'Ave Maria »*. Montpellier : Imprimerie Aristide Quillet, 1943, p. 35)
(ii) [ESP] *Mereçer no lo hedes*, ca esto es aguisado (*Cantar de Mío Cid*. éd. R. Menéndez Pidal, Madrid, Espasa Calpe, 1980[5], v. 197)
(iii) [OCC] Domna, dis el, eu lai irai / Si-us voletz, e *adur lo-us ai* (*Le roman de Jaufre*, dans *Les Troubadors. L'œuvre épique*, Paris, Desclée de Brouwer, 2000, v. 3325)
(iv) [PORT] e nós *gradecer-vo-lo-emos* (Fernão Lopes, *Crónica do rei D. Pedro I*, éd. G. Macchi, Paris : CNRS, 1985, p. 26)

cette séparabilité ne semble pas exister en ancien français, si l'on compare à l'occitan.

Mais cette affirmation n'est que le résultat de l'observation d'une régularité dans les textes et n'a pas (ne peut avoir) le degré de certitude que possède l'affirmation de l'inexistence de cette construction dans les langues romanes modernes. Il suffit de songer qu'elle est bien plus facile à trouver en espagnol ou en portugais médiévaux qu'en occitan ou catalan[18], pour envisager la possibilité qu'elle puisse apparaître dans un texte italien ou français qui n'a pas été dépouillé à ces fins. Il ne s'agit pas de dire que cela doit être le cas, mais tout simplement d'insister sur le fait que rien ne garantit qu'un n+1 texte ne nous oblige pas à revenir sur notre conviction et qu'il nous place dans une situation comme celles envisagées dans les paragraphes précédents. Nous sommes face à une illustration parfaite du problème de Hume : la régularité nous pousse à faire une prédiction sur ce que nous n'avons pas encore vu, mais cette prédiction n'a pas de justification conceptuelle.

Le problème du n+1 texte comporte plus que le risque d'avoir à revenir sur une certitude empirique. Il peut aussi comporter le risque d'annuler un raisonnement explicatif. Les auteurs ayant remarqué l'absence d'interposition de pronom dans le futur de l'ancien français et de l'italien ont tenté, bien entendu, des explications plus ou moins raisonnables à cette particularité (cf., par exemple, Teckavcic 1972 : 2/306-308). Ce qui me semble important, c'est qu'une quelconque tentative d'explication causale peut devenir la seule base logique de la certitude empirique : la garantie qu'une séquence n'existe pas vient finalement de l'explication que l'on a trouvée à son possible inexistence et non pas de la certitude empirique de son inexistence. Ce qui revient à dire que le raisonnement déductif a fini par s'imposer et que tout le château risque de s'écrouler grâce à un n+1 texte[19].

---

[18] Preuve en est que l'affaire est traitée, lorsqu'elle l'est (Anglade 1921, par exemple, n'en fait pas mention), de façon très anecdotique dans les grammaires historiques du catalan (par exemple, Duarte & Alsina 1986 : 2/108) et de l'occitan (par exemple, Jensen 1994 : 243 qui donne les deux mêmes exemples que dans Jensen 1990 : 351), tandis que, pour l'espagnol, il y a une considérable bibliographie à propos de l'alternance entre les constructions avec et sans interposition. D'ailleurs, les exemples catalan et occitan de la note précédente sont les seuls de chacun des textes cités.

[19] Si je me permets cette remarque c'est parce que j'ai fait moi même la mauvaise expérience d'une erreur de ce genre et que, bien qu'anecdotique, elle vaut la peine d'être racontée. Ainsi, en consultant le CORDE, j'avais « constaté » que les séquences avec double datif de l'espagnol comme *se me ha olvidado* (lit. « Il se m'a oublié », 'j'ai oublié') n'apparaissaient pas au XIIIe siècle. J'ai donc organisé un raisonnement (qui me semblait fort acceptable) pour expliquer leur origine à partir d'autres constructions attestées dans ce même siècle, et en faisant la prédiction que, sans ces dernières, la séquence à double datif ne pouvait pas exister. Passer de là à la conviction inébranlable qu'elles n'existaient donc pas a été facile (cf. Barra Jover 2003). Or, la consultation ultérieure d'un autre corpus, celui de Davis (www.corpusdelespagnol.org/) montrait qu'elles existaient bel et bien au XIIIe et que

d) Le dernier problème devant être signalé concerne la disponibilité de l'information. En diachronie, seul ce qui a été retenu et stocké comme donnée est accessible au chercheur qui aborde un sujet. Ce qui peut avoir l'air d'emblée d'une lapalissade ne l'est pas, tout simplement parce que le diachronicien risque toujours de croire avoir vu des choses qui n'existent pas ou de croire à l'inexistence de choses qu'il a vues. Il s'agit, en effet, d'une des grandes frustrations dans ce domaine : on a beau avoir lu des centaines de textes, on est incapable de répondre avec certitude de l'existence ou de l'inexistence d'une quelconque construction si l'on n'en a pas retenu des exemples en tant que pièce d'un fichier (ou par cœur, bien sûr).
Partons du fait qu'il n'y a pas d'hypothèses issues uniquement des données car le traitement des données est déjà « coloré » par certaines hypothèses. Ceci revient à dire que la sélection de ce qui doit être retenu comme donnée est, en soi même, une hypothèse et que, en tant que telle, elle peut être entièrement erronée. Nous pouvons donc avoir affaire à un fichier « faux » à cause des critères ayant guidé sa constitution. L'important ici est qu'une erreur de fichier peut produire la conviction sur l'inexistence de quelque chose que l'on a pourtant vu une centaine de fois[20]. Des situations comme la précédente sont fréquentes et personne n'en est à l'abri.
On peut aussi croire avoir vu ce que l'on n'a pas vu. Pour n'en donner qu'un exemple anecdotique, il a été admis que la conjonction *para que* de l'espagnol procède d'un ancien *pora que*. L'idée est raisonnable dans la mesure où *pora* semble être l'ancêtre de *para* dans les contextes de rection nominale et, du coup, *pora que* a été souvent introduit dans le répertoire des conjonctions de l'ancien espagnol comme un membre de plein droit dans la liste de l'époque. Or, les cas de *pora que* sont presque inexistants, toujours dialectaux et chronologiquement tardifs (cf. Barra Jover 2002 : 210 et ss. pour plus de détails) et *para que* s'est formée indépendamment de *pora*.

J'ai voulu dans cette section exposer des problèmes d'induction dont tout diachronicien a fait l'expérience, et qui nous aident à comprendre jusqu'à quel point la certitude fondée sur l'observation (le problème de Hume) peut nous conduire à des erreurs de jugement ainsi qu'à des certitudes qui, tout en ayant l'air d'avoir une source empirique, ont une source conceptuelle. Il est temps de passer aux solutions possibles.

3. Solutions

tout mon raisonnement était complètement faux. Evidemment, cette communication n'a jamais pris la forme d'article.

[20] On trouvera dans Barra Jover (2002 : 100 et ss.) des remarques sur un travail ponctuel et non imaginaire où ce type d'erreur apparaît.

Tout ce qui a été dit dans la sous-section §2.2 peut être résumé ainsi : en l'absence de locuteurs compétents, nous ne pouvons pas avoir accès aux données nécessaires pour reconstituer la grammaire d'une époque ni pour expliquer les changements d'une époque à l'autre. Nous n'avons accès qu'à une partie de ce qui était possible mais nous n'avons aucune certitude sur ce qui était impossible, ce qui semble nous condamner à la simple constatation. Dans cette section, seront présentées les différentes options que nous laisse une telle limitation.

3.1. Solutions à partir de l'élargissement de la base empirique : les corpus informatisés

Comme je l'ai dit, certains des problèmes évoqués plus haut proviennent d'un traitement de corpus « artisanal ». Cette stratégie a été adoptée afin de mieux souligner le type de solution que l'arrivée des corpus informatisés peut proposer aux problèmes d'induction. Ainsi, Habert *et al*. 1997 : 131-132), dans leur section consacrée à la contribution des corpus informatisés à la recherche diachronique, insistent sur le fait que certains résultats, comme, par exemple, ceux obtenus par Marchello-Nizia (1995) à propos de l'évolution des démonstratifs, doivent leur solidité à l'énorme masse de données ayant pu être manipulée. Ceci revient à dire qu'il y a une solution quantitative au problème de l'induction, dans la mesure où les données se constituent en « hypertexte » statistiquement représentatif de la production langagière de telle ou telle époque et, partant, garantissent que toute hypothèse fondée sur la régularité est légitime. Il me semble intéressant de prendre le temps de vérifier si, et comment, l'hypertexte informatique[21] répond aux problèmes évoqués en §2.2.

Quant au problème de la représentation restreinte dans l'échantillonnage, autrement dit, le fait qu'une construction donnée n'apparaisse que dans un genre de textes (par exemple, les termes à valeur minimale) et qu'il est impossible d'interpréter son implantation dans « la langue » de l'époque, on ne voit pas en quoi l'élargissement empirique, lorsqu'il ne fait que confirmer cette distribution, peut nous aider. Il est pourtant plus intéressant de se demander si un grand hypertexte peut changer notre situation face aux

[21] La notion d'hypertexte est utilisée dans un sens large étant donné qu'elle peut désigner d'emblée deux types de choses : les textes informatisés que l'on peut lire en continu (sur le web ou non) et les bases de données que l'on peut traiter avec un programme de concordances. De même, ces bases de données peuvent être d'accès lexical ou être plus ou moins « étiquetées ». Il m'est impossible d'entrer ici dans les détails (pour le français, cf. Kuntsmann 2000 ; pour les langues romanes, cf. le site Menestrel :www.ext.upmc.fr/urfist/mediev.htm), mais les romanistes ne disposent pas encore de corpus étiquetés du point de vue syntaxique et seules les francistes et les hispanistes disposent de corpus permettant directement l'accès à un programme de concordances pour un mot ou une séquence de mots.

occurrences distribuées d'une façon chronologiquement incompréhensible (le cas de *nadi(e)* en ancien espagnol). En effet, nous pourrions nous attendre à ce qu'un volume très supérieur de données puisse tracer un fil chronologique plus lisible. Cependant, si nous comparons les données sur *nadi(e)* obtenues artisanalement (celles évoquées en §2.2) à celles obtenues avec le CORDE, nous constatons que les problèmes d'interprétation, pour ne pas dire la perplexité, ne font qu'augmenter car les 135 occurrences obtenues pour la période 1200-1399 (126 de *nadi* et 9 de *nadie*) ne font que souligner le problème de la distribution[22] et nous laissent également démunis face à la question « *nadi(e)* appartenait-il à la grammaire de l'ancien espagnol ? ».

En ce qui concerne le n+1 texte, je ne vois pas non plus en quoi l'élargissement de la masse empirique pourrait résoudre le problème d'un point de vue logique. Et ceci pour deux raisons : d'une part, la représentativité statistique d'un corpus diachronique, lorsqu'il s'agit de décrire des états de langue pendant des siècles, est fragile tant du point de vue quantitatif que du point de vue de l'échantillonnage ; d'autre part, même dans le cas idéal d'un contenant tous les textes d'une époque, il s'agit toujours et uniquement de textes connus ou accessibles. Je me permets d'insister : il existe toujours la possibilité logique que d'autres apparaissent et qu'ils aillent à l'encontre des certitudes obtenues avant leur apparition. Ceci n'empêche pas que l'hypertexte apporte une certitude « psychologique », mais ce type de certitude a toujours existé.

C'est, cependant, à propos du quatrième point évoqué que les corpus informatiques semblent apporter un changement qualitatif que j'aimerais appeler la « versatilité » du fichier. Toujours dans les limites des questions que l'on peut poser à un corpus informatisé, des lacunes produites par une mauvaise planification du fichier sont aisément réparables, tandis que les lacunes résultant de dépouillements manuels mal planifiés ne l'étaient pas, tant le coût en temps était considérable. C'est, de mon point de vue, le progrès qualitatif le plus important car le nombre d'informations erronées (ou de données imaginaires), ainsi que le nombre de points « vite expédiés » face à l'absence d'informations ont nettement diminué. Nous allons aussi voir plus loin que la conséquence majeure de ce pouvoir empirique d'aller à l'encontre de l'imagination accorde aux corpus informatisés un rôle essentiel dans la falsification des hypothèses obtenues (trop) déductivement.

Jusqu'à présent, je me suis borné à évaluer les solutions que les corpus informatisés pouvaient apporter aux quatre problèmes posés par la question « la séquence X appartient-elle à la langue Y dans le temps Z ? ». Il y a, bien entendu, d'autres remarques à faire mais je voudrais m'arrêter uniquement

[22] D'emblée, il n'y a qu'une occurrence pouvant être placée au XIV siècle (*Crónica de los estados penínsulares*, daté de 1305-1328). Parmi les 134 restantes, seules 23 apparaissent dans des textes littéraires, les autres dans des textes légaux. Qui plus est, 88 d'entre elles sont concentrées sur deux textes.

sur un aspect : l'influence que la configuration de la technologie disponible peut avoir sur le type de questions que l'on se pose. Comme il a déjà été dit, il y a des limites aux questions que l'on peut poser à un corpus informatisé. Ces limites peuvent, parfois, être provisoires car les corpus diachroniques finiront, tôt ou tard, par être étiquetés syntaxiquement de façon à accepter des questions sur tout type de séquences[23]. Or, il y aura toujours des questions auxquelles le corpus informatisé le plus sophistiqué ne pourra jamais répondre. Par exemple, la conjonction *sans que* n'apparaît en français qu'à partir du XVe siècle. Il y a lieu de se poser la question suivante : existait-il auparavant d'autres constructions pouvant apparaître dans les mêmes contextes que *sans que* ? Un autre exemple : on n'a guère de connaissances sur une construction romane où un nom nu suivi d'une subordonnée adjective (pas nécessairement relative) thétique constitue une séquence indépendante qui reprend un topique (par exemple, « le président nous a proposé de commencer les cours à 5h00 du matin, *idée* (*qui* a été) très mal accueillie ». Il est pertinent de se poser des questions sur son origine et son implantation, mais on voit mal comment chercher dans un corpus, si bien étiqueté soit-il. J'ai bien peur que ce genre de questions ne disparaisse devant l'impuissance de l'informatique à y répondre, au point que les limitations de l'outil informatique amènent à ne pas traiter des facteurs pertinents pour élucider un problème donné. Un autre risque me semble exister dans la perception atomisée du texte que produisent, par exemple, les concordanciels. Le fait de pouvoir cibler avec précision et vitesse notre objectif est très positif en soi mais nous empêche, du même coup, de promener notre regard sur d'autres espaces du texte où de nouvelles questions ou des éléments insoupçonnés de réponse nous attendent.

3.2. L'approche sceptique et l'acceptation de la portée restreinte : les traditions discursives

Une solution possible aux problèmes d'induction posés tout au long de cet article consiste à déclarer non pertinente la question « la séquence X appartient-elle à la langue Y dans le temps Z ? ». Dans un moment donné de l'histoire d'une langue, il existe une « constellation » de traditions discursives, chacune fortement ancrée dans des conditions pragmatiques (cf. Schlieben-Lange 1983, Koch 1997 et Oesterreicher 1997, pour le concept de « tradition discursive », qui ne doit pas être confondu avec l'idée de « genre »). Les résultats obtenus à partir de l'examen d'une ou plusieurs de ces traditions ne sont pas extrapolables et encore moins susceptibles de nous amener à une description de la « langue » de l'époque. Cette approche est actuellement et particulièrement développée par une partie de la romanistique allemande jusqu'au point de mettre en question la possibilité d'avoir des

[23] Sur l'étiquetage des corpus diachroniques, voir Habert *et al*. (1997 : 121 et ss.)

connaissances générales sur des choses comme la formation des temps composés romans (cf. Jacob 2001). Les innovations linguistiques seraient le produit des besoins pragmatiques de telle ou telle tradition, la possibilité d'une diffusion à d'autres traditions existant toujours (cf. Kabatek 2001). Il ne s'agit pas, bien entendu, de se borner à affirmer que la langue parlée à une époque passée nous est difficilement accessible (cela est admis par n'importe qui) ; il s'agit de nier toute possibilité inductive et d'accepter le scepticisme humien comme limite à nos connaissances sur la ou les grammaires d'une époque donnée.
La pertinence de la notion de tradition discursive ne sera pas débattue ici[24]. La seule chose qui sera mise en question est l'impossibilité d'accès, ne serait-ce que partiel, à ce que nous pouvons nommer « la grammaire » de telle ou telle époque (cf. §4).

3.3. Le recours au raisonnement déductif et ses limites

Comme cela a été dit dans notre présentation, le terme « déduction » est pris ici dans un sens informel. Au lieu de vouloir répondre à des questions concernant *un* changement linguistique, on peut théoriser sur le problème *du* changement linguistique, les hypothèses énoncées n'étant pas le résultat direct de l'interprétation des données recueillies, mais la conséquence d'un ensemble de propositions théoriques préalablement établi. Dans ce cas, il serait plus adéquat de parler de « justificationnisme » dans le sens accordé par Lakatos (1978 : *passim*) à ce terme : les propositions théoriques doivent être « légitimées » par des faits empiriques. Il est évident que le chercheur procède, dans ce cas, à un tri sélectif inévitablement guidé par ses objectifs et que l'exemple sélectionné est l'exemple « heureux » (parfois de seconde main) auquel on accorde la valeur de preuve, indépendamment de ce que la masse des données disponibles analysée en tant qu'ensemble puisse révéler. Dans les termes de notre introduction : soit un quelconque exemple-occurrence est promu exemple-représentation sans aucun dispositif protocolaire de validation de la promotion, soit on propose directement un exemple-représentation dont l'isomorphisme avec des occurrences n'a pas été vérifié. Un exemple du premier type de démarche nous est fourni par Harris & Campbell (1995 : 66) lorsqu'ils proposent comme exemple de réanalyse l'évolution des interrogatives françaises. D'après eux, des séquences comme *dit-il ?* ont donné lieu à une particule interrogative *ti* qui devient générale en français. Le seul problème est que les exemples à l'appui ne sont pas représentatifs de ce qui arrive en français, mais marginaux pour ne pas dire

[24] Le rôle joué par ces traditions en tant que sources de certaines innovations est accepté dans Barra Jover (2007a et 2007b).

anecdotiques[25]. Un exemple typique de la deuxième démarche peut être l'évolution suivante (emprunté à Tekavcic 1972 : 251) : latin classique DUX MILITES HORTATUR UT SE DEFENDANT > latin tardif DUX MILITES EXHORTAT QUOD EUM/ILLU DEFENDANT > italien *Il comandante esorta i soldati a difenderlo*. Il va de soi que, pour ce qui est du latin classique ou tardif, tout au moins, les exemples proposés sont des représentations postulées et non attestables en tant qu'occurrences.

Le « justificationnisme » n'est pas l'apanage ni le fardeau d'une approche théorique quelconque. Il devient presque inévitable lorsque le problème du changement se pose en termes universels. Ainsi, et sans aspirer à l'exhaustivité, il est présent dans des approches structuralistes (cf. Anderson 1973), fonctionnalistes (cf. Hopper & Traugott 1993, en termes de grammaticalisation ; Harris & Campbell 1995, en termes typologiques), générativistes paramétriques (cf. Roberts 1993, Lightfoot 1999, Kroch 2003[26]), générativistes lexicales (cf. Vincent 2001), psycholinguistiques (cf. Berg 1998), computationnelles (cf. Niyogi 2007), « écologiques » (cf. Mufwene 2001), téléologiques (cf. Keller 1990 et sa « main invisible ») et même dans des approches à couleur sociolinguistique (cf. Weinreich *et al*. 1968). Il n'est pas dans les objectifs du présent article de juger le bien-fondé des cadres théoriques, mais d'évaluer le rôle des données en tant que source ou en tant que régulateur des hypothèses à propos d'un changement ponctuel qui, soit est présenté comme preuve validant une théorie universelle du changement, soit fait l'objet d'une étude spécifique dans et pour un cadre donné.

De ce point de vue, il y a lieu de retenir les interrogations de certains auteurs face au problème de la représentativité et il me semble possible de le faire de façon synthétique en ciblant une notion qui est, *mutatis mutandi*, partagée par la plupart des approches, à savoir, la réanalyse. Posé de la façon la plus neutre possible, il s'agit d'une situation évolutive où une configuration morphosyntaxique donnée se voit attribuer, pour des raisons stipulées de façon différente par chaque théorie, une analyse nouvelle impliquant des changements qualitatifs dans la grammaire de la langue en question. Quelques exemples : un verbe qui régit une phrase subordonnée peut être réanalysé comme un auxiliaire qui perd ses propriétés de sélection

---

[25] Pour éviter tout malentendu, voici les propos des auteurs mêmes : « This *ti* came to be reanalyzed as a marquer for questions involving third person masculine pronoun subjects, and then later was extended, gradually becoming a general interrogative particle, as in :
(13) les filles sont ti en train de dîner ?
(14) tu vas ti ?

[26] Fuss & Trips (2004) proposent une bonne synthèse de l'approche paramétrique, mais ils omettent les réserves méthodologiques de Kroch (2003) qui, comme nous allons le voir plus loin, doute de la valeur des protocoles à l'œuvre dans la syntaxe diachronique générativiste.

thématique ; un adjectif démonstratif peut être réanalysé comme un article défini ou comme un pronom clitique ; une prédication seconde peut être réanalysée comme un prédicat principal ou un datif peut être réanalysé comme un sujet dont les marques de cas ne sont plus interprétables et disparaissent. Ce type de problèmes est le plus souvent étudié dans des cadres justificationnistes (grammaticalisation, principes et paramètres, par exemple) qui en font des illustrations d'énoncés théoriques préalables[27] et qui doivent faire face à la causalité (qui sera toujours l'affaire majeure de la diachronie). Ce qui semble intéressant pour notre discussion est le fait que les arguments empiriques avancés, autrement dit, les exemples « heureux » se heurtent, dans ce cas, au problème de la représentativité aussi bien dans sa dimension quantitative que dans sa dimension qualitative.

Du côté quantitatif, l'on peut proposer l'existence de séquences « ambiguës » se prêtant à la réanalyse tout en gardant la possibilité d'être encore interprétées dans la grammaire ancienne. On a ici affaire au postulat de la gradualité et à l'idée qu'un changement peut être justifié par la fréquence des séquences interprétables en « double grammaire ». C'est la position de Lightfoot (1979 et 1999, par exemple) tant pour ce qui est de la disparition du cas que pour ce qui est du passage de l'ordre V2 à l'ordre SVO en moyen anglais. Or, comme Kroch (1989) l'a bien montré, tous les arguments déductifs avancés par Lightfoot sont falsifiés par des études de corpus dans la mesure où la fréquence des structures hypothétiquement à la source du changement n'a pas changé pendant la période en question. Dans le même ordre d'idées, il est possible que l'on postule théoriquement qu'un changement est la pré-condition d'un autre, ce qui demande une chronologie stricte. Par exemple, Lightfoot (1979) avance que la condition pour la réanalyse des auxiliaires en moyen anglais est qu'ils apparaissent sans objet direct. Or, Warner (1983) a montré quantitativement que les deux choses se produisent en même temps.

Du côté qualitatif, on peut postuler l'existence des séquences « ambiguës » sans aucun soutien empirique. C'est ce qui arrive, en théorie de la grammaticalisation, pour la formation de certaines périphrases verbales romanes. Pour en donner un exemple négatif, il est, à ma connaissance, fort improbable que l'on trouve un exemple de la séquence *compter* + INF de *je compte le faire*, où le verbe *compter* garde les caractéristiques sémantiques et sélectionnelles de son emploi lexical[28].

---

[27] Ceci dit, certains auteurs, tout en adoptant une démarche justificationniste et tout en se plaçant dans un programme de recherche externe à la diachronie, font ressortir leurs résultats de l'analyse approfondie d'un corpus. Il y a lieu, à ce propos, de citer le travail de Rouveret (2004) sur la position des clitiques en ancien français.

[28] Voir, à ce propos, les exemples de Harris & Campbell (1995 : 70) qui figurent parmi les rares auteurs qui se prononcent explicitement contre la nécessité des séquences ambiguës.

Il y a, évidemment, une leçon à tirer de tout cela : les intuitions valables en synchronie sont trompeuses en diachronie, le risque de postuler (en dehors, bien entendu, de toute tentative explicite de reconstruction) le non attesté est toujours présent et le rôle des études sur corpus est, heureusement mais seulement, celui d'un élément rectificateur. On aura toujours affaire au problème de la représentativité dans la mesure ou nous manquons d'intuitions sur ce qui est possible ou impossible. Il y a pourtant peu de chercheurs qui réagissent explicitement face à ces limites et qui proposent ne serait-ce que des ébauches de méthodologies aptes à pallier les difficultés inhérentes au manque de preuves négatives essentielles pour répondre aux questions du type « X appartient à la grammaire de Y au moment Z ? ».

Kroch (2003), qui est un bon exemple de diachronicien placé dans un programme théorique formel mais bien rodé au travail sur corpus, affiche un grand scepticisme par rapport aux informations qu'on peut tirer des *scripta*. Il ne trouve comme solution méthodologique que la quête de situations semblables dans des langues modernes afin de projeter leur évidence négative et leur représentativité quantitative sur les langues anciennes. Faarlund (1990), plus optimiste, propose la méthodologie des « occasions manquées » (*missed opportunities*) pour compenser les jugements de grammaticalité. Autrement dit, si l'on repère régulièrement des contextes où la construction X pourrait apparaître et qu'elle ne le fait jamais, on peut se permettre d'interpréter cela comme une preuve de (a)grammaticalité. Ce sont des exemples de propositions qui ont recours à l'anachronisme, c'est-à-dire à la projection des propriétés d'une grammaire moderne sur une ancienne, mais il me semble que leur légitimité doit dépendre de leur aptitude à être explicitées d'une façon protocolaire nous permettant de revenir à la démarche inductive.

Pour ma part, je pense qu'il existe la possibilité de développer des protocoles inductifs où l'anachronique soit compensé par la réponse des données[29]. « Anachronique » ne doit pas être interprété ici de façon péjorative et je suis loin de critiquer les deux propositions citées, ne serait-ce que parce que je propose dans Barra Jover (2007c) que l'étude des fautes écrites de pluriel nominal en français actuel peut nous aider à mieux interpréter les *scripta* par rapport à la disparition du cas morphologique en ancien français car les deux changements sont commensurables. D'ailleurs, il me semble difficile qu'un bon dépouillement à visée diachronique ne soit pas orienté par le critère des « occasions manquées »

4. Une autre possibilité : l'accès à une grammaire individuelle

Dans les pages précédentes ont été soulevés de nombreux problèmes concernant la représentativité et le rôle des données dans l'élaboration et la validation des hypothèses diachroniques. Il est peut-être temps d'envisager

des solutions, si limitées soient-elles. Pour ce faire, nous allons nous poser de façon abstraite un problème qui semble concentrer la plupart des difficultés exposées : soit une innovation dans une langue donnée, je me pose la question de son origine et de son implantation à telle ou telle époque. Autrement dit, on a affaire au problème de la représentativité qualitative et quantitative des données, ainsi qu'au problème de la causalité. C'est le cas, par exemple, de Lightfoot (1979) à propos de la corrélation entre la perte de l'objet direct et la réanalyse de certains verbes en tant qu'auxiliaires. Comme nous l'avons vu, les hypothèses de l'auteur se heurtent à des difficultés empiriques majeures et il serait souhaitable de chercher une méthodologie permettant d'établir des protocoles de validation inductifs.

J'aimerais insister sur l'idée qu'il s'agit ici d'une affaire de procédure et non d'une discussion sur les hypothèses sous-tendant les programmes de recherche. D'ailleurs, le fait que ce qui est développé dans cette section puisse recevoir une interprétation « ontologique » (cf. §5) est secondaire.

D'emblée, il me semble qu'il faut renoncer, en l'absence de locuteurs compétents, à répondre à des questions comme « la construction X appartient à la langue Y dans le temps Z ? ». En d'autres termes, j'accepte le scepticisme de Hume concernant la certitude logique obtenue à partir de la seule observation des données. Cependant, il est possible de se poser d'autres questions en partant d'un principe qui semble facile à accepter : nous pouvons avoir accès, grâce aux *scripta*, à la production d'un locuteur donné et nous pouvons supposer que cette production est le résultat d'une grammaire intériorisée cohérente. Ceci étant admis, nous pouvons formuler une question différente, à savoir, « dans quelles conditions le locuteur X peut produire l'innovation Y ? ». Nous allons voir, je l'espère, que le fait d'accepter que cette question est la seule abordable, a des conséquences méthodologiques majeures car nous sommes en mesure de procéder à une démonstration empirique par rapport à la causalité. En effet, nous pouvons, dans ce cas, revenir à la solution de Bacon citée plus haut (« Mais toute interprétation plus vraie de la nature s'obtient à l'aide d'instances et d'expériences convenables et appropriées. Là, les sens jugent de l'expérience seule, l'expérience, de la nature et de la chose même »).

Commençons par une modélisation très large d'un type de raisonnement diachronique, qui semble valable aussi bien pour la phonologie que pour la morphosyntaxe et qui peut être attribuée à n'importe quel cadre théorique (« état de choses » fait, dans ce qui suit, référence à des facteurs internes comme la réalisation d'un allophone ou d'un phonème nouveau, l'adoption d'un ordre de constituants, l'extension de la sous-catégorisation d'un item, la perte d'un trait morphologique, etc.) :

- Il y a un état de choses A que je juge être la cause d'un état de choses B.
- L'état de choses B est une innovation non triviale. Une innovation non triviale produit une séquence qui serait agrammaticale dans un autre temps ou

dans une autre langue possédant le même trait concerné, c'est-à-dire qu'elle implique un changement substantiel dans la grammaire.
- L'état de choses A est une innovation triviale : une séquence pouvant être grammaticale à n'importe quel moment et dans n'importe quelle langue possédant le trait concerné car elle n'implique qu'un changement accidentel.
Les arguments conceptuels stipulant la relation causale entre A et B ne nous concernent pas ici et dépendent, bien entendu, du cadre de raisonnement dans lequel se place le chercheur.
Quelques exemples du passage du latin au roman peuvent nous aider à mieux cerner la notion de changement non trivial : le fait qu'un nom puisse apparaître dans tous les contextes sans marque de cas à partir d'un moment donné, qu'un verbe inaccusatif (IRE, VENIRE) avec un réseau thématique propre puisse devenir un auxiliaire, que la séquence verbe+objet+prédication seconde puisse apparaître sans accord entre les deux derniers, que les prépositions et les adverbes puissent régir une proposition-*que*, qu'un allophone d'un phonème X devienne l'allophone d'un phonème Y, etc.
Pour ce qui est du changement trivial, on peut donner comme exemples : la fréquence plus grande de telle où telle configuration, l'élargissement des propriétés sélectionnelles sémantiques (non catégorielles) d'un verbe ou d'une préposition (par exemple, qu'un verbe ou une préposition qui ne le faisaient pas avant puissent sélectionner des noms d'événement ou abstraits), l'extension d'une entrée lexicale au détriment d'une autre, la réalisation phonétique différente d'un morphème, l'apparition d'un nouvel allophone pour un phonème, etc.
Nous avons donc une hypothèse ayant la forme : « B apparaît parce que A » et nous devons la prouver grâce au dépouillement exhaustif d'un corpus placé dans une tranche chronologique précise (par exemple, 30 textes d'une centaine de pages, d'auteur différent et, par exemple, étalés au long d'un siècle). Si nous procédons de la manière habituelle, autrement dit, en traitant le corpus comme un hypertexte et sans accorder un rôle majeur au fait qu'il est composé de la production de locuteurs différents, nous pouvons obtenir quatre résultats dont aucun ayant le rang d'une réfutation ou d'une corroboration empirique :
1) Il n'y a pas d'exemples de B et il y a des exemples de A.
2) Il n'y a pas d'exemples de A et il y a des exemples de B.
3) Il n'y a ni exemples de A ni de B.
4) Il y a des exemples de A et de B.
Les résultats 1 et 3 ne sont pas informatifs. Ils veulent tout simplement dire que la tranche chronologique est mal choisie. Le résultat 2 peut être interprété comme une réfutation, mais il est facile de trouver des arguments pour la surmonter (les occurrences A appartenaient à un registre informel, le corpus est trop petit, il s'agit de copies modernisées, etc.). Le résultat 4 peut être interprété comme un argument favorable même s'il ne constitue pas une démonstration parce que le lien causal entre A et B n'est pas prouvé. Mais si

nous songeons à tous les problèmes de représentativité évoqués en §2, les résultats 1 à 4 peuvent être réinterprétés et manipulés à loisir ou peuvent devenir un élément périphérique dans l'argumentation[30].

Revenons donc à ce qui a été proposé plus haut : je me limite à me demander dans quelles conditions un locuteur X peut produire une séquence et j'applique l'aphorisme de Bacon à la lettre. Dans ce cas, je procède de façon protocolaire dans la mesure où :

a) L'hypothèse « B apparaît parce que A » devient le conditionnel expérimental « si un locuteur X produit B, alors ce même X doit produire A ». L'implication inverse, tout en possédant un contenu informationnel remarquable, ne fait pas partie du protocole.

b) Le conditionnel est testé sur chaque locuteur. Autrement dit, je procède à 30 expériences.

La toute première conséquence est que le dépouillement du corpus peut produire, cette fois, six résultats :

1) Je ne trouve ni A ni B dans aucun texte.
2) Je ne trouve que A dans certains textes.
3) Je ne trouve que B dans certains textes.
4) Je trouve A dans certains textes et B dans d'autres, ils ne coïncident pas nécessairement.
5) Je trouve A dans certains textes et B dans une partie de ces mêmes textes.
6) Je trouve A et B dans les mêmes textes.

Dans le premier et le deuxième cas, l'hypothèse n'est ni confirmée ni infirmée. Dans le troisième et le quatrième cas, elle est directement réfutée. Dans le cinquième et le sixième cas, elle est provisoirement confirmée. En réalité, on procède d'une façon très simple : l'hypothèse n'est jamais définitivement corroborée. Bien au contraire, elle risque à chaque test d'être refusée. Si l'on élargit le corpus, on ne fais qu'ajouter des tests qui risquent de la rejeter. Plus elle tient, plus on aura des garanties qu'elle est vraisemblable (il n'est jamais dit qu'elle est vraie parce qu'aucune hypothèse ne peut jamais être déclarée vraie).

J'ai évoqué l'aphorisme 50 de Bacon car l'intérêt de ce conditionnel repose sur son caractère protocolaire et non sur son interprétation ontologique. On peut donner nombreuses raisons pour justifier un résultat négatif, car il existe toujours la possibilité que, tout au long d'une centaine de pages, un locuteur pouvant utiliser une construction ne le fasse pas. Mais ceci fait partie du caractère mécanique de l'expérience et tout résultat peut être ultérieurement pondéré en s'accordant, par exemple, une marge d'erreur justifiée en termes probabilistiques.

Il me semble légitime de proposer cette méthodologie car elle a été suivie avec de bons résultats dans un travail antérieur. Barra Jover (2002 : chap. 4 et

[30] Preuve en est que Lightfoot (1999) trouve toujours des arguments contre les faits empiriques qui lui sont présentés.

5) propose une description ainsi qu'une explication des conditions d'apparition des séquences Prép+proposition-*que* et Adv+proposition-*que* dans l'évolution du latin à l'espagnol. Il n'est pas possible ici de reproduire tout le développement, mais quelques exemples peuvent être suffisamment éloquents. Les changements triviaux permettant l'apparition de chacune de ces constructions sont stipulés par projection des hypothèses obtenues à partir de l'observation de la syntaxe de l'espagnol moderne. Les voici :

(16) [Prép [proposition-*que*]] est possible lorsque Prép peut sélectionner un DP défini dont la tête nominale implique une dimension temporelle (p. ex. *la chute*).

(17) [Adv [proposition-*que*]] est possible lorsque Adv a des emplois anaphorique et qu'il peut apparaître dans la périphérie gauche de la phrase[31].

A partir de là, le conditionnel formulé est:

(18) Si, dans un texte donné, on trouve X+proposition-*que*, dans ce même texte X doit apparaître au moins une fois dans les conditions décrites dans (16) et (17).

Le test a été appliqué à une trentaine de textes allant du XIIIe au XIXe siècles avec des résultats positifs pour toutes les prépositions et adverbes (la marge d'erreur étant presque nulle). Comme résultat parallèle non décisif, il y a eu de nombreux textes où la condition n'était pas remplie et qui ne présentaient pas non plus l'innovation X+proposition-*que*. Ce qui m'intéresse de signaler ici, à l'aide d'un exemple de chaque type, c'est la façon dont ces résultats doivent être interprétés.

Pour une séquence comme *para que* ('pour que') les premiers exemples arrivent au XIVe siècle. Mais ceci ne nous autorise pas à affirmer que la séquence n'était pas possible avant (*para* apparaît, bien entendu, dans d'autres contextes dès le XIIIe siècle) ni qu'elle « appartient » à la grammaire du XIVe siècle. Elle apparaît dans un texte du début du siècle et dans un autre de la fin ; tous les deux affichent la condition (16), mais elle n'apparaît pas dans deux textes du milieu du siècle (les deux auteurs, d'ailleurs, ne remplissent pas la condition (16)).

Pour une séquence comme *siempre que* ('chaque fois que') les premières occurrences arrivent dans deux textes du XVIe siècle, tandis que *siempre* ('toujours') est attesté dans des emplois non liés depuis le XIIIe. Ces deux textes remplissent la condition (17) ce qui n'est pas le cas des précédents.

Deux choses sont à signaler. La première est que le fait que des textes antérieurs ne remplissent pas la condition n'est pas protocolairement décisif mais c'est un atout complémentaire. La deuxième est que ces résultats

[31] Un emploi anaphorique implique qu'un adverbial a une portée déterminée par un intervalle temporel introduit préalablement. Par exemple, en français, (i) est un cas d'adverbial anaphorique apparaissant dans la périphérie gauche et (ii) d'adverbial non anaphorique ne pouvant pas le faire :

(i) Il a promis qu'il le ferai et *après* il n'a rien fait

(ii) Il a promis de le faire et il l'a fait *vite*.

n'aspirent pas à dire à partir de quand les séquences concernées « appartiennent » à l'espagnol. Il n'est pas affirmé (il ne saurait l'être) que *siempre que* n'est pas possible au XIVe siècle. Il est tout simplement dit que s'il apparaît dans un texte du XIVe, ce texte doit remplir la condition (17). Cela change substantiellement la portée de nos affirmations diachroniques.

Voyons, à présent, ce que nous gagnons et ce que nous perdons en utilisant ce procédé. Commençons par le côté positif :

a) Nous pouvons parler légitimement de démonstration empirique, c'est-à-dire de procédure inductive.

b) La procédure est réplicable, donc toutes les hypothèses ainsi formulées sont falsifiables.

c) Même si l'hypothèse « B parce que A » est obtenue spéculativement, qu'elle est la conséquence d'un programme de recherche indépendant ou que c'est, tout simplement, une projection anachronique, elle est validée par les données de l'époque et seulement par les données de l'époque.

d) Tous les problèmes de représentativité quantitative et qualitative disparaissent, dans la mesure où la question « la séquence X appartient à la langue Y au moment Z ? » ne fait plus sens. Autrement dit, 30 textes devenus le terrain de 30 expériences fournissent plus d'information que 300 traités comme hypertexte.

Quant à ce que nous perdons, il faut signaler les limites qui nous sont imposées :

a) Nous devons renoncer à décrire un état de langue générale et nous contenter de décrire les propriétés d'une série d'idiolectes écrits. Nous renonçons, par conséquent, à avoir le degré de certitude et la portée de la description synchronique.

b) Il y a des problèmes qui ne sont pas faciles à traiter pour deux raisons : la première, d'ordre quantitatif, touche aux séquences dont la probabilité d'apparition dans un texte est très réduite. Dans ce cas, le recours inévitable à un ensemble croisé d'idiolectes pour tester un conditionnel rend les résultats uniquement spéculatifs. La deuxième, d'ordre conceptuel, concerne des problèmes pour lesquels un énoncé « B parce que A » ne peut pas être réduit à un conditionnel dont le terme A est un observable ou pour lesquels sa réduction entraîne un éloignement excessif du contenu de l'hypothèse.

c) Nous ne pouvons pas traiter des informations venant de *scripta* très précieux, notamment, tous les documents légaux qui ne sont pas des copies et dont la datation est sûre mais qui sont de courte extension et d'auteur toujours différent.

Mais il va donc de soi que les idées exposées ne sont pas *la solution* aux problèmes inductifs de la diachronie et qu'elles doivent être interprétées comme une composante à développer en dehors de toute exclusivité. Ces précautions prises, elles nous invitent à réfléchir, en guise de conclusion, à la nature de la description grammaticale.

5. La description grammaticale en général : la grammaire de Jones ou celle de Simone[32] ?

La méthodologie présentée dans la section précédente implique que la description diachronique soit réduite au cumul de résultats obtenus à partir de plusieurs grammaires individuelles (idiolectes, dans la terminologie la plus traditionnelle) d'une autre époque, et que les affirmations sur une séquence donnée doivent être formulées comme la ou les conditions qu'un idiolecte remplit pour pouvoir la produire et non comme les conditions que la langue X possède à une époque donnée. J'aimerais finir le présent travail en proposant au lecteur une brève réflexion sur les conséquences d'une interprétation ontologique de cette méthodologie. La théorie Principes et Paramètres accepte explicitement que l'objet décrit par le linguiste est la « langue interne », mais il s'agit d'une expression galvaudée. Il y a donc lieu de rappeler ici le cadre dans lequel cette notion doit être placée :

(19) Définition d'Langue-I (Chomsky & Lasnik 1995 : 15) :
When we say that Jones has de language L, we now mean that Jones's language faculty is in the state L, which we identify with a generative procedure embedded in performance systems. 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 ». The concept of language is internal, in that it deals with an inner state of Jones's mind/brain, independent of other elements in the world. It is individual in that it deals with Jones, and with language communities only derivately, as groups of people with similar I-languages. It is intensional in the technical sense that the I-language is a function specified in intension, not extension : its extension is the set of S[tructural] D[escriptions] (what we might call the structure of the I-language).

Il s'agit de savoir si les descriptions grammaticales que nous faisons synchroniquement, que ce soit d'une construction spécifique ou d'une langue, constituent des objets pouvant être implémentés dans un locuteur réel (Jones) ou si elles produisent des objets exigeant un locuteur virtuel ressemblant à cette Simone dont les capacités ne peuvent jamais être réunies en un seul acteur réel. Ce que nous avons vu sur la diachronie nous permet d'émettre des réserves à deux versants qui impliquent que les frontières entre langue-I et langue-E risquent d'être mal fixées du point de vue méthodologique :
a) L'isomorphisme substantiel (les variantes sont des accidents négligeables) entre les grammaires individuelles d'un même espace langagier peut être une illusion fondée sur la compétence passive et non sur la performance. Autrement dit, le fait qu'un locuteur accepte un énoncé comme appartenant à

[32] « Simone » fait ici allusion à l'actrice virtuelle du film homonyme réalisé par Andrew Niccol (2002). Un logiciel permet au réalisateur joué par Al Pacino de faire faire à Simone tout ce que n'importe quel acteur réel pourrait faire.

sa langue ne veut peut-être pas dire qu'il possède les mécanismes pouvant le produire.

b) Les objets décrits par les linguistes pourraient, dans ce cas, constituer des répertoires désordonnés de variantes et non des algorithmes répondant à ceux qu'un locuteur donné peut avoir intériorisés. Autrement dit, les difficultés toujours existantes pour trouver, pour un phénomène donné, une représentation qui ne produise pas d'anomalies (énoncés prédits comme grammaticaux et qui ne le sont pas ou vice-versa) pourraient venir du fait que le linguiste cherche un seul mécanisme pour produire ce qui est produit par plusieurs mécanismes qui ne sont pas isomorphiques.

C'est pourquoi j'aimerais finir ce texte en invitant, à nouveau, le lecteur à réfléchir au sens tant méthodologique qu'ontologique de l'expression « la séquence X appartient à la langue Y ».

REFERENCES BIBLIOGRAPHIQUES

ANDERSON, James M. (1973). *Structural aspects of linguistic change*. London : Logman.

ANGLADE, Joseph (1921). *Grammaire de l'ancien provençal*. Paris : Klincksieck.

BACON, Francis (1620/2004). *Novum Organum*. Paris : PUF.

BARRA JOVER, Mario (1992). *La quantification indéfinie dans les langues romanes*. Thèse de doctorat, Université de Strasbourg.

BARRA JOVER, Mario (2000a). Constatation et induction face aux corpus diachroniques : le problème du n+1 texte. *Les Cahiers FORELL*, 14 : 7-21. Version espagnole revue et amplifiée : « Corpus diacrónico, constatación e inducción ». Dans Jacob & Kabatek (éds). 177-197.

BARRA JOVER, Mario (2000b). Métaphores et représentations : le passage de la boîte noire à l'auto-organisation. Dans *Représentation(s)*. Poitiers MSHS, 15-21.

BARRA JOVER, Mario (2002). *Propiedades léxicas y evolución sintáctica. El desarrollo de los mecanismos de subordinación en español*. La Coruña : Toxosoutos.

BARRA JOVER, Mario (2003). Del pronombre personal latino a la morfología verbal en *se/le* del español. Communication présentée au VI Congreso Internacional de Historia de la Lengua Española (Madrid 29 de Septiembre-4 de Octubre de 2003).

BARRA JOVER, Mario (2007a). Tradición discursiva, creación y difusión de innovaciones sintácticas : la cohesión de los argumentos nominales a partir del siglo XII. Dans Kabatek, J. (éd). *Sintaxis histórica del español: Nuevas perspectivas desde las Tradiciones Discursivas*. Frankfurt / Madrid : Vervuert / Iberoamericana (Lingüística Iberoamericana).

BARRA JOVER, Mario (2007b). Cambios en la arquitectura de la prosa española y romance : sintaxis y cohesión discursiva por correferencia nominal. *Revista de Filología Española* 87.

BARRA JOVER, Mario (2007c). Comment évolue un trait grammatical : le pluriel en français dans une perspective romane. *Romance Philology* 60.

BERG, Thomas (1998). *Linguistic structure and change*, Oxford, Oxford University Press.

BILBER, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8 : 243-257.

BUTT, Miriam ; KING, Tracy H. (éds) (2001). *Time over Matter : Diachronic Perspectives on Morphosyntax*. Stanford, California : CSLI Publications.

CARDINALETTI, Anna (2007). Pronomi obliqui. Dans Salvi & Renzi (2007).

CHOMSKY, Noam ; LASNIK, Howard (1995). The Theory of Principles and Parameters. Dans Chomsky, Noam. *The Minimalist Program*. Cambridge Mass. : MIT Press, 13-127.

CORRADI-FIUMARA, Gemma (1995). *The Metaphoric Process*. London-New York : Rouledge.

DUARTE, Carles ; ALSINA, Alex (1986). Gramàtica històrica del català. Barcelona : Curial.

FAARLUND, Jan Terje (1990). *Syntactic Change. Toward a Theory of Historical Syntax*. Berlin / New York : Mouton de Gruyter.

FRANK, Barbara ; HAYE, Thomas ; TOPHINKE, Doris (éds.) (1997). *Gattungen mittelalterlicher Schriftlichkeit*. Tübingen : Narr.

FOUCHE, Pierre (1967). *Morphologie historique du verbe français*. Paris : Klincksieck.

FUSS, Eric ; TRIPS, Carola (éds). (2004). *Diachronic Clues to Synchronic Grammar*. Amsterdam / Philadelphia : John Benjamins.

HABERT, Benoît ; NAZARENKO, Adeline et SALEM, André (éds.) (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

HARRIS, Alice C.; CAMPBELL, Lyle (1995). *Historical syntax in cross-linguistic perspective*. Cambridge : Cambridge University Press.

HOPPER, Paul J. ; TRAUGOTT, Elizabeth Closs (1993). *Grammaticalization*. Cambridge ; Cambridge University Press.

HUME, David (1758 / 1983). *Enquête sur l'entendement humain*. Paris : Flammarion.

JACOB, Daniel (2001). ¿Representatividad lingüística o autonomía pragmática del texto antiguo ? El ejemplo del pasado compuesto. Dans Jacob et Kabatek 2001 : 153-176.

JACOB, Daniel ; KABATEK, Johannes (éds) (2001). *Lengua medieval y tradiciones discursivas en la Península Ibérica*. Frankfurt-Madrid : Vervuert-Iberoamericana.

JENSEN, Frede (1990). *Old French and Comparative Gallo-Romance Syntax*. Tübingen : Niemeyer.

JENSEN, Frede (1994). *Syntaxe de l'ancien occitan*. Tübingen : Niemeyer.

KABATEK, Johannes (2001). ¿Cómo investigar les tradiciones discursivas medievales ? El ejemplo de los textos jurídicos castellanos. Dans Jacob et Kabatek (éds.) (2001) : 97-132.

KANT, Emmanuel (1783/1993). *Prolégomènes à toute métaphysique future qui pourra se présenter comme science*. Paris : Vrin.

KELLER, Rudi. (1990). *Sprachwandel. Von der unsichtbaren Hand in der Sprache*. Tübingen : Francke.

KOCH, Peter (1997). Diskurstraditionen : zu ihrem sprachtheoretischen Status und ihrer Dynamik. Dans Frank ; Haye ; Tophinke (éds) (1997) : 42-79.

KROCH, Anthony (1989). Reflexes of grammar in patterns of langage change. Language Variation and Change 1 : 199-244.

KROCH, Anthony (2003). Syntactic change. Dans Baltin, Marc ; Collins Chris (éds.) (2003). *Handbook of Contemporary Syntactic Theory*. Oxford : Blackwell, 699-729.

KUNTSMANN, Pierre (2000). Ancien et moyen français sur le web : textes et bases de données. *Revue de Linguistique Romane* 64 : 17-42.

LAKATOS, Imre (1978). *The methodology of scientific research programmes*. Cambridge : Cambridge University Press.

LIGHTFOOT, David (1979), *Principles of diachronic syntax*, Cambridge, Cambridge University Press.

LIGHTFOOT, David (1999), *The development of language. Acquisition, change, and evolution*, Oxford, Blackwell.

LLORENS, Eduardo. L. (1929), *La negación en el español antiguo con referencias a otros idiomas*, Anejo 11 de la *Revista de Filología Española*. Madrid : CSIC.

MCENERY, Tony ; WILSON, Andrew (2001). *Corpus Linguistics*. Edinburgh : Edinburgh University Press.

MALKIEL, Yacob (1945). Old Spanish *nadi(e), otri(e)*. *Hispanic Review*, 13 : 204-240.

MALKIEL, Yacob (1948). Hispanic *algu(i)en* and related formations. *University of California Publications in Linguistics*, 1/ 9 : 357-442.

MARCHELLO-NIZIA, Christiane (1995). *L'evolution du français : ordre des mots, démonstratifs, accent tonique*. Paris :Armand Colin.

MÖHREN, Frankwalt (1980). *Le renforcement affectif de la négation par l'expression d'une valeur minimale en ancien français*. Tübingen : Max Niemeyer.

MONTGOMERY, Thomas (1965). A datum for the history of castilian *Alguien and Nadie*. *Hispanic Review*, 33 : 52-57.

MUFWENE, Salikoko S. (2001). *The Ecology of Language Evolution*. Cambridge : Cambridge University Press.

NIYOGI, Partha (2007). *The Computational Nature of Language Learning and Evolution*. Cambridge Mass. : MIT Press.

OAKES, Michael Philip. (1998). *Statistics for corpus linguistics*. Edinburgh : Edinburgh University Press.

OESTERREICHER, Wulf (1997). Zur Fundierung von Diskurstraditionen. Dans Frank ; Haye ; Tophinke (éds) (1997) : 19-41.

POPPER, Karl (1990). *Le réalisme et la science*. Paris : Hermann.

POPPER, Karl (1991). La connaissance conjecturale : ma solution du problème de l'induction. Dans *La connaissance objective*. Paris : Aubier, pp. 39-78.

RAMAT, PAOLO ; BERNINI, Giuliano (1990). Area influence versus typological drift in Western Europe : the case of negation. Dans Bechert, Johannes ; Bernini, Giuliano ; Buridant, Claude (éds). *Towards a Typology of European Languages*. Berlin-New York : Mouton de Gruyer, 25-46.

RASTIER, François (2005). Enjeux épistémologiques de la linguistique de corpus. Dans Williams 2005 : 31-46.

ROBERTS, Ian (1993). *Verbs and diachronic syntax*. Dordrecht : Kluwer.

ROUVERET, Alain (2004). Les clitiques pronominaux et la périphérie gauche en ancien français. *Bulletin de la Société de Linguistique de Paris* 94 : 181-237.

SAÉZ DURAN, Juan (1996). Castellano medieval *esse* en textos literarios. Dans *Actas del III CIHLE*. Madrid : Arco/Libros, vol. I : 555-566.

SCHLIEBEN-LANGE, Brigitte (1983). *Traditionen des Sprechens*. *Element einer pragmatischen Sprachgeschischtsschreibung*. Stuttgard : Kohlmammer.

SALVI, Giampaolo ; RENZI, Lorenzo (coord.) (2007). *Grammatica dell'italiano antico*. http://geocities.com/gpsalvi/konyv/

SQUARTINI, Mario (2007). Il verbo. Dans Salvi & Renzi (2007).

SUTHERLAND, Kathryn (éd.) (1997). *Electronic Text*. Oxford : Clarendon Press.

TOGNINI-BONELLI, Elena (2001). *Corpus Linguistics at Work*. Amstredam/Philadelphia : John Benjamins.

TEKAVCIC, Pavao (1972). *Grammatica storica dell'italiano*. Bologna : Il Mulino

VAN FRAASSEN, Bas C. (1994). *Lois et symétrie*. Paris : Vrin.

VINCENT, Nigel (2001). LFG as a Model of Syntactic Change. Dans Butt & King (éds.) :1-42.

WARNER, Anthony (1983). Review of David Lightfoot, *Principles of Diachronic Syntax*. *Journal of Linguistics* 19 : 187 : 209.

WEINREICH, Uriel ; LABOV, William ; HERZOG, Marvin I. (1968). Empirical foundations for a theory of language change. Dans Lehmann, Winfred P. ; Malkiel, Yacob (éds.). *Directions for historical linguistics*. Austin : University of Texas Press, 95-195.

WILLIAMS, Geoffrey (sous la direction de) (2005). *La linguistique de corpus*. Rennes : PUR.

WITTGENSTEIN, Ludwig (1918/1961). *Tractatus logico-philosophicus*. Paris : Gallimard.