

## *Exemples attestés et exemples construits dans la pratique du lexique-grammaire*

Éric LAPORTE  
Université Paris-Est  
Institut Gaspard-Monge  
5, bd Descartes  
77454 Marne-la-Vallée CEDEX 2  
eric.laporte@univ-paris-est.fr

Une des grandes controverses épistémologiques en linguistique est l'opposition entre linguistique introspective et linguistique de corpus. Outre son intérêt théorique, cette question a des enjeux dans le traitement automatique des langues. Croft (1993, 1998) contribue à ce débat en proposant une opposition entre méthode expérimentale et méthode observationnelle. Dans cet article, nous nous proposons d'examiner cette opposition à la lumière de l'expérience apportée par l'application à grande échelle d'une méthode de description de langues vivantes par des locuteurs natifs, le lexique-grammaire.

### **1. Manipulation et observation en linguistique**

On a souvent opposé deux abordages méthodologiques de la linguistique, tous les deux caricaturés dans un célèbre article de Fillmore (1992) sous les noms de linguistique dans un fauteuil et linguistique de corpus. Dans un contexte où il traite de typologie des langues, Croft (1993, 1998) propose une opposition entre méthode expérimentale et méthode observationnelle, qui recouvre très étroitement l'opposition précédente.

La « méthode expérimentale », qu'on pourrait peut-être plus proprement qualifier de manipulatoire, consiste à émettre une hypothèse linguistique, à forger des exemples dans lesquels on fait varier systématiquement et indépendamment les paramètres pertinents, à soumettre ces formes à des jugements introspectifs d'acceptabilité, et à en déduire des règles. Cette méthode peut théoriquement être appliquée en l'absence de tout corpus d'exemples préexistant à l'étude.

La « méthode observationnelle » consiste, quant à elle, à observer les formes qui figurent dans un corpus préexistant à l'étude, puis à formuler des généralisations. Elle constitue pour Croft « une alternative légitime à la méthode expérimentale ».

Prenons comme exemple un problème déjà utilisé pour discuter de ces deux méthodes par Boons *et al.* (1976), celui de la préposition régie par le verbe *abonder* : *en* ou *de* ? Appliquons la « méthode expérimentale » et formulons par exemple l'hypothèse que seule l'une des deux prépositions est employée, ou qu'il s'agit de variantes libres. Les exemples forgés seront tels que les suivants :

- (1) *La forêt abonde en champignons*
- (2) *La forêt abonde de champignons*
- (3) *Sa conversation abonde en jeux de mots idiots*
- (4) *Sa conversation abonde de jeux de mots idiots*

etc., de façon à faire varier les valeurs lexicales des arguments du verbe, des déterminants et de tout autre paramètre jugé pertinent par l'expérimentateur. Observant que toutes ces formes sont acceptables, celui-ci validera l'hypothèse d'une règle selon laquelle il s'agit de variantes libres.

Appliquons maintenant la « méthode observationnelle » : nous nous munissons d'un corpus et nous comptons par exemple le nombre d'occurrences de chacune des séquences *abonder en* et *abonder de*. Avec un corpus de 820 000 mots constitué

d'articles du *Monde*, nous n'avons obtenu qu'une occurrence de *abonder en*. Nous ne pouvons en déduire qu'une chose : le corpus est trop petit pour se prononcer. En prenant comme corpus la collection des pages Web écrites en français, nous avons 98 000 occurrences de *abonder en* et 35 000 de *abonder de*. On peut en déduire que les deux constructions sont en usage.

Indépendamment des questions de fond, il faut bien reconnaître que certains des termes retenus par les différents auteurs sont peu appropriés. Nous ne visons pas ici l'expression *linguistique dans un fauteuil*, qui est ouvertement polémique et stigmatise les excès des linguistes peu empressés de vérifier si leurs théories sont en accord avec l'usage effectif, mais les termes de « méthode expérimentale » et de « méthode observationnelle », aussi contestables l'un que l'autre. En effet, chacune des deux méthodes que nous venons d'illustrer se fonde sur des observations empiriques. De plus, dans les deux cas, nous avons répondu à une question préalablement formulée en appliquant une stratégie qui recourt à une confrontation avec la réalité, ce qui fait de l'opération une expérience. Toute expérimentation comporte d'ailleurs une part d'observation.

Dans cet article, nous nous proposons d'examiner différents aspects de cette opposition à la lumière, d'une part, des arguments avancés par Croft, et d'autre part, de l'expérience apportée par l'application à grande échelle d'une méthode de description syntactico-sémantique, le lexique-grammaire. Nous nous limiterons au cas de l'étude d'une langue vivante par des locuteurs natifs. Dans ce cadre, nous argumenterons en faveur d'une exploitation combinée des deux types de méthodes.

## **2. Le parallèle avec les sciences expérimentales**

Le parallèle avec les sciences expérimentales est souvent mis à contribution dans le débat qui nous occupe. Il peut en effet être assez éclairant. Nous nous pencherons en particulier sur les pratiques de la recherche médicale, de la biologie et de la physique.

### **2.1. Expérimentation et observation en recherche médicale**

Croft (*ibid.*) emprunte en fait l'opposition entre expérimental et observationnel à la médecine. Ici encore, ces termes sont peu adaptés, mais ils sont traditionnels. En médecine, la distinction entre étude expérimentale et étude observationnelle consiste en une différence de degré dans le contrôle des paramètres. Dans les deux cas, il s'agit d'évaluer les corrélations entre des symptômes et des facteurs éventuels tels que des traitements médicaux, des habitudes de vie ou des configurations génétiques.

Une étude expérimentale est une étude dans laquelle l'expérimentateur contrôle tous les paramètres qui peuvent se révéler être des facteurs potentiels des symptômes étudiés. Pour chacun de ces paramètres, soit on connaît sa valeur (par exemple, on sait de chaque sujet s'il est végétarien ou s'il mange de la viande), soit on s'assure que les résultats statistiques de l'étude seront indépendants de lui, en constituant soigneusement les cohortes de sujets à cet effet. Par exemple, pour rendre l'étude indépendante de l'inclination personnelle des sujets quant à la consommation de viande, on constitue une cohorte de végétariens et une cohorte de mangeurs de viande en leur demandant d'adopter le régime alimentaire correspondant, et non de suivre librement leur choix. Ainsi, le choix qui aurait été le leur spontanément, même s'il n'est pas enregistré formellement comme paramètre, sera sans influence sur les résultats statistiques, du moment que les cohortes sont suffisamment grandes.

Dans une étude observationnelle, on n'exige pas que l'expérimentateur contrôle tous les paramètres. Par exemple, on peut comparer les symptômes d'une cohorte de végétariens et d'une cohorte de mangeurs de viande sans se préoccuper de la façon dont

ils ont déterminé leur régime alimentaire, ce qui est beaucoup plus facile à réaliser. Cependant, les résultats d'une telle étude pourront être biaisés par la confusion entre deux paramètres *a priori* distincts : le régime alimentaire effectif et celui que le sujet aurait adopté spontanément. Par exemple, elle ne permettra pas de prédire les symptômes d'un sujet personnellement enclin au végétarisme, mais à qui on prescrirait de manger de la viande.

Le protocole expérimental est considéré comme plus à même de permettre la démonstration statistique de relations de cause à effet, alors qu'une étude observationnelle fournit des indications. Ainsi, contrairement à ce que suggère Croft pour le domaine de la typologie des langues, la méthode observationnelle en recherche médicale n'est en général pas « une alternative légitime à la méthode expérimentale ».

## **2.2. Expérimentation et observation en physique et en biologie**

Les correspondances évoquées par Croft (*ibid.*) entre la physique et la « méthode expérimentale », puis entre la biologie et la « méthode observationnelle », nous semblent moins convaincants que le parallèle avec la recherche médicale. La physique comporte une partie d'observation, par exemple l'observation systématique du spectre des corps. Même si ces observations sont effectuées au moyen d'expériences, un de leurs objectifs est l'observation pure et simple du monde qui nous entoure. Quant à la biologie, elle peut recourir à une démarche observationnelle : par exemple, un naturaliste se rend dans un milieu naturel et y observe la présence d'espèces animales ou végétales qu'il y rencontre ; mais cela n'exclut pas la réalisation d'expériences, par exemple sur l'influence de facteurs internes ou externes sur le comportement d'un animal.

On observe en fait, dans les sciences expérimentales en général, une combinaison de pratiques comparables à ce que Croft appelle les méthodes expérimentale et observationnelle. Ces deux types de pratiques correspondent à des objectifs immédiats distincts, elles sont perçues comme complémentaires par les scientifiques, et elles se fournissent des résultats mutuellement. Par exemple, il est naturel que des observations, même réalisées en dehors d'un protocole expérimental strict et coûteux, suscitent des hypothèses qui sont ensuite testées lors d'expériences rigoureuses ; inversement, certains résultats obtenus par des expériences, notamment des mesures de grandeurs, peuvent être considérées comme des observations empiriques. Ces situations sont abondamment illustrées par l'histoire des sciences expérimentales.

Nous pensons en fait qu'il en va de même pour la linguistique, et que la querelle entre le linguiste introspectif et le linguiste de corpus, tels qu'ils sont caricaturés par Fillmore (1992), n'a d'autre raison d'être que leur réticence à faire collaborer deux abordages méthodologiques pourtant compatibles. C'est ce que nous tentons de montrer dans ce qui suit.

## **3. Manipulation et observation dans la pratique du lexique-grammaire**

Le lexique-grammaire (Gross, 1975, 1981, 1994) désigne à la fois une méthodologie et une pratique effective de description manuelle syntaxico-sémantique. Cette méthodologie et cette pratique se sont d'ailleurs développées simultanément à partir de la fin des années 1960, se nourrissant mutuellement. Nous allons délimiter ce qu'elles doivent aux méthodes « expérimentale » et « observationnelle ».

### **3.1. Le lexique-grammaire : principes, résultats**

La base théorique sur laquelle se fonde le lexique-grammaire est le distributionnalisme de Harris (1964, 1976). Les principes méthodologiques qui s'y ajoutent (Gross, 1975, 1981, 1994) peuvent être vus comme l'adoption de certaines priorités dans un programme de description syntaxico-sémantique des langues.

L'interaction entre le lexique et la syntaxe est ainsi considérée comme une clé incontournable. La recherche exclusive de règles de syntaxe générales, indépendantes du matériel lexical qu'elles manipulent, est dénoncée comme une impasse. Inversement, la description du vocabulaire d'une langue est vue comme l'étude des façons dont chaque élément lexical s'insère dans des phrases. En d'autres termes, l'unité minimale prise comme contexte pour la description d'un mot est la phrase élémentaire.

Le lexique-grammaire pose également une exigence de formalisation. Les résultats de la description doivent être suffisamment formels pour permettre :

- une vérification par confrontation à la réalité de l'usage,
- une application au traitement automatique des langues.

Cette contrainte de formalisation se manifeste par l'adoption d'un modèle discrétisé de la syntaxe. Ainsi, l'acceptabilité est modélisée par une notion binaire : pour les besoins de la description, une phrase est considérée soit comme acceptable, soit comme inacceptable. De même, l'ambiguïté lexicale est représentée par la séparation d'un mot en un nombre entier d'entrées lexicales, qui sont distinctes les unes des autres au même titre que deux entrées de mots morphologiquement différents. Les propriétés syntaxico-sémantiques sont identifiées par des intitulés qui représentent des structures de phrases, intitulés assez informels tels que «  $N_0 V N_1 W = N_1 V W$  » (voir section 3.2), mais qui forment une liste systématiquement confrontée à toutes les entrées. Enfin, seules sont retenues les propriétés pour lesquelles on trouve une procédure permettant de déterminer de façon suffisamment fiable si une entrée donnée la possède ou non : les propriétés sont donc modélisées comme binaires et non comme des continums.

Les résultats obtenus par l'application de ces principes méthodologiques par quelques dizaines de linguistes pendant quelques dizaines d'années font du lexique-grammaire une entreprise sans précédent. Pour se limiter au français, environ 13 000 entrées verbales, 10 000 entrées nominales, 12 000 entrées de phrases figées, 11 000 entrées adverbiales, donc un total de 75 000 entrées ont été établies, et 98 % d'entre elles croisées avec plusieurs centaines de propriétés syntaxico-sémantiques. Plus de la moitié de ces entrées sont mises en ligne gratuitement sur <http://infolingu.univ-mlv.fr>, et constituent une base d'informations syntaxico-sémantiques pour le traitement des langues sans équivalent dans le monde par son volume, la richesse des phénomènes linguistiques qu'elle prend en compte, et son degré de formalisation. Étant donné les principes ci-dessus, les résultats de la description prennent naturellement la forme de tableaux à double entrée, ou tables, qui croisent des entrées lexicales avec les propriétés syntaxico-sémantiques (figure 1). Plusieurs milliers d'autres entrées n'ont été publiées que sur papier, mais suivant le même modèle. Enfin, des descriptions plus ou moins substantielles, toujours suivant le même modèle, existent pour une dizaine d'autres langues, les mieux représentées étant l'italien, le portugais, le grec moderne et le coréen.

	N0					N0			
	=:					V			
N0 =:	N-			N0		N1			N1
Nhum	hum	Ppv	<ENT>	V	<ENT>Det1	Prep	<ENT>N1		=:
						N2			Npc
									[passif]

+	+	<E>détendre	- la	-	atmosphère	-	+
+	-	<E>détenir	- la	-	vérité	-	+
					hache de		
+	-	<E>déterrer	- la	-	guerre	-	+
+	+	<E>détourner	- la	-	conversation	-	+
+	-	<E>détourner	- la	-	tête	+	-
+	-	<E>devancer	- le	-	appel	-	+
+	-	se dévisser	- le	-	cou	+	-
+	+	<E>dévorer	- les	-	distances	-	-
+	+	<E>dévorer	- les	-	kilomètres	-	-
+	+	<E>dévorer	- la	-	route	-	-

Figure 1. Extrait d'une table d'expressions verbales figées (M. Gross)

### 3.2. Précautions méthodologiques pour la reproductibilité

Pour explorer l'interaction entre le lexique et la syntaxe, il est naturel de combiner systématiquement les entrées lexicales à toutes les structures de phrases observées, et d'analyser les séquences ainsi forgées : sont-elles acceptables ? quelles sont leurs particularités distributionnelles et sémantiques ?

La qualité des résultats dépend donc des capacités des linguistes à analyser les exemples construits. L'expérience montre que la manière la plus efficace d'effectuer cette exploration et cette analyse est de recourir massivement à l'introspection. Cependant, on est alors confronté à trois risques d'erreurs.

Le premier est celui d'une capacité insuffisante du linguiste à analyser les séquences, et notamment à juger de leur acceptabilité. On exclut qu'un linguiste applique la méthode du lexique-grammaire à un autre idiome que sa propre langue maternelle, y compris avec l'aide d'un informateur. Mais même ainsi, le jugement d'acceptabilité est un talent dont nous ne sommes pas tous également dotés, comme d'ailleurs c'est le cas pour à peu près toute activité humaine.

Le deuxième risque est celui d'une différence entre la langue décrite et l'idiolecte du descripteur. Nous en avons eu un exemple concret à la suite d'une relecture d'un de nos articles qui citait l'expression adverbiale *au petit bonheur la chance* sous la forme *au petit bonheur de la chance*, la seule que connaisse notre idiolecte personnel.

Le troisième risque est celui d'un préjugé inconscient du linguiste, influencé par un désir de voir vérifiée une de ses hypothèses. On a par exemple une tendance naturelle à régulariser un phénomène. Lors de l'étude de la relation de nominalisation établie entre les deux phrases suivantes :

*Luc atterrit* = *Luc fait un atterrissage*

on peut ainsi être tenté de surestimer l'acceptabilité de la séquence (5) :

*Luc embraye* = (5) ? *Luc fait un embrayage*

Tous ces problèmes sont bien connus des linguistes qui mènent une activité descriptive régulière. Ils se sont posés dès le début de la construction du lexique-grammaire. Ils ont leur équivalent dans toute science expérimentale : il s'agit des difficultés pratiques susceptibles de faire obstacle à la reproductibilité d'une expérience ou de la mesure d'une grandeur. Une expérience, une mesure, n'ont d'intérêt scientifique que si elles sont reproductibles, c'est-à-dire si un expérimentateur qui la pratique à nouveau obtient les mêmes résultats.



consiste à comparer deux différences sémantiques. Par exemple, la différence entre (6) et (7) :

(6) *Luc plie la branche* = (7) *La branche plie*

peut être comparée à celle entre (8) et (9) :

(8) *Luc pèse le sac* = (9) *Le sac pèse*

Le lecteur percevra probablement, comme nous, que la différence entre (6) et (7), qui semble avoir trait à la causalité du procès, ne se retrouve pas du tout entre (8) et (9). Cette observation est beaucoup plus reproductible que celle qui consisterait par exemple à caractériser en quoi (6) et (7) diffèrent sémantiquement. De plus, la notion d'évaluation sémantique différentielle est au cœur de la notion harrissienne de transformation : la relation  $N_0 V N_1 W = N_1 V W$  ne sera considérée comme une transformation que si la différence sémantique entre (6) et (7) se retrouve dans un nombre suffisant de paires reproduisant les deux mêmes structures avec un autre matériel lexical.

Le lexique-grammaire fait appel à des critères sémantiques absolus dans les cas exceptionnels où ils donnent des résultats jugés suffisamment reproductibles. Ainsi, certaines propriétés syntaxico-sémantiques des verbes distributionnels formalisées dans le lexique-grammaire mettent en jeu le fait qu'une phrase exprime un déplacement d'une entité, dénotée par un des arguments, par rapport à un lieu, dénoté par un autre. Par exemple, la propriété  $N_0 V N_1 Loc N_2$  n'est vérifiée que s'il y a déplacement de l'entité dénotée par  $N_1$  par rapport au lieu dénoté par  $N_2$  :

(10) *Luc glisse l'enveloppe dans le tiroir*

Pour rendre suffisamment reproductible l'application de ce critère, notamment dans le cas d'un déplacement plus ou moins abstrait, ou de phrases dont l'interprétation met en jeu un autre procès parallèle au déplacement, il a été nécessaire de l'encadrer en mettant au point le processus suivant (Guillet et Leclère, 1992) : on forge deux phrases locatives, dont l'une est la négation de l'autre, en l'occurrence

(11) *L'enveloppe n'est pas dans le tiroir*

(12) *L'enveloppe est dans le tiroir*

et on vérifie que l'interprétation de (10) suppose que celle de (11) est vérifiée avant le procès, et celle de (12) après.

C'est à ce prix que des propriétés syntaxico-sémantiques ont pu être considérées comme définies avec suffisamment de précision pour que cela eût un sens de les confronter à l'ensemble du lexique.

Les tables de lexique-grammaire réalisées dans les années 1970 et 1980 l'ont été grâce aux précautions méthodologiques que nous venons d'exposer. Comme le lecteur a pu s'en rendre compte, elles ne font appel à aucun moment à l'utilisation de corpus, se limitant ainsi à une linguistique exclusivement introspective et manipulatoire.

En effet, à cette époque, les collections de textes disponibles sur format électronique étaient trop limitées en taille pour apporter une amélioration du processus. Les concordanciers disponibles n'étaient pas suffisamment élaborés pour permettre de produire des concordances lemmatisées à partir de texte non annoté (c'est d'ailleurs toujours le cas des concordanciers utilisés par la majorité des linguistes). Enfin, il n'existait pratiquement pas de corpus de textes annotés ni lemmatisés.

c) Au début des années 1990, cette situation a changé, ce qui a permis aux contributeurs du lexique-grammaire de recourir de plus en plus facilement à une troisième précaution méthodologique : la prise en compte d'exemples attestés dans des corpus. En effet, d'une part, avec la création du système Intex (Silberstein, 1993), il a été possible de rechercher dans de grandes collections de textes des structures linguistiques spécifiées par leur contenu lexical et morpho-syntaxique (lemmes,

catégories grammaticales, traits flexionnels), et de produire les concordances correspondantes<sup>2</sup>, beaucoup plus utiles que celles produites par les concordanciers sans lexiche. D'autre part, avec la création du web, puis du moteur de recherche Google (1999) et du système Webcorp (Renouf, 2003), d'immenses collections de textes sont devenues accessibles. Ainsi, le projet BFQS recourt fréquemment à des formes attestées sur le web (figure 2). Ce contrôle supplémentaire par l'observation de corpus se substitue partiellement au contrôle mutuel évoqué en a) ci-dessus, avec l'avantage qu'il met en jeu plus de locuteurs. Cependant, il ne saurait remplacer purement et simplement l'ensemble des précautions méthodologiques que nous avons exposées. Les raisons en sont bien connues et ont été présentées bien des fois dans le débat sur les mérites respectifs de la linguistique introspective et de la linguistique de corpus. Contentons-nous de les rappeler brièvement :

- L'observation de corpus ne fournit pas d'analyses des différences de sens ou des différences entre variantes d'une langue.
- Elle ne fournit pas, à elle seule, de formalisation des faits observés.
- Elle n'atteste pas les inacceptabilités : par exemple, l'absence de l'expression *abonder de* dans un corpus de 820 000 mots ne prouve en rien que cette expression soit inusitée.

Expression	B	F	Q	S	Paraphrase	Exemple
Amuser à des riens (s')	+	+	+	+	Se distraire avec des futilités	Il est comme un petit enfant, il s'amuse à des riens.
Amuser à un rien (s')	+	!	-	+	Se distraire avec des futilités	
Amuser bien (s')	+	-	-	-	Se plaire quelque part	Est-ce que tu t'amuses bien dans ton nouvel appartement ?
Amuser la galerie	+	+	+	+	Distraire l'assistance	"Lorsqu'il était petit, il amusait la galerie avec ses mimiques, ses blagues : un acteur était né. " (www)
Amuser le tapis	-	+	-	+	Distraire l'assistance	"Raffarin veut-il amuser le tapis ? Après tout, pourquoi pas, mais la situation dramatique de la France mérite mieux." (www)
Amuser le temps	-	-	+	-	Faire passer le temps	Pierre n'a rien fait de la journée. De plus en plus, j'ai l'impression qu'il amuse le temps.

Figure 2. Un extrait du dictionnaire BFQS

Pour toutes ces raisons, bien que le travail descriptif du lexique-grammaire ait recouru de façon croissante à un contrôle par l'observation de corpus dans les années 1990 et 2000, cela n'a en aucun cas amené à abandonner les précautions

<sup>2</sup> Le système Unitex (Paumier, 2006), disponible gratuitement (<http://univ-mlv.fr/~unitex>), propose la même fonctionnalité. Le système Glossanet fait de même en explorant le texte de pages web renouvelées quotidiennement, gratuitement aussi (Fairon et Singler, 2006).

méthodologiques mises au point lors de la période précédente ; cette nouvelle précaution s'est simplement ajoutée aux précédentes, faisant du lexique-grammaire une méthode qui relève à la fois de la linguistique introspective et de la linguistique de corpus, un peu comme le préconisait Fillmore (1992). Les projets américains FrameNet (Baker *et al.*, 2003) et VerbNet (Kipper-Schuler *et al.*, 2006) témoignent d'ailleurs d'une relative convergence vers des objectifs proches de ceux du lexique-grammaire.

Examinons par exemple la procédure utilisée dans le cadre du projet BFQS pour détecter les expressions verbales figées dont l'emploi n'est pas uniforme dans les quatre variantes du français. Les représentants de chaque variante établissent d'abord quatre listes séparées. Celles-ci sont comparées dans un deuxième temps. Or, pour découvrir, par exemple, qu'une expression de la liste B (pour Belgique) est inusitée dans la variante F (pour France), il faut un contact entre un représentant de la variante B et un représentant de la variante F. En effet, si une expression de la liste B est absente de la liste F, l'auteur de cette dernière peut tout simplement ne pas l'avoir remarquée. De plus, si une expression figure à la fois dans les listes B et F, cela ne signifie pas nécessairement qu'elle soit commune aux deux variantes : il faut dans ce cas confronter les interprétations, et si elles sont différentes, il s'agit, d'un point de vue lexicologique, de deux expressions distinctes, chacune étant usitée dans une des variantes et inusitée dans l'autre.

On le voit, cette procédure nécessite des analyses de nature introspective. Elle est loin de se limiter à la comparaison informatique de deux bases de données, et encore moins de deux corpus.

d) Les tables de lexique-grammaire étant pour l'essentiel publiées, il est loisible à chacun de juger si les différentes précautions prises ont joué leur rôle, en vérifiant si les résultats obtenus sont en accord avec les jugements d'acceptabilité que peuvent émettre les locuteurs du français. Un tel examen montre qu'une certaine proportion des marques indiquant si les entrées possèdent ou non les propriétés sont erronées. Cependant, des deux principales causes apparentes de ces erreurs sont :

- d'une part, l'existence de colonnes correspondant à des propriétés mal définies, introduites à titre expérimental par les linguistes dans l'attente de l'opinion de leurs pairs, mais qu'il vaudrait mieux ne pas considérer comme des solutions satisfaisantes des problèmes descriptifs auxquels elles correspondent ; par exemple, les propriétés des verbes distributionnels mettant en jeu les noms de parties du corps (*Npc*) ;

- d'autre part, des erreurs informatiques de transfert de données d'un système à un autre<sup>3</sup>.

Ainsi, les informations linguistiques formalisées dans les tables de lexique-grammaire dans leur état actuel possèdent un intérêt scientifique et technique de premier ordre, sans être exemptes d'erreurs. Mais la méthode de construction de ces tables présente un intérêt plus important encore du point de vue linguistique ; de plus, elle permet de corriger les erreurs et de construire les tables manquantes pour d'autres parties du lexique ou d'autres langues.

### 3.3. Mise à l'épreuve d'hypothèses

Dans la section précédente, nous avons essentiellement évoqué des techniques d'analyse d'exemples construits. Tournons-nous maintenant vers l'art de forger les exemples. Pour pouvoir tirer des conclusions valides de leur analyse, il faut en effet les construire de façon rigoureuse et organisée. Considérons par exemple le problème de la relation

---

<sup>3</sup> Pendant les vingt premières années d'existence du lexique-grammaire, les outils informatiques de manipulation des bases de données contenant du texte étaient déficients, et les normes de représentation du texte chaotiques.

entre les constructions en *abonder de* et *abonder en* que nous avons mentionnées dans la section 1, et une troisième construction illustrée par :

(13) *Les champignons abondent dans la forêt*

L'identité du matériel lexical et la ressemblance sémantique font penser à une transformation qui intervertirait le sujet et le complément. Mettons donc à l'épreuve l'hypothèse d'une transformation  $N_0$  *abonder* *Loc*  $N_1 = N_1$  *abonder* (*de + en*)  $N_0$  (Boons et al., 1976 ; Salkoff, 1983).

Cette notation informelle ne fait pas apparaître les déterminants et les modificateurs du nom, mais le déterminant de  $N_0$  ne concorde pas dans nos exemples (1) et (13). Dans d'autres transformations intervertissant les arguments syntaxiques d'un verbe distributionnel (dites croisements), les déterminants peuvent être conservés :

*Luc saupoudre un peu de sucre sur le gâteau*

*Luc saupoudre le gâteau d'un peu de sucre*

*Luc va saupoudrer ce sucre sur le gâteau*

*Luc va saupoudrer le gâteau de ce sucre*

Testons donc l'hypothèse, généralisante donc simplifiante, d'une conservation du déterminant dans la transformation hypothétique qui nous occupe. Une expérimentation adaptée à la vérification de cette hypothèse consiste à faire varier le déterminant de  $N_0$  dans (1) et (13) avant d'analyser les séquences ainsi forgées :

(13) *Les champignons abondent dans la forêt*

= (14) \* *La forêt abonde (des + en les) champignons*

(15) \* *Des champignons abondent dans la forêt*

= (16) *La forêt abonde (de + \* en des) champignons*

Remarquons que la construction des exemples n'est pas triviale : elle nécessite d'appliquer les contractions, celles qui sont connues (*de les = des*) comme celles qui sont moins souvent citées (*de des = de*) par les linguistes. Les diverses interdictions marquées dans les exemples (14)-(16) ci-dessus invalident l'hypothèse qui les a suscitées, et suggèrent que la réalité de l'usage est plus complexe. Toutefois, la prise en compte d'autres déterminants, au contraire, suggère que l'hypothèse serait valide pour certains déterminants, ici *ce type de* et *un certain type de* :

(17) (*Ce + Un certain*) *type de champignons abonde dans la forêt*

= (18) *La forêt abonde (de + en) (ce + un certain) type de champignons*

Ceci suggère à l'expérimentateur de passer en revue les différentes catégories de déterminants et de mener des expériences indépendantes en fonction de cette typologie.

Sans pousser plus loin cette étude de cas, nous pouvons observer que les exemples qui nous ont permis de trouver des réponses partielles à nos questions successives ont la particularité de faire varier chaque paramètre indépendamment. Entre (1) et (2), c'est la préposition qui varie. Entre (1) et (3), comme entre (2) et (4), ce sont les arguments qui varient indépendamment de la préposition. Entre (1) et (13), nous avons fait varier deux paramètres à la fois : la position des arguments et le déterminant de  $N_0$ . Ce manquement à la rigueur est corrigé par la construction des exemples (14) à (16) puis (17) et (18), destinés à séparer les deux paramètres en question.

En d'autres termes, comme dans toute science expérimentale, on imagine des expériences destinées à mettre en évidence séparément les effets liés aux différents paramètres qui peuvent se révéler être des facteurs des phénomènes observés. On parvient ainsi à valider les différentes hypothèses sous-jacentes à ces expériences, ou à imaginer d'autres hypothèses.

#### 4. Les critiques de Croft

Dans son article, Croft (*ibid.*) préconise implicitement l'emploi de la « méthode observationnelle », « une alternative légitime à la méthode expérimentale », en formulant deux critiques à l'encontre de cette dernière. Examinons-les maintenant à la lumière de l'expérience apportée par le lexique-grammaire. Gardons cependant à l'esprit que Croft se place dans le contexte de la typologie des langues, un domaine qu'on ne saurait identifier à celui de la description syntaxico-sémantique.

#### 4.1. Objectivité et subjectivité

La première critique de Croft contre la linguistique introspective est classique : comme l'expérience introspective porte sur le jugement d'acceptabilité de l'expérimentateur lui-même, ce dernier est objet de sa propre expérience, d'où un risque de biais, que nous avons d'ailleurs déjà évoqué dans la section 3.2. Comme le dit Croft, « ce sont là des conditions qu'un psychologue, par exemple, rejetterait immédiatement ». La linguistique de corpus est exempte de ce risque, car l'observateur est indépendant des auteurs du corpus.

Remarquons tout d'abord que dans d'autres sciences expérimentales, le risque lié au fait que l'expérimentateur est en partie objet de sa propre expérience est considéré, dans certaines conditions, comme un risque contrôlé qui n'entache en rien la validité des résultats obtenus. Ainsi, lorsqu'un naturaliste observe l'odeur d'un champignon (un des éléments essentiels à la détermination des espèces en pratique), on n'exige pas qu'il le fasse sans savoir d'où vient l'odeur, ni sans savoir où il était quand il a trouvé le champignon. Ce sont des conditions qu'un biologiste rejetterait immédiatement, car elles sont aussi inutiles qu'impraticables. Les biologistes ont d'ailleurs le bon sens de ne pas rechercher l'aide de psychologues pour déterminer les espèces de champignons. En sciences expérimentales, la notion de reproductibilité est considérée comme plus pertinente que celle d'objectivité.

Pour revenir à la linguistique, nous renvoyons à la section 3.2 de cet article pour rappeler que les auteurs du lexique-grammaire, conscients du risque de biais en question, se sont entourés d'un luxe de précautions qui les ont obligés à de la rigueur et qui reposent sur des notions non pas psychologiques, mais bien linguistiques.

Croft n'y fait aucune allusion, bien qu'elles aient été mises en pratique sur des données d'une large couverture lexicale et grammaticale, et que la majeure partie des résultats obtenus aient été publiés. Peut-être juge-t-il la linguistique introspective à l'aune de son représentant le plus connu, la grammaire générative.

Il est vrai que la critique sur le manque d'objectivité est assez juste en ce qui concerne ce mouvement, grand producteur de structures abstraites dont il est difficile en pratique de vérifier la conformité avec des faits observables ; en d'autres termes, des hypothèses infalsifiables au sens de Popper (1959). La grammaire générative n'applique pas de méthodes particulièrement élaborées en matière d'observation. La notion d'observation est d'ailleurs considérée comme relativement triviale dans les traditions « culturelles » de la grammaire générative. Ainsi, l'« adéquation observationnelle » occupe le niveau le plus bas dans la hiérarchie des trois niveaux d'adéquation d'une représentation grammaticale : l'adéquation observationnelle, descriptive et explicative. Cela en dit long, étant donné l'omniprésence de la notion de prestige au sein de ce mouvement. Cette position collective peut être vue comme une réaction à la position méthodologique de Harris, résolument orientée, quant à elle, vers la surface directement observable.

Cependant, la linguistique introspective ne mérite pas d'être évaluée à travers ceux de ses représentants qui se montrent désinvoltes en ce qui concerne l'observation des faits, fussent-ils les plus connus.

Quant à la linguistique de corpus, son exigence d'objectivité est parfois aussi excessive qu'elle est prise à la légère par la grammaire générative. Le développement spectaculaire de la linguistique de corpus présente d'ailleurs, lui aussi, certains aspects d'une révolution face à une position méthodologique antérieure, vue comme la « linguistique dans un fauteuil », notamment au Royaume-uni où la linguistique de corpus a presque étouffé les autres abordages de la linguistique.

Mais nous nous égarons en direction de la sociologie des mouvements scientifiques. Revenons à des arguments scientifiques.

#### **4.2. Formulation d'hypothèses**

La deuxième critique de Croft s'applique encore moins bien au domaine qui nous intéresse. Dans l'application de la « méthode expérimentale », déplore-t-il, « aucune généralisation préformulée n'est testée comme doit toujours le faire un expérimentateur pour mener une expérience ». Voilà une raison bien peu convaincante de renoncer à la linguistique introspective.

Premièrement, la formulation d'hypothèses préalables aux expériences est au coeur de la pratique effective de la linguistique introspective (cf. section 3.3). On l'emploie d'ailleurs pour se prémunir contre la complexité des faits qui, selon Croft, empêche de l'appliquer. Il est vrai que la typologie des langues, qui suscitait les propos de Croft, met en jeu un niveau de complexité supplémentaire.

Deuxièmement, l'observation de faits indépendamment de la formulation explicite d'une hypothèse est une activité scientifique légitime. Les sciences expérimentales en fournissent de nombreux exemples, de la recherche médicale (les études observationnelles, justement) à la physique (l'observation et le recensement systématiques des corps célestes, ou des propriétés des éléments) en passant par la biologie (l'observation des espèces vivantes peuplant les biotopes).

Troisièmement, la « méthode observationnelle » préconisée par Croft ne comporte en général, quant à elle, pas de formulation d'hypothèses.

Rejeter la linguistique introspective au motif qu'on n'y formule pas d'hypothèses reviendrait donc à se priver de cet outil en invoquant le fait qu'il est souhaitable de s'en servir.

#### **5. La règle et l'exemple**

L'article de Croft (*ibid.*), comme on l'a vu, emprunte les termes, sinon les notions, de « méthode expérimentale » et de « méthode observationnelle » à la recherche médicale, où on considère consensuellement, contrairement à ce que Croft préconise implicitement, que la première fournit plus d'informations que la deuxième. Pouvons le parallèle avec la recherche médicale un peu plus loin que le simple emprunt de termes. Le défaut d'une étude « observationnelle » est de ne fournir qu'une collection de cas, alors qu'une étude « expérimentale » est conçue pour que la collection de cas ait les propriétés statistiques nécessaires à ce qu'on puisse en déduire des relations de cause à effet, autrement dit des règles. On a donc deux types d'études : les unes moins coûteuses, les autres plus à même de mettre en évidence des règles. Or cette complémentarité se retrouve entre la linguistique de corpus et la linguistique introspective.

Pour le montrer, rappelons d'abord quelques notions de bon sens sur les notions de règle et d'exemple. En mathématiques, une règle a plus de valeur que des exemples si elle est plus générale. Ainsi, comparons une règle : « Aucun entier pair supérieur à 2 n'est premier », à deux exemples : « 6 et 14 ne sont pas premiers ». Il est naturel de considérer la règle comme plus intéressante, car on peut en déduire les deux exemples.

Cette préférence a ses limites. En effet, si la règle est fautive, elle n'a plus aucune valeur. De même, si les exemples énumèrent la totalité des possibilités couvertes par la règle, la préférence en faveur de celle-ci ne va plus de soi.

Dans la suite, nous examinons ce que ces notions apportent au débat sur la linguistique introspective et la linguistique de corpus, puis sur les principaux abordages du traitement automatique des langues.

### 5.1. En linguistique

En principe, la linguistique introspective et manipulative a la capacité de découvrir et formaliser des règles, dont l'accumulation peut former une grammaire. Nous renvoyons à la section 3 pour des exemples. En revanche, la linguistique de corpus pure se limite à des exemples tirés d'un corpus, et donc produit des résultats d'une portée moins générale, à moins d'une généralisation plus ou moins hasardeuse. En d'autres termes, elle ne résout pas, à elle seule, le problème de la formalisation. C'est là la raison principale qui nous conduit à soutenir la persistance d'abordages de la linguistique qui ne relèvent pas exclusivement de la linguistique de corpus.

Cependant, ce raisonnement a ses limites.

Premièrement, les règles produites par la linguistique introspective peuvent être fausses, notamment si la confrontation avec la réalité linguistique a été insuffisante ou a manqué de rigueur : c'est la « linguistique dans un fauteuil ».

Deuxièmement, même si aucun corpus ne peut recouvrir toutes les possibilités d'une langue, le web vu comme collection de textes d'une langue donnée tend d'une certaine façon à se rapprocher de cet idéal par son volume et sa diversité, au moins pour des langues telles que le français ou l'anglais, et malgré ses défauts : on y trouve par exemple beaucoup d'erreurs.

Ces deux réserves sont justement ce qui a motivé l'élaboration du savoir-faire méthodologique accumulé au cours du développement du lexique-grammaire, parallèle à celui de la linguistique de corpus. D'une part, des précautions méthodologiques appropriées ont permis de trouver des solutions au manque de rigueur dans l'observation qui a suscité la révolution des corpus ; d'autre part, la source d'information croissante constituée par les collections de textes a été de plus en plus utilisée.

Les résultats obtenus par l'application de ce corpus de méthodes ont surpris plus d'un linguiste. Il s'agit bien de règles, mais les différences entre entrées lexicales et entre constructions donnent l'image d'un chaos d'irrégularités beaucoup plus important qu'on ne pouvait le prévoir. Le modèle du lexique-grammaire permet évidemment de représenter le fait que deux éléments lexicaux possèdent exactement les mêmes propriétés syntaxico-sémantiques, mais lorsque cela se produit, l'examen d'autres propriétés permet presque toujours de trouver des différences entre les deux entrées. Or aucune théorie linguistique n'avait prévu une telle diversité, qui est contraire à l'intuition à peu près unanime des linguistes et des locuteurs. Une vaste entreprise de collection d'observations était donc aussi nécessaire qu'elle l'a été en physique ou en biologie, et l'est toujours.

De plus, ce résultat étonnant conduit à relativiser la notion de règle. Il apparaît en effet que nous avons une nette tendance intuitive à exagérer la généralité des règles en matière de syntaxe et de sémantique, puisque nous sommes si nombreux à être surpris de leur voir tant d'exceptions. Ceci démontre à nouveau la nécessité de prendre en compte des données observables susceptibles de contrebalancer cette tendance en mettant en évidence des contre-exemples aux règles qui nous semblent prévaloir.

De ce point de vue, une des traditions « culturelles » de la grammaire générative constitue un handicap : l'idéalisation de la notion de règle et la valorisation extrême de la généralité des règles. Le danger de cette tendance est de faire perdre de vue qu'une règle générale, mais qui n'est pas en conformité avec la réalité, est une généralisation hâtive, sans valeur scientifique.

## 5.2. En traitement automatique des langues

Bien que le traitement automatique des langues ne se soit pas structuré sur le modèle de la linguistique, on y trouve une opposition entre deux abordages méthodologiques qui correspondent d'une certaine façon à l'opposition entre linguistique introspective et linguistique de corpus.

Dans l'abordage dit symbolique, on utilise un modèle formel dans lequel on représente (par des symboles) les notions linguistiques ou cognitives pertinentes, leurs relations, et les règles à appliquer lors des traitements. Le modèle et les règles de manipulation sont construits manuellement. Par exemple, pour la traduction automatique, ces règles peuvent décrire des correspondances de langue à langue, ou des constructions syntaxiques dans une des langues. La linguistique introspective, dans la mesure où elle produit des résultats formels, nourrit l'abordage symbolique.

Dans l'abordage dit probabiliste, qui est majoritaire, on part de corpus servant d'exemples du traitement à effectuer, et on produit automatiquement, par analyse statistique, des données numériques à l'aide desquelles des programmes reproduisent d'aussi près que possible le comportement illustré par le corpus. Par exemple, pour la traduction automatique, on se procure un corpus de textes dans une langue et leur traduction dans l'autre, et on les fournit à un logiciel d'apprentissage automatique. La production de règles est automatisée sous la forme d'un traitement statistique du corpus d'apprentissage. Il s'agit donc d'une « industrialisation » de ce qui reste artisanal dans l'abordage symbolique.

L'abordage hybride qui consiste à combiner les deux précédents est peu développé et consiste souvent en quelques gouttes symboliques dans une mer probabiliste.

Les arguments qui ont assuré la popularité de l'abordage probabiliste auprès des ingénieurs sont liés à l'automatisme du processus, qui, d'une part, apporte une garantie d'objectivité, et d'autre part limite le coût de réalisation des produits. Ces deux arguments font écho à une partie du débat entre linguistique introspective et linguistique de corpus.

L'argument de l'objectivité reprend une des critiques classiquement émises par les linguistes de corpus (cf. section 4.1). Ici encore, les adeptes de l'abordage probabiliste parlent d'objectivité, mais jamais de reproductibilité ; ils ne se prononcent jamais sur les précautions méthodologiques prises par certains adeptes de la linguistique introspective, ou de l'abordage symbolique, pour s'assurer que leurs résultats décrivent bien une langue, et non un expérimentateur. Peut-être ignorent-ils ces précautions ; mais l'ignorance ne justifie pas un choix scientifique.

L'argument du coût rappelle l'opposition entre études observationnelles et études expérimentales en recherche médicale : celles-ci nécessitent plus de rigueur, et sont susceptibles de donner des résultats plus précis, mais le prix à payer pour cette différence de qualité est qu'elles sont plus coûteuses. De même, élaborer manuellement des modèles formels de la syntaxe, de la sémantique et du lexique est plus coûteux que d'en faire un modèle numérique par des programmes d'analyse statistique. Cependant, on ne peut aborder la question du coût sans celle de la qualité : le rapport qualité/coût est un critère d'évaluation plus pertinent que celui du coût. Or les produits

de l'abordage probabiliste ne donnent que des résultats rudimentaires, qui ne sont suffisants que pour les applications les plus élémentaires, comme les moteurs de recherche. En effet, l'abordage probabiliste est techniquement incompatible avec la manipulation de structures complexes, c'est-à-dire comportant de nombreux paramètres dont chacun peut prendre de nombreuses valeurs. Et c'est bien le cas des objets de base de la syntaxe : entrées lexicales, constructions syntaxiques, contextes. L'argument du coût est donc loin d'être décisif.

Citons un troisième et dernier argument, fréquemment cité dans les publications qui relèvent de l'abordage probabiliste : l'étude descriptive de la langue y est présentée comme « longue et ennuyeuse », raison invoquée pour exclure l'abordage symbolique. Cette expression est si souvent employée qu'elle fait presque partie du jargon du domaine. Voilà à nouveau un argument bien peu convaincant. Si un auteur invoquant cet argument s'est tourné vers la recherche en informatique, on peut croire sans difficulté qu'il a peu de goût pour la description linguistique, mais en quoi cela fait-il obstacle à ce qu'il utilise des données linguistiques construites par d'autres chercheurs qui, eux, sont passionnés par cette activité ? S'il n'aime pas faire la cuisine, il ne va pas pour autant recourir au jeûne : de même, si la réalisation d'un système de qualité le nécessite, pourquoi ne pas utiliser des résultats issus d'une autre discipline ? En supposant que notre auteur est de bonne foi, il manque singulièrement d'imagination. Le fait que des revues et colloques parmi les plus prestigieux publient des centaines d'articles dont les auteurs répètent cet argument nous semble même honteux pour leurs comités de sélection.

Si les mérites scientifiques et techniques de l'abordage probabiliste sont aussi discutables, comment expliquer sa popularité ? Elle permet à ses adeptes de faire l'économie d'une collaboration entre deux disciplines, l'informatique appliquée et la linguistique descriptive. Si cette explication est valable, le fond du problème serait le même que la querelle entre linguistique introspective et linguistique de corpus : une réticence à faire collaborer deux abordages méthodologiques pourtant compatibles.

### **Conclusion**

L'opposition de Croft (1993, 1998) entre « méthode expérimentale » et « méthode observationnelle » renouvelle le vieux débat entre linguistique introspective et linguistique de corpus, en suscitant un parallèle avec les sciences expérimentales, auxquelles Croft emprunte ces termes. L'exemple du lexique-grammaire, une méthode de description syntaxico-sémantique dont les fondements se réfèrent explicitement aux sciences expérimentales, est particulièrement éclairant dans ce débat et dans ce parallèle. Rappelons l'essentiel des enseignements que nous avons proposé d'en tirer.

- La formulation de règles conformes à la réalité de l'usage d'une langue est une technique qui ne se résume pas à une simple observation d'exemples.

- Cependant, elle nécessite bien une observation intensive d'exemples, ainsi que des précautions méthodologiques rigoureuses dans cette activité d'observation.

- Les traditions apparemment opposées de la linguistique introspective et de la linguistique de corpus sont donc complémentaires et de nature à se combiner pour favoriser le succès d'une telle entreprise ; il est donc contre-productif d'exclure l'une ou l'autre.

- La méthodologie du lexique-grammaire en fournit un exemple concret et productif de résultats.

Ces réflexions invitent les linguistes à surmonter leur réticence à combiner les deux types de méthodes.

De même, en traitement automatique des langues, la majeure partie de la communauté en reste à l'abordage probabiliste, renonçant à faire collaborer l'informatique appliquée avec la linguistique descriptive.

## Références

- Baker, Collin F., Charles J. Fillmore, Beau Cronin. 2003. The Structure of the Framenet Database, *International Journal of Lexicography* 16.3, pp. 281-296.
- Boons, Jean-Paul, Alain Guillet, Christian Leclère. 1976. *La structure des phrases simples en français. 1. Constructions intransitives*, Genève : Droz.
- Croft, William. 1993. "Functional-typological theory in its historical and intellectual context", *Sprachtypologie und Universalienforschung* 46, pp. 15-26.
- Croft, William. 1998, La théorie de la typologie fonctionnelle dans son contexte historique et intellectuel, *Verbum* 1998/3, pp. 289-307.
- Fairon, Cédric, John V. Singler. 2006. "I'm like, 'Hey, it works!': Using GlosaNet to find attestations of the quotative (be) like in English- language newspapers", in A. Renouf and A. Kehoe (eds). *The Changing Face of Corpus Linguistics*. Language and Computers 55. Rodopi, Amsterdam/New York, pp. 325-336.
- Fillmore, Charles. 1992. "'Corpus linguistics' vs. 'Computer-aided armchair linguistics'". *Directions in Corpus Linguistics*, Mouton de Gruyter, pp. 35-60. (Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm.)
- Gross, Maurice. 1975. *Méthodes en syntaxe*, Paris : Hermann.
- Gross, Maurice. 1981. « Les bases empiriques de la notion de prédicat sémantique », *Langages* 53, pp. 7-52, Paris : Larousse.
- Gross, Maurice. 1984. "A linguistic environment for comparative Romance syntax". In *Papers from the XIIth Linguistic Symposium on Romance Languages*, P. Baldi (ed.), pp. 373-446, Amsterdam/Philadelphia: John Benjamins.
- Gross, Maurice. 1994. "Constructing Lexicon-grammars". *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford University Press, pp. 213-263.
- Guillet, Alain, Christian Leclère. 1992. *La structure des phrases simples en français. 2. Constructions transitives locatives*, Genève : Droz.
- Harris, Zellig. 1964. "The Elementary Transformations", *Transformations and Discourse Analysis Papers* 54, in Harris, Zellig S. 1970, *Papers in Structural and Transformational Linguistics*, Dordrecht: Reidel, pp. 482-532.
- Harris, Zellig. 1976. *Notes du cours de syntaxe*, Paris : Seuil.
- Kipper-Schuler, Karin, Anna Korhonen, Neville Ryant, Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa.
- Labelle, Jacques. 1990. "Norms and variants in French", *Linguisticae Investigationes* 13:2, pp. 281-306, Amsterdam/Philadelphia: John Benjamins.
- Lamiroy, Béatrice, Christian Leclère, Jean René Klein, Jacques Labelle. 2003. "Expressions verbales figées et variation en français: le projet BFQS", *Cahiers de lexicologie* 83-2, pp. 153-172.
- Paumier, Sébastien. 2006. *Unitex Manual*. <http://univ-mlv.fr/~unitex>.
- Popper, Karl. 1959. *The Logic of Scientific Discovery*, Basic Books, New York.
- Renouf, Antoinette. 2003. 'WebCorp: providing a renewable data source for corpus linguists', in S. Granger & S. Petch-Tyson (eds.), *Extending the scope of corpus-based research: new applications, new challenges*. Amsterdam: Rodopi.
- Salkoff, Morris. 1983. Bees are swarming in the garden: a systematic synchronic study of productivity. *Language* 59, pp. 288-346.

Silberztein, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris : Masson.

### **Résumé**

L'opposition de Croft (1998) entre « méthode expérimentale » et « méthode observationnelle » renouvelle le vieux débat entre linguistique introspective et linguistique de corpus, en suscitant un parallèle avec les sciences expérimentales, auxquelles Croft emprunte ces termes. L'exemple du lexique-grammaire, une méthode de description syntactico-sémantique dont les fondements se réfèrent explicitement aux sciences expérimentales, confirme, d'une part, que la formulation de règles conformes à la réalité de l'usage d'une langue ne se résume pas à une simple observation d'exemples, et d'autre part qu'elle nécessite bien une observation intensive d'exemples, ainsi que des précautions méthodologiques rigoureuses dans cette activité d'observation. Les traditions apparemment opposées de la linguistique introspective et de la linguistique de corpus sont donc complémentaires et de nature à se combiner pour favoriser le succès d'une telle entreprise. Ces réflexions invitent les linguistes à surmonter leur réticence historique à combiner les deux types de méthodes.

De même, en traitement automatique des langues, la majeure partie de la communauté en reste à l'abordage probabiliste, renonçant à faire collaborer l'informatique appliquée avec la linguistique descriptive.

### **Abstract**

Croft (1993) contrasts an 'experimental method' with an 'observational method', thus renewing the discussion between introspective linguistics and corpus linguistics, by suggesting a parallel with experimental sciences, which these terms come from. The example of lexicon-grammar, a method of syntactic-semantic description constructed with explicit reference to experimental sciences, confirms (i) that formulating rules in accordance with the real usage of a language is not only a matter of observing examples, and (ii) that it does require intensive observation of examples, as well as rigorous methodological precautions in this observation. Thus, the apparently opposed traditions of introspective linguistics and of corpus linguistics are complementary and should be combined for the success of such an enterprise. These thoughts are an invitation for linguists to overcome their historical resistance to combining both types of methods.

Similarly, in natural language processing, most of the community sticks to the stochastic approach, which amounts to giving up co-operation between computer technology and descriptive linguistics.