Verbum XXX, 2008, 4

DE L'INTÉRÊT DES CORPUS DIVERSIFIÉS POUR LES DESCRIPTIONS EN (MORPHO)SYNTAXE. RÉFLEXIONS ET ILLUSTRATION AVEC LE PRONOM RELATIF *LEQUEL*

Mireille BILGER
Université de Perpignan-via-Domitia

RÉSUMÉ

*Cet article a pour objectif de rappeler quelques aspects concernant ce que l'on peut appeler la linguistique sur corpus, entre autres, ceux qui touchent à la dimension quantitative des données et aux différents types de productions. Une petite étude sur les occurrences du pronom relatif "*lequel*" et ses variantes, dans des corpus variés, servira d'illustration pour montrer l'intérêt qu'il y aurait à développer ce type de description afin de parfaire nos connaissances sur la langue.*

ABSTRACT

*This article sets out to revisit certain aspects concerning what we call corpus linguistics, in particular those pertaining to the quantitative approach to data and to production types. A small study of occurrences of the relative pronoun "*lequel*" and its variant forms, in different corpora, will serve as an illustration to show the interest for developing this type of description in order to gain a better understanding of language.*

Si l'on peut dire aujourd'hui que le débat sur l'utilité des corpus est clos, il n'en demeure pas moins vrai qu'il convient de rester vigilant face à ce consensus qui peut cacher certaines divergences, entre autres, sur **la fonction accordée au corpus** (va-t-il être un simple réservoir d'exemples ou le matériau source pour la description ?, selon la distinction *corpus–driven* versus *corpus based* utilisée par la tradition anglo-saxonne ou encore sur **la "pertinence" du corpus étudié**. Il va de soi que selon les objectifs de la recherche, cette notion de "pertinence" risque de varier[1]. Autrement dit, sans

1 Par exemple, les corpus conçus pour des études sur l'interaction vont présenter des caractéristiques bien différentes des corpus destinés à des études syntaxiques ou lexicales.

vouloir remettre en cause l'intérêt des corpus, il convient de s'interroger sur leur utilisation, de faire un tri entre les "services" qu'ils rendent et les "mirages" qu'ils peuvent faire naître. De fait, la portée - ou encore la validité - de la description basée sur ce type de données risque d'être fort différente, *voire discutable*, selon la façon dont le corpus a été conçu et les raisons pour lesquelles ce dernier l'a été.

En ce qui concerne le domaine (morpho)syntaxique, nombreux sont les travaux qui illustrent l'intérêt de s'appuyer sur des corpus variés et échantillonnés afin d'en parfaire la description, de renouveler l'analyse de certaines structures ou d'en proposer une nouvelle présentation. L'exemple le plus abouti reste encore l'ouvrage *Longman Grammar, Spoken and Written English* (Biber & alii, 1999), qui fournit des indications très précises sur la répartition des faits langagiers dans cinq types de productions (fiction, presse et prose académique / oral de conversation et non conversationnel)

L'article proposé ici a pour objectif de rappeler comment l'exploitation de corpus, variés et équilibrés, permet d'affiner la description du français et conduit à revenir sur des oppositions souvent présentées comme fondamentales, telles que celles que l'on a pu poser entre oral et écrit, lexique et grammaire, ou encore entre "système" et "usages" (Halliday, 1991). En nous appuyant sur une petite étude concernant le pronom relatif *lequel* et ses variantes, nous montrerons que travailler sur des grandes masses de données (par exemple les corpus d'oral et d'écrit) peut donner certains résultats, mais que la prise en compte d'un découpage plus fin (comme Presse et Littérature, oral de parole privée, oral de parole publique) peut en révéler d'autres qui obligeront à réexaminer les premiers. La description est donc étroitement dépendante des données utilisées et donc de la représentativité du corpus.

Ceci dit, avant d'entrer dans le vif du sujet, il est sans doute intéressant de revenir brièvement sur deux aspects concernant ce que l'on peut appeler **la linguistique sur corpus** : la dimension quantitative et la dimension "variationnelle" du corpus.

1. LA DIMENSION QUANTITATIVE

Il va de soi que la taille du corpus est un aspect important en ce qui concerne la représentativité des résultats, même si la notion de "gros corpus" ne renvoie pas aux mêmes réalités selon qu'il s'agit de productions écrites ou de productions orales (cf., entre autres, Habert, 2000). Par ailleurs, l'approche quantitative à l'aide de logiciels nécessite un certain nombre de précautions, en particulier s'il s'agit de corpus oraux. Rappel de quelques difficultés :

Certaines conventions de transcription peuvent compliquer le comptage automatique dans la mesure où la continuité de la séquence étudiée, comme la séquence *il y a* dans (1) ou la séquence *avoir la chance* dans (2), peut être interrompue par les signes marquant une pause (+) ou une indication de prise de parole :

(1) *il + y a quelques années*

ou par les conventions signalant la prise de parole d'un locuteur (L2) :

(2) *L1 quand on a*
L2 ça ne ça ne fait rien bien sûr
L1 la chance d'avoir euh travaillé (P06)

Ces différents éléments de contexte immédiat peuvent fortement jouer lors des décomptes automatiques et la recherche d'une forme peut en être affectée. Il en est de même d'ailleurs pour les productions écrites avec les différents signes de ponctuation.

Les transcriptions de l'oral comportent de nombreuses répétitions, des accidents de performance qui compliquent les comptages, du moins qui nécessitent une prise de position raisonnée. Ainsi, si l'on décide de compter les sujets "*je*" (ou "*j'*"), quel résultat proposer pour :

(3) *je je ne me rappelle plus de euh la date exacte (P06)*

Toutes les formes se prêtent a priori à une description quantifiée. Cependant, pour certaines d'entre elles, le nombre d'occurrences est très élevé (*de*, *à*, *et*, etc.) ce qui risque d'en alourdir l'examen. En ce cas, il est souvent préconisé d'effectuer des sélections de type "sondage" (par exemple, une forme sur cinq). Cette méthode présente l'avantage de réduire le nombre d'exemples à traiter tout en permettant de rendre les fréquences comparables et de juger de la pertinence des différences rencontrées. Mais elle présente aussi des inconvénients :

- d'une part, elle ne peut illustrer que de grandes tendances distributionnelles, susceptibles d'être modifiées par une étude exhaustive ultérieure,

- d'autre part, des exemples intéressants mais trop rares risquent de ne pas être pris en compte.

La deuxième grande limite, concernant "les comptages", est liée au fait que bon nombre de phénomènes grammaticaux intéressants à étudier ne sont pas forcément introduits par des morphèmes particuliers. Les requêtes automatiques deviennent alors difficiles, voire impossibles à formaliser. C'est le cas, par exemple, des séquences nominales de type "gapping", comme "*dans les campagnes sur l'agriculture*" dans (4) :

(4) *L'économie de la région elle est basée sur le tourisme + dans les campagnes sur l'agriculture* (oral)

Par ailleurs, certaines formes ou structures qui existent pourtant dans la langue ne se retrouvent pas forcément dans les corpus (sans doute parce que les situations ne s'y prêtent pas suffisamment), c'est le cas par exemple :

- des exclamatives du type (*comme il est beau*)

- de certaines tournures interrogatives indirectes du type (*je sais pas c'est qui / je sais pas c'est quoi*)

Enfin, pour terminer sur cet aspect "quantitatif", il est à signaler que la plupart des travaux sur corpus, menés à l'aide de concordanciers, mentionnent souvent des résultats en nombre d'occurrences ou sous forme de pourcentages - ce qui permet déjà de donner une idée de la variation – mais rares

sont ceux qui utilisent l'arsenal des outils statistiques (par exemple le Khi2) pour vérifier si les résultats sont pertinents ou la variation significative. Cette lacune, qui peut être "générationnelle" ou liée à la formation donnée en France, mériterait sans aucun doute une réflexion plus ample, même si l'on peut toujours se satisfaire d'indices permettant de révéler des grandes tendances de fonctionnement, susceptibles d'être approfondies et précisées par la suite.

2. LA DIMENSION VARIATIONNELLE DU CÔTÉ DES PRODUCTIONS

L'opposition la plus générale passe, encore aujourd'hui, entre l'oral et l'écrit. Pourtant, cela fait bien une trentaine d'années que de nombreux travaux ont contesté cette vision naïve d'une syntaxe différente entre ces deux modes de productions (entre autres, Blanche-Benveniste, 1993 ; Gadet, 2007). Plus de trente ans donc que l'on sait que du point de vue qualitatif, il y a en fait peu de différence. En revanche, ce qui peut être très différent entre l'écrit et l'oral, ce sont les faits de distribution et l'importance quantitative de certains phénomènes. Mais là encore, ces différences vont pouvoir être modulées selon les types de productions orales ou écrites sur lesquelles on travaille. Autrement dit, cette simple opposition "binaire" (oral/écrit) risque de s'avérer rapidement insuffisante et d'autres découpages risquent d'influer tout autant sur les résultats de l'analyse.

Les approches pour décrire ou classifier les types de productions sont nombreuses. Les notions de *genres* (Malrieu & Rastier, 2001 ; Rastier, 2002 ; Kerbrat-Orecchioni & Traverso, 2004 ; *Langage et Société*, 1999), de *registre (*Biber, 1988) ou de *style* (Gadet & Tyne, 2004 ; Bilger & Cappeau, 2004), même si elles sont loin d'être totalement interchangeables (Branca, 1999), cherchent à conceptualiser ces faits de variation liés aux types de productions.

Ces classifications font ressortir un point important : elles montrent que pour de nombreux phénomènes (lexicaux et morphosyntaxiques) l'opposition oral / écrit n'est pas adaptée ou pas suffisante et que d'autres sortes de découpages internes (identifiés comme des types de production) permettent de mieux rendre compte des fonctionnements. Pour l'écrit, une longue tradition nous a appris à reconnaître qu'il existe des différences importantes selon "les genres", (entre autres, Bronckart et alii, 1985). Pour l'oral, cette dimension doit aussi être sollicitée.

Cette problématique en relation avec la typologie des productions est d'ailleurs, aujourd'hui, le domaine qui suscite le plus grand nombre de réflexions (Rastier, 2000 ; Biber, 2006), d'autant plus que l'on peut donner à la liste des "genres", ou des types de productions, une extension plus ou moins large, ou plus ou moins limitée[2].

2 Ainsi, la Presse va pouvoir se décliner en sous-types : Sciences, Politique, People, etc., tout comme la Littérature (Fiction) : romans, nouvelles, théâtre, etc., sans oublier les "genres traditionnels" (récits, descriptions, etc.). Il en est de même pour l'oral qui va pouvoir se décliner en fonction de la situation (parole privée/publique)

Si la plupart des distinctions retenues se prêtent aisément aux critiques[3] et demanderaient encore à être affinées il n'en demeure pas moins qu'elles permettent de montrer comment les généralisations linguistiques se distribuent en fonction de la variété discursive des productions, variété discursive qui elle-même est définie à partir des différentes fonctions du langage. Les travaux qui en découlent donnent une idée de ce que pourrait être une linguistique fondée sur une méthode empirique, et ils ont de fait contribué à repenser l'opposition entre "système" et "usages" du système.

En ce sens, pour que la description proposée ait une portée générale ou soit autre chose que la description d'une production particulière, il est nécessaire de s'appuyer sur un corpus suffisamment varié, qui regroupe différents types de productions échantillonnées (écrites et orales).

Avoir à sa disposition un corpus de ce type, que l'on pourrait appeler **corpus de référence**, serait d'ailleurs utile, quelle que soit la recherche envisagée, dans la mesure où il permettrait de contraster les résultats obtenus et de mieux calculer ce qui peut être particulier ou spécifique dans le corpus étudié. Ceci dit, nous n'avons pas encore à notre disposition ce type de corpus[4] qui pourrait servir d'étalon à l'ensemble des recherches.

3. UNE ILLUSTRATION : LE PRONOM RELATIF "*LEQUEL*"

Des travaux menés précédemment ont montré que la plupart des catégories sont sensibles aux types de productions, par exemple : la préposition "*contre*", la forme adjectivale "*évident*", le joncteur "*et*" (Bilger & Cappeau, 2007-2009), *les adverbes en "–ment"* (Bilger, 2004), la forme "*clair*" (Bilger & Cappeau, à paraître). L'impact du type de corpus sur les descriptions fournies se révèle donc indéniable et il est nécessaire de le prendre en compte. Pour illustrer une nouvelle fois l'intérêt de cette démarche, nous proposons une petite étude sur le pronom "*lequel*" et ses variantes.

Cette étude a été menée à l'aide du concordancier "Contextes" élaboré par J. Véronis et à partir de différents types de corpus :

- 2 corpus écrits d'un million de mots chacun (*Presse* et *Littérature*),
- 3 corpus oraux :
    - 1 corpus de type "archive" d'un million de mots (*Corpaix*)
    - 1 corpus édité par P. Cappeau "Hommes Politiques" (*HP*) se caractérisant par une situation de parole (parole publique) et une unité thématique, 600.000 mots.

---

mais aussi selon des thématiques, telles que : parole professionnelle, récits de vie, débat politique, etc.)

3 Par exemple, l'opposition entre écrit et oral dans les travaux de Biber est critiquée par certains spécialistes de traduction car elle ne tiendrait pas suffisamment compte des "genres" (Canon-Roger et Chollier, 2008).

4 Si ce n'est le *Corpus de Référence du Français Parlé*, mais celui ne concerne que l'oral et ne répond pas entièrement aux caractéristiques de ce que devrait être un véritable corpus de référence.

- Le Corpus de Référence du Français Parlé (*CRFP*) de 440.000 mots se caractérisant, entre autres, par 3 situations de parole (privée, professionnelle, publique)

Ne seront étudiées ici que les formes "*lequel*" et ses variantes (*auquel, duquel, auxquels, desquels, laquelle, lesquelles, auxquelles, desquelles*) dans des emplois de relatifs. Les cas où ces formes sont employées comme "adjectif relatif" :

(5) *laquelle violence reste toujours une responsabilité* (oral, *HP*),

(6) *lequel enfouissement est désastreux* (oral, parole Publique, *CRFP*)

(7) *auquel cas c'est un mauvais coup porté au gouvernement* (oral, *HP*),

ou comme pronoms interrogatifs, rares dans les corpus étudiés, si ce n'est dans les interrogations indirectes :

(8) *il sait pas laquelle choisir* (oral *Corpaix*)

ne seront pas pris en compte.

Il ne s'agira pas non plus ici de revenir sur les problèmes généraux posés par les pronoms relatifs qui ont donné lieu à de nombreuses études[5], ni même de présenter une monographie sur cette forme de pronom en particulier, mais de faire en sorte de mieux connaître la façon dont elle apparaît dans les productions.

Si on compare les deux ensembles de corpus "oral-écrit", on note que les formes en "*lequel*" sont plus nombreuses à l'écrit qu'à l'oral, près de 20% en plus. (Écrit : 1163 occurrences, Oral : 760 occurrences). De même, la forme sujet est plus souvent usitée à l'écrit. Cela correspond à 8,5% des emplois contre 3,6% à l'oral.

"*Lequel*" sujet	98 (1163 occurrences à l'écrit)	28 (760 occurrences à l'oral)

Ces résultats semblent conformes aux intuitions que l'on peut avoir sur l'emploi de cette forme et à ce que de nombreux auteurs ont régulièrement signalé, par exemple :

> "***Lequel*** *reste une forme savante ; il est à peu près absent de la langue poétique et familière et les grammairiens classiques, eux-mêmes, vont le proscrire en dehors des cas obliques (duquel, auquel). On ne s'étonnera pas, dans ces conditions, que cette forme, pourtant si utile, ait été complètement ignorée par la langue populaire*." (Guiraud, 1966, 47)

Cependant, au regard des résultats, ce dernier propos n'est pas tout à fait exact : les emplois de "*lequel*" pronom relatif sont, somme toute, plus fréquents dans les productions orales que ce que l'on pouvait supposer. De même, s'il est d'usage de signaler les erreurs d'accord en genre et/ou en nombre comme étant un phénomène bien représenté :

(9) *il y a beaucoup de domaines* ***dans lequel*** *le GATT…* (oral, HP)

---

5 Cf., entre autres, les travaux de Guiraud (1966), Gadet (1992), Muller (2003).

l'étude des corpus oraux révèle au contraire un pourcentage d'erreur plus faible - moins de 2% - que celui auquel on pouvait s'attendre. L'intérêt suscité par ce type d'erreur est cependant suffisamment grand pour qu'un site Web[6] lui soit dédié. Le site en question est conçu pour recevoir les erreurs entendues dans les médias (à ce jour près de 150 exemples) ou celles relevées avec le moteur de recherche Google. Certains taux de fréquence d'erreur sont également signalés, par exemple :

- *la raison pour lequel* (fréquence d'erreur 1/1000)
- *une chose sur lequel* (fréquence d'erreur 2%)
- *une situation dans lequel* (fréquence d'erreur 1/4000)

En revanche, rares sont les informations concernant les différentes prépositions qui peuvent introduire cette forme "pronom" et leurs répartitions selon le type de productions. Ces informations sont cependant intéressantes car elles révèlent que ce sont **4** prépositions, aussi bien à l'oral qu'à l'écrit, qui se partagent la très grande majorité des emplois du pronom "*lequel*", mais ces dernières ne sont pas forcément les mêmes ou ne présentent pas les mêmes rangs de fréquence :

- Pour l'écrit, 80% des occurrences du pronom "*lequel*" sont introduites par les prépositions "***à***"[7] (33%), "***de***"[8] (16,5%), "***dans***" (16,5%) et "***sur***" (13,5%).
- Pour l'oral, 90% des occurrences sont introduites par les prépositions "***dans***" (30%), "***à***"[9] (24%), "***sur***" (19%) et "***pour***" (16%).

Les deux modes de production se distingueraient donc, entre autres, sur l'usage quasi inexistant de la préposition "*de*" à l'oral et de la préposition "*pour*" à l'écrit. Cependant, ces résultats méritent d'être affinés, car si on tient compte des différents types de productions : *Presse*, *Littérature*, oral "tout venant" (*Corpaix*), oral parole publique (*HP*), cf. les résultats du tableau suivant :

	ÉCRIT	**ÉCRIT**	**ORAL**	**ORAL**
	LITT(544)	**PRESSE(521)**	**CORPAIX(226)**	**HP(367)**
à	**182 (33%)**	**172 (33%)**	44 (19%)	**100 (28%)**
avec	46 (8%)	42 (8%)	15	8
dans	76 (14%)	**99 (19%)**	**61 (27%)**	**107 (29%)**

6 http://philippe.gambette.free.fr/danlekel/

7 La répartition entre les différentes formes est la suivante : "*auquel*" (30,50%), "*auxquels*" (20,62%), "*à laquelle*" (30,79%), "*auxquelles*" (18%).

8 "*duquel*" (43,42%), "*desquels*" (12,57%), "*de laquelle*" (32,57%), "*desquelles*" (11,42%). A noter que plus des 3/4 des occurrences de "*duquel*" se retrouvent dans le seul corpus "*Littérature*".

9 La répartition entre les différentes formes est : "*auquel*" (43,75%) (plus des 3/4 des occurrences se relèvent dans le seul corpus *HP*), "*auxquels*" (11,80%), "*à laquelle*" (26,38%), "*auxquelles*" (18,05%)

de	**115 (21%)**	60 (11,5%)	9	22
pour	17	39 (7%)	41 (18%)	56 (15%)
selon	0	38 (7%)	1	8
sur	94 (17%)	51(10%)	**51 (23%)**	66 (18%)

Les prépositions les plus fréquentes introduisant "*lequel*"[10]

On note que c'est seulement dans le corpus *Littéraire* que les prépositions "***à***" et "***de***" introduisent 54% des occurrences. Viennent ensuite 3 autres prépositions "***sur***", "***dans***", et "***avec***". La préposition "***pour***" continue à être rare et la préposition "***selon***" totalement absente.

En revanche, dans le corpus *Presse,* les deux prépositions "***à***" et "***dans***" introduisent 52% des occurrences. Viennent ensuite 5 autres prépositions "***de***", "***sur***", "***avec***", "***pour***" et "***selon***", dont les taux de fréquence sont quasi identiques. Ce corpus présente donc une variété de formes plus ample que le précédent.

Une étude plus poussée de la préposition "***de***" dans ces deux types de corpus écrits révèle par ailleurs une autre différence : "***de***" est plutôt introduit par une locution "temporelle" (*au cours de*) dans le corpus *Presse*, cela correspond à 37% des emplois, alors que dans le corpus *Littéraire*, cette préposition est plutôt locative (*au milieu de, autour de, au fond de*) ; cela correspond à près de 34% des emplois.

En ce qui concerne les corpus oraux, on note que proportionnellement, c'est dans le corpus *HP* que l'on relève un plus grand nombre de formes "*lequel*" (3 fois plus que dans *Corpaix*), et que ce sont essentiellement les taux de fréquences des prépositions, et non leur forme, qui les différencient. Ainsi, dans le corpus *HP*, ce sont les prépositions "***dans***" et "***à***" qui introduisent 57% des occurrences, viennent ensuite "***sur***" et "***pour***", alors que dans *Corpaix*, ce sont les prépositions "***dans***", et "***sur***" qui introduisent 50% des occurrences, viennent ensuite "***à***", et "***pour***" avec des taux de fréquence quasi identiques.

Le corpus oral *HP* présente de ce fait certaines similitudes avec le corpus écrit *Presse*, concernant notamment les rangs de fréquence de "***dans***" et "***à***", mais il s'en distingue aussi par la rareté des prépositions "***de***" et "***selon***" - cette dernière paraissant être spécifique de l'écrit journalistique.

Ces deux types de corpus *Presse* et corpus oraux partagent une autre particularité, celui de développer un certain nombre de **collocations**, phénomène que l'on ne relève pas dans le corpus *Littéraire*. On note, par exemple, que dans le corpus Presse, le terme "**condition(s)**" introduit 50% des occurrences de la préposition "*dans*" (*condition(s) dans la(les)quelles*) et que le terme "***raison(s)***" introduit lui aussi 50% des occurrences de la préposition "*pour*" (*raison(s) pour la(les)quelles*). Ce phénomène se retrouve accentué

10 Les prépositions telles que *parmi*, *entre* et *sans* ne se retrouvent que dans les deux corpus de productions écrites et à des taux de fréquence négligeables.

dans le corpus oral des *HP*, puisque ce lexème introduit cette fois près de 80% des séquences.

Pour conclure cette étude, si on tient compte des données fournies par la totalité des corpus examinés, on relève que le pronom relatif en fonction *sujet* correspond à 6,55% de l'ensemble des occurrences, le reste est introduit par une préposition à des degrés divers, par ordre décroissant :

- "***à***" (30%), "***dans***" (20,68%), "***sur***" (15,80%), "***de***" (12,42%), "***pour***" (9%), "***avec***" (6,69%)

L'ensemble de ces résultats mériterait d'être vérifié sur un plus grand nombre de données, mais il semble que l'on peut déjà les interpréter comme étant révélateurs d'une certaine distribution qui s'opère dans l'usage des locuteurs en fonction de la situation (oral/écrit) et du genre dans lesquels ils évoluent.

La prise en compte d'informations distributionnelles et quantifiées permet de renouveler la présentation que l'on peut donner des phénomènes linguistiques étudiés, et notamment d'éviter de mettre en avant des séquences peu fréquentes, comme cela se produit souvent dans les manuels de grammaire ou dans les dictionnaires.

Tous ces arguments, qui militent en faveur de ce type d'approche, montrent cependant la nécessité de disposer de corpus qui allient une taille "critique" (c'est-à-dire qui permet de se livrer à des relevés significatifs) et une diversité suffisante.

Cette question de variété est d'ailleurs essentielle, puisque, comme nous venons de le voir, c'est ce paramètre-là qui permet de dépasser la frontière entre l'oral et l'écrit, du moins qui oblige à redéfinir l'opposition entre ces deux termes.

4. CONCLUSION

Ce type de description et d'approche, concernant les faits de langue, se révèle utile aussi bien d'un point de vue pratique que théorique.

En ce qui concerne le domaine pratique, et notamment l'enseignement de la langue, on note que, si l'existence de cette variabilité du côté des productions se vérifie facilement dans les corpus, ces phénomènes sont, cependant, rarement mentionnés dans les descriptions proposées. Comme le signalent Debaisieux & Boulton (2008), les outils pédagogiques sont lents à tirer profit des recherches effectuées et des résultats obtenus. Ce constat s'explique sans doute pour plusieurs raisons, entre autres :

- les études sur corpus, et notamment sur corpus diversifiés et échantillonnés, semblent être moins avancées pour le français que dans d'autres langues ;

- l'idée (fortement ancrée) de la norme et du bon usage a brouillé la relation entre français écrit et français parlé et des associations étroites (*écrit* avec *littéraire* et *oral* avec *familier*) ont été installées.

- le fait que la grammaire du français que nous connaissons est avant tout celle du français écrit, et même d'un français écrit particulier (le français

littéraire), et non la grammaire de la langue française prise dans son ensemble.

Cette distorsion se retrouve par exemple dans les dictionnaires (cf. le *TLFI*) où l'entrée *lequel* accorde une place largement dominante à la fonction *sujet* et où les exemples prépositionnels font apparaître des formes peu fréquentes "*de*", "*pour*", "*par*", "*sous*", "*parmi*", sans rapport avec leur poids dans la distribution effective. Il en est de même dans les manuels de grammaire LM ou FLE (cf. "*La grammaire des premiers temps*, vol.2, PUG") dans lesquels on peut, certes, trouver une liste de prépositions, mais sans que ces dernières soient hiérarchisées et sans que les phénomènes de collocations soient mentionnés.

Certes, il ne s'agit pas ici de rouvrir le débat sur ce que devrait être la norme : l'usage le plus fréquent ou l'usage le plus adéquat en fonction du type de production, mais de fait, la prégnance de certaines représentations de la langue est encore forte. La prise en compte de la variation du point de vue des productions semble avoir encore beaucoup de mal à s'imposer et à transparaître dans les manuels destinés au public.

On rêverait pour le français d'une grammaire aussi aboutie que celle de Biber et alii (1999) pour l'anglais, mais pour l'instant, un tel projet ne semble pas prêt à voir le jour, il reste néanmoins la possibilité de montrer à travers de petits exemples comment notre connaissance de certains points de langue pourrait être modifiée par rapport aux descriptions habituelles et l'intérêt que cela aurait pour l'apprentissage de la langue française, maternelle ou FLE/FLS.

En ce qui concerne le domaine théorique, la description fondée sur corpus met en jeu la conception même de la grammaire, il n'y aurait pas d'un côté un "système abstrait" et de l'autre des "réalisations individuelles", mais bien comme le dit Halliday (1991), un système linguistique qui serait composé du cumul de l'ensemble des usages. C'est ce que suggère aussi Rastier :

> *"Plutôt que d'une linguistique du texte (autrement dit, d'une linguistique de la parole), nous avons besoin d'une linguistique tout court, qui fasse droit à tous les paliers de complexité de son objet, du mot à la phrase et au texte, puis du texte au genre, au discours, au corpus"* (Rastier, 2001, 9)

Il va de soi que le développement d'études à partir de corpus reposant sur des paramètres à la fois quantitatifs et qualitatifs devrait permettre de dégager une typologie des genres et donc d'accéder à cette linguistique "tout court".

BIBLIOGRAPHIE

BIBER D., JOHANSSON S., LEECH G., CONRAD S. & FINEGAN E. (1999), *Longman grammar of spoken and written English*, London, Pearson.

BIBER D. (2006), *University Language. A corpus-based study of spoken and written registers*, Amsterdam, John Benjamins Publishing Company.

BILGER M. (2004), "Quelques données sur les adverbes en *–ment"*, *Recherches sur le français parlé*, 18, publication du GARS, Université de Provence, 63-80.

BILGER M. (2007), "Réflexions sur un obscur objet de désir ; le corpus", Les Cahiers de *l'Association for French Language Studies*, 13-1, (http://www.afls.net/cahier

BILGER M. & CAPPEAU P. (2007-2009), "De la constitution des corpus oraux à l'analyse : exemples en syntaxe", in Bruxelles S. & alii (éds), *Grands corpus de français parlé, bilan historique et perspectives de recherche*, Cahiers de linguistique, 33/2, 163-182.

BILGER M. & CAPPEAU P. (à paraître), "Les données des corpus ou comment dépasser certaines représentations", in Abecassis M. & alii (éds), actes du colloque *Les voix du français : usages et représentations*, AFLS, 3-5 septembre 2008, Oxford.

BLANCHE-BENVENISTE C. (1993), "Une description linguistique du français parlé", *Le Gré des langues*, 12, 8-28.

BRANCA-ROSOFF S. (1977), "Quel *lequel* ? À propos des formes en *lequel/laquelle* en français de Montréal", *Recherches Sur le Français Parlé*, publication de l'Université de Provence, 170-185.

CANON-ROGER F. & CHOLLIER C. (2008), *Des genres aux textes. Essais de sémantique interprétative en littérature de langue anglaise*, Arras, Artois Presses Université.

CAPPEAU P. & GADET F. (2007), "Maître-mot et pierre philosophale : l'exploitation sociolinguistique des grands corpus", *RFLA*, XII-1, 99-110.

CORI M. & DAVID S. (2008), "Les corpus fondent-ils une nouvelle linguistique ?". *Langages*, 171, 111-129.

CRFP : Corpus de Référence de Français Parlé (ESA6060-DELIC)(1998-2004) (http://www.up.univ-mrs.fr/delic/crfp), *Recherches Sur le Français Parlé*, 18, Presses Universitaires d'Aix-en-Provence.

DEBAISIEUX J.-M. & BOULTON A. (2007), "Alors la question c'est...? Questions pragmatiques et annotation pédagogique des corpus", *Cahiers 13.2.*, AFLS.

GADET F. (1992), *Le français populaire*, Paris, Puf, collection Que Sais-je ?

GADET F. (2007), *La variation sociale en français*, Paris/Gap, Ophrys, nouvelle édition.

GADET F. & TYNE H. (éds) (2004), "Le style comme perspective sur la dynamique des langues", *Langage et société*, 109.

GUIRAUD P. (1966), "Le système des relatifs en français populaire", *Langages*, 1-3, 40-48

HABERT B. (2000), "Des corpus représentatifs : de quoi, pour quoi, comment ?", in Bilger M. (éd.), "Linguistique sur corpus – Études et réflexions", *Cahiers*, 31, P.U. de Perpignan, 11-58.

HALLIDAY M.A.K. (1985), *Spoken and written Language*, Oxford, Oxford University Press.

HALLIDAY M.A.K. (1991), "Corpus studies and probalistik grammar", in Aijmer K. & Altenberg B. (eds), *English Corpus Linguistics*, Londres/Ney-York, Longman, 30-43.

HUNSTON S. (2002), *Corpora in Applied Linguistics*, Cambridge, Cambridge University Press.

KERBRAT-ORECCHIONI C. & TRAVERSO V. (2004), “Types d’interactions et genres de l’oral”, *Langages*, 153, 41-51.

Langage et Société (1999), 87, Types, modes et genres.

MCENERY T., XIA R. & TONO Y. (2006), *Corpus-Based Language Studies. An advanced resource book*, New York, Routledge Applied Linguistics.

MULLER C. (2003), “Réflexions sur les relatives”, *Cahiers de Grammaire*, 30, 319-357.

RASTIER F. (2001), *Art et sciences des textes*, Paris, PUF.

RASTIER F. (2002), *L’accès aux banques textuelles - des genres à la doxa*. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Acces. html>.

RASTIER F. (2005), “Enjeux épistémologiques de la linguistique de corpus”, in Williams G. (éd.), *La linguistique de corpus*, Rennes, PUR, 31-45.

RASTIER, F. & MALRIEU D. (2001), “Genres et variations morphosyntaxiques”, *Traitement Automatique des Langues*, 42-2, 548-577.