French Language Studies **18** (2008), 1–2, © Cambridge University Press
doi:10.1017/S0959269507003134

Préface: Le français à la lumière des corpus

JACQUES DURAND

Université de Toulouse-Le Mirail and CNRS CLLE-ERSS

Il y a encore quelques années, un numéro du *Journal of French Language Studies* sur le thème de ce numéro spécial aurait exigé une introduction détaillée justifiant une approche fondée sur les corpus et la comparant systématiquement à une approche jugée plus traditionnelle. De fait, la situation a fortement évolué sur le plan national et international. On ne compte plus les ouvrages, les articles, les revues, les numéros spéciaux de revues consacrés aux corpus en linguistique. Ce numéro spécial ne saurait donc prétendre à une singularité qui le détacherait comme figure sur un fond sans valeur. Il n'en offre pas moins une perspective très intéressante. Tout d'abord, même si la plupart des linguistes s'accordent pour reconnaître une valeur heuristique aux corpus et pour saluer l'importance de données qualitatives et quantitatives plus fiables, tous ne sont pas d'accord sur le type de linguistique à construire. L'article de Bernard Laks, 'Pour une phonologie de corpus', se refuse à considérer que les linguistiques de corpus ne sont rien d'autre que de simples dispositifs techniques au service d'une linguistique descriptive, empirique ou herméneutique. En offrant une perspective sur la longue durée, il réanalyse l'opposition classique entre sciences de l'*exemplum* et sciences du *datum*, et cherche à démontrer que la linguistique, et singulièrement la phonologie, se sont construites, contre la grammaire, comme des sciences empiriques ayant pour objet la modélisation des observables linguistiques. Comprendre l'ancienneté de la notion de corpus et savoir construire son historicité permettent à ses yeux de frayer le chemin d'une linguistique de la parole, condition *sine qua non*, comme Saussure l'a souvent dit d'une linguistique de la langue.

L'intérêt des autres articles de ce numéro est qu'ils offrent un panorama représentatif (mais, bien sûr, non exhaustif) des divers domaines où les approches de corpus ont permis des avancées significatives depuis une vingtaine d'années. Dans 'French liaison in the light of corpus data', Jacques Durand et Chantal Lyche présentent une analyse de la liaison à partir de résultats obtenus au sein du projet fédératif PFC (*Phonologie du français contemporain: usages, variétés et structure*). Ils démontrent que les données extraites de codages portant sur cent à deux cent locuteurs, selon les phénomènes, modifient ou précisent les analyses traditionnelles. Ils passent en revue les contextes où les liaisons sont catégoriques ou variables, enchaînées ou non, et en examinent les conséquences pour les systèmes intériorisés par les locuteurs. On ne tire pas les mêmes conclusions selon que l'on a fait des observations à la volée ou examiné plus de 28000 codages avec des outils de fouille systématique. Dans 'Extensive data for morphology: using the World Wide

Web', Nabil Hathout, Fabio Montermini et Ludovic Tanguy présentent des études récentes en morphologie fondées sur des données extensives. Tout en tenant compte des limites méthodologiques de l'utilisation de la toile, ils démontrent que le recours à de grandes masses de données permet d'aller bien au-delà des approches traditionnelles s'appuyant sur des listes extraites de dictionnaires. La thèse des auteurs est développée et argumentée à travers une étude des formes suffixales: *-esque*, *-este*, *-able*, *-ment*. Dans le domaine de la syntaxe, Cécile Fabre et Didier Bourigault abordent le problème classique du rattachement des syntagmes prépositionnels au sein des structures prédicatives. La thèse avancée est que l'exploitation de corpus annotés catégoriellement et syntaxiquement, et la mise au point de méthodes de quantification, confirme l'absence d'une distinction tranchée entre arguments et circonstants. En effet, la méthode met au jour des positions médianes qui mettent en évidence des configurations récurrentes propres au corpus, au comportement intermédiaire entre arguments et circonstants prototypiques.

Les deux autres articles qui complètent ce volume vont au-delà de l'analyse grammaticale traditionnelle. Dans 'Annotating an oral corpus using the Text Encoding Initiative. Methodology, problems, solutions', Janice Carruthers se penche sur ces 'écrivains d'oralité' que sont les nouveaux conteurs et explore les conventions de transcription et de codage que propose la TEI (Text Encoding Initiative). C'est donc un exercice de méthode qui nous est proposé. Enfin, dans 'Tool-assisted analysis of interactional corpora', le groupe ICOR (M. Bert, S. Bruxelles, C. Etienne, L. Mondada, V. Traverso) présente la base de données CLAPI. Ce Corpus de LAngue Parlée en Interaction, qui a été développé dans le laboratoire ICAR (Université de Lyon et CNRS), a l'originalité d'être multimodal et de s'attaquer à l'analyse des énoncés en contexte d'interaction. La démonstration porte ici sur le segment 'voilà' qui est étudié à partir de contextes précis abondamment décrits par les auteurs.

Le lecteur, s'il prend la peine d'examiner ce numéro de bout en bout, pourra juger sur pièce des avancées réalisées. En tant que coordinateur de ce numéro, je tiens à remercier les collègues suivants qui ont aidé à expertiser les articles à une vitesse record: Luc Baronian, Mireille Bilger, Philippe Blache, Georgette Dal, Sylvain Detey, Julien Eychenne, Wyn Johnson, Yuji Kawaguchi, Stephanie Kelly, Laurence Labrune, Jacques Lamarche, Chantal Lyche, Christiane Marchello-Nizia, Salah Mejri, Jesse Tseng, Anne Violin-Wigent, Douglas Walker. Sans leurs efforts, les revues qui s'efforcent de jouer le jeu de l'évaluation anonyme ne pourraient survivre.

https://doi.org/10.1017/S0959269507003134 Published online by Cambridge University Press