

CORPUS **Corpus**
5 | 2006
Corpus et stylistique

Benoît Habert — *Instruments et ressources électroniques pour le français*. Paris /Gap : Ophrys (« L'Essentiel Français »), 2005, 176 pages, 10 euros

Etienne Brunet



Édition électronique

URL : <http://journals.openedition.org/corpus/571>
ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 1 décembre 2006
ISSN : 1638-9808

Référence électronique

Etienne Brunet, « Benoît Habert — *Instruments et ressources électroniques pour le français*. Paris /Gap : Ophrys (« L'Essentiel Français »), 2005, 176 pages, 10 euros », *Corpus* [En ligne], 5 | 2006, mis en ligne le 30 août 2007, consulté le 21 avril 2019. URL : <http://journals.openedition.org/corpus/571>

Ce document a été généré automatiquement le 21 avril 2019.

© Tous droits réservés

Benoît Habert — *Instruments et ressources électroniques pour le français*. Paris /Gap : Ophrys (« L'Essentiel Français »), 2005, 176 pages, 10 euros

Etienne Brunet

- 1 *Instruments et ressources électroniques pour le français* est le dernier numéro d'une longue série dirigée par Catherine Fuchs, sous le titre *L'essentiel français*, aux éditions Ophrys. C'est aussi le dernier d'une longue série d'ouvrages publiés par Benoît Habert.
- 2 Bilan des ressources électroniques disponibles pour l'analyse du français, ce livre est lui-même une ressource : la plus complète, la plus objective, la plus précise, la plus claire. C'est aussi la plus pratique puisque les multiples adresses qu'elle contient sont disponibles sur le site de l'auteur¹. Il est donc inutile de recopier les dix pages de liens rassemblés à la fin de l'ouvrage, puisqu'on les retrouve sur internet, disposés dans le même ordre et sans risque d'erreur. Habert fournit 150 liens, dûment choisis et classés. Google en livrerait cent fois plus, en s'appuyant seulement sur les mots du titre, mais dans un fouillis inextricable où l'accessoire et le futile étoufferaient l'essentiel et l'utile. Il serait pourtant imprudent de s'engager trop vite la carte en mains sur les itinéraires proposés. Les voyages sur internet, même avec un bon guide, cela se prépare. Et la lecture de l'ouvrage est un préalable indispensable à l'exploration linguistique assistée par ordinateur.
- 3 La contrainte de l'espace (la collection limite la taille du volume en deçà de 200 pages) aurait pu conduire à un simple catalogue de produits ou à une compilation d'adresses. Or Habert livre une véritable synthèse, un état des lieux, dense et sobre.
- 4 Cela ne va pas sans sacrifice : dans un domaine où le progrès est rapide, et où les instruments sont souvent dépassés avant d'avoir servi, Habert juge superflu d'évoquer le passé et son panorama historique tient en deux pages. Habert s'abstient aussi de

s'engager dans des considérations théoriques sur le bien-fondé de la recherche linguistique conduite « aux instruments ». Son ouvrage précédent *Les linguistiques de corpus* (Armand Colin, 1997) fait le point sur cette question. Habert évite aussi de s'écarter du lieu géométrique qu'il a choisi et où l'informatique croise la linguistique. Sur ce carrefour débouchent aussi d'autres voies qu'il aurait pu emprunter et qui mènent d'un côté à la littérature, de l'autre aux mathématiques. Paradoxalement Habert utilise un prétexte littéraire, le sonnet de Rimbaud *Le dormeur du val* pour exemplifier les méthodes utilisées, tout en étant conscient qu'un objet aussi particulier et aussi étroit ne peut guère aboutir à un résultat exploitable (pareillement les méthodes instrumentées appliquées au *Bateau ivre* lors d'un colloque à l'université de Bologne en 2004 ont souffert de la comparaison avec l'intuition et la mémoire humaines, plus à l'aise dans les petits espaces). Quant aux statistiques, elles ne sont qu'évoquées brièvement car ce sont des outils d'exploitation qui supposent la maîtrise préalable des ressources et conduisent à l'obtention et à l'interprétation des résultats.

- 5 Or ce n'est pas la phase finale de la recherche qui fait l'objet de cet ouvrage mais la phase initiale, celle où se trouve le chercheur qui se préoccupe d'assembler les matériaux de la recherche, les textes à réunir, le codage à établir, les normes à respecter, les bases ou dictionnaires à consulter, les ressources à télécharger, les logiciels à utiliser. On admire la précision pédagogique avec laquelle l'auteur explique la façon dont certains logiciels comme Cordial ou TreeTagger s'acquittent, plus ou moins bien, des opérations d'étiquetage et de lemmatisation. On le voit démonter avec virtuosité la structure du codage de présentation html, celle du codage logique xml, et même les ponts qui vont de l'un à l'autre. Que ce soit dans l'écrit, l'oral ou le lexique (trois chapitres sont consacrés à ces trois niveaux), Habert pèse l'intérêt de chaque traitement et la valeur ajoutée de chaque annotation, de chaque analyse, de chaque enrichissement.
- 6 Nul souci d'établir un palmarès des produits présentés: ceux qu'il choisit ont surtout vertu d'exemple. Les noms de personne sont rares. Les adjectifs absents. L'étude est précise, documentée et puissamment ancrée dans l'actualité, mais elle reste neutre et impersonnelle. Elle est pourtant animée d'une foi conquérante. Habert croit dans les vertus de la normalisation, du travail collectif, de la transparence. Il applaudit aux efforts de la TEI, du XML, et de l'open source. Il s'emploie à faire le lien entre les initiatives dispersées qui s'ignorent. Il mobilise les forces pour affronter le grand enjeu : celui de la sémantique. Professeur de sémantique à Nanterre, il a conscience en effet que les traitements numériques butent encore devant le sens des mots et que les ressources électroniques s'épuisent devant l'obstacle. Il sait aussi qu'il faut accepter encore d'avoir les mains sales et qu'on est encore forcé de les plonger parfois dans le code, pour accorder les sorties d'un programme aux entrées de l'autre, pour mettre en œuvre les scripts, pour combler les défaillances des métadonnées, combattre l'égoïsme des formats propriétaires et surtout veiller à ne pas dégrader la chaîne de l'information, denrée périssable et sensible aux ruptures, comme les aliments soumis à la chaîne du froid.

NOTES

- 1.. <http://www.limsi.fr/Individu/habert/Publications/publis-bh/index.html>