Habert, Benoît, *Instruments et ressources électroniques pour le français*

Gap/Paris : Ophrys, 2005

Sarah Leroy

Édition électronique

URL : <http://journals.openedition.org/praxematique/673>

DOI : 10.4000/praxematique.673

ISSN : 2111-5044

Éditeur

Presses universitaires de la Méditerranée

Édition imprimée

Date de publication : 1 décembre 2006

Pagination : 187-202

ISSN : 0765-4944

Référence électronique

Sarah Leroy, « Habert, Benoît, *Instruments et ressources électroniques pour le français* », *Cahiers de praxématique* [En ligne], 46 | 2006, mis en ligne le 01 décembre 2009, consulté le 23 septembre 2020.
URL : <http://journals.openedition.org/praxematique/673> ; DOI : <https://doi.org/10.4000/praxematique.673>

Benoît HABERT

INSTRUMENTS ET RESSOURCES ÉLECTRONIQUES POUR LE FRANÇAIS
Gap/Paris : Ophrys, 2005, 170 p.

Sept ans après *De l'écrit au numérique*¹, premier ouvrage français à présenter aux linguistes des « boîtes à outils » pour constituer et manier des corpus électroniques, BH récidive avec *Instruments et ressources électroniques pour le français*, un manuel bref mais dense qui marque une nouvelle étape dans ce domaine de l'instrumentation en linguistique². Les deux ouvrages sont sensiblement différents, tant par leur niveau de technicité que par le public qu'ils visent, mais on peut cependant constater avec le plus récent une réelle évolution des pratiques.

On y mesure ainsi les changements survenus dans ce domaine en progression rapide : des standards émergents en 1998 sont désormais la norme et sont présentés comme tels ; c'est le cas par exemple de la TEI pour la structuration textuelle ou d'Unicode pour l'encodage des caractères. Mais on y perçoit surtout certaines mutations au sein de la communauté des linguistes, où la pratique de l'instrumentation se répand et se banalise, à mesure que les outils et les ressources se stabilisent et se simplifient. Un linguiste lambda peut désormais, avec des compétences n'excédant guère celles de la bureautique, utiliser des instruments et des ressources électroniques pour une recherche dans son domaine, quel qu'il soit. De ce fait, la question de l'utilisation de l'informatique en linguistique dépasse désormais la communauté transdisciplinaire du Traitement Automatique des Langues pour concerner toute approche empirique de tel ou tel aspect d'une langue. Cet ouvrage paru dans une collection à vocation didactique et pratique, et limité au domaine français, propose donc de vraies clés au linguiste non informaticien et laisse à l'arrière-plan les discussions et les discordes entre ce qu'on peut appeler les « linguistiques de corpus »³ (courant dans lequel il s'inscrit néanmoins) et les approches formalistes, également très représentées en TAL.

Les instruments et ressources disponibles pour le français étant de natures, d'origines, de maturation, d'accès divers, BH a délaissé une présentation par type de tâche ou de donnée au profit d'une organisation sui-

1. Habert B., Fabre C., Isaac F., 1998, *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Paris : Masson.

2. Domaine que l'auteur a largement contribué à développer, comme en témoignent plusieurs publications, dont en particulier Habert B., 2005, « Portrait de linguiste(s) à l'instrument », *Texte!* 10-4 [www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html].

3. Cf. Habert B., Nazarenko A., Salem A., 1997, *Les Linguistiques de corpus*, Paris : Armand Colin.

vant les niveaux et les problématiques linguistiques. Quatre chapitres y sont consacrés, encadrés par deux fois deux chapitres plus théoriques.

Après une rapide introduction fixant quelques repères terminologiques, techniques et historiques, les chapitres I et II introduisent la notion fondamentale d'annotation des données, à l'aide d'un exemple récurrent¹ qui permet de montrer concrètement ce qu'est une annotation. On a ainsi un aperçu d'annotations basiques (étiquetage et lemmatisation), puis plus spécifiques (comme l'annotation de relations syntaxiques ou d'analyse métrique), généralement réalisées à la suite de la première étape. Ce panorama entraîne des remarques sur les limites de l'annotation et les choix que le chercheur est amené à faire.

Les chapitres III à VI constituent le cœur de l'ouvrage, présentant conjointement les ressources (corpus, dictionnaires) et les instruments permettant de les exploiter. On parcourt ainsi trois niveaux d'analyse (joliment organisés en « textes », « paroles » et « mots ») avant de faire une incursion du côté de la variation. Le chapitre consacré aux textes fait évidemment la part belle à Frantext, tant comme ressource que du point de vue des instruments proposés, mais mentionne également d'autres bases textuelles moins connues et évoque (sans citer le moteur Glossanet, couramment utilisé) l'« aspiration » de données sur le Web. Ces ressources peuvent s'enrichir d'annotations au niveau de la phrase (arborescence syntaxique, liens de corréférence) ou au niveau du texte (segmentations thématique ou structurelle, résumé automatique). Le chapitre IV est consacré aux corpus oraux, beaucoup plus difficiles à obtenir que leurs correspondants écrits, pour des raisons diverses (préjugés envers l'oral, obstacles juridiques, complexité et coût du recueil de données). De ce fait, si les principaux corpus sont mentionnés, ce chapitre est surtout consacré au problème majeur de la transcription de l'oral, des choix qu'elle implique (niveaux de transcription, finesse et précision) et des outils qui permettent de la réaliser. Les difficultés de l'étiquetage de l'oral, qui sont rapidement abordées, restent importantes, et les progrès passeront, semble-t-il, par l'exploitation d'états intermédiaires entre oral et écrit. Les mots qui sont traités au chapitre V sont ceux des versions électroniques des dictionnaires, dont l'informatisation renouvelle considérablement les structures et les modes de consultation, mais également ceux qui se trouvent dans les lexiques, bases lexicales donnant des informations syntaxiques, morphologiques ou de fréquence, qui peuvent être exploitées dans différentes perspectives (étiquetage de textes, repérage de termes...). Les mots sont aussi parfois ceux d'un corpus (voire leurs étiquettes) ; ils font alors l'objet d'analyses statistiques visant à repérer les spécificités de différents groupes textuels ou à faire appa-

1. Il s'agit du *Dormeur du val*, que l'on retrouve tout au long de l'ouvrage.

raître des phraséologies d'ordre idio- ou sociolectal. Les développements en direction de la sémantique lexicale, qu'il s'agisse de classes sémantiques ou de lexiques sémantiques organisés, semblent être la prochaine étape à atteindre. Enfin, le chapitre vi est consacré à des domaines ou des perspectives émergentes à l'instrumentation, soit que celle-ci rende possible des recherches jusqu'alors très difficiles à mener (c'est le cas de l'analyse multimodale de la communication), soit au contraire qu'elle gagne des domaines pour lesquels elle était auparavant inadaptée (c'est le cas des états anciens ou des variétés diatopiques du français).

Les deux derniers chapitres constituent un retour théorique et méthodologique qui reprend des questions évoquées dans les deux premiers, à la lumière des connaissances nouvelles du lecteur. Le chapitre vii est le plus technique, il aborde clairement les difficultés concrètes, mais aussi les astuces d'installation, de préparation et d'échange des données, et globalement d'environnement informatique (caractères, outils bureautiques), des instruments et ressources électroniques. Il débouche sur une réflexion, au chapitre viii, sur les méthodes d'instrumentation en linguistique et sur le devenir des ressources accumulées. L'instrumentation constitue un véritable enjeu scientifique, car elle a pour but de permettre la reproductibilité et la cumulativité des résultats ; elle doit donc être évaluable et réutilisable. Dans ce souci de partage et de progression des données, BH insiste sur les métadonnées et la nécessaire documentation des corpus, rappelant qu'un corpus, pour être utile, n'est pas un « amas textuel » mais se constitue en fonction d'objectifs et d'hypothèses précis, qui doivent être clairement indiqués.

La lecture ne s'achève pas avec ces développements ; outre les classiques index et glossaire¹, une bibliothèque de liens² reprenant toutes les références données dans l'ouvrage (selon l'ordre des chapitres successifs), est fournie en fin de volume, avant une double bibliographie.

Cet ouvrage est donc tout simplement précieux, car il rassemble une somme d'informations quasi-exhaustive (pour une période donnée), présente les objets dans le détail avec un rare souci pédagogique et réussit pourtant à ne pas être une compilation, un guide des données et des logiciels. On y trouve en effet à la fois des « bons tuyaux » et des réflexions, voire des positionnements, sur l'utilisation d'instruments en linguistique. La volonté

1. Classiques, mais néanmoins extrêmement bien faits : l'index est hiérarchisé et spécifie à quelle page se trouve la définition du terme, le glossaire se trouvant de ce fait réduit à l'essentiel.

2. Une version mise à jour et cliquable en est disponible sur le Web :
www.limsi.fr/Individu/habert/Publications/InstrumentsEtRessourcesElectroniquesPourLeFrancais.html.

de partage et de visibilité des pratiques, qui semblent encore parfois ésotériques à certains, apparaît notamment avec la constante illustration des propos, à l'aide de tableaux, de figures, dont il faut souligner l'importance et la pertinence : sorties d'analyseurs, copies d'écran, etc., tout est fait pour que le lecteur se rende compte clairement des résultats escomptables, y compris des défauts de certains outils et des scories qu'ils peuvent produire. En ce sens, l'ouvrage de BH non seulement se coule à merveille dans la collection L'Essentiel français, mais est destiné à constituer quelque temps (jusqu'au prochain ?) un des ouvrages de référence sur la question.

Sarah LEROY

UMR 7114 MoDyCo, Université Paris X