Comptes rendus

Langue française

Etienne Brunet: *Le Vocabulaire français de 1789 à nos jours* d'après les données du *Trésor de la langue française*. Préface de Paul Imbs. Genève, Slatkine, et Paris, Champion 1981. 3 vol., 852, 518 et 454 p.

De nombreux linguistes en France et à l'étranger ont déjà pu tirer profit de l'immense corpus établi à Nancy, en vue de la rédaction du dictionnaire *Trésor de la langue française* (TLF). On y a cherché des exemples et des données concernant un certain nombre de vocables ou de constructions. Dans le cas de l'ouvrage récent d'E. Brunet, *Le Vocabulaire français de 1789 à nos jours*, il ne s'agit plus d'exemples et de données partielles mais du corpus TLF entier, à savoir soixante-dix millions d'occurrences tirées d'un millier de textes publiés de 1789 à 1964. Par son ampleur déjà, l'ouvrage d'E. Brunet nous impressionne. Il comprend trois volumes dont le premier, illustré de près de trois cents figures et tableaux, fournit le plan d'ensemble, le deuxième énumère par ordre alphabétique les 6700 vocables les plus fréquents et le troisième trace en fonction du temps les courbes des 907 vocables majeurs. En fait, cet ouvrage se range, comme plusieurs autres ouvrages du même auteur, dans le cadre de la linguistique quantitative. Il est pourtant d'une tout autre dimension que la plupart des œuvres que nous avons vues jusqu'ici: par la taille des matériaux traités ainsi que par les problèmes posés.

Le corpus qu'E. Brunet a choisi d'analyser se compose de textes essentiellement littéraires, écrits par environ 350 auteurs différents. Sélectionné pour représenter le français écrit soutenu, ce corpus avait été enregistré sur bandes perforées au Centre de recherche pour un Trésor de la langue française et il a été soumis à une première exploitation quantative pour le *Dictionnaire des fréquences* (Klincksieck, Paris, 1969-1971).

En constituant son corpus définitif quinze ans plus tard, E. Brunet adopte les principes originairement établis. Ainsi exclut-il des comptes les noms propres et les mots étrangers; il garde les divisions effectuées à l'intérieur du corpus, c'est-à-dire divisions (1) en quinze tranches chronologiques, (2) en quatre genres littéraires (prose littéraire, prose technique, prose poétique et vers) et (3) en trois genres selon la personne dominante dans les parties de texte. Certes, cette prise de position a ses avantages mais aussi ses inconvénients: avantages, car elle facilite la comparaison des différents résultats obtenus sur le corpus, inconvénients, parce qu'il y a certains points faibles dans la partition du corpus, ce dont E. Brunet est lui-même bien conscient. C'est que les quinze tranches chronologiques ne sont pas

de la même longueur, ni en temps (27 ans pour la tranche la plus longue et 5 ans pour la plus courte) ni en nombre d'occurrences (variation de 4 millions à 6 millions). De plus, il existe une certaine imprécision dans le classement des textes en genres littéraires. En effet, dans le *Répertoire des textes littéraires et techniques des XIX^e-XX^e siècles, enregistrés sur ordinateurs à l'ILF* (1980), on a reconsidéré les textes afin d'obtenir un classement plus précis et plus fin. Cette révision étant postérieure à l'établissement de son corpus, E. Brunet n'a pas pu en profiter. Lorsqu'on se sert des données présentées dans l'ouvrage, on doit pourtant retenir ces faits pour ne pas tirer de conclusions hâtives. Cela vaut en particulier pour le tome 3, où les neuf cents et quelques courbes ne rendent pas compte du déséquilibre qu'il y a entre les tranches chronologiques.

Avant de soumettre le corpus TLF à de nouveaux traitements automatiques, E. Brunet a dû remplir certaines lacunes et rectifier certains chiffres. Comme plus d'une décennie s'est écoulée entre les premiers traitements des textes et le travail de Brunet, on ne s'étonnera pas de trouver un petit décalage entre les données des deux phases. Il y a également quelques divergences en ce qui concerne la partition des valeurs entre les différentes formes homographes, car n'ayant à sa disposition ni les textes originaux ni un fichier informatisé et désambiguïsé couvrant le corpus entier, l'auteur a été contraint d'estimer la fréquence des homographes à partir des données accessibles, datant uniquement du XX^e siècle.

E. Brunet confie le corpus corrigé et mis à jour à son collaborateur principal, l'ordinateur. C'est grâce à ce moyen inappréciable qu'il peut le traiter à plusieurs niveaux différents: au niveau des signes de ponctuation, des lettres, des suffixes et des préfixes ainsi qu'au niveau des vocables. Pour chaque élément, il fait enregistrer la fréquence dans les quinze tranches et les sept genres, ce qui donne 128 sous-fréquences, y compris les totaux. Ensuite, afin d'interpréter les chiffres bruts, l'auteur se sert de nombreux modèles statistiques. Grâce aux explications claires complétées de renvois aux manuels de la statistique ou de la linguistique quantitative, en particulier aux ouvrages de Charles Muller, même les lecteurs non-initiés pourront comprendre les calculs. Dans le premier chapitre du tome 1, E. Brunet développe les mérites de la loi hypergéométrique, lors de l'interprétation des faits linguistiques; il précise les avantages de cette loi, maintenant connue en lexicométrie grâce à l'Unité de recherche de lexicologie et textes politiques de Saint-Cloud, sur les modèles les plus souvent employés, tels que la loi normale et la loi binomiale. Cependant, dans le cas de son corpus très étendu, il préfère la loi normale à cause de son application comparativement simple et peu coûteuse. Pour comparer l'emploi des différents vocables, l'écart réduit sera donc l'instrument utilisé de préférence. On le retrouve aux tomes 2 et 3 de l'ouvrage: dans le deuxième tome, l'écart réduit de chacun des 6700 vocables majeurs est calculé pour chaque genre et chaque tranche. Et à partir de ces données sont établis les coefficients de Spearman et de Bravais-Pearson. Par conséquent, c'est à ce tome qu'il faut se référer pour connaître la valeur du vocable dans les différents genres et dans le temps ainsi que sa tendance chronologique. Dans le troisième tome, certaines valeurs de l'écart réduit sont reprises et présentées cette fois-ci en graphiques afin d'illustrer l'évolution temporelle.

Les tomes 2 et 3, exposant presque exclusivement des résultats sortis de l'ordinateur, constituent donc des volumes de référence sans texte. Par contre, le premier tome se prête à une lecture attentive. C'est là que l'on trouve décrits et discutés les principes et les méthodes qui ont guidé le travail. E. Brunet y rend aussi compte des résultats concernant l'ensemble du corpus ainsi que les différents sous-ensembles. Après avoir illustré

les distributions des lettres et des signes de ponctuation à travers les genres et le temps, il examine l'emploi fluctuant des vocables dans les catégories grammaticales. Un chapitre entier est consacré aux mots grammaticaux qui, comme nous l'avons déjà vu, par exemple, dans l'ouvrage de Mosteller, F. & Wallace, D., *Inference and Disputed Authorship: The Federalists* (Addison-Wesley, Reading, Mass., 1964), ne sont certainement pas sans importance et révèlent peut-être plus d'un style que les mots dits "pleins". Les deux derniers chapitres du tome 1 traitent d'abord les vocables caractéristiques des tranches chronologiques, puis des genres littéraires.

A l'exception des mots grammaticaux, E. Brunet étudie un vocabulaire lemmatisé, c'est-à-dire un vocabulaire dans lequel les formes d'un verbe sont réunies à l'infinitif, les formes d'un substantif au singulier, les formes de l'adjectif au masculin singulier, etc. Pour la lemmatisation, il a suivi les principes appliqués au *Dictionnaire des fréquences*, et il aurait difficilement pu faire autrement, car ses données ne lui permettent pas de remonter aux textes originaux. D'ailleurs, même si ceux-ci lui avaient été accessibles, on ne s'imagine guère comment il aurait pu lemmatiser une seconde fois cette masse d'occurrences. Avant d'entamer une tâche de lemmatisation de cette ampleur, on doit attendre le résultat des tentatives de l'analyse automatique des textes qui se font actuellement dans plusieurs groupes de recherche. Lorsque le logiciel sera mieux développé, on pourra certainement en tirer profit pour la lemmatisation des textes.

Dans son ensemble, *Le Vocabulaire français de 1789 à nos jours* est si riche et abondant en données quantitatives ainsi qu'en perspectives nouvelles qu'il nous ouvre de nombreuses voies à exploiter dans l'avenir. Les chercheurs puiseront avec profit des renseignements et des idées dans cet ouvrage pendant longtemps, tout en se réjouissant du style humoristique, plein de métaphores, propre à Etienne Brunet.

Gunnel Engwall
Stockholm

Jane-Odile Halmøy: *Le gérondif. Eléments pour une description syntaxique et sémantique.* Thèse de doctorat présentée devant l'Université de Trondheim le 8 mai 1982. Tapir. Université de Trondheim, 1982.

La thèse qu'a soutenue Jane-Odile Halmøy (J-O. H.) devant l'Université de Trondheim le 8 mai 1982 est une contribution fort importante à l'étude du gérondif français moderne. Son volume impressionnant (451 pages) ne doit aucunement faire peur à personne, car, malgré son érudition profonde, c'est un livre relativement facile à lire, bien pensé, bien pesé et bien présenté, et dont la qualité la plus remarquable est la synthèse de l'analyse scientifique et de la présentation pédagogique, synthèse, en effet, difficilement réalisable et rarement réalisée.

L'organisation formelle de la thèse est d'une clarté exemplaire. Une courte "Introduction" (p. 1-6) précède un chapitre détaillé et très instructif appelé: "Le gérondif vu par les grammairiens et les linguistiques. Exposé et commentaire" (p. 7-47). Suivent, après ce *Stand der Forschung*, les trois chapitres centraux de la thèse: "Description morphologique" (p. 48-68), "Description syntaxique" (p. 69-219) et "Description sémantique" (p. 220-386). La thèse