Compte-rendu de : Gunnel Engwall: "Vocabulaire du roman français" (1962-1968)" et "Dictionnaire des fréquences"

Étienne Brunet

Review Article

Gunnel Engwall: *Vocabulaire du roman français (1962–1968). Dictionnaire des fréquences,* (Data linguistica n°17, Almqvist & Wiksell International, Stockholm, 1984, LXVIII, 427pp. + 43 microfiches, chacun de 207 pp.)

Voici l'un des joyaux de la collection dirigée par Sture Allen. Les joyaux se cisèlent lentement. L'auteur - qui est orfèvre en la matière - aura mis plus de vingt ans à parfaire son chef-d'oeuvre alors même que les premiers résultats étaient livrés au public dès 1971 (G. Engwall, « Etude sur la fréquence des mots dans quelques romans français », *XIII^e Congrès international de linguistique et philologie romane,* Québec) et qu'une thèse, issue de la même entreprise, était soutenue en 1974 *(Fréquence et distribution du vocabulaire dans un choix de romans français,* Scriptor, Stockholm, 1974, 187 pp.). Il est vrai qu'entretemps Gunnel Engwall s'est penchée sur les manuscrits français d'August Strindberg (comme le *Plaidoyer d'un fou),* et qu'elle participe activement à la monumentale édition du grand écrivain suédois. Doit-on se réjouir ou s'affliger d'avoir un peu attendu? A l'évidence il faut s'en féliciter.

Car depuis dix ans l'informatique éditoriale a fait de gros progrès et il suffit de comparer les listes grossières parues dans la thèse de 1974 (en majuscules, avec des chiffres en guise d'accents) avec la présentation somptueuse que permet la photocomposition et qui n'est pas un luxe lorsqu'on doit produire des chiffres et des tableaux. Cela est vrai de l'ouvrage proprement dit, mais aussi des mirofiches qui accompagnent le texte et dont la réalisation est très soignée. Alors que l'informatique a suscité tant de publications hâtives, où de médiocres logiciels de traitement de texte associés à de pauvres imprimantes permettent à peine d'atteindre le niveau technique de la dactylographie (c'est ce qu'on appelle la qualité courrier!), voilà un travail de professionnel qui fait plaisir à l'oeil.

La clarté n'est pas seulement dans la typographie. Elle est aussi dans le choix des objectifs et la disposition des résultats. L'auteur ne ménage pas les explications sur le but poursuivi - l'étude de la langue romanesque contemporaine - et le corpus choisi. Rarement on a vu tant de scrupules pour fixer les critères de sélection et retenir les textes qui satisfaisaient à toutes les contraintes de genre, de lieu, de temps, de sujet, de style, de représentativité. Une véritable étude de bibliométrie a précédé l'ouverture du chantier, semblable à ces études de marché que font les entreprises avant de se lancer dans la production. En choisissant précautionneusement les best-sellers de la période 1962-

1968[1], l'auteur circonscrivait un domaine voisin de celui du *Trésor de la langue française* (qui ne dépasse pas l'année 1964) et se donnait les moyens de compléter - au point sensible de l'actualité littéraire - la vaste enquête de Nancy.

Quant aux résultats, leur abondance faisait problème. Mais Gunnel Engwall a su en tirer le meilleur parti, en économisant le papier de l'imprimeur et la fatigue de l'utilisateur, en cultivant la richesse et la clarté de l'information sans tomber dans la redondance. Elle est parvenue à ses fins en variant les filtres, utilisant tour à tour des tris hiérarchiques, catégoriels, alphabétiques, inverses. Et loin d'entasser le Pélion sur l'Ossa, elle suit un fil chronologique, passant progressivement des matériaux bruts aux produits finis:

- La première liste rend compte des formes graphiques telles que la machine, guidée par des instructions simples et solides relativement à la segmentation, les a distinguées dans le texte, puis classées selon la fréquence. Ici l'auteur parle de *types,* en empruntant le terme à Herdan (mais le mot a-t-il les mêmes connotations en anglais et en français?).
- La seconde est un dictionnaire inverse des *formes.* Des types aux formes la différence est énorme, car les formes sont désambiguïsées et pourvues d'un code grammatical. L'homographie est la terreur des lexicomètres. La plupart prennent la fuite, sans gloire, ou cherchent des faux-fuyants moins glorieux encore. Gunnel Engwall a osé affronter l'hydre aux mille têtes bourgeonnantes. Manuellement'. Méthodiquement. En examinant, dans leur contexte (dans la concordance), toutes les occurrences suspectes où le verbe et le substantif ont le même vêtement unisexe (la *danse* et il *danse),* où l'adjectif se pare des plumes du nom (le *bon,* la *bonne),* et où, plus généralement, la fonction (et donc la nature) diffère d'un emploi à l'autre— ce qui se produit précisément dans beaucoup des mots fonctionnels. Si les grammairiens étaient équitables ou pitoyables, ils devraient interdire aux mots le cumul abusif des fonctions. Si encore G. Engwall s'était contentée de catégories grossières, si en particulier elle avait admis un grand réceptacle pour recueillir indifféremment tous les mots grammaticaux, la tâche eût été moins rude. Mais, héroïque jusqu'au bout, l'auteur a passé au tamis les 10146 occurrences de *le,* pour distinguer l'article du pronom, les 7987 emplois de *les,* les 5067 *est* (pour y relever quatre fois le point cardinal), les 5031 *pas* (où la négation cède 146 fois le pas au substantif), les 3235 *ce,* les 2942 *s',* les 1547 *si,* et les 2583 *sur* (en n'y trouvant aucun adjectif, aucun fruit *sur).* Aucune enquête de cette taille (le corpus rassemble 500000 mots) n'a jamais atteint tant de minutie et de précision philologique. Ici il faut s'arrêter pour admirer, et s'agenouiller. Voilà enfin ouverts des domaines jusqu'ici

interdits par l'homographie: celui de l'article défini', qui compte 32708 occurrences (dont 5995 *l',* 10912 *la,* 8623 *le,* 7129 *les, 1 el,* 13 *L',* 16 *La,* 13 *Le* et 6 *Les),* et celui de la troisième personne. Voilà enfin des bases à peu près sûres pour la statistique des parties du discours—mais les épigones auront-ils plus de courage que les devanciers et suivront-ils la pente raide où Gunnel Engwall s'est engagée?

- L'étape de la désambiguïsation étant franchie, rien ne s'oppose à la lemmatisation proprement dite dont rendent compte les listes 3 (sous forme hiérarchique) et 4 (dans l'ordre alphabétique). Ce dernier index, gros de 153 pages, est le plus important de l'ouvrage. Il restitue non pas les références comme le titre le laisserait entendre[5]— mais la fréquence et la répartition de chaque lemme et de chaque forme rattachée au lemme. Comparées aux fréquences relevées dans le *Trésor de la langue française,* ces données[6] fournissent de précieuses informations sur l'évolution et l'actualisation de l'usage littéraire et donnent un aperçu de ce que sera le *Supplément,* qui est en projet, du *Trésor*.

- La cinquième liste est comme la seconde un dictionnaire inverse, appliqué cette fois aux lemmes. Cet outil est destiné aux recherches qui portent sur la finale des mots et principalement à l'étude de la suffixation[7]. L'outil est précieux, puisqu'il permet de ne rien oublier, mais peu précis, puisqu'il ne permet pas de distinguer les vrais et les faux suffixes. Mais le départ entre les uns et les autres est nécessairement une décision linguistique, à prendre cas par cas, et il serait imprudent de déléguer ce pouvoir à la machine. Observons ici l'embarras qu'engendre l'accentuation du français dans les tris alphabétiques: la neutralisation des accents produit des séries comme *pâte, pâté, insecte, été, bête, méchanceté, fête, diète,* dans lesquelles le suffixe *té* est difficile à isoler. Certaines finales, plus délicates à appréhender, auraient peut-être mérité des programmes spécifiques, ou simplement un tri multicritère, qui prenne en compte les codes grammaticaux (au moins les mieux représentés: *nf, nm, a d j)*[8]
- La sixième liste produit précisément le résultat d'une sélection multicritère, où le tri primaire envisage la catégorie, et le tri secondaire la fréquence. *Petit* et *grand* viennent en tête des adjectifs, comme dans le corpus du *Trésor* (mais en inversant les places), le *jour* et *l'homme* occupant les deux premiers rangs parmi les substantifs (là aussi il y a inversion dans le TLF)[9].

On passera vite sur les tableaux de chiffres reproduits aux pages 415-424 (et commentés clairement dans l'introduction XLIX–LIII). Qu'il s'agisse de la fréquence des caractères

(dans les types, les formes et les lemmes), de la longueur des mots, de la distribution des classes de fréquences, de la distribution des formes dans les lemmes, ou de celle des catégories grammaticales, on ne considère ici que des effectifs (ou des pourcentages) abstraits et décharnés, où l'identité des mots a disparu. Certains se féliciteront de la discrétion de ces données purement quantitatives, d'autres, moins nombreux et plus spécialisés, auraient peut-être souhaité quelques détails supplémentaires (concernant les 25 textes individuels du corpus). Mais tous approuveront la netteté de ces chiffres.

Mais l'essentiel de l'ouvrage n'apparaît qu'à la dernière page, quand on ferme le livre et que les doigts découvrent glissées dans la couverture, les 43 microfiches qui contiennent la concordance intégrale des 500000 mots du corpus. Chaque fiche comporte 207 pages de 60 lignes (au total c'est donc un volume de près de 9000 pages). Aucun risque de se perdre, chaque fiche contenant en très gros caractères le premier mot traité (et le premier contexte)[10]. La concordance, d'un modèle éprouvé (de type KWIC), place la forme traitée en position centrale avec un contexte d'environ 60 caractères de part et d'autre et, en marge, les références (du texte, de la page et de la ligne). Le tri ne s'exerce pas seulement sur la forme elle-même, mais aussi sur le contexte de droite, ce qui réunit visuellement les expressions, les constructions, les cooccurrences[11]. On a donc le moyen de contrôler tous les choix de G. Engwall en matière de lemmatisation, et de procéder soi-même à d'autres regroupements ou d'autres distinctions. Mais on dispose surtout d'un outil, léger et puissant à la fois, qui autorise les recherches les plus diverses sur le fonctionnement du discours littéraire. On peut certes imaginer d'autres systèmes documentaires qui livreraient l'information par d'autres canaux: l'Institut National de la langue française a choisi la voie télématique avec sa base de données *Frantext*: tôt ou tard la technologie du disque optique s'imposera aussi, en association avec un système local d'interrogation. Mais, en l'état actuel des équipements documentaires, la solution offerte par la microfiche l'emporte de loin pour l'économie, la disponibilité et - comme on dit maintenant - la convivialité. Les chercheurs - ou convives - que G. Engwall convie aimablement à sa table de lecture ne seront pas déçus.

Etienne Brunet,
Institut National de la langue française

Notes

1. Voici les 25 auteurs retenus: C. Aubry, M. Bataille, Y. Berger, J. Cabanis, J. P. Chabrol, M. Droit, J. Dutourd, C. Etcherelli, B. et F. Groult, J. Husson, S. Japrisot, J. M. G. Le Clézio, P. Moinot, I. Monési, F. Nourrissier, J. P. Oliver, C. Paysan, G. Perec, R. V. Pilhes, B. P. Delpech, H. F. Rey, C. de

Rivoire, R. Sabatier, F. Sagan, R. Vrigny. Notons que l'exclusion d'écrivains célèbres s'explique par tel ou tel des critères de sélection: naissance à l'étranger (M. Duras, J. Kessel, G. Simenon, H. Troyat, M. Yourcenar), cadre en dehors des frontières (A. Cohen, H. de Montherlant, P. de Mandiargues, A. Robbe-Grillet, R. Vaillant), époque en dehors des limites (L. Aragon, J. Giono, J. Green, M. Pagnol, R. Queneau). Il n'y a pas lieu d'avoir trop de regrets, car les écrivains éliminés au moins les plus vieux—se trouvent déjà dans le corpus du *Trésor.*

2. C'est ce qui s'est produit en effet. L'entreprise de Gunnel Engwall a été récemment associée à l'Institut National de la langue française, et un protocole de collaboration vient d'être signé par le CNRS.
3. L'auteur n'a pas disposé des programmes de lemmatisation automatique, qui à l'heure actuelle peuvent prendre en charge une grande part du travail.
4. L'article indéfini est moins bien loti et ne se dégage pas des numéraux. Et le pluriel *des* n'a subi aucun traitement. Il est vrai que dans bien des cas les valeurs de l'indéfini et du numéral se croisent dans *un* et *une,* comme l'indéfini et le partitif dans beaucoup d'emplois de *des.*
5. Les références sont données dans la concordance, en même temps que le contexte.
6. Elles ont déjà servi de toile de fond pour l'étude de quelques romans français, et de contrepoint ou de référence pour le contrôle des corpus antérieurs. Quand les "normes" divergent, le dictionnaire des fréquences de G. Engwall peut prétendre au statut de mètre-étalon, vu la rigueur exemplaire de sa fabrication. Voir là-dessus un article à paraître de Paul Fortier (Colloque de Liège 1987, *La lettre et le chiffre).*
7. Le tri inverse appliqué aux formes dans la liste 2 sert plutôt à l'étude de la flexion verbale. Charles Muller en a extrait des remarques judicieuses sur l'emploi des temps et des modes: « Données quantitatives en morpho-syntaxe: l'imparfait du subjonctif dans quelques romans contemporains », in *Le français moderne,* 1986/2, pp. 220-230. Un compte-rendu très favorable du présent ouvrage figure, sous la même signature, dans le même numéro, pp. 234-237.
8. Le suffixe *-té* apparaîtrait sans scories si la liste inverse était réduite aux seuls noms féminins et de la même façon les deux suffixes *-eur* se trouveraient décantés.
9. La *femme,* qui vient en troisième position dans le roman contemporain, recule d'une place dans le corpus du *Trésor,* pour laisser passer le *temps*—lequel n'a que le rang 6 dans la liste de G. Engwall. Comme ces mots ne souffrent guère de l'homographie et que les procédures de lemmatisation ne peuvent être mises en cause, on voit que la stabilité des fréquences laisse à désirer et que les corpus, si vastes ou si soigneux soient-ils, sont des 'normes' mouvantes qui auraient besoin elles-mêmes d'une norme.
10. De plus une table d'orientation figure en bas et à droite de la microfiche, qui dirige l'utilisateur dans les lignes et les colonnes.
11. Quand on s'intéresse à un syntagme, la démarche consiste donc à repérer le premier élément constitutif.