

R. Harald BAAYEN – *Word Frequency Distributions*.
Text, Speech and Language Technology n°18,
Dordrecht : Kluwer Academic Publishers, 2001, 334
p. + 1 CD-Rom

Benoît Habert et Michèle Jardino

**Édition électronique**

URL : <http://journals.openedition.org/corpus/42>

DOI : 10.4000/corpus.42

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 décembre 2003

ISSN : 1638-9808

Référence électronique

Benoît Habert et Michèle Jardino, « R. Harald BAAYEN – *Word Frequency Distributions*. Text, Speech and Language Technology n°18, Dordrecht : Kluwer Academic Publishers, 2001, 334 p. + 1 CD-Rom », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/42> ; DOI : <https://doi.org/10.4000/corpus.42>

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

R. Harald BAAYEN – *Word Frequency Distributions. Text, Speech and Language Technology n°18*, Dordrecht : Kluwer Academic Publishers, 2001, 334 p. + 1 CD-Rom

Benoît Habert et Michèle Jardino

- 1 H. Baayen (désormais HB) a publié récemment, généralement en partenariat, de nombreux articles importants en statistique linguistique, sur la notion de productivité morphologique tout particulièrement [Baayen *et al.*, 2000 ; Baayen & Schreuder, 2000]¹ L'ouvrage de synthèse qu'il présente est une étude statistique approfondie des distributions des fréquences des mots. A travers ce sujet clairement affiché dans le titre mais à première vue rebattu, HB semble, au moins pour les distributions des mots (par opposition aux autres « événements linguistiques » que l'on peut compter : parties du discours, étiquettes syntaxiques ou sémantiques, etc.), vouloir avancer vers des modèles quantitatifs rendant mieux compte de la difficulté à extrapoler d'un fragment langagier à un ensemble plus vaste d'énoncés ou, à l'inverse, à interpoler de données volumineuses à un ensemble restreint.
- 2 Le premier chapitre décrit les lois empiriques (en particulier la fameuse loi de Zipf ou les formules de Herdan) qui visent à rendre compte de la variation des fréquences de mots. Il souligne que, contrairement à la situation habituelle en statistique, l'augmentation en taille des données langagières utilisées n'aboutit pas forcément à une plus grande précision de l'estimation de certains paramètres (comme le montrent par exemple p. 17 les ajustements nécessaires pour adapter les prédictions de Zipf à différentes tailles de texte). Il essaie également de montrer les conséquences d'une simplification souvent faite : les mots arriveraient « au hasard » dans les textes. L'utilisation de la distribution log-normale, où la transformation logarithmique permet dans certains cas de ramener à la distribution normale des distributions biaisées, n'est

pas non plus satisfaisante pour le lexique : les fréquences manifestent des déviations importantes dans les basses fréquences, les paramètres de la distribution log-normale varient également avec la taille de l'échantillon.

- 3 Les chapitres deux à quatre explicitent les modèles statistiques mis en œuvre pour caractériser les distributions des fréquences lexicales comme des LNRE (*Large Number of Rare Events distributions*) : ce sont des modèles² qui prennent en compte les très nombreux mots qui ont des probabilités d'apparition extrêmement faibles et qui s'éloignent par là même des modèles statistiques « classiques » (binomial).
- 4 Le chapitre deux est consacré aux modèles non paramétriques. Après des rappels de probabilités, il présente le modèle d'urne, puis le modèle binomial et son approximation poissonnienne, adaptée aux événements rares. HB, « tirant » 90 millions de mots du *British National Corpus* (BNC), souligne que la taille du vocabulaire continue à augmenter avec celle de l'échantillon (en revanche, plus de la moitié des types – formes ou chaînes de caractères – trouvés dans le BNC ont une fréquence relative de 0.0000001). Quand on estime les probabilités des types à partir d'un échantillon, ces probabilités par définition ont pour somme 1. Le fait que les données langagières relèvent en fait des LNRE oblige alors en quelque sorte à « réserver » une masse de probabilité pour les types non encore rencontrés si l'on prend un échantillon plus large ou différent. HB présente p. 57 une technique d'estimation (Good-Turing) des probabilités des types à partir d'un échantillon qui opère cette « épargne probabiliste ». Il montre ensuite p. 69 les limites de l'extrapolation des probabilités pour des échantillons substantiellement plus volumineux.
- 5 Le chapitre trois entend enrichir les expressions non paramétriques du chapitre précédent avec des hypothèses paramétriques sur la structure de la distribution. Trois familles de modèles LNRE sont présentées, pp. 82-117. Ces modèles sont tous fondés sur un paramètre libre correspondant à la taille de l'échantillon à partir duquel les modèles interpolent ou extrapolent. L'ensemble des paramètres doit être évalué par ajustement des courbes théoriques et expérimentales. Une étude comparative sur *Alice au pays des merveilles*, *De l'autre côté du miroir* (L. Carroll également), *Le chien des Baskerville* (C. Doyle), *La guerre des mondes* (Wells) et la partie du BNC d'oral « situationnel » (lié à des contextes d'interaction bien délimités) montre en fait que le modèle optimal varie selon les textes. Le nombre de paramètres à évaluer pour chaque méthode aurait d'ailleurs pu être signalé clairement, en séparant par exemple dans le tableau de résultats 3.3 les paramètres évalués des valeurs connexes. C'est en effet un élément important de choix du modèle.
- 6 Le chapitre quatre aborde les « assemblages »³ de distributions (*mixture distributions*) nécessaires pour rendre compte de l'hétérogénéité interne des corpus (soit parce qu'ils regroupent des textes d'auteurs, de styles ou de genres distincts, soit parce que leurs mots sont de structure distincte : mots simples comme *arbre* / mots complexes, *i.e.* en plusieurs mots comme *banque d'arbres* ou construits morphologiquement comme *arboré*). A partir de la base lexicale CELEX, HB met en évidence, par exemple p. 135 pour le néerlandais, les différences de comportement fréquentiel des noms simples et des noms construits en *-heid* (équivalent de *-ness* en anglais, dans *fitness*, 'fait d'être en forme') : les premiers présentent peu d'hapax au contraire des seconds. HB raffine ensuite cette analyse p. 145 en montrant que les deux sens du suffixe *-heit* (rôle abstrait : 'le fait de' / rôle anaphorique) ont eux-mêmes deux distributions distinctes : le sens abstrait correspond à des mots plus fréquents que ceux qui relèvent du rôle anaphorique.

L'objectif d'HB est d'être capable de repérer les distributions distinctes à l'œuvre (d'où peut-être le pluriel du titre), un peu à la manière dont nous sommes capables, lorsque nous assistons à un concert, d'« isoler », de l'oreille aidée du regard, un instrument ou un groupe d'instruments du rendu global. Le suffixe *-ness* donne ainsi lieu à des observations divergentes quand on oppose partie d'oral situationnel du BNC et partie d'écrit : des distributions distinctes sont à l'œuvre, en fonction de « variables cachées » (les mots relèveraient d'axes distincts, comme la partie du discours et le mode de construction morphologique, intervenant de manière croisée dans leur comportement quantitatif global) et il faut se donner les moyens de les distinguer.

- 7 Le chapitre cinq montre les limites de l'hypothèse d'une distribution aléatoire des mots, liées en particulier à la cohésion thématique ou discursive des textes et à l'arrivée « en rafales » des mots « pleins ». Ces arrivées groupées, qui correspondent à des spécialisations thématiques ou à des changements de thème, contribuent, aux côtés ou à la place d'autres causes, évoquées par HB, à la difficulté à extrapoler d'un fragment à un échantillon plus large⁴. HB fournit deux techniques d'ajustement des modèles LNRE pour pallier partiellement ces limites. La première revient à partitionner les mots en deux ensembles : « agglutinants » / « éparpillés » et à formuler un sous-modèle distinct par ensemble⁵. La seconde paramètre les modèles LNRE examinés au chapitre trois en fonction de la taille du texte.
- 8 Le dernier chapitre est consacré à des exemples d'applications (parmi lesquels, pour montrer la généralité des modèles LNRE, quatre exemples soit non limités aux mots isolés - structures consonne-voyelle, paires de mots - soit non linguistiques comme les années citées dans trois journaux pendant une année) : lien entre taille des mots et longueur de l'échantillon, rapport entre fréquence des lemmes (ou formes canoniques) et des flexions, productivité morphologique et « registres », attribution d'auteur et style, distributions de paires de mots. Il se termine par des indications techniques (y compris pré-traitement des textes) et méthodologiques sur la mise en œuvre concrète des approches préconisées.
- 9 *Word Frequency Distributions* offre, c'est une de ses forces et une de ses originalités, au moins par rapport à la littérature statistique linguistique en français, plusieurs niveaux de lecture clairement signalés comme tels. Les sections commencent par des résumés explicites, détachés, qui permettent au profane de saisir les enjeux (à défaut de pouvoir discuter les choix faits). De nombreux graphes, sous formes de tables ou de courbes, illustrent et étayent les modèles présentés. On admirera tout particulièrement les titres des figures (repris p. ix-xx), petits paragraphes de cinq lignes en moyenne, qui fournissent toutes les clés de lecture souhaitables. Une formation mathématique beaucoup plus poussée est cependant nécessaire pour examiner les démonstrations détaillées qui sont également fournies et en particulier pour discuter sur le fond les propositions avancées. Exercices et solutions sur le versant plus mathématique figurent également. La bibliographie du domaine (y compris française), de 120 références, commentée avec précision, ainsi que les « états de l'art » partiels qui accompagnent les développements donnent la possibilité au lecteur, quel que soit son niveau de compréhension, d'explorer d'autres pistes et positionnements (d'autant que la littérature rassemblée et mise en perspective par HB s'échelonne sur plusieurs dizaines d'années et n'est pas toujours aisément accessible, en raison soit de son ancienneté soit de sa technicité soit de son origine, de nombreux travaux cités provenant de l'ex-URSS). HB fournit enfin LEXSTAT, un logiciel pour Unix/Linux sous licence GPL (GNU

Public Licence⁶) qui permet d'opérer sur ses propres données les calculs décrits dans le livre (les sous-programmes sont clairement documentés). L'interopérabilité est assurée avec les environnements statistiques « libres » ou non que sont R (<http://cran.r-project.org/welcome.html>) et SPLUS. Les commandes du logiciel sont documentées (dans l'esprit du *man* d'Unix).

- 10 On peut regretter que les études exhaustives réalisées ici se soient appuyées le plus souvent sur des textes de petite taille (*Alice au pays des merveilles*, 26 505 « mots », sert d'exemple récurrent, même si les 100 millions de mots du BNC sont aussi utilisés). Néanmoins le logiciel fourni devrait permettre de vérifier si les modèles présentés rendent également compte des fréquences des mots dans les grands corpus électroniques constituables à partir des archives disponibles (comme celles du journal *Le Monde*).
- 11 On aurait sans doute aimé également que l'ouvrage, sans faire pour autant de concessions quant aux démonstrations mathématiques, lie davantage les propositions de fond, qu'HB sait présenter de manière presque transparente, à ce qu'elles permettent de dire sur l'interaction des axes de description des mots : morphologie, syntaxe, sémantique. Le dernier chapitre est en effet décevant. Consacré à des exemples d'application, il en survole sept en une vingtaine de pages, ce qui fait bien peu par application. Il s'agit d'illustrations, d'esquisses presque, plutôt que de mises en œuvre fructueuse du paradigme choisi. C'est aux articles auxquels HB a contribué qu'il faut alors se reporter, même si *Word Frequency Distributions* leur fournit un cadre unificateur stimulant. On ne peut pour terminer que conseiller la lecture de ces articles.

BIBLIOGRAPHIE

Baayen R. H. & Schreuder R. (2000). « Towards a psychological computational model for morphological parsing », *Philosophical Transactions : Mathematical, Physical and Engineering Sciences* 358 : 1281-1293

Baayen R. H., Schreuder R. & Sproat, R. (2000). « Morphology in the mental lexicon : a computational model for visual word recognition », in Eynde, F. V. & Gibbon, D. (eds), *Lexicon Development for Speech and Language Processing*. Text, Speech and Language Technology n°12, Dordrecht : Kluwer Academic Publishers, pp. 267-293.

Lafon P. (1981). « Statistiques des localisations des formes d'un texte », *MOTS* 2 : 157-188.

NOTES

- 1.. La question est d'ailleurs reprise p. 154 et p. 203.
- 2.. Introduits par Khmaladze en 1987.
- 3.. En vinification, pour certaines appellations, obtention d'un vin par mélange de cépages distincts, dans des proportions prédéfinies et/ou ajustées à la dégustation. Le

rapport entre les distributions d'ensemble et les distributions sous-jacentes « assemblées » au sein d'un corpus est probablement à peu près aussi complexe que celui entre le vin obtenu et ses composants.

4.. Le fait d'enlever ces mots « agglutinants » aboutit à un meilleur ajustement de la taille attendue du vocabulaire par rapport à la taille observée en fonction de la taille de l'échantillon.

5.. A partir d'un modèle d'ajustement de l'interpolation binomiale proposé par P. Hubert et D. Labbé en 1988.

6.. Licence qui régit les droits de distribution du logiciel « libre » de type Emacs, Linux, etc.