Études métalexicographiques par analyse d'échantillons : toute représentativité est-elle illusoire ?

Franck Sajous

Études métalexicographiques par analyse d'échantillons : toute représentativité est-elle illusoire ?

Franck Sajous
CLLE, CNRS & Université de Toulouse 2

Résumé

Le dictionnaire général étant trop volumineux pour être étudié dans son intégralité, son analyse passe généralement par l'examen d'un échantillon et par la généralisation à l'ensemble du dictionnaire des résultats observés sur cet échantillon. Si beaucoup d'énergie est consacrée à l'analyse des échantillons, les métalexicographes semblent, en revanche, accorder peu d'importance aux mécanismes de sélection des échantillons eux-mêmes. Après avoir rappelé les principes de l'échantillonnage et les différentes techniques de sélection des échantillons, cet article montre, à travers différentes expériences, que l'examen de zones contiguës, qui a quasi-systématiquement la préférence des métalexicographes, est à proscrire, d'autres méthodes d'échantillonnage pratiquées sur l'ensemble du dictionnaire étant toujours préférables. L'article montre également que même la méthode de sélection la moins problématique ne peut garantir totalement la représentativité d'un échantillon. Les résultats obtenus par analyse d'échantillons doivent donc être considérés avec circonspection.

Mots-clés : Métalexicographie, méthodes d'échantillonnage, représentativité

Abstract

As general dictionaries are too voluminous to be studied in their entirety, their analysis generally require sample analysis. Contrary to the analysis of the selected samples, which concentrate much of metalexicographers efforts, the sampling methods are often neglected. After recalling sampling principles and sample selection techniques, this article shows, through various experiments, that the examination of one-stretch samples, which is almost systematically preferred by metalexicographers, is to be avoided, as sampling of the entire dictionary by using other methods always produce more reliable estimations. The article also shows that even the least problematic technique cannot fully guarantee the representativeness of a sample. Results obtained from sample analysis should therefore be treated with caution.

Keywords: Metalexicography, sampling methods, representativeness

Pour citer cet article :
SAJOUS, Franck (2024). Études métalexicographiques par analyse d'échantillons : toute représentativité est-elle illusoire ? Dans Nicolas SORBA et Nadine VINCENT (dir.), *Les Dictionnaires numériques dans l'espace francophone, des ressources porteuses de culture et d'idéologies*. Sherbrooke : Éditions de l'Université de Sherbrooke, pp. 239–272.

Études métalexicographiques par analyse d'échantillons : toute représentativité est-elle illusoire ?

Franck Sajous
CLLE, CNRS & Université de Toulouse 2

1 Introduction

L'étude des caractéristiques de la structure et du contenu du dictionnaire général, lorsqu'elle est réalisée manuellement, passe généralement par l'analyse d'un échantillon et par la généralisation à l'ensemble du dictionnaire des résultats observés sur cet échantillon, l'ensemble étant trop volumineux pour être étudié dans son intégralité [1]. Bukowska (2010) observe que, si beaucoup d'énergie est consacrée à l'analyse des échantillons, très peu de réflexion semble être accordée par les métalexicographes aux mécanismes de sélection des échantillons eux-mêmes. La méthodologie d'échantillonnage mérite pourtant réflexion, sous peine d'introduire des biais importants de sélection et, partant, d'analyse. Quelques métalexicographes sont bien conscient·e·s de certains écueils à éviter, comme l'absence de représentativité d'un échantillon de taille trop modeste. Dans la critique des dictionnaires qu'il pratique, Corbin (1984 : 113) s'impose un ensemble de contraintes méthodologiques, notamment celle de mener des « études extensives sur des échantillons suffisamment importants pour être estimés significatifs ». Nous souscrivons sans réserve à cette autodiscipline. Rappelons néanmoins qu'en statistiques, la *significativité* est un concept qui a un sens spécifique [2], tout comme la *représentativité* : pour qu'un échantillon puisse être estimé *représentatif* (avec une probabilité donnée), sa taille minimale (« suffisamment importante ») se calcule. Mais, plus que la taille de l'échantillon, c'est l'importance capitale de la méthode d'échantillonnage qui semble être largement méconnue des métalexicographes.

En menant des expériences d'échantillonnage sur le dictionnaire *Usito* [3], nous montrons dans cet article que l'examen de zones contiguës, qui a quasi-systématiquement la préférence des métalexicographes, est à proscrire et que

1. Cette prémisse souffre quelques exceptions notables. Martinez (2013), par exemple, parcourt intégralement les éditions ou millésimes successifs des *Petit Larousse illustré* (PLI), *Petit Robert* (PR) et *Dictionnaire de l'Académie française* (DAF) afin de recenser les changements dans la macrostructure de ces dictionnaires. Corbin (1990) parcourt, à l'aide de 45 étudiant·e·s, l'intégralité des nomenclatures de cinq grands dictionnaires généraux à la recherche de noms de végétaux dérivés en *-ier* et de leur base apparente.

2. Par exemple, des tests statistiques permettent de calculer la significativité d'éventuelles différences mesurées entre deux échantillons.

3. Dictionnaire du français québécois développé par l'Université de Sherbrooke, accessible en ligne à l'adresse : `https://usito.usherbrooke.ca/`.

d'autres méthodes d'échantillonnage pratiquées sur l'ensemble du dictionnaire lui sont préférables. Nous montrerons toutefois qu'on ne peut, en aucun cas, garantir totalement la représentativité d'un échantillon. De telles conclusions peuvent revêtir une certaine valeur d'évidence pour les spécialistes des techniques d'échantillonnage. Nous pensons toutefois que l'étude empirique présentée dans ce chapitre se justifie par sa vocation pédagogique et par le public spécifique visé (les métalexicographes). En effet, un parcours de la littérature sur les analyses métalexicographiques par analyse d'échantillons suggère qu'il est nécessaire d'illustrer, par des expériences concrètes portant sur l'objet d'étude des métalexicographes (les dictionnaires), les enjeux du choix des méthodes d'échantillonnage dans le domaine spécifique de la métalexicographie et les précautions à prendre lors de l'interprétation des résultats observés. [4]

2 Échantillon et échantillonnage : principes

Les statistiques inférentielles consistent à quantifier les caractéristiques d'une population restreinte (un échantillon) dans le but d'estimer celles de la population générale. Concernant l'analyse des dictionnaires, il s'agit de sélectionner un ensemble d'observables (pages, articles, définitions, exemples ou toute délimitation cohérente d'un segment textuel relativement au cadre de l'étude menée) et de quantifier certaines caractéristiques pour la sélection choisie. Les valeurs *réelles* des caractéristiques *mesurées* sur l'échantillon analysé servent à *estimer* les valeurs *probables* de ces caractéristiques pour l'ensemble du dictionnaire.
On peut recourir à l'analyse d'échantillons en synchronie, pour décrire les caractéristiques d'un dictionnaire, comparer plusieurs dictionnaires distincts ou différentes parties d'un même dictionnaire, ou en diachronie, pour comparer différentes éditions ou différents tomes (rédigés à des périodes plus ou moins éloignées) d'un même dictionnaire. Selon les cas, il s'agit donc d'analyser un seul échantillon ou de comparer les similitudes ou les différences entre plusieurs échantillons. Mais quelle que soit la configuration, la manière de constituer les échantillons (taille et méthode de sélection) a une incidence déterminante sur la fiabilité des résultats observés et sur celle de l'estimation qui en découle.

L'analyse d'échantillons porte le plus souvent sur des propriétés mesurables numériquement. C'est de ce type d'analyse dont il est question à partir de la section 2.4. Mais on trouve également des études qualitatives menées sur des corpus d'observables relativement vastes, qualifiés d'*échantillons* par les auteur·rice·s de ces études. Nous donnons trois exemples de ces cas particuliers

4. Nous nous inscrivons en cela dans la lignée de Bukowska, dont l'article ne constitue pas un travail fondateur ou « de référence » en statistiques inférentielles, mais est l'un des rares, à notre connaissance, à problématiser la mise en œuvre et la pertinence de différentes méthodes d'échantillonnage dans le domaine spécifique de la métalexicographie. C'est la raison pour laquelle nous référerons à ce travail au long de ce chapitre.

en section 2.2. La section 2.3, consacrée aux méthodes d'échantillonnage, concerne les analyses aussi bien quantitatives que qualitatives, tandis que la section 2.4, qui aborde la question du contrôle de la représentativité, porte spécifiquement sur les analyses quantitatives (chiffrées).

2.1 Intérêt de l'échantillonnage selon le support des dictionnaires

Nous avons justifié, en introduction, l'intérêt de l'échantillonnage par la taille des dictionnaires généraux qui, le plus souvent, dissuade d'envisager leur étude exhaustive. C'est assurément le cas pour les dictionnaires imprimés. En revanche, l'intérêt de recourir à l'échantillonnage pour l'étude des dictionnaires numériques ne s'impose pas nécessairement à l'esprit, ceux-ci se prêtant aux études computationnelles qui permettent de les analyser dans leur globalité. Deux situations justifient cependant d'échantillonner sur ces dictionnaires :

— lorsqu'aucun moyen (API, archive, etc.) n'est proposé pour accéder automatiquement au contenu des dictionnaires et/ou lorsque des contraintes légales empêchent leur analyse automatique ;

— lorsque l'analyse à mener est trop difficilement automatisable, comme c'est le cas pour l'étude de certains phénomènes qualitatifs. À supposer que les études mentionnées en section 2.2 aient porté sur des dictionnaires accessibles automatiquement et légalement permissifs, les analyses qualitatives menées manuellement ne seraient pas automatisables pour autant. [5]

Dans ces situations, l'étude des dictionnaires numériques est donc manuelle et requiert la constitution d'un échantillon à analyser.

2.2 Le cas particulier des études qualitatives : quantifier n'est pas – nécessairement – chiffrer (ni généraliser)

Lehmann (1995) et Corbin (1995) étudient, dans le même numéro de la revue *Lexique*, les transformations que subissent les citations littéraires d'un dictionnaire à l'autre : du *Grand Robert* (GR) au PR pour Lehmann et du PR au *Micro Robert* (MR) pour Corbin. Lehmann choisit d'analyser les articles de la lettre N, Corbin ceux de la lettre F. Afin de mettre en évidence « [l]e poids des contraintes dictionnairiques sur l'évolution des marqueurs », Martinez (2011 : 44–47) se penche sur « le sort de la marque *littéraire* » à travers les dix révisions des millésimes 1997 à 2007 du PLI. Pour mener son étude, il analyse les articles de la lettre E. On trouve, dans les trois articles, des commentaires sur le choix des échantillons. Corbin (1995 : 125–126), qui parle d'une « étude

5. Pour un aperçu des problèmes que pose ce type d'analyse, lorsqu'il est appliqué aux dictionnaires (ou encyclopédies), voir dans ce collectif la section 7 de l'article intitulé « Pour une analyse qualitative *et* quantitative, manuelle *et* computationnelle, synchronique *et* diachronique, des dictionnaires numériques ».

menée sur un corpus extensif dont la présente contribution va donner un aperçu nécessairement simplifié et fragmentaire, l'analyse en vraie grandeur requérant une publication plus développée », consacre un paragraphe à l'« échantillon lexicographique étudié » :

> L'étude porte sur l'ensemble des entrées commençant par *F*, soit près d'un vingtième du texte lexicographique des deux dictionnaires. En valeur relative, cet échantillon paraît suffisant pour un ensemble de données très copieux à étudier : 1 847 entrées dans le *PR*, 1 304 dans le *MR*. Le choix de la lettre *F* n'a pas d'autre justification qu'une certaine familiarité avec cet échantillon, sur lequel a déjà été menée antérieurement une étude comparative du même ordre concernant l'utilisation des marques d'usage. (Corbin, 1995 : 126)

Lehmann consacre également une section, intitulée « Le corpus », au choix et à la description de son échantillon :

> Pour avoir un échantillon suffisamment vaste, on a procédé au dépouillement exhaustif d'une lettre, la lettre *N*, dans les deux dictionnaires [...]. Cette lettre a été retenue parce qu'à ce stade de la fabrication du *GR*, l'équipe rédactionnelle est constituée de manière durable et que cette même équipe est, pour l'essentiel, responsable de l'élaboration du *PR*. (Lehmann, 1995 : 108)

Elle écrit plus bas que la lettre N du PR contient 804 articles, dont 243 comportent des citations. Ces citations comportent environ un tiers de syntagmes signés (contre deux tiers de phrases), proportion qui lui semble en accord avec l'intention décrite dans la préface du dictionnaire. À propos d'une potentielle généralisation de son observation, elle écrit :

> Mais l'exhaustivité du dépouillement d'une lettre reste fallacieuse et n'autorise pas une exploitation d'ordre statistique dont l'intérêt est, d'ailleurs, bien secondaire. On peut, en revanche, considérer qu'elle permet un repérage qualitatif certainement complet des différentes procédures des manipulations de la citation, l'abrègement en syntagme illustrant une de ces figures. (Lehmann, 1995 : 109)

Martinez commente lui aussi le choix de son échantillon :

> [l]a lettre E du dictionnaire, pas plus que n'importe quel échantillon limité à une tranche alphabétique, n'est représentative de l'ensemble du texte. Nous l'avons choisie parce que les marqueurs *litt.* y foisonnent. La seule démonstration tentée étant celle du caractère désordonné et aléatoire des changements de marqueurs, les données observées et les résultats déduits ne fluctuent pas en fonction du choix de l'échantillon. (Martinez, 2011 : 45)

Martinez parcourt l'ensemble des pages de la lettre E de 11 millésimes (des 64 pages du PR1997, comprenant 170 articles marqués aux 63 pages du PR2007 comprenant 197 articles marqués).

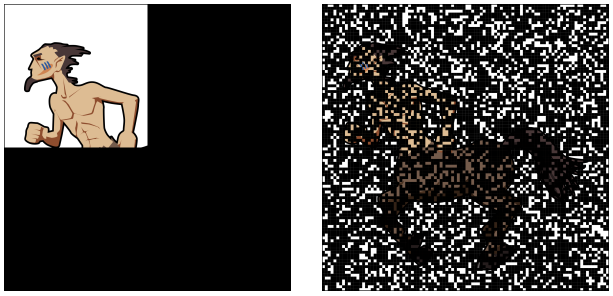
Les trois métalexicographes ont en commun de reconnaître l'arbitraire du choix de la tranche sélectionnée, qui ne repose sur aucun calcul statistique, tout en fournissant les justifications de ce choix : une certaine familiarité pour Corbin, la stabilisation de l'équipe rédactionnelle pour Lehmann et le foisonnement des marques étudiées pour Martinez. Concernant la taille de l'échantillon analysé, Lehmann parle d'échantillon suffisamment vaste et Corbin d'échantillon suffisant et d'un ensemble de données très copieux à étudier. On peut accorder aux deux métalexicographes que les échantillons sont vastes, et le matériau à étudier copieux, mais comment doit-on comprendre *suffisant* et *suffisamment vaste* (*i.e. suffisant* pour quoi faire ?). En statistiques, la *représentativité* d'un échantillon est une notion spécifique (cf. section 2.4) qui concerne les analyses quantitatives chiffrées. Lehmann et Corbin en ont bien conscience et ne prétendent pas démontrer mathématiquement la représentativité de leurs échantillons. Lehmann s'interdit une exploitation d'ordre statistiques. Quant à Corbin, il écrit que sa contribution « va donner un aperçu nécessairement simplifié et fragmentaire » et estime que son échantillon « paraît suffisant pour prétendre à une certaine représentativité ». Il n'affirme donc pas que son échantillon *est* représentatif. Pour autant, dire que l'assertion précédente le sous-entend ne paraît pas exagéré. Les études de Lehmann, Corbin et Martinez, dont le but est avant tout de mettre au jour les mécanismes et les motivations qui sous-tendent certains traitements, et d'étudier les effets de ces traitements, ne tombent pas dans la catégorie des analyses quantitatives chiffrées. Lehmann écrit par ailleurs que l'intérêt d'une exploitation statistique de son échantillon serait « bien secondaire ». Après avoir précisé la taille de leur corpus (en nombre d'articles examinés), et le nombre d'observables identifiés (citations, marques), Lehmann, Corbin et Martinez mènent une étude qualitative des traitements lexicographiques, qu'ils estiment plus ou moins récurrents. Corbin, par exemple, exprime la fréquence des différentes catégories de transformation qui affectent les citations du PR1967 lors de leur insertion dans le MR1971 par des quantités relatives ou des proportions non chiffrées : « la plupart des modifications », la/les modification(s) « la plus apparente », « nombreuses et variées », « assez nombreuses », « très nombreuses », les substitutions « les plus fréquentes ». Cela confirme bien que l'établissement des proportions exactes des catégories respectives des différents mécanismes observés n'est pas l'objet prioritaire de ces études. Comme nous l'avons écrit plus haut, il s'agit avant tout de démontrer l'existence d'un phénomène qui, si ses occurrences sont « très nombreuses » dans un corpus « copieux », a de bonnes chances de ne pas être marginal. On peut néanmoins regretter, dans ces études, que le chiffrage des phénomènes observés, immédiatement disponible pour les métalexicographes qui ont collecté et annoté leur matériau dictionnairique, ne soit pas communiqué aux lecteur·rice·s. En effet, publier des données

chiffrées n'empêcherait aucunement les métalexicographes de commenter les résultats avec les mêmes formulations que celles rapportées plus haut, tout en permettant au lectorat d'être en mesure de se forger sa propre opinion sur la base des quantifications fournies. Une autre clarification souhaitable de la démarche adoptée dans ces études qui n'ont pas de prétention statistique serait de renoncer à emprunter la terminologie de ce domaine (*significativité*, *représentativité*) et à abandonner le terme d'*échantillon* au profit de *sous-corpus*.

2.3 Méthodes d'échantillonnage

Avant de passer en revue les différentes familles de méthodes d'échantillonnage existantes, tentons d'en donner une idée intuitive avec une illustration par l'image. Imaginons que l'on présente à des sujets différents échantillons composés de 25% des pixels d'une image et qu'on leur demande de se faire une idée de ce que représente l'image d'où chaque échantillon est issu. Les sujets à qui l'on soumet les échantillons représentés dans les figures 1a et 1b se feront probablement des idées assez similaires de l'image globale et ils n'auront qu'à moitié raison. Celui à qui l'on présente l'image 1c aura du mal à imaginer quoi que ce soit. Face à l'échantillon 1d, le sujet aura probablement en tête une image radicalement différente de celle que se figurent les sujets à qui l'on

(a) quart inférieur gauche (b) quart inférieur droit (c) quart supérieur gauche

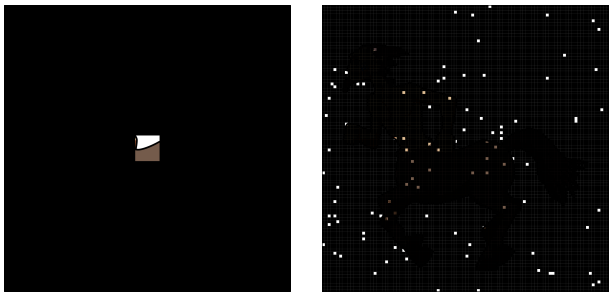


(d) quart supérieur droit (e) tirage aléatoire

FIGURE 1 – Échantillons issus d'une même image (25% des pixels)

a présenté les échantillons 1a et 1b et, comme eux, n'aura qu'à moitié raison. L'échantillon 1e, qui correspond à un tirage aléatoire d'un nombre de pixels identique à celui des autres échantillons, donne une idée plus globale, mais néanmoins imprécise, de l'image réelle [6].

La première leçon à tirer de cette illustration est que la méthode d'échantillonnage choisie a une incidence certaine sur l'estimation que l'on peut se faire de la réalité après examen d'un échantillon. La seconde est que même un échantillon de taille inhabituellement importante (les échantillons sont constitués ici de 25% de la totalité des observables) n'offre pas la garantie de fournir une estimation fiable du tableau général. Notons qu'en métalexicographie, la taille des échantillons analysés se situe généralement entre 1 et 2% de celle du dictionnaire [7]. Deux échantillons, constitués chacun de 1% des pixels de notre image d'illustration sont représentés en figure 2. Qu'ils soient constitués par sélection contiguë (fig. 2a) ou par tirage aléatoire (fig. 2b), il est difficile de prétendre qu'un quelconque échantillon de cette taille puisse être *représentatif* et que l'on puisse généraliser les observations effectuées.



(a) zone contiguë (b) tirage aléatoire

FIGURE 2 – Échantillons issus d'une même image (1% des pixels)

2.3.1 Sélection par zone contiguë

La sélection d'articles contigus au sein d'une tranche donnée du dictionnaire, comme celle mise en œuvre dans les études qualitatives mentionnées en section 2.2, est également couramment pratiquée dans les analyses quantitatives chiffrées ayant vocation à généraliser les résultats mesurés sur échantillon. Pourtant, cette méthode d'échantillonnage est celle qui implique les biais les plus

6. L'image utilisée provient du site `https://publicdomainvectors.org`. Merci à l'auteur·rice qui l'a placée dans le domaine public.

7. On trouve néanmoins des études d'une plus grande ampleur. Dans son analyse du « recyclage des définitions » mentionnée plus haut, Corbin (1995) analyse par exemple 4,6% du PR (89 pages sur 1 938) et 4,8% du MR (56 pages sur 1 155). On comprend dans ce cas que, même sans en faire la démonstration mathématique, Corbin puisse être tenté de « prétendre à une certaine représentativité ».

importants, qu'ils émanent des propriétés inhérentes au lexique de la langue étudiée ou d'artéfacts liés au processus rédactionnel. Dans les projets au long cours, différents événements peuvent advenir, comme des changements de direction éditoriale, de l'équipe des rédacteur.rice.s, de corpus ou de pratiques lexicographiques. Ainsi, généraliser à l'ensemble d'un dictionnaire les observations menées sur son premier (ou dernier) tome est hasardeux. C'est la raison pour laquelle certain·e·s métalexicographes considèrent que, pour étudier un dictionnaire conçu sur une période étendue et rédigé selon une chronologie suivant l'ordre alphabétique, les lettres du milieu de l'alphabet seraient les mieux adaptées pour servir d'échantillon. Les lexicographes adopteraient en effet un *modus operandi* régulier au moment où cette zone est atteinte, entre rodage initial et accélération finale due à la pression financière et temporelle. En particulier, les traitements opérés dans les articles situés en fin d'alphabet seraient moins fouillés qu'au début du dictionnaire et la sélection des mots pour inclusion dans la nomenclature deviendrait plus sévère à mesure de l'avancement du projet. C'est ce qu'Osselton (2007) nomme « *alphabet fatigue* ». Il observe ce phénomène dans tous les dictionnaires britanniques du XVII^{e} siècle, mais l'explique finalement par une autre raison qu'une évolution du processus éditorial : le fait que ces dictionnaires aient tous bâti leur nomenclature à partir d'une même liste de mots. À l'inverse de l'*alphabet fatigue*, pour les dictionnaires historiques du XIX^{e} et XX^{e} siècles (*e.g.* l'*Oxford English Dictionary*, mais également un dictionnaire allemand et un autre danois), l'assurance d'un financement institutionnel pérenne a permis aux lexicographes d'affiner les traitements avec l'avancée de la rédaction. L'évolution supposée du processus lexicographique n'est donc pas systématique et ne va pas toujours dans le même sens. Par ailleurs, Bukowska (2010) montre à travers l'étude du *Webster's Revised Unabridged Dictionary* (1913) que les lettres du milieu de l'alphabet ne sont pas nécessairement le reflet de pratiques rédactionnelles stabilisées (elle observe, pour la lettre M, une nette baisse du nombre de citations). Cette croyance en une stabilisation du processus éditorial en milieu d'alphabet – ou à mi-projet – est pourtant tenace, comme en témoigne la justification du choix de la lettre N dans une étude récente : « N was chosen as a range well on in the alphabet, at which point editorial practices, in the still pervasive alphabetical editing order, can be considered as stabilized » (Ferrett et Dollinger, 2021). Clarifions notre propos. Il ne s'agit pas de nier le fait que des changements interviennent au cours de la conception d'un dictionnaire et que, notamment, la densité de certaines informations varie parfois à mesure de l'avancement du projet.[8] Ce qu'il convient selon nous de réfuter à la suite de Bukowska, c'est qu'il existerait une zone de l'alphabet à partir de laquelle il serait particulièrement propice de tirer un échantillon.

8. Cela vaut non seulement pour la densité, mais également pour la nature des informations, qui peut varier tout au long du processus lexicographique, comme le montre Radermacher (2004) dans son étude du *Trésor de la langue française*.

2.3.2 Tirage aléatoire

La constitution d'un échantillon par tirage aléatoire consiste à sélectionner au hasard le nombre d'observables requis pour atteindre la taille d'échantillon souhaitée. Cette sélection peut se faire dans une zone particulière (*e.g.* une lettre donnée ou un tome particulier) ou sur l'ensemble du dictionnaire. La mise en œuvre peut s'appuyer, dans les dictionnaires papier, sur la sélection aléatoire d'un numéro de page, puis, au sein de la page, sur celle de l'article, etc. Par exemple, pour comparer automatiquement [9] les styles définitoires de six dictionnaires anglais édités entre la fin du XVIII^{e} siècle et le début du XXI^{e} siècle, Kamiński (2015) constitue, pour chaque dictionnaire, des échantillons en se fondant sur un tirage aléatoire des numéros de pages. Les dictionnaires numériques présentent différents cas de figure. *Usito* et le *Trésor de la langue française informatisé* (TLFi) donnent accès à l'ensemble de leur nomenclature [10], à partir de laquelle un tirage peut être effectué. D'autres, comme le *Wiktionnaire*, disposent d'un lien *Page au hasard* [11]. Pour les dictionnaires dépourvus de ces fonctionnalités, d'autres stratégies doivent être envisagées, en fonction des contraintes légales qu'ils imposent.

2.3.3 Échantillonnage probabiliste stratifié

L'échantillonnage probabiliste stratifié est un raffinement du tirage aléatoire qui consiste d'abord à diviser le dictionnaire en zones non chevauchantes (les strates) correspondant par exemple aux lettres initiales des vedettes ou à des parties rédigées par des éditeurs différents, puis à produire aléatoirement des échantillons qui respectent certaines proportions, relativement à celles du dictionnaire (*e.g.* parties du discours, marquage) ou à celles calculées en corpus (*e.g.* parties du discours, rangs de fréquence). Sa mise en œuvre n'est pas toujours possible, dans la mesure où l'on ne connaît pas nécessairement les proportions de telle ou telle caractéristique pour l'ensemble du dictionnaire (*e.g.* proportions respectives des parties du discours, de la lettre initiale des vedettes…). Les expériences menées par Bukowska (2010) montrent que cette méthode n'améliore pas systématiquement les résultats obtenus par tirage aléatoire.

2.3.4 Échantillonnage systématique

L'échantillonnage systématique consiste à sélectionner un observable tous les N à partir d'un point de départ choisi arbitrairement ou tiré aléatoirement.

9. La comparaison, fondée sur le calcul des fréquences de n-grammes de mots dans les définitions, et sur un clustering de ces n-grammes, est quantitative.

10. Cf. section 3 pour *Usito* et la page `http://atilf.atilf.fr/tlf.htm` pour le TLFi.

11. Le *Wiktionnaire* met en libre accès une archive de l'intégralité de ses articles (*dump*) et peut donc être analysé automatiquement dans son ensemble, sans recourir à l'échantillonnage.

Selon Freeman (1963), cité par Bukowska (2010), cette méthode pose un problème d'ordre théorique, que nous n'étayerons pas ici : le fait que la théorie des probabilités et les statistiques inférentielles ont peu à dire sur la confiance que l'on peut accorder à un échantillonnage non aléatoire. Elle est néanmoins utilisée (quoique rarement à notre connaissance) en métalexicographie. Par exemple, Cormier et Fernandez (2005) la mettent en œuvre afin d'évaluer la potentielle influence de la nomenclature anglaise du *Great French Dictionary* (1688) de Guy Miège sur celui du *Royal Dictionary* (1699) d'Abel Boyer. L'échantillon étudié, qui représente 5% de la partie anglais-français du dictionnaire royal, est constitué en sélectionnant une page sur vingt, à partir de la page 5, tirée au hasard.

2.3.5 Pratique dominante

Malgré les mises en garde de Bukowska contre la constitution d'échantillons par zones contiguës, les métalexicographes semblent continuer de privilégier cette méthode d'échantillonnage. Par exemple, pour étudier les mots composés dans deux dictionnaires du norvégien nynorsk et bokmål, Paulsen (2023) sélectionne cinq échantillons composés de zones contiguës du début de l'alphabet. Francœur (2021), qui mène une étude quantitative et qualitative sur deux dictionnaires bilingues du XVII^{e} siècle afin d'étudier un possible plagiat portant sur la partie anglais-français, travaille sur les 1 061 articles de la lettre F du dictionnaire le plus récent. Le choix de cette tranche, utilisée pour son étude quantitative de la nomenclature, n'est pas justifiée par l'autrice. L'étude qualitative (chiffrée), qui consiste à comparer les éléments microstructurels des articles des deux dictionnaires, porte sur les 100 premiers articles de la lettre F. Il est probable que ces pratiques ne soient pas guidées par une remise en cause du travail de Bukowska, ou plus généralement d'un consensus en statistiques inférentielles, mais reflètent plutôt la méconnaissance de l'un et de l'autre. La sélection par zone contiguë relève plus d'un réflexe hérité d'une tradition métalexicographique que d'un choix conscient. Le seul cas d'échantillonnage que nous avons trouvé où la méthode de sélection par zone contiguë est justifiée par les auteur·rice·s est celui de Ferrett et Dollinger (2021), cité en section 2.3.1, qui réfèrent à l'article de Bukowska et justifient leur renoncement à un échantillonnage sur l'ensemble des zones du dictionnaire par la complexité de mise en œuvre et une possible inadéquation avec l'étude menée.

Dans ce contexte, la sélection d'échantillon par tirage aléatoire fait figure d'exception. Podhajecka (2015), qui tente de mettre au jour le processus éditorial (en particulier les sources utilisées) d'un dictionnaire bilingue anglais-polonais du XIX^{e} siècle, réfère également à Bukowska et sélectionne un échantillon constitué aléatoirement à partir de chaque lettre de l'alphabet. Il est cependant étonnant de ne trouver aucune information sur la taille de l'échantillon, la proportion d'observables tirés de chaque lettre, etc. L'échan-

tillon n'apportant pas de preuve probante de l'utilisation d'autres sources que celles mentionnées par l'auteur du dictionnaire, elle parcourt manuellement l'ensemble du dictionnaire à la recherche d'indices qui pourraient attirer son attention. La conclusion de l'autrice à ce stade de l'article et la démarche adoptée par la suite sont quelque peu déroutantes : « the sample alone did not provide me with strong enough evidence. As combining two different methods has invalidated the quantitative results, only the qualitative ones will be referred to in the sections below ». L'abandon de l'échantillon analysé ne plaide pas contre l'échantillonnage par tirage aléatoire mais trahit un rapport problématique aux données et aux statistiques : soit, une fois une méthodologie définie, on se fie à l'échantillon constitué et analysé, soit la méthodologie doit être remise en cause. On ne peut en effet justifier l'abandon d'un échantillon par le fait qu'il ne fournit pas les preuves qui valideraient une hypothèse de départ.

2.4 Représentativité et estimation par intervalles

Bukowska (2010) critique l'absence de contrôle des métalexicographes sur la représentativité des échantillons :

> Most of the samples in current metalexicographic research are judgmental one-stretch samples based on what metalexicographers intuitively consider reliable and representative, usually without having tested this representativeness in any way.

Outre la méthode de sélection, le critère proposé par Bukowska pour « contrôler » la fiabilité d'un échantillon est le calcul de l'intervalle de confiance. Ce concept statistique popularisé par les sondages électoraux permet d'estimer, avec un certain niveau de confiance, que la valeur réelle d'une caractéristique de la population est comprise dans un intervalle situé autour de la valeur de cette même caractéristique mesurée sur l'échantillon. L'amplitude de l'intervalle dépend de la marge d'erreur. Lorsque la caractéristique à observer est une proportion (*e.g.* proportion d'articles marqués, intention de votes pour un candidat…), on peut formaliser l'énoncé précédent de la façon suivante [12] :

$$p_r \in [p_o - m_e ; p_o + m_e]$$

$$m_e = z \times \text{erreur-type} = z \times \sqrt{\frac{p_o(1-p_o)}{n}}$$

12. Ce calcul s'applique lorsque la distribution de la variable étudiée est normale. Lorsqu'elle ne l'est pas, le calcul s'applique également sous les conditions suivantes : $n \geq 30$, $np_o \geq 5$ et $n(1-p_o) \geq 5$. Lorsque la caractéristique observée est une moyenne, l'erreur-type s'obtient en divisant l'écart-type des moyennes de l'échantillon par la racine carrée de sa taille : $\frac{\sigma}{\sqrt{n}}$.

Avec :

— p_o : proportion observée dans l'échantillon ;

— p_r : proportion réelle dans la population ;

— n : taille de l'échantillon ;

— m_e : marge d'erreur ;

— z : coefficient correspondant au niveau de confiance souhaité dans la table de la loi normale centrée réduite (*e.g.* 1,96 pour un niveau de confiance de 95% ou 2,58 pour un niveau de confiance de 99%).

Nous illustrons l'estimation de la représentativité par intervalles de confiance en section 3.4.

3 Expériences : marquage FIG. et FAM. dans *Usito*

Afin de comparer les pertinences respectives des différentes méthodes d'échantillonnage décrites en section 2, nous nous intéressons ci-après à la proportion d'articles contenant les marques FAM. (familier) et FIG. (figuré) dans le dictionnaire *Usito*. Ces marques ne sont pas spécialement plus intéressantes que d'autres phénomènes qui se prêteraient à des expériences de quantification, mais elles présentent l'avantage d'être suffisamment répandues dans le dictionnaire et de constituer un observable factuel, donc aisément identifiable automatiquement.

En septembre 2022, la section « Tous les mots du dictionnaire [13] » comptait 46 364 vedettes. Parmi elles, nous avons restreint notre corpus aux 31 310 noms, non pour limiter les calculs à effectuer, mais pour éliminer un potentiel facteur explicatif des divergences observées entre les échantillons, notamment entre ceux issus de différentes tranches alphabétiques (cf. section 3.1). En effet, à supposer que la probabilité qu'un article soit marqué dépende de la partie du discours de sa vedette (que cette hypothèse soit fondée ou non), on pourrait expliquer les différences de proportions d'articles marqués observées entre plusieurs échantillons issus de tranches différentes par l'inégale répartition des parties du discours dans l'alphabet. Or, dans la présente étude, seule l'incidence de la méthode d'échantillonnage sur les résultats observés nous intéresse « toutes choses égales par ailleurs ».

13. Cette rubrique ne recense pas les sous-entrées : on y trouve *affirmative*, mais pas *dans l'affirmative*, sous-entrée présente sous *affirmative*. La liste des vedettes est disponible à l'URL suivante : `https://usito.usherbrooke.ca/index/mots`. Étonnamment, c'est la rubrique intitulée « Tous les articles du dictionnaire », disponible à l'URL `https://usito.usherbrooke.ca/index/articles` qui recense les 75 109 vedettes, sous-entrées et renvois. Corbin (2020) relève, en juillet 2020, 73 000 items indexés dans la rubrique « Tous les mots du dictionnaire », alors accessible à l'URL `https://usito.usherbrooke.ca/index/mots`. Il semble donc que les listes correspondant aux deux rubriques aient été interverties – à tort – entre les deux périodes.

Pour chacune des deux marques, il s'agit dans les expériences qui suivent, après avoir calculé automatiquement la proportion d'articles marqués pour l'ensemble du corpus, de générer automatiquement un certain nombre d'échantillons, selon différentes méthodes (en faisant varier un seul facteur à la fois), puis de mesurer pour chaque échantillon la proportion d'articles marqués. En simulant ainsi automatiquement, de façon répétée, les résultats que l'on pourrait obtenir de manière manuelle, nous pourrons nous faire une idée des méthodes d'échantillonnage qui permettent d'atteindre les résultats les plus proches de la réalité (*i.e.* de la proportion réelle des articles marqués parmi les noms de l'ensemble du dictionnaire).

Nous comparons en section 3.1 les résultats des méthodes d'échantillonnage par zone contiguë et par tirage aléatoire, lorsque la sélection est opérée dans une tranche alphabétique donnée. Nous répétons cette expérience en section 3.2 en appliquant ces deux mêmes méthodes à une sélection pratiquée sur l'ensemble du dictionnaire afin de comparer les résultats obtenus à ceux des échantillons issus des tranches alphabétiques. Nous testons en section 3.3 les échantillonnages systématique et probabiliste stratifié, que nous comparons à la meilleure des méthodes identifiées précédemment. Enfin, en section 3.4, nous discutons de manière empirique de la possibilité de « contrôler » la représentativité des échantillons.

3.1 Sélection par zone contiguë *vs* tirage aléatoire

Partant du constat que les métalexicographes tirent généralement leurs échantillons au sein d'une tranche donnée, nous avons, afin de reproduire leur démarche, divisé *Usito* en tranches, préalablement à la génération des échantillons (contigus et aléatoires), chaque tranche étant composée des noms commençant par la même lettre initiale. Le tableau 1 donne le nombre d'articles que contient chaque tranche et le pourcentage que ce nombre représente par rapport à la totalité du corpus. Sans surprise, les tranches sont de tailles très variables. Une première question, pour un·e métalexicographe menant une étude manuelle, serait de savoir de quelle tranche tirer un échantillon, et si ce choix a une incidence sur la fiabilité (la représentativité) des résultats observés (*i.e.* un échantillon issu d'une tranche donnée est-il plus susceptible d'être représentatif de l'ensemble du dictionnaire qu'un échantillon issu d'une autre tranche ?).

Le diagramme en barres de la figure 3 représente, pour chaque tranche, la proportion d'articles comportant la marque FIG. La boîte à moustaches représente la variabilité de ces proportions pour l'ensemble des tranches. Parce que nous traitons automatiquement le dictionnaire dans sa globalité, nous pouvons calculer la proportion réelle d'articles marqués dans notre corpus. Sur les 31 310 articles, 2 663 contiennent la marque, soit 8,51%. Cette proportion est représentée par la ligne horizontale pointillée dans la figure 3. Les proportions d'articles marqués sont très variables d'une tranche à l'autre puisqu'elles fluctuent de

Tranche	Nb articles	% articles	Tranche	Nb articles	% articles
a	3 350	10,70	**n**	647	2,07
b	1 909	6,10	**o**	700	2,24
c	3 855	12,31	**p**	3 248	10,37
d	1 826	5,83	**q**	172	0,55
e	1 004	3,21	**r**	1 669	5,33
f	1 234	3,94	**s**	2 305	7,36
g	1 156	3,69	**t**	1 667	5,32
h	842	2,69	**u**	154	0,49
i	1 061	3,39	**v**	799	2,55
j	290	0,93	**w**	61	0,19
k	165	0,53	**x**	24	0,08
l	945	3,02	**y**	46	0,15
m	2 084	6,66	**z**	97	0,31

Tableau 1 – Répartition par tranches des 31 310 noms (vedettes) d'*Usito*

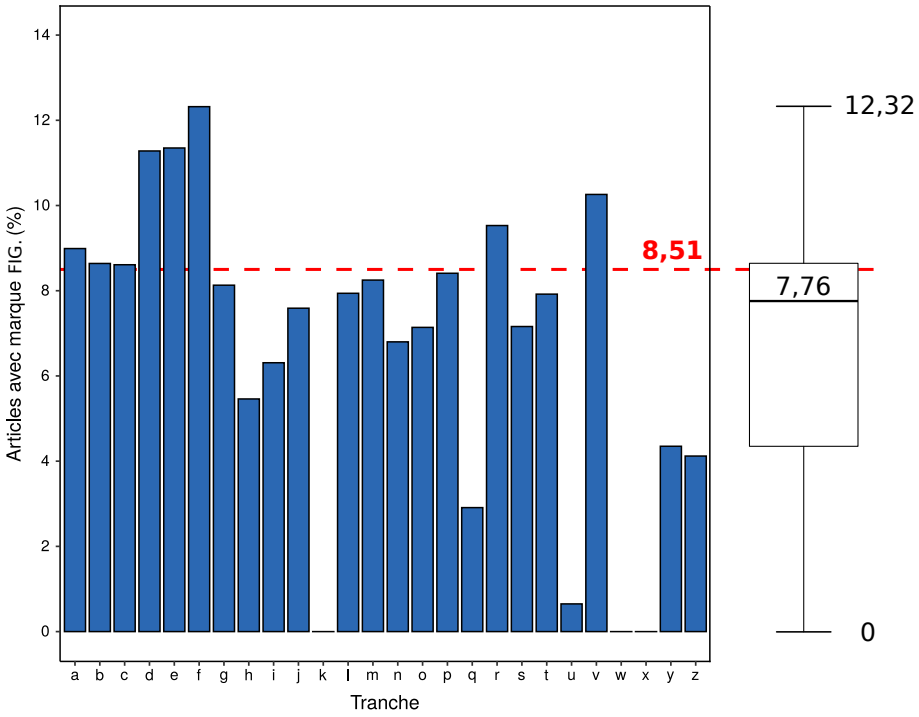


FIGURE 3 – Proportion d'articles du corpus contenant la marque FIG.

0 à plus de 12%. Les tranches affichant une proportion de 0% sont celles qui comportent le moins d'articles. Une grande variabilité s'observe néanmoins entre les « grosses » tranches. On peut donc supposer que le choix de la tranche à partir de laquelle est tiré un échantillon aura un impact sur le résultat. Une lecture intuitive de ce diagramme indique que, pour ce dictionnaire et pour le phénomène observé, les tranches les plus représentatives (et donc, celles dont il conviendrait de tirer les échantillons) sont les tranches *b*, *c* et *p*. Or, d'une part, les métalexicographes qui mènent une étude manuelle n'ont pas accès à cette information. D'autre part, cette lecture est également quelque peu naïve ou, du moins, vaine car non recyclable. Peut-on en effet imaginer que ces tranches soient également les plus représentatives quels que soient le dictionnaire et le phénomène étudiés ? Comparons, dans *Usito*, avec la marque FAM. : 3 106 noms sont marqués, sur les 31 310 du corpus, soit 9,92%. Le diagramme en barres de la figure 4 montre, pour chaque tranche, la proportion d'articles marqués FIG. et FAM., ainsi que les proportions réelles calculées sur la globalité des noms du dictionnaire. Comme pour la marque FIG., on constate que les proportions d'articles marqués FAM. fluctuent énormément d'une tranche à l'autre : entre 3% et 18,5% hors « petites tanches », *i.e.* entre moins du tiers et un peu moins du double de la valeur réelle calculée sur l'ensemble du dictionnaire. On constate surtout que les « bonnes » et les « mauvaises » tranches, *i.e.* celles

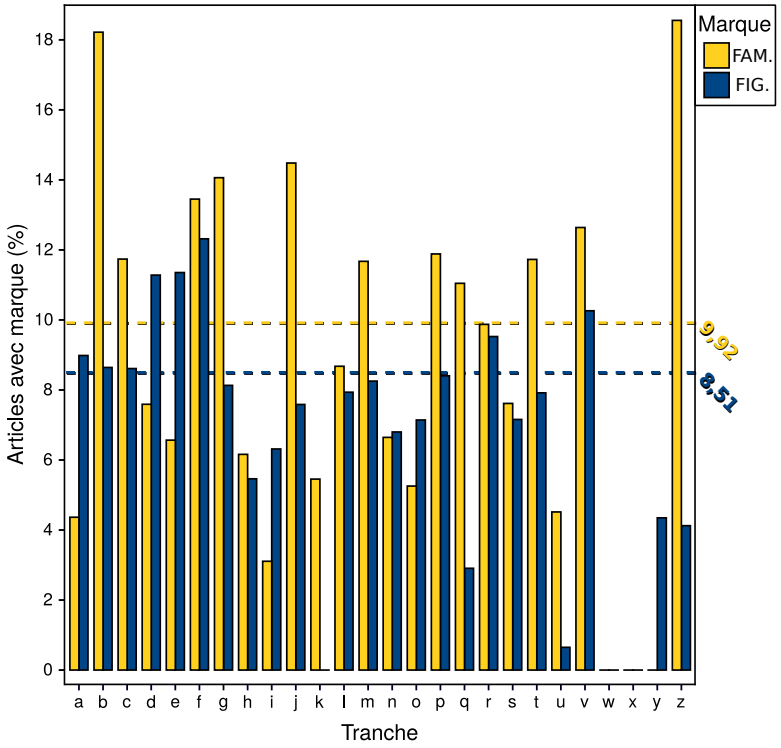


FIGURE 4 – Proportion, globale et par tranche, d'articles marqués FIG. et FAM.

dont les proportions sont respectivement les plus proches et les plus éloignées de la valeur globale réelle, ne sont pas les mêmes que pour la marque FIG. Par exemple, la tranche *b*, qui est pour la marque FIG. l'une de celles dont la proportion d'articles marqués est la plus proche de celle du dictionnaire dans sa globalité, est la tranche dont la proportion est la plus éloignée de la valeur réelle pour le marquage FAM. Cette observation confirme qu'il n'existe pas de « bonne tranche » qui serait plus représentative que les autres pour tous les phénomènes observables.

Afin de comparer les méthodes d'échantillonnage, l'étape suivante consiste à produire les échantillons. Pour chaque tranche, sont générés automatiquement :

— 100 échantillons de 500 articles (maximum [14]) contigus, l'article de départ étant tiré aléatoirement ;

— 100 échantillons de 500 articles (maximum), tous tirés aléatoirement.

Le nombre et la taille des échantillons choisis sont arbitraires mais nous semblent raisonnables. Un échantillon d'une taille de 500 articles à analyser (pour 31 310 noms, soit 1,6% du corpus) semble en effet cohérent avec l'ordre de grandeur des échantillons constitués dans les analyses quantitatives manuelles que l'on trouve dans la littérature. Un nombre de 100 échantillons par tranche, qui correspond à une expérience où l'on demanderait à un·e métalexicographe de répéter 100 fois son analyse, avec un nouvel échantillon à chaque fois, assure en outre une variabilité suffisante pour mener les comparaisons entre méthodes d'échantillonnage. [15] Pour chaque échantillon, la proportion d'articles marqués est ensuite mesurée. À titre d'exemple, le diagramme en barres de la figure 5 représente les proportions d'articles portant la marque FIG. dans les 100 échantillons composés d'articles contigus de la tranche *a* (la ligne pointillée rouge d'ordonnée 8,51 représente la proportion pour le dictionnaire ; la ligne jaune en pointillés mixtes d'ordonnée 8,99 représente la proportion de la tranche *a*).

Notons pour cet exemple à quel point les valeurs obtenues fluctuent d'un échantillon à l'autre (moins de 5% à plus de 15%, alors que les valeurs réelles pour cette tranche et pour l'ensemble du dictionnaire sont de 8,99% et 8,51%), bien qu'ils soient tirés de la même tranche et par la même méthode. C'est le simple fait de sélectionner dans la tranche un point de départ différent pour constituer les échantillons d'articles contigus qui est responsable de cette importante variabilité.

14. Lorsqu'une tranche contient moins de 500 articles, l'échantillon est constitué de l'ensemble de la tranche. Dans ce cas, les 100 échantillons, tous identiques, afficheront la même proportion d'articles marqués, *i.e.* celle de la tranche entière.

15. Les expériences ont été réitérées en faisant varier le nombre et la taille des échantillons. Comme on peut s'y attendre, plus la taille des échantillons est importante, plus les proportions obtenues sont proches de la proportion réelle, et plus le nombre d'échantillons est élevé, plus la variabilité des distributions est grande. Pour le reste, les résultats obtenus vont toujours dans la même direction que celle des résultats présentés ci-après.

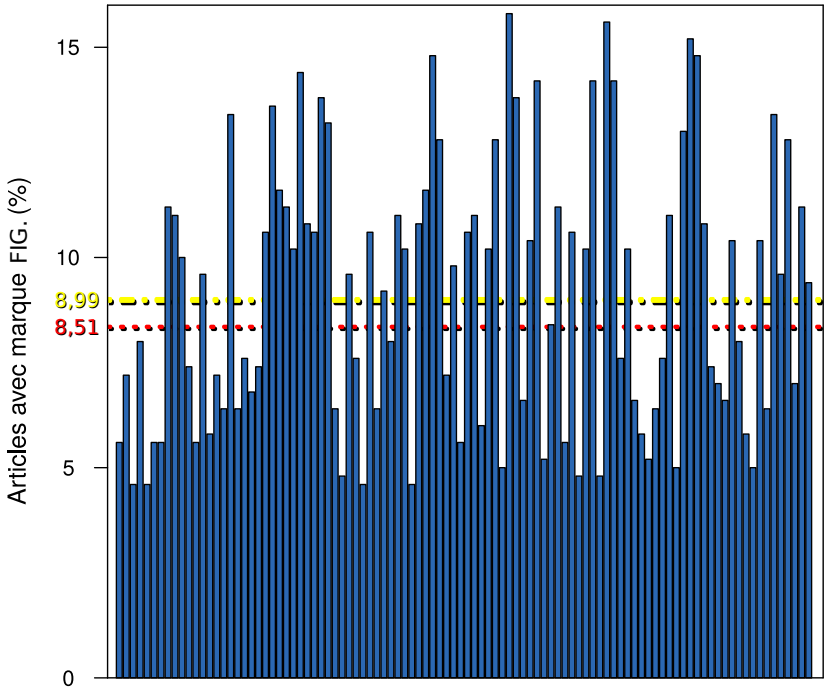


FIGURE 5 – Proportion d'articles marqués FIG. pour les 100 échantillons de 500 articles contigus de la tranche *a*

Il reste maintenant à comparer les résultats obtenus par les deux méthodes d'échantillonnage. Dans la figure 6, qui reprend la figure 3, les boîtes à moustaches représentent, pour chaque tranche, la variabilité des proportions calculées pour les échantillons d'articles contigus (en marron, boîte de gauche) et pour ceux obtenus par tirages aléatoires (en jaune, boîte de droite).

Chaque boîte à moustaches représente donc la variabilité des proportions pour les 100 échantillons générés par une méthode donnée au sein d'une tranche donnée. Par exemple, c'est la variabilité des 100 échantillons dont les proportions sont représentées en figure 5 (pour rappel : 100 échantillons de 500 articles contigus issus de la tranche *a*) qui est représentée par la boîte à moustaches marron la plus à gauche de la figure 6. Les tranches pour lesquelles les boîtes à moustaches sont réduites à un unique trait horizontal (ce qui indique une absence de variabilité) sont celles dont la taille est inférieure à 500 (et dont les échantillons extraits sont tous identiques).

Concernant les méthodes d'échantillonnage, on observe que les boîtes à moustaches marrons, qui correspondent à la sélection contiguë, sont plus étalées verticalement que les boîtes à moustaches jaunes (sélection aléatoire). Les résultats sont donc moins homogènes pour la première méthode, ce qui signifie qu'en tirant deux échantillons d'articles contigus au sein d'une même tranche, la différence entre leurs proportions respectives observées aura tendance à

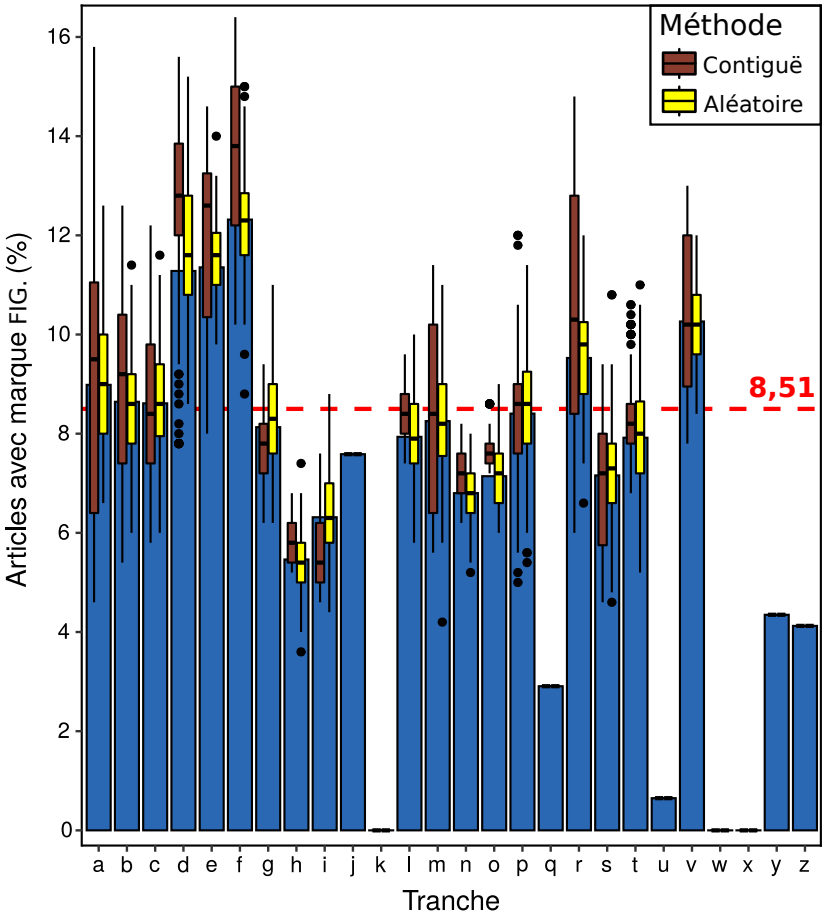


FIGURE 6 – Variabilité des proportions d'articles contenant la marque FIG. dans les échantillons, par tranche et par méthode d'échantillonnage

être plus marquée qu'entre deux échantillons tirés aléatoirement. La méthode d'échantillonnage par zone contiguë est donc moins robuste que celle par tirage aléatoire. Plus important encore, on constate que les boîtes à moustaches correspondant à la sélection contiguë sont moins centrées sur les valeurs réelles des tranches et offrent donc des estimations moins fiables (*i.e.* moins proches de la réalité). Ces deux observations nous conduisent à abandonner la méthode par zone contiguë dans la suite des expériences.

3.2 Source des échantillons : tranche *vs* globalité du dictionnaire

Si l'on se concentre sur la meilleure des deux méthodes étudiées en section 3.1 (le tirage aléatoire), on observe que, même pour les « grosses » tranches, la variabilité des proportions mesurée entre les différents échantillons reste importante (cf. fig. 6). La proportion obtenue pour un échantillon pourra donc

conduire à une estimation éloignée de la réalité, qu'on la compare à la proportion réelle de la tranche dont est issu l'échantillon ou à la proportion réelle calculée sur l'ensemble du dictionnaire : alors que cette dernière est de 8,51%, les valeurs obtenues pour les échantillons tirés aléatoirement, pour les « grosses » tranches, oscillent entre moins de 4% et plus de 15%, c'est-à-dire entre quasiment la moitié et le double de la valeur réelle.

Nous avons jusqu'ici sélectionné chaque échantillon au sein d'une même tranche car, comme nous l'avons écrit en section 3.1, c'est la manière dont les métalexicographes procèdent la plupart du temps. Nous proposons de comparer les résultats déjà observés à ceux obtenus en générant, pour chacune des deux marques FIG. et FAM., 100 échantillons supplémentaires sélectionnés aléatoirement à partir de l'ensemble des noms du dictionnaire, indépendamment de leur initiale [16]. Les caractéristiques (quartiles, valeurs minimales, médianes, moyennes et maximales) des distributions correspondantes sont données dans le tableau 2.

Marque	**% réel global**	**Distribution des échantillons**					
		Min.	**Q1**	**Médiane**	**Moyenne**	**Q3**	**Max.**
FIG.	8,51	5,80	7,60	8,00	8,31	9,00	12,00
FAM.	9,92	6,40	9,15	9,90	9,82	10,60	12,40

Tableau 2 – Tirage aléatoire sur l'ensemble du dictionnaire : caractéristiques des distributions

Les figures 7 et 8 représentent quant à elles la variabilité des proportions d'articles marqués dans les échantillons ainsi générés à partir de la totalité du dictionnaire, en comparaison de la variabilité de ceux tirés au sein des différentes tranches. Les boîtes à moustaches bleues, à droite des figures, représentent les distributions dont les caractéristiques sont données dans le tableau 2. Pour la marque FIG., 5 des 26 tranches (*b*, *c*, *g*, *m* et *p*) produisent des échantillons dont la distribution est plus centrée sur la proportion réelle d'articles marqués (cf. fig. 7). Les 21 autres tranches de la marque FIG. et l'intégralité des tranches de la marque FAM. (cf. fig. 8) produisent des échantillons dont la distribution est moins bien centrée sur la valeur réelle. Ce résultat illustre qu'il est préférable de constituer un échantillon par tirage aléatoire sur l'ensemble du dictionnaire qu'au sein d'une tranche donnée.

3.3 Tirages aléatoire, probabiliste stratifié et systématique

Après avoir déterminé, en section 3.1, que l'échantillonnage par tirage aléatoire est préférable à la sélection par zone contiguë et, en section 3.2, qu'un

16. Dans le cas de l'analyse manuelle d'un dictionnaire papier, une mise en œuvre possible est de tirer aléatoirement un numéro de page, puis l'article dans la page, et d'itérer autant de fois que d'observables à sélectionner.

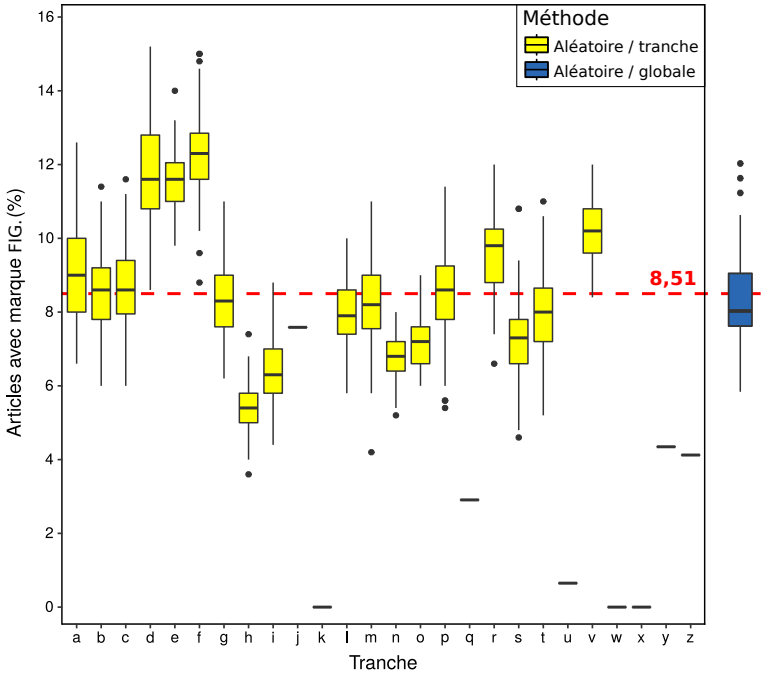


FIGURE 7 – Marque FIG. : échantillonnage par tranche *vs* sur tout le dictionnaire

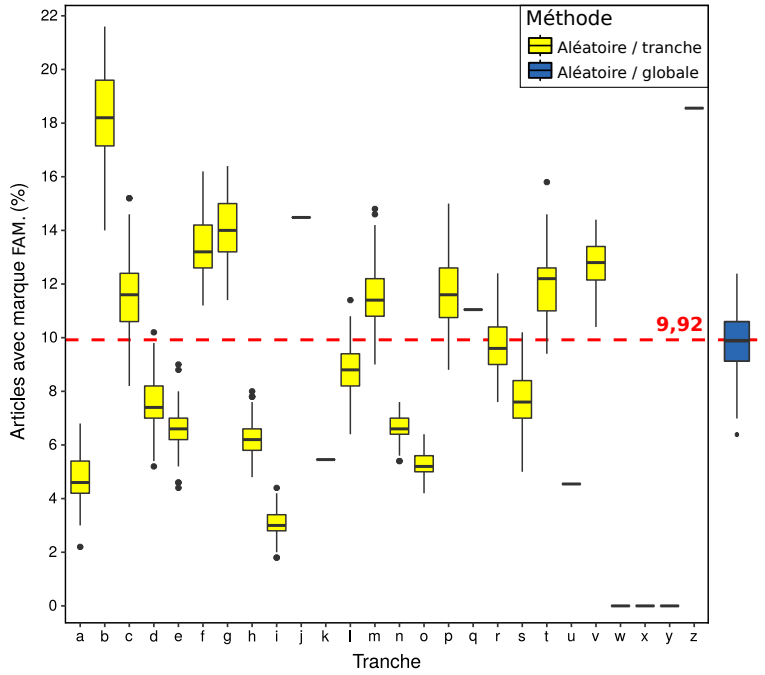
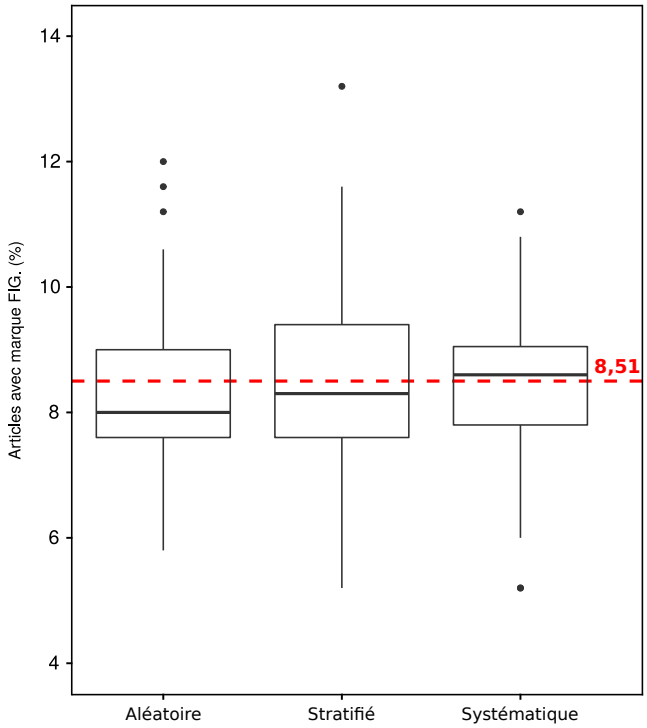


FIGURE 8 – Marque FAM. : échantillonnage par tranche *vs* sur tout le dictionnaire

tirage aléatoire sur l'ensemble du dictionnaire est préférable à ce même tirage pratiqué au sein d'une tranche donnée, nous testons ici les échantillonnages systématique et probabiliste stratifié. Pour chacune de ces deux méthodes, 100 nouveaux échantillons de 500 articles sont générés. Pour l'échantillonnage systématique, cela consiste à sélectionner un nom tous les 62 (= 31 310 / 500), en décalant d'un article le point de départ pour chaque échantillon. Pour la méthode probabiliste stratifiée, l'échantillon est constitué en pratiquant un tirage aléatoire qui respecte les proportions de noms par lettre initiale données dans le tableau 1, section 3.1. [17] Les distributions des proportions d'articles marqués dans ces nouveaux échantillons sont représentées, à côté de celle obtenue pour les échantillons générés par tirage aléatoire, dans la figure 9 pour la marque FIG. et dans la figure 10 pour la marque FAM.

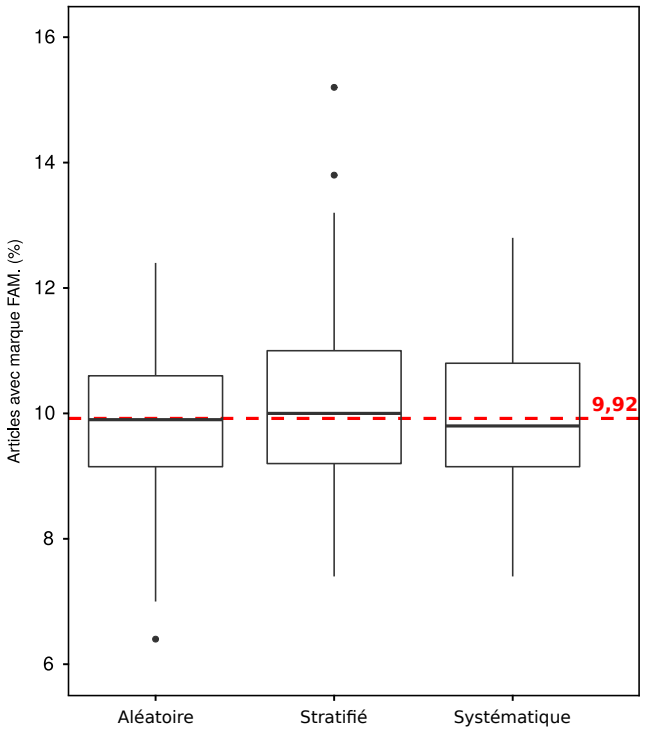


Tirage	Min	Q1	Médiane	Moyenne	Q3	Max
Aléatoire	5.80	7.60	8.00	8.31	9.00	12.00
Stratifié	5.20	7.60	8.30	8.51	9.40	13.20
Systématique	5.20	7.80	8.60	8.52	9.05	11.20

FIGURE 9 – Marque FIG. : comparaison des méthodes d'échantillonnage sur tout le dictionnaire

17. Il s'agit donc de tirer aléatoirement 54 ($500 \times 10,7\%$) articles de la tranche *a*, 30 ($500 \times 6,1\%$) articles de la tranche *b*, etc.

Avec une médiane de 8,3 %, l'échantillonnage probabiliste stratifié a globalement un léger avantage sur le tirage aléatoire (médiane de 8 %) pour la marque FIG. (valeur réelle : 8,51%), mais dégrade très légèrement les résultats pour la marque FIG. (médianes respectives de 10 % pour l'échantillonnage probabiliste stratifié et de 9,9 % pour le tirage aléatoire, avec une valeur réelle de 9,92 %). Dans ses expériences, Bukowska (2010) constate de la même manière que l'échantillonnage probabiliste stratifié n'améliore pas systématiquement le tirage aléatoire. Quant à l'échantillonnage systématique, il produit les meilleurs résultats pour la marque FIG. et les moins bons pour la marque FAM. si l'on considère uniquement les valeurs médianes, les résultats obtenus avec les trois méthodes étant néanmoins, pour cette dernière marque, très proches de la proportion réelle.



Tirage	Min	Q1	Médiane	Moyenne	Q3	Max
Aléatoire	6,40	9,15	9,90	9,82	10,60	12,40
Stratifié	7,40	9,20	10,00	10,17	11,00	15,20
Systématique	7,40	9,15	9,80	9,97	10,80	12,80

FIGURE 10 – Marque FAM. : comparaison des méthodes d'échantillonnage sur tout le dictionnaire

Les résultats empiriques de cette comparaison ne permettent pas de conclure de manière définitive quant à la préférence à accorder à l'une ou l'autre des trois méthodes d'échantillonnage. Notons que s'il existe une réserve d'ordre théorique sur l'échantillonnage systématique, que nous avons rapportée en section 2.3.4, c'est, pour les deux marques, la méthode d'échantillonnage qui produit les distributions avec les variabilités les plus réduites. C'est donc, pour cette expérience, la méthode la plus robuste. À l'inverse, c'est l'échantillonnage probabiliste stratifié qui produit, pour les deux marques, les distributions affichant une plus grande variabilité et qui semble donc, pour cette expérience, la moins robuste. Rappelons enfin une considération pratique : nous avons pu tester l'échantillonnage probabiliste stratifié, en choisissant comme strates les tranches alphabétiques correspondant aux lettres initiales des vedettes, parce que nous pouvions accéder (automatiquement) à la nomenclature du dictionnaire étudié. La mise en œuvre de l'échantillonnage probabiliste stratifié est conditionnée par la possibilité d'accéder à ce type d'information, ce qui n'est pas toujours possible.

3.4 Contrôle de la représentativité des échantillons

Dans l'expérience décrite en section 3.2, la méthode la plus fiable (tirage aléatoire sur tout le corpus) pour estimer le taux de marquage FAM. produit une distribution dont la médiane (9,9%) se confond quasiment avec la valeur effective de la proportion (9,92%) calculée sur l'ensemble des noms du dictionnaire. Cependant, parmi les 100 échantillons de 500 articles générés aléatoirement, ceux affichant les proportions minimales et maximales (respectivement 6,4% et 12,4%, cf. tableau 2) donneraient une vision relativement éloignée de la réalité à un·e métalexicographe malchanceux·euse qui les aurait tirés (rappelons qu'un·e métalexicographe ne constitue généralement qu'un seul échantillon pour estimer une caractéristique du dictionnaire). Mais peut-on réellement parler de malchance, *i.e.* quelle est la probabilité de tirer un échantillon conduisant à une estimation aussi inexacte ? L'histogramme de la figure 11 représente le nombre d'échantillons répartis par tranches de pourcentages d'articles marqués. [18]

Sans surprise, la distribution est normale et centrée sur la valeur réelle : la plupart des échantillons fournissent des estimations relativement proches de la réalité et quelques-uns des estimations éloignées. Mais les métalexicographes n'ont pas de moyen de déterminer avec certitude si la proportion observée (mesurée) p_o sur leur échantillon est plus ou moins fiable. La réponse de Bukowska (2010) est celle qui s'impose à toute personne pratiquant l'échantillonnage statistique : le recours aux intervalles de confiance, mentionnés en section 2.4. Le tableau 3 donne, pour les deux échantillons présentant les proportion minimales

18. Il se lit comme suit : parmi les 100 échantillons générés aléatoirement, un seul affiche une proportion d'articles marqués comprise entre 5,5 et 6,5 ; 2 échantillons ont une proportion comprise entre 6,5 et 7,5 ; 12 ont une proportion comprise entre 7,5 et 8,5 ; etc.

et maximales observées (p_{min} et p_{max}), ainsi que pour un échantillon (fictif) dont la proportion observée serait strictement égale à la proportion réelle (p_r), les marges d'erreur au niveau de confiance 95% et les intervalles de confiance correspondant.

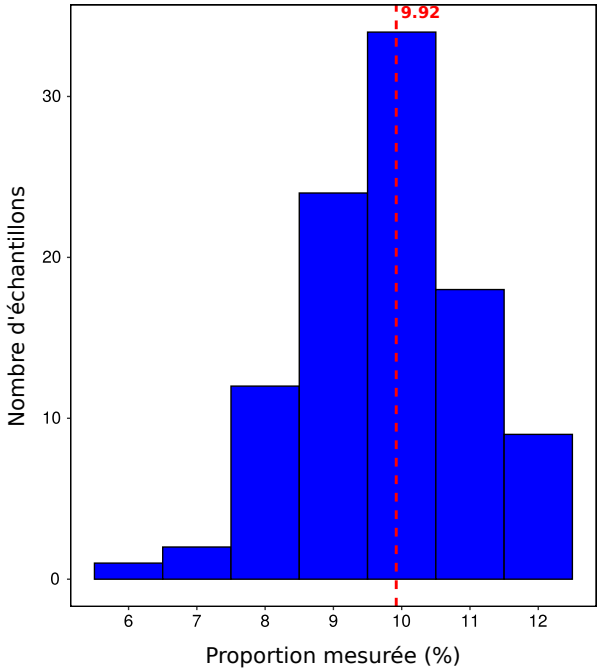


FIGURE 11 – Tirage aléatoire sur tout le dictionnaire - nombre d'échantillons par tranche de pourcentages d'articles marqués FAM.

proportion observée (%)	**marge d'erreur (%)**	**intervalle de confiance (%)**
6,40 ($p_o = p_{min}$)	2,15%	**[4,25; 8,55]**
9,92 ($p_o = p_r$)	2,62%	[7,3; 12,54]
12,40 ($p_o = p_{max}$)	2,89%	[9,51; 15,29]

Tableau 3 – Marque FAM. - marges d'erreur au niveau de confiance 95% et intervalles de confiance des échantillons

Ces données se lisent de la façon suivante (un·e métalexicographe ne formulerait qu'un seul énoncé correspondant à l'échantillon tiré) :

— ($p_o = p_{min}$) il y a 95% de chances que la proportion de noms marqués FAM. dans *Usito* soit comprise entre 4,25% et 8,55%

— ($p_o = p_r$) il y a 95% de chances que la proportion de noms marqués FAM. dans *Usito* soit comprise entre 7,3% et 12,54%

— ($p_o = p_{max}$) il y a 95% de chances que la proportion de noms marqués FAM. dans *Usito* soit comprise entre 9,51% et 15,29%

On constate que, pour l'échantillon présentant la valeur minimale de 6,4%, l'intervalle de confiance ne contient pas la valeur réelle (9,92%). Une question se pose : combien d'échantillons sont dans ce cas, *i.e.* pour combien d'échantillons la valeur réelle est-elle en dehors de l'intervalle de confiance ? La figure 12 représente, pour les 100 échantillons de 500 articles générés pour notre expérience, la proportion observée pour chaque échantillon et l'intervalle de confiance au niveau de confiance 95%. La ligne horizontale d'ordonnée 9,92 correspond à la proportion réelle calculée sur tout le corpus. On voit que la valeur réelle est en dehors des intervalles de confiance de seulement 3 échantillons (représentés en jaune). Pour cette expérience, les chances (97%) de tirer un échantillon dont l'intervalle de confiance contient la valeur réelle sont bien supérieures ou égales au niveau de confiance choisi (95%). Peut-on dire pour autant que l'on « contrôle » ainsi la représentativité d'un échantillon ? D'une part, les métalexicographes n'ont aucun moyen de savoir si leur échantillon fait partie des 95 « bons » pourcents théoriques ou des 5 « mauvais » [19]. D'autre part, même dans le cas idéal où la valeur observée correspond exactement à la valeur réelle, *i.e.* 9,92% (information à laquelle les métalexicographes n'ont pas accès), doit-on se réjouir de pouvoir affirmer que la valeur réelle est probablement comprise entre 7,3% et 12,54% ? Pour cette taille d'échantillon analysé, et, dans une moindre mesure, pour cette proportion observée, la marge d'erreur et l'amplitude et l'intervalle de confiance qui en résulte nous paraissent bien étendus (on s'en rend compte en comparant l'empan vertical des intervalles de confiance à celui des barres représentant les proportions mesurées). Pour diminuer la marge d'erreur, au même niveau de confiance (95%), la solution est d'augmenter la taille de l'échantillon examiné. Pour la proportion $p_o = 9,92\%$, obtenir une marge d'erreur qui n'excède pas 1% nécessiterait, au même niveau de confiance, de sélectionner un échantillon constitué de 3 500 articles [20] (ce qui n'empêcherait pas le risque d'obtenir un intervalle de confiance ne contenant pas la proportion réelle avec une probabilité de 5%). Pour avoir 99% de chances d'obtenir une marge d'erreur qui n'excède pas 1%, c'est un échantillon de près de 6 000 articles qu'il faudrait constituer.

Revenons à nos 100 échantillons d'une taille plus réaliste de 500 articles chacun. Afin d'estimer, dans notre expérience, l'erreur probable entre la proportion mesurée sur un échantillon e et la proportion réelle, on définit le « taux d'erreur » [21] de la manière suivante :

$$T_e = \frac{|p_o - p_r|}{p_r}$$

19. Ou, dans la pratique, pour cette expérience, des 97 « bons » pourcents ou des 3 « mauvais ».
20. Cf. la formule donnée en section 2.4.
21. Les guillemets signalent ici que cette notion, comme les représentations qui en sont données en figures 13 et 14 ne sont pas standards en statistiques.

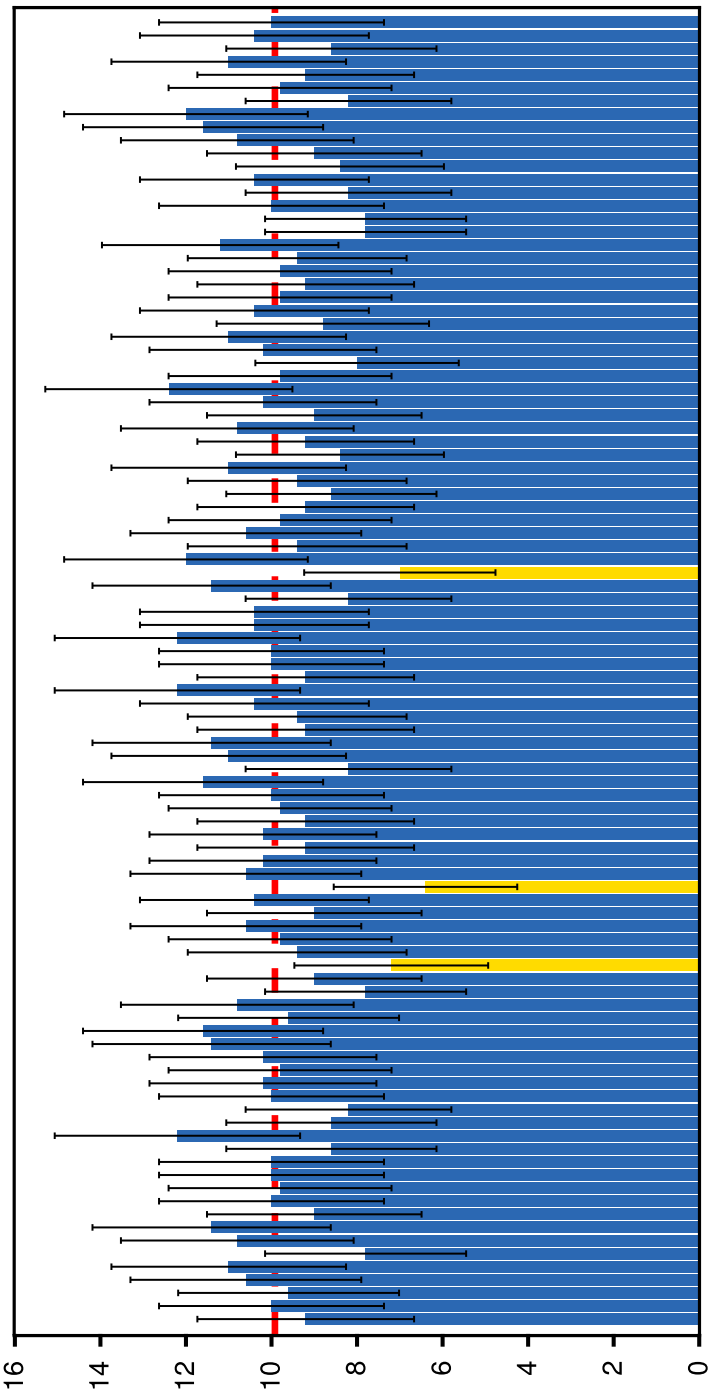


FIGURE 12 – Marque FAM. - intervalles de confiance au niveau de confiance 95% des 100 échantillons de 500 articles

Ce taux d'erreur exprime le rapport entre l'erreur d'estimation [22] et la valeur réelle. La boîte à moustaches de la figure 13 représente la variabilité de ce taux d'erreur pour les 100 échantillons. Elle montre que le taux d'erreur fluctue, selon les échantillons, entre 0,8% et 35,5%, avec une valeur médiane de 7,26%.

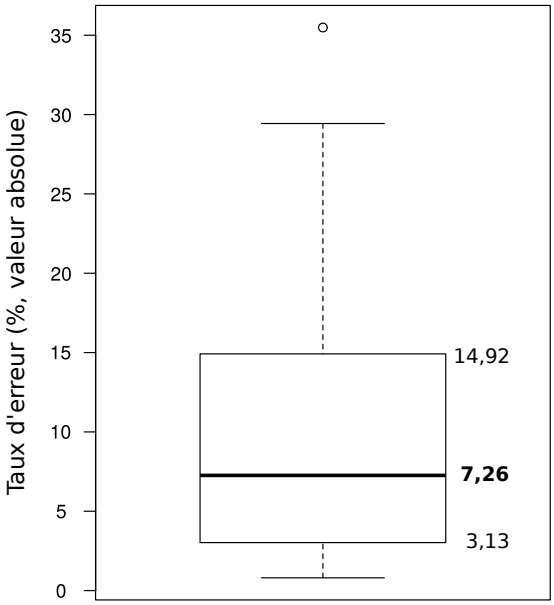


FIGURE 13 – Marque FAM., tirage aléatoire sur tout le dictionnaire - variabilité des taux d'erreur réels des échantillons

Une autre représentation de ces résultats est proposée en figure 14, qui illustre la probabilité qu'un des échantillons, parmi les 100 tirés aléatoirement dans notre expérience, affiche un taux d'erreur inférieur à un seuil donné. L'interprétation que l'on peut donner de la courbe représentée est qu'un échantillon tiré au hasard parmi les 100 échantillons générés a :

— 10% de chances d'afficher un taux d'erreur inférieur à 1% ;

— 19% de chances d'afficher un taux d'erreur inférieur à 2% ;

— 34% de chances d'afficher un taux d'erreur inférieur à 5% ;

— 61% de chances d'afficher un taux d'erreur inférieur à 10% ;

— 87% de chances d'afficher un taux d'erreur inférieur à 20% ;

— 100% de chances d'afficher un taux d'erreur inférieur à 36%.

Cette lecture est celle qui consiste à voir le verre à moitié plein. Une autre lecture possible consiste à donner l'interprétation réciproque qu'un échantillon a, dans cette expérience :

22. Différence, en valeur absolue, entre la valeur mesurée sur l'échantillon et la valeur réelle sur l'ensemble du corpus.

— 90% de chances d'afficher un taux d'erreur supérieur à 1% ;
— 81% de chances d'afficher un taux d'erreur supérieur à 2% ;
— 66% de chances d'afficher un taux d'erreur supérieur à 5% ;
— 39% de chances d'afficher un taux d'erreur supérieur à 10% ;
— 13% de chances d'afficher un taux d'erreur supérieur à 20% ;
— aucune chance d'afficher un taux d'erreur supérieur à 36%.

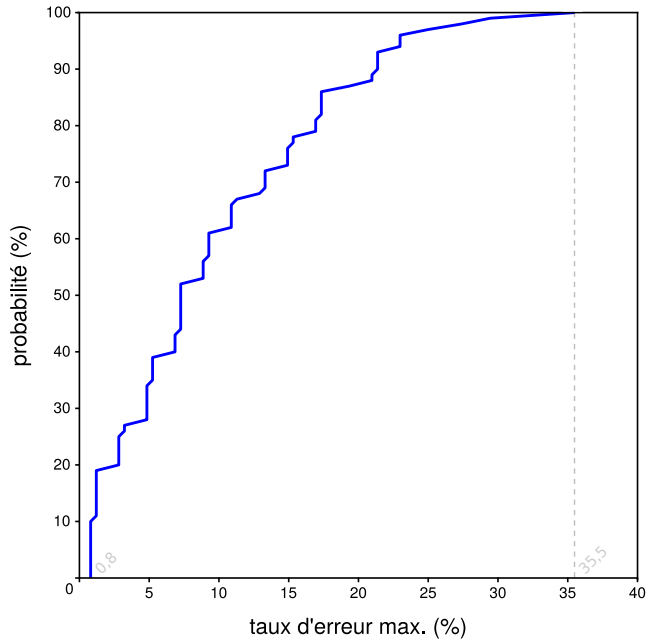


FIGURE 14 – Probabilité, pour un des 100 échantillons tirés, d'afficher un taux d'erreur inférieur à un seuil donnée

Ce « risque » pourra être jugé plus ou moins acceptable, mais gardons à l'esprit que les résultats présentés ici à titre d'illustration ne valent que pour l'expérience menée. La courbe représentée en figure 14 n'est pas une loi de probabilité générale : elle n'est pertinente que pour le phénomène étudié (marquage FAM.) observé dans le dictionnaire choisi (*Usito*) avec les paramètres de l'échantillonnage mis en œuvre (méthode de sélection et taille des échantillons), mais aussi pour le « hasard du moment » : un autre tirage aléatoire reproduisant l'ensemble de ces conditions expérimentales conduirait à des résultats différents. Or si, dans une certaine limite (à définir), un taux d'erreur peut être jugé raisonnable, c'est bien le manque de contrôle sur ce taux (et donc sur la représentativité des échantillons sélectionnés et analysés) qui ne l'est pas. Bukowska se satisfait de l'estimation par intervalles de confiance comme moyen de « contrôler » la représentativité. Dans la mesure où il n'existe pas d'autre moyen statistique qui en permette un meilleur contrôle, nous devons

bien nous en contenter aussi. Mais, avant de se lancer dans une estimation quantitative par échantillonnage, il convient de bien mesurer les efforts à fournir pour analyser un échantillon dont la taille permettra de pouvoir prétendre à une certaine représentativité, cette représentativité étant – seulement – plus ou moins probable.

4 Synthèse et discussion

Cet article, destiné avant tout aux métalexicographes, se fixait deux objectifs : d'une part, celui de montrer empiriquement, à travers des expériences portant spécifiquement sur l'analyse de dictionnaire, l'importance du choix des méthodes d'échantillonnage ; d'autre part, de relativiser le « contrôle » que l'on peut exercer sur les échantillons et, partant, sur les estimations que l'on peut effectuer à partir des résultats mesurés sur ces échantillons. Des expériences décrites en sections 3.1 à 3.4, nous retiendrons que :

1. la méthode d'échantillonnage par zone contiguë est la plus problématique. Un tirage aléatoire lui est dans tous les cas préférable. L'échantillonnage probabiliste stratifié, plus complexe – et pas toujours possible – à mettre en œuvre, de même que l'échantillonnage systématique, peuvent, selon les cas, améliorer ou dégrader les résultats ;
2. sauf à vouloir comparer deux parties spécifiques d'un dictionnaire, échantillonner sur l'intégralité du dictionnaire est plus fiable qu'échantillonner au sein d'une tranche donnée ;
3. même les méthodes d'échantillonnage les moins problématiques n'offrent pas de garantie absolue que l'échantillon observé soit représentatif : le calcul des intervalles de confiance n'apporte pas de solution totalement satisfaisante pour « contrôler » cette représentativité.

Faut-il alors, à l'aune de cette dernière assertion, renoncer aux analyses par échantillonnage en métalexicographie ? Sur la question de la représentativité des échantillons, on peut prêter aux métalexicographes chevronné·e·s une certaine expérience de la discipline et une certaine familiarité avec les dictionnaires étudiés qui pourraient les alerter lorsqu'un échantillon leur paraît trop atypique de la représentation qu'ils se sont forgée du dictionnaire. Mais c'est bien le danger : on aura toujours tendance à faire confiance à un échantillon qui va dans le sens d'une intuition fondée sur une observation récurrente. Serait-il en revanche plus sage, en tirant un échantillon qui infirme son intuition, de récuser l'échantillon (et d'en tirer de nouveaux, jusqu'à ce que l'un d'entre eux se décide à rentrer dans le rang) ou de remettre en cause son intuition ? À qui pencherait en faveur de la première option, nous demanderions quel est alors l'intérêt de recourir aux statistiques.

Nous avons mentionné, en section 2.2, des études menées sur des échantillons dont la priorité n'est pas, à proprement parler, de quantifier des phé-

nomènes, mais de montrer leur existence et leurs effets. Ces études entrent selon nous dans la catégorie des études qualitatives (quand bien même certaines portent sur une masse de donnée extrêmement importante). Concernant les études quantitatives, devant l'ensemble des problèmes méthodologiques que posent les techniques d'échantillonnage, nous concluons que les analyses menées sur l'intégralité des observables sont à privilégier à chaque fois que leur mise en œuvre est possible [23]. C'est le cas pour les dictionnaires disponibles au format numérique et dont le statut légal est suffisamment permissif. Dans le cas contraire, il convient de mettre en œuvre les méthodes les moins problématiques (tirage aléatoire, probabiliste stratifié ou systématique sur l'intégralité du dictionnaire d'un échantillon de taille suffisante selon le niveau de confiance souhaité) et, malgré ces précautions, de considérer les résultats obtenus avec toute la circonspection qu'ils méritent.

Remerciements

L'auteur tient à remercier les évaluateur·rice·s anonymes pour leurs suggestions et commentaires constructifs.

Bibliographie

BUKOWSKA, Agnieszka A. (2010). Sampling techniques in metalexicographic research. Dans DYKSTRA, Anne et SCHOONHEIM, Tanneke (dir.), *Proceedings of the 14th EURALEX International Congress*, p. 1258–1269. Leeuwarden/Ljouwert, The Netherlands.

CORBIN, Pierre (1984). « Lexicographe-conseil ». *Lez Valenciennes. Cahiers de l'UER Froissart*, n° 9, p. 113–121.

CORBIN, Pierre (1990). « Le monde étrange des dictionnaires (7). Logique linguistique et logique botanique : problèmes posés par la définition d'une classe de mots dérivés français ». *Cahiers de lexicologie*, n° 56-57, p. 75–108.

CORBIN, Pierre (1995). « Le monde étrange des dictionnaires (8). Du *Petit Robert* (1967) au *Micro Robert* (1971) : le recyclage de citations ». *Lexique*, n° 12/13, p. 125–145.

CORBIN, Pierre (2020). « Les dictionnaires monolingues généraux du français actuel gratuits en ligne : évolutions récentes (2020) ». *Academic Journal of Modern Philology*, n° 9, p. 65–77.

CORMIER, Monique C. et FERNANDEZ, Heberto (2005). « From the Great French Dictionary (1688) of Guy Miège to the Royal Dictionary (1699) of

23. Cette conclusion rejoint celle formulée en conclusion de l'article intitulé « Pour une analyse qualitative *et* quantitative, manuelle *et* computationnelle, synchronique *et* diachronique, des dictionnaires numériques », dans ce même collectif.

Abel Boyer: Tracing Inspiration ». *International Journal of Lexicography*, n° 18(4), p. 479–507.

FERRETT, Emma et DOLLINGER, Stefan (2021). « Is digital always better? Comparing two English print dictionaries with their digital counterparts ». *International Journal of Lexicography*, n° 34(1), p. 66–91.

FRANCŒUR, Aline (2021). « Forging a New Path in English-French Lexicography: Guy Miège in Relation to Robert Sherwood ». *International Journal of Lexicography*, n° 34(4), p. 437–452.

FREEMAN, Harold (1963). *Introduction to statistical inference*. Rearing, MA, Addison-Wesley Publishing Company.

KAMIŃSKI, Mariusz Piotr (2015). « In Search of Lexical Discriminators of Definition Style: Comparing Dictionaries through N-Grams ». *International Journal of Lexicography*, n° 29(4), p. 403–423.

LEHMANN, Alise (1995). « Du *Grand Robert* au *Petit Robert* : les manipulations de la citation littéraire ». *Lexique*, n° 12/13, p. 105–124.

MARTINEZ, Camille (2011). « Le poids des contraintes dictionnairiques sur l'évolution des marqueurs dans les *Petit Larousse* (1997–2007) ». Dans BAIDER, Fabienne, LAMPROU, Efi, et MONVILLE-BURSTON, Monique (dir.), *La marque en lexicographie : états présents, voies d'avenir*, p. 39–50. Limoges, Lambert-Lucas.

MARTINEZ, Camille (2013). « La comparaison de dictionnaires comme méthode d'investigation lexicographique ». *Lexique*, n° 21, p. 193–220.

OSSELTON, Noel E. (2007). « Alphabet Fatigue and Compiling Consistency in Early English Dictionaries ». Dans CONSIDINE, John et IAMARTINO, Giovanni (dir.), *Words and Dictionaries from the British Isles in Historical Perspective*, p. 81–90. Newcastle, Cambridge Scholars Publishing.

PAULSEN, Mikkel Ekeland (2023). « Wheat or Chaff? A Compound Selection Model Based on Look-Up Data ». *International Journal of Lexicography*, n° 36(3), p. 306–324.

PODHAJECKA, Mirosława (2015). « Erazm Rykaczewski's A Complete Dictionary English and Polish... (1849): Uncovering the Compilation Process ». *International Journal of Lexicography*, n° 29(1), p. 1–30.

RADERMACHER, Ruth (2004). *Le* Trésor de la Langue Française*. Une étude historique et lexicographique*. Thèse de doctorat, Université Marc Bloch, Strasbourg.