Vers une estimation robuste des proportions lexicales

Guillaume Guex, Aris Xanthos

Université de Lausanne – {guillaume.guex, aris.xanthos}@unil.ch

Abstract

Many methodological papers concerned with the measurement of lexical diversity have studied the dependence between variety (number of different words) and sample size. The question is also relevant for the related concept of lexical proportion (or relative type frequency), i.e. the proportion of word types that possess a given property, since it is defined as a ratio between two varieties. Yet the issue has not been specifically addressed in the literature and the use of "raw" lexical proportions is attested in recent works performing statistical analysis of textual data in various disciplines. The present contribution seeks to answer the following questions: (i) what is the extent and nature of the dependence of raw lexical proportion on sample size? (ii) to what extent does an approach based on the resampling principles in use for lexical diversity measurement make it possible to estimate lexical proportion in a way that is less dependent on sample size? From an applied perspective, the goal of this research is to provide statistical textual data analysis practioners with a mathematically sound and empirically tested method for estimating lexical proportions in a robust fashion.

Keywords: lexical proportion, type frequency, lexical diversity, sample size, estimate, robustness.

Résumé

Dans le champ des recherches sur la mesure de la diversité lexicale, la problématique de la dépendance entre variété (nombre de mots distincts) et taille d'échantillon a fait l'objet de nombreux travaux. Comme le concept voisin de proportion lexicale, soit la proportion des mots d'un lexique manifestant une propriété donnée, se définit comme un rapport de variétés, la question de sa relation avec la taille d'échantillon se pose de façon aussi pressante que pour la variété elle-même. Elle a pourtant fait couler beaucoup moins d'encre et l'usage de la proportion lexicale « brute » reste attesté dans des travaux récents d'analyse de données textuelles dans divers domaines. Dans ce contexte, cette contribution cherche en particulier à répondre aux questions suivantes : (i) dans quelle mesure et de quelle façon l'estimation de la proportion lexicale brute est-elle liée aux variations de taille d'échantillon ? et (ii), dans quelle mesure une approche basée sur les principes de rééchantillonnage qui sous-tendent les méthodes en usage dans le domaine de la diversité lexicale permet-elle d'obtenir une estimation de proportion lexicale plus robuste face aux variations de taille d'échantillon ? Du point de vue applicatif, cette recherche vise ultimement à mettre à disposition des praticiens et praticiennes en analyse de données textuelles une méthodologie mathématiquement fondée et empiriquement éprouvée pour l'estimation robuste des proportions lexicales.

Mots clés : proportion lexicale, diversité lexicale, taille d'échantillon, estimation, robustesse.

1. Introduction

Dans le champ des recherches sur la mesure de la diversité lexicale, la problématique de la dépendance des indices relativement aux variations de taille d'échantillon a été largement thématisée et a fait l'objet d'approches aussi bien théoriques qu'empiriques. Le constat fondamental en la matière est que la variété w (nombre de types) d'un échantillon textuel ne peut que croitre avec sa taille n (nombre de tokens). La variété et ses diverses transformations, dont le ratio types-tokens (RTT) w/n, ne permettent donc pas la comparaison directe d'échantillons de tailles variables (Tweedie et Baayen, 1998).

Curieusement, le concept voisin de proportion lexicale,[1] soit la proportion des types manifestant une propriété donnée (par exemple contenir une consonne géminée, posséder un préfixe dérivationnel, présenter le trait sémantique « animé », etc.), ne semble pas avoir fait l'objet d'une réflexion comparable dans la littérature, et l'usage de la proportion lexicale « brute » reste attesté dans des travaux récents d'analyse de données textuelles dans divers domaines (pour n'en citer que quelques exemples : Alcaraz Mármol, 2011 ; Brugidou, 2011 ; Rodriguez, 2012 ; Gilles, 2017). Cet indice se définit pourtant comme un rapport de variétés : la variété des types possédant la propriété considérée et la variété globale de l'échantillon. La question de la dépendance entre proportion lexicale « brute » et taille d'échantillon se pose donc de façon tout aussi pressante que pour la variété elle-même.

Pour la plupart, les propositions récentes visant à surmonter le problème de dépendance entre variété et taille d'échantillon sont des variantes plus ou moins complexes de l'approche initialement formulée par Johnson (1944), consistant à diviser l'échantillon considéré en segments de m tokens adjacents et calculer la moyenne du RTT de ces segments (parfois appelée RTT *segmental moyen*). L'intuition qui sous-tend la méthode est que, si cette moyenne dépend naturellement de la valeur choisie pour le paramètre m, elle ne dépend toutefois pas de la longueur de l'échantillon ; c'est ce qui justifie de comparer le degré de diversité lexicale de deux ou plusieurs échantillons de tailles variables sur la base de leur RTT segmental moyen pour une valeur de m fixée.

Dubrocard (1988) met en œuvre une méthode apparentée, mais substituant au découpage de l'échantillon en segments contigus le tirage aléatoire de b sous-échantillons de m tokens sans remise.[2] La méthode VOCD de Malvern et Richards (1997) adopte aussi le principe du calcul du RTT moyen dans des sous-échantillons de m tokens, avec la particularité de faire varier m sur un intervalle de valeurs croissantes (35, 36, …, 50 tokens) pour dresser une courbe de RTT que ces auteurs qualifient d'*empirique* ; il s'agit ensuite d'approximer cette courbe au moyen d'une courbe *théorique* dépendant d'un paramètre unique, dont la valeur optimale (identifiée par une technique d'ajustement de courbe) constitue la diversité lexicale mesurée pour l'échantillon considéré.

En dépit de sa remarquable popularité, la méthode VOCD présente le défaut d'avoir été développée sans tirer parti de la possibilité, démontrée par Serant (1988), de calculer analytiquement l'espérance de la variété sur *tous* les sous-échantillons possibles de m tokens tirés sans remise selon le processus indiqué ci-dessus.[3] McCarthy et Jarvis (2007) soutiennent d'ailleurs que la procédure d'ajustement de courbe intervenant dans l'algorithme de VOCD n'a pas d'autre utilité que d'atténuer le bruit résultant de l'échantillonnage aléatoire, et qu'en ce

[1] aussi désigné sous le nom de *(relative) type frequency* en anglais, notamment dans le domaine de l'étude de l'acquisition des langues, où il joue un rôle particulier dans la perspective de l'évaluation du caractère plus ou moins productif des aspects de la compétence linguistique des apprenants (voir par exemple Bybee, 1985).

[2] Un token donné ne peut donc apparaitre qu'une seule fois dans un sous-échantillon particulier, mais il peut en revanche être présent dans plusieurs sous-échantillons différents.

[3] Un développement mathématique différent de celui de Serant (1988) est proposé dans l'annexe A. L'annexe B démontre par ailleurs que, sous l'hypothèse que le processus ayant généré un échantillon donné obéit à une loi multinomiale, un sous-échantillon tiré sans remise de cet échantillon peut être considéré comme ayant été généré par le même processus avec un nombre de tokens réduit ; dans ces conditions, l'espérance de la variété ne dépend effectivement que de la taille des sous-échantillons et non de celle de l'échantillon.

sens cette approche fournit essentiellement une approximation de l'espérance de la variété des sous-échantillons – au même titre que la moyenne empirique de la variété calculée selon la méthode de Dubrocard (1988) mentionnée plus haut.

Revenant du concept de variété à celui de proportion lexicale, cette contribution cherche en particulier à répondre aux questions suivantes, qui font respectivement l'objet des sections 2 et 3 : dans quelle mesure et de quelle façon l'estimation de la proportion lexicale « brute » est-elle liée aux variations de taille d'échantillon ? et dans quelle mesure une approche basée sur le calcul de l'espérance de la variété dans des sous-échantillons permet-elle d'obtenir une estimation de proportion lexicale plus robuste face à ces variations ? En guise de conclusion, la section 4 propose une synthèse de nos observations ainsi qu'une perspective sur des développements possibles de cette recherche.

2. Dépendance entre proportion lexicale et taille d'échantillon

La proportion des mots du lexique d'un échantillon textuel possédant une propriété quelconque dépend généralement de la taille de l'échantillon – sauf dans le cas trivial d'une propriété commune à tous les mots ou au contraire qu'aucun mot ne possède. En outre, contrairement à la variété, qui ne peut que diminuer lorsqu'on l'estime sur un nombre de tokens réduit, la proportion lexicale peut aussi bien diminuer qu'augmenter dans les mêmes circonstances. Pour s'en convaincre, il suffit de considérer la bipartition du lexique en deux sous-ensembles complémentaires, par exemple les mots possédant un suffixe et ceux n'en possédant pas ; comme l'union des deux couvre le lexique entier, si l'une des deux proportions diminue en réduisant le nombre de tokens, l'autre doit nécessairement augmenter.

L'expérience suivante montre comment la relation entre la propriété considérée et la distribution de fréquence des mots détermine l'effet de la réduction du nombre de tokens sur la proportion lexicale correspondante. Dans le cadre d'une comparaison entre deux échantillons textuels de tailles différentes, il n'est pas possible de savoir *a priori* si la proportion obtenue sur l'un des échantillons est sur- ou sous-évaluée par rapport à l'autre, à moins de connaitre la relation entre la propriété en question et la fréquence des mots.

Pour illustrer ce propos, nous examinons trois propriétés lexicales particulières au sein d'un échantillon constitué par le texte intégral du roman *De la Terre à la Lune* de Jules Verne, tel que mis à disposition sur le site du Projet Gutenberg.[4] Après réduction des majuscules et suppression des blancs, ponctuations et symboles spéciaux, l'échantillon contient 57'140 tokens formant un lexique de 8'311 types. Les propriétés considérées sont le fait d'être :

1. un « stopword », c'est-à-dire d'appartenir à une liste prédéfinie de mots considérés comme non pertinents pour les systèmes de recherche d'information[5] (319 types, soit 3.8% du lexique) ;
2. un mot plus long que la moyenne, qui s'élève à 7.8 lettres dans l'échantillon (4'265 types, soit 51.3% du lexique) ;
3. un mot de longueur paire (4'196 types, soit 50.5% du lexique).

[4] http://www.gutenberg.org/files/799/799-0.txt

[5] Il s'agit typiquement de prépositions, articles, conjonctions, etc. La liste de stopwords utilisée ici est celle mise à disposition pour le français par Jacques Savoy sur sa page *IR Multilingual Resources at UniNE* (http://members.unine.ch/jacques.savoy/clef/frenchST.txt).

Ces propriétés ont été choisies pour refléter trois types de relation avec la distribution de fréquence des mots : typiquement, les stopwords sont des mots plutôt fréquents, les mots longs sont plutôt rares et on ne s'attend pas à observer une forte dépendance entre parité de longueur et fréquence. La Figure 1 ci-dessous, qui présente la répartition de chacune des trois propriétés en fonction du rang des mots de l'échantillon classés par fréquence décroissante, confirme bien ces prédictions.

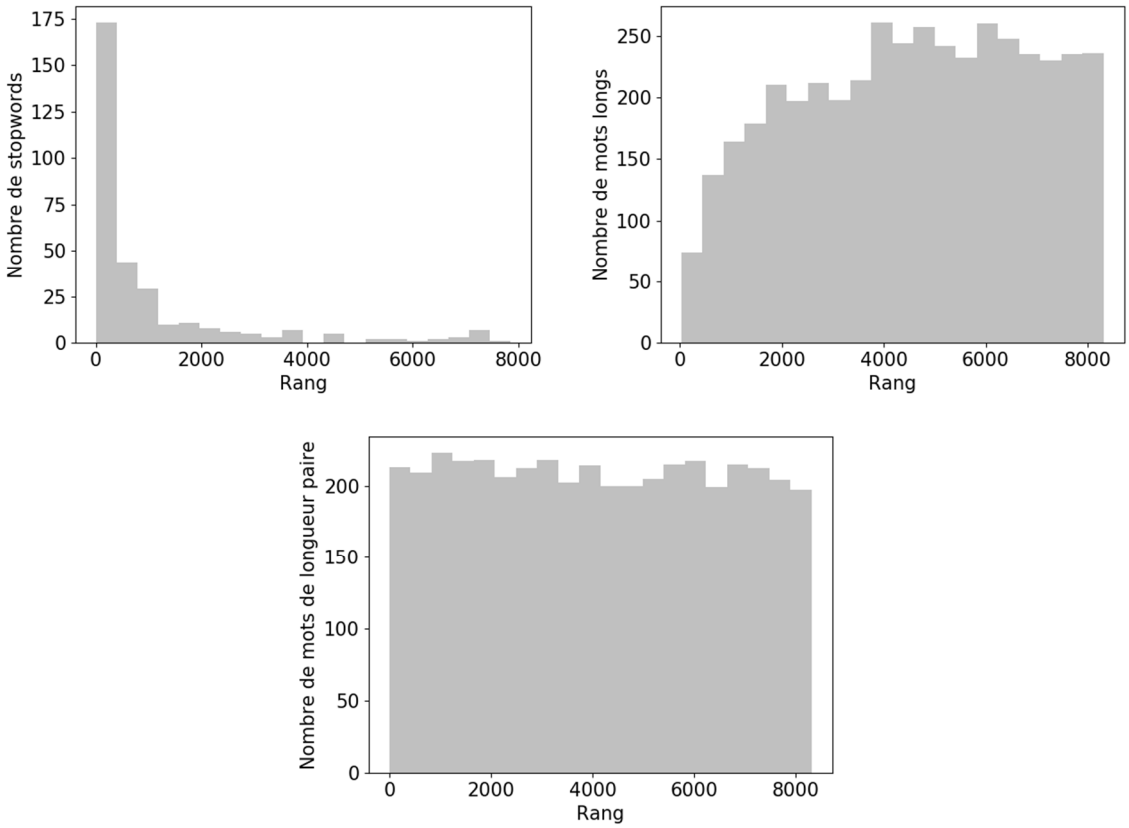

Figure 1. Histogramme du nombre de mots possédant la propriété en fonction du rang des mots, classés par fréquence décroissante. En haut à gauche : stopwords. En haut à droite : mots longs. En bas : mots de longueur paire.

Pour examiner l'effet de la variation du nombre de tokens sur la proportion lexicale de chacune des trois propriétés, nous tirons 100 sous-échantillons sans remise[6] de 5'000 tokens, 100 sous-échantillons de 10'000 tokens, 15'000 tokens, ..., 55'000 tokens. Pour les trois propriétés, la Figure 2 ci-dessous présente les proportions lexicales observée dans les 100 sous-échantillons de chaque taille, ainsi que dans l'échantillon entier (57'140 tokens).[7] La comparaison entre cette figure et la précédente montre bien comment chaque proportion lexicale est différemment affectée par la relation entre la propriété correspondante et la

[6] Cf. note 2 ci-dessus.

[7] Notons ici que nous considérons, à défaut de mieux, les sous-échantillons tirés sans remise comme des nouveaux échantillons issus d'une loi identique au texte initial. L'annexe B permet de montrer que c'est bien le cas si l'on suppose l'échantillon initial comme issu d'une loi multinomiale. Bien cette dernière hypothèse soit une simplification drastique, elle nous apparait comme peu gênante lorsque l'on s'intéresse à des indicateurs qui ne prennent pas en compte l'ordre des mots.

distribution de fréquence des mots. Ainsi, une propriété surreprésentée dans les mots fréquents, comme le fait d'être un stopword, verra typiquement sa proportion lexicale réduire en augmentant le nombre de tokens. Inversement, la proportion lexicale d'une propriété sous-représentée dans les mots fréquents, comme le fait d'être un mot long, aura tendance à augmenter avec le nombre de tokens. Enfin, une propriété à peu près uniformément répartie, telle que la parité de la longueur, donnera lieu à une mesure de proportion lexicale relativement stable (sauf peut-être pour un nombre de tokens très réduit).

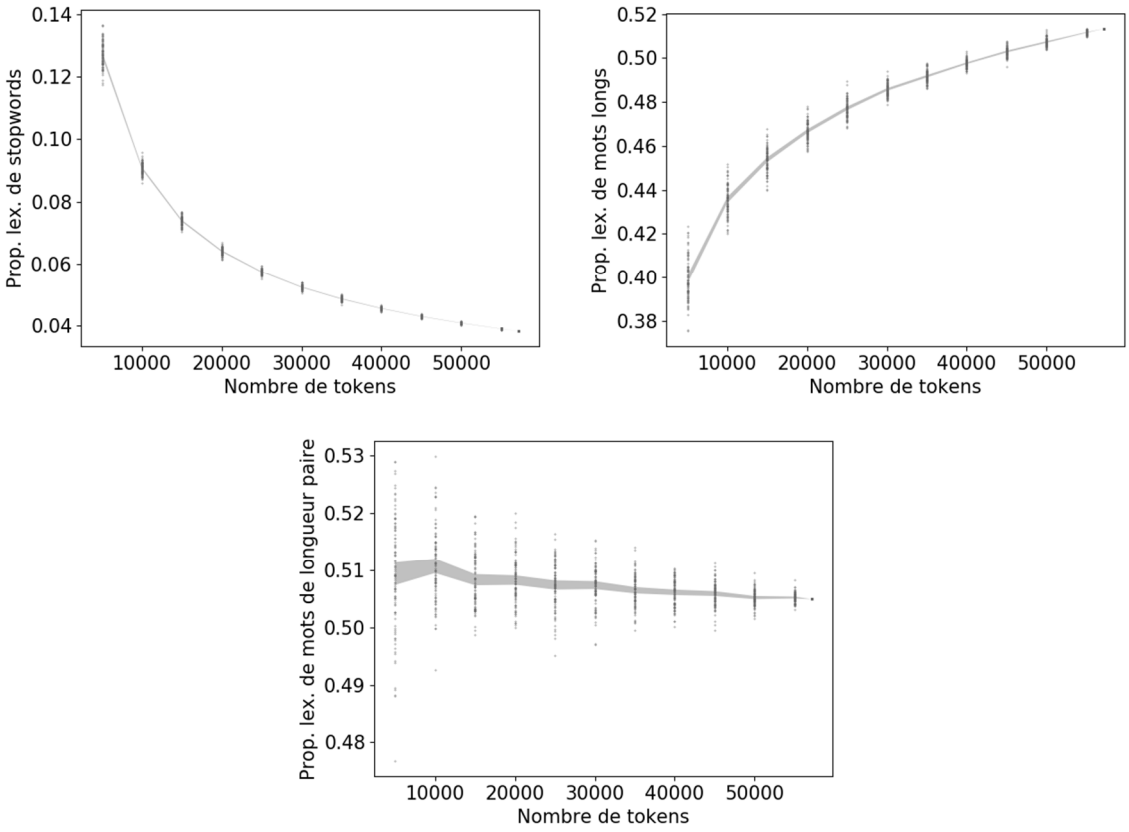

Figure 2. Proportion lexicale des différentes propriétés en fonction du nombre de tokens. La surface grisée représente l'intervalle de confiance à 95% autour de la moyenne. En haut à gauche : les stopwords. En haut à droite : les mots longs. En bas : les mots de longueur paire.

3. Espérance de la proportion lexicale sous-échantillonnée

Les résultats de la section précédente montrent que la proportion lexicale brute dépend de façon complexe de la taille d'échantillon et il s'ensuit qu'elle est inadéquate pour la comparaison directe d'échantillon de tailles différentes. Dans cette section, nous nous intéressons en particulier à la possibilité de transposer au domaine de l'estimation de la proportion lexicale les principes de rééchantillonnage en usage pour la mesure de la diversité lexicale.

Il est bien sûr envisageable d'adopter sans modification l'approche de Dubrocard (1988) mentionnée en section 1. Dans chaque échantillon à comparer, cela reviendrait à estimer la proportion lexicale de la propriété considérée par la moyenne des proportions lexicales correspondantes observées dans b sous-échantillons de m tokens tirés sans remise de l'échantillon en question (la valeur du paramètre m étant la même pour tous les échantillons).

Le principal désavantage de cette approche est de reposer sur un tirage aléatoire et donc de présenter un certain degré de variabilité. Certes, l'impact de ce facteur peut être atténué en tirant un nombre b plus élevé de sous-échantillons, mais dans ce cas c'est sur le plan computationnel que l'approche devient désavantageuse. Dans l'idéal, il serait préférable de pouvoir calculer analytiquement l'espérance de la proportion lexicale dans *tous* les sous-échantillons de taille m donnée, de façon analogue au développement mathématique proposé par Serant (1988) pour la variété lexicale.

Dans cette perspective, définissons un échantillon textuel, de taille n, comme un vecteur $\mathbf{n} = (n_1, \dots, n_w)$, où $n_i \geq 0 \ \forall i$ représente le nombre d'occurrences du type i (on a donc $\sum_i n_i = n$). Le lexique de l'échantillon est représenté par l'ensemble $\mathcal{W}$, de taille w. Soit maintenant un sous-ensemble de mots $\mathcal{P} \subseteq \mathcal{W}$ possédant une propriété particulière, de taille w_p. En tirant, sans remise, un échantillon de taille $m \leq n$, représenté par un vecteur aléatoire $S = (M_1, \dots, M_w)$ où M_i représente le nombre de tokens tirés de type i, avec $\sum_i M_i = m$, on peut définir la variété, c'est-à-dire le nombre de types présents dans S, avec la variable aléatoire suivante :

$$V := \sum_{i \in \mathcal{W}} I\,(M_i > 0) \tag{1}$$

où $I(\cdot)$ dénote la fonction indicatrice. Le nombre de types avec propriété présents dans S peut quant à lui être défini avec la variable aléatoire :

$$V_p := \sum_{i \in \mathcal{P}} I\,(M_i > 0) \tag{2}$$

On trouve alors que (voir Annexe A) :

$$\mathbb{E}(V) = w - \sum_{i \in \mathcal{W}} \pi_i \qquad \mathbb{E}(V_p) = w_p - \sum_{i \in \mathcal{P}} \pi_i \tag{3}$$

$$\mathrm{Var}(V) = \sum_{i \in \mathcal{W}} \sum_{j \in \mathcal{W}} (\pi_{ij} - \pi_i \pi_j) \qquad \mathrm{Var}(V_p) = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} (\pi_{ij} - \pi_i \pi_j) \tag{4}$$

$$\mathrm{Cov}(V, V_p) = \sum_{i \in \mathcal{W}} \sum_{j \in \mathcal{P}} (\pi_{ij} - \pi_i \pi_j) \tag{5}$$

où π_i est la probabilité de ne pas tirer le type i et π_{ij} la probabilité de ne tirer ni i, ni j. Etant donné qu'il s'agit d'un tirage sans remise et grâce à la loi hypergéométrique, on a :

$$\pi_i := \mathbb{P}(M_i = 0) = \begin{cases} \frac{\binom{n-n_i}{m}}{\binom{n}{m}} & \text{si } n_i \leq n - m \\ 0 & \text{sinon} \end{cases} \tag{6}$$

$$\pi_{ij} := \mathbb{P}(M_i = 0 \cap M_j = 0) = \begin{cases} \frac{\binom{n-n_i-n_j}{m}}{\binom{n}{m}} & \text{si } n_i + n_j \leq n - m \\ 0 & \text{sinon} \end{cases} \tag{7}$$

Nous sommes spécifiquement intéressés ici par le calcul analytique de $\mathbb{E}(V_p/V)$, l'espérance de la proportion lexicale. Cette espérance s'avère difficile à obtenir directement mais on peut

recourir à un développement de Taylor[8] pour approximer sa valeur (voir par exemple Benaroya, Han et Nagurka, 2005, p. 169) :

Approximation du premier ordre :
$$\mathbb{E}\left(\frac{V_p}{V}\right) \approx \frac{\mathbb{E}(V_p)}{\mathbb{E}(V)}$$

Approximation du deuxième ordre :
$$\mathbb{E}\left(\frac{V_p}{V}\right) \approx \frac{\mathbb{E}(V_p)}{\mathbb{E}(V)} - \frac{\mathrm{Cov}(V, V_p)}{\mathbb{E}(V)^2} + \frac{\mathbb{E}(V_p)\mathrm{Var}(V)}{\mathbb{E}(V)^3}$$

La Figure 3 ci-dessous reprend l'exemple du roman de Jules Verne, *De la Terre à la Lune*, introduit dans la section précédente, et superpose aux résultats obtenus précédemment (cf. Figure 2) l'approximation du premier ordre de l'espérance de la proportion lexicale :

$$\mathbb{E}\left(\frac{V_p}{V}\right) \approx \frac{\mathbb{E}(V_p)}{\mathbb{E}(V)} = \frac{w_p - \sum_{i \in \mathcal{P}} \pi_i}{w - \sum_{i \in \mathcal{W}} \pi_i} \tag{8}$$

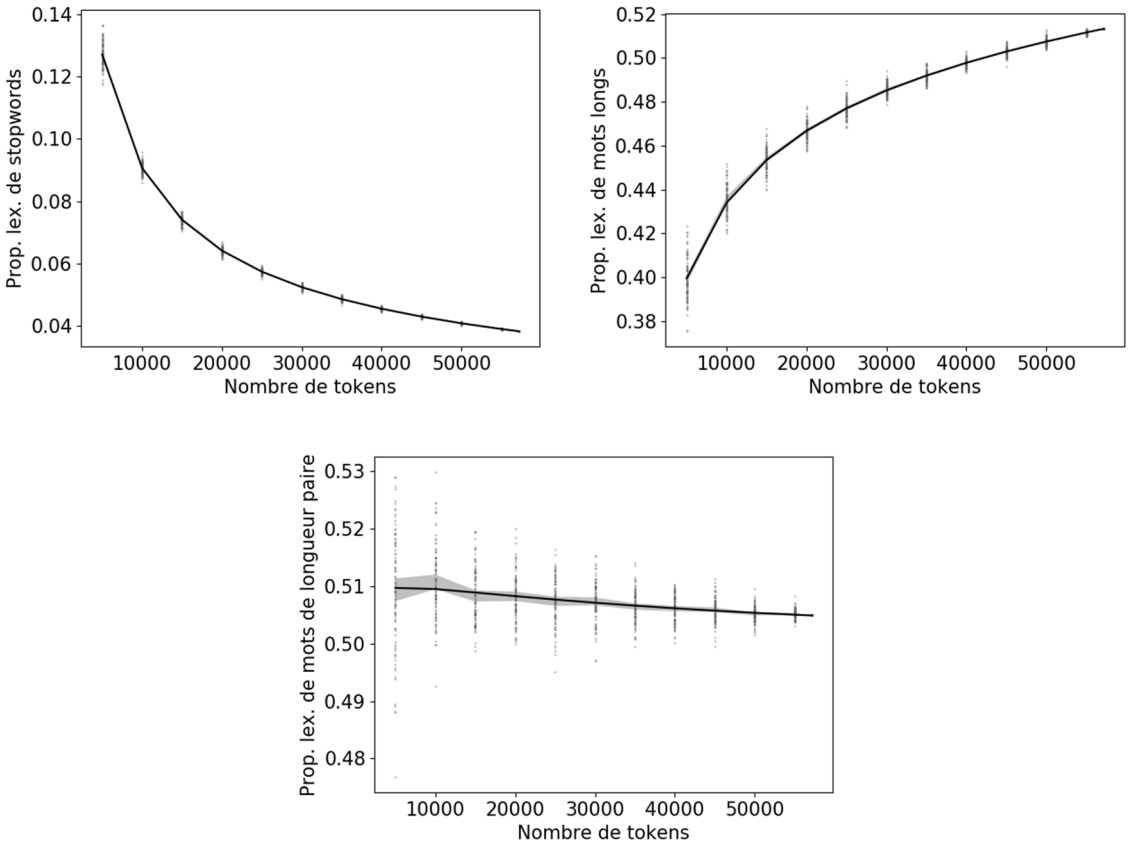

Figure 3. Proportion lexicale des différentes propriétés en fonction de la taille d'échantillon. La surface grisée représente l'intervalle de confiance à 95% autour de la moyenne. La ligne noire continue représente la proportion théorique (équation (8)). En haut à gauche : stopwords. En haut à droite : mots longs. En bas : mots de longueur paire.

[8] Il s'agit en réalité d'un développement de Taylor de la fonction $f(V, V_p) = V_p/V$ autour du point $\mathbb{E}(V_p)/\mathbb{E}(V)$ et à la suite duquel on calcule l'espérance $\mathbb{E}\big(f(V, V_p)\big)$. Notons que l'espérance du développement d'ordre zéro est égale à l'espérance du développement du premier ordre.

L'approximation du premier ordre se situe presque toujours dans l'intervalle de confiance à 95% de la moyenne empirique et constitue donc un très bon estimateur de cette dernière.

La différence entre l'approximation du premier ordre et celle du deuxième ordre, qui inclut les termes avec le carré et le cube de l'espérance de la variété au dénominateur, est trop faible pour être utilement représentée sur les diagrammes. Le gain lié au passage au deuxième ordre, tel que mesuré par l'erreur quadratique moyenne entre espérance approximée et moyenne empirique, est minime dans le cas des stopwords et des mots longs, et négatif pour les mots de longueur paire (cf. Tableau 1 ci-dessous). En revanche, la complexité du calcul de l'approximation du deuxième ordre est nettement plus élevée que celle du premier ordre (quadratique au lieu de linéaire en fonction du nombre de types n).

Propriété	**Approximation du 1er ordre**	**Approximation du 2ème ordre**
Stopwords	3.56×10^{-9}	3.31×10^{-9}
Mots longs	1.44×10^{-7}	1.36×10^{-7}
Mots de longueur paire	9.95×10^{-8}	1.03×10^{-7}

Tableau 1. Erreur quadratique moyenne entre les approximations du premier et deuxième ordre de l'espérance de la proportion lexicale des sous-échantillons et sa moyenne empirique.

5. Conclusion

Cette contribution a été motivée par le constat que la question de la robustesse de l'estimation de la proportion lexicale, au contraire de l'estimation de la diversité lexicale, n'a pas fait l'objet d'une réflexion méthodologique spécifique. Dans ce contexte, nous avons cherché à mieux comprendre la nature de la relation entre proportion lexicale « brute » et taille d'échantillon, d'une part, et à obtenir une estimation de proportion lexicale plus robuste face aux variations de taille d'échantillon en nous inspirant des principes de rééchantillonnage en usage pour la mesure de la diversité lexicale, d'autre part.

Concernant le premier objectif, les expériences présentées dans cette étude ont montré que l'effet de la réduction du nombre de tokens sur la proportion lexicale est déterminé par la relation entre la propriété en question et la distribution de fréquence des mots: lorsque la taille d'échantillon diminue, la proportion lexicale tend à diminuer pour une propriété sous-représentée dans les mots fréquents (par exemple être un mot long), tandis qu'elle tend à augmenter pour une propriété surreprésentée dans les mots fréquents (par exemple être un stopword). En ce qui concerne le second objectif, nous nous sommes intéressés en particulier à la possibilité de donner pour l'espérance de la proportion lexicale dans des sous-échantillons de taille fixe une expression analytique analogue à celle formulée par Serant (1988) pour l'espérance de la variété lexicale. Bien qu'un calcul exact de l'espérance de la proportion lexicale reste difficile à atteindre, nos expériences ont montré que l'approximation de Taylor du premier ordre fournit un très bon estimateur de la moyenne empirique avec une complexité algorithmique qui reste d'ordre linéaire.

A l'aune de ces résultats et à ce stade de nos recherches, notre recommandation concrète à l'intention des praticiens et praticiennes en analyse de données textuelles souhaitant comparer la proportion lexicale d'une propriété au sein d'échantillons textuels de tailles variables est donc de l'estimer au moyen de l'équation (8) ci-dessus, en adoptant pour tous les échantillons une taille de sous-échantillon commune m inférieure à celle du plus petit des échantillons à

comparer. Les estimations de proportion lexicale obtenues par cette façon de procéder sont naturellement dépendantes de la valeur choisie pour le paramètre m, mais non de la taille originale des échantillons (du moins sous l'hypothèse que le processus ayant généré chacun des échantillons obéit à une loi multinomiale, cf. annexe B ci-dessous).

L'une des perspectives qu'ouvre cette recherche est en lien avec la méthode de mesure de diversité lexicale VOCD (Malvern et Richards, 1997). Bien que cette méthode puisse être améliorée en tirant parti de la possibilité de calculer analytiquement l'espérance de la variété des sous-échantillons (cf. section 1), elle présente l'intérêt de reposer sur l'observation de la variété – ou plus exactement du TTR – sur une plage de tailles de sous-échantillons, plutôt que pour une taille de sous-échantillon unique. A l'issue de la présente étude se pose la question de la possibilité et de l'opportunité d'adapter cette approche à l'évaluation de la proportion lexicale, ce qui impliquerait en premier lieu de construire un modèle de la relation entre proportion lexicale et taille d'échantillon, dépendant idéalement d'un unique paramètre.

Sur un tout autre plan, la recherche présentée dans cette contribution soulève la question de la diffusion de l'innovation méthodologique qu'elle promeut. La nécessité d'un changement de pratique dans un paradigme scientifique (fût-il très spécifique comme dans le cas d'espèce) constitue sans doute une condition nécessaire à son adoption par une portion substantielle de la communauté, mais en aucun cas une condition suffisante. Il faut encore et peut-être surtout œuvrer au développement logiciel et en particulier à la conception d'interfaces qui devront permettre aux praticiens et praticiennes du domaine concerné de s'approprier la nouvelle méthodologie d'un point de vue tant opérationnel que conceptuel (voir notamment Leblanc 2010; Xanthos 2014). Dans le cas d'espèce, le tout premier pas dans cette direction a consisté à publier sous licence open source le code python permettant de répliquer les expériences conduites dans cette étude.[9] Les étapes suivantes consisteront d'abord à rendre les fonctionnalités correspondantes accessible aux programmeurs par le biais d'un package, puis de permettre aux utilisateurs non spécialistes de mettre en œuvre ce calcul par le biais d'une interface graphique. En toute vraisemblance, c'est à ce prix seulement que nos propositions pour une estimation robuste des proportions lexicales auront une chance de rejoindre, pour un temps au moins, le corps des bonnes pratiques en analyse de données textuelles.

Remerciements

Les auteurs remercient Marianne Kilani-Schoch, qui la première a attiré leur attention sur les questions méthodologiques auxquelles cet article tente d'apporter des réponses, ainsi que deux relecteurs anonymes pour leurs remarques constructives.

Annexe A : développements mathématiques

Preuve de (3)

$$\begin{aligned}\mathbb{E}(V) &= \mathbb{E}\left(\sum_{i\in\mathcal{W}} I\left(M_i > 0\right)\right) = \sum_{i\in\mathcal{W}} \mathbb{E}\left(I(M_i > 0)\right) \\ &= \sum_{i\in\mathcal{W}} \mathbb{P}\left(M_i > 0\right) = \sum_{i\in\mathcal{W}} \left(1 - \mathbb{P}(M_i = 0)\right) = w - \sum_{i\in\mathcal{W}} \pi_i\end{aligned}$$

[9] https://github.com/sliunil/lexical_proportion_GGX_AXS

Pour $\mathbb{E}(V_p)$, le développement est similaire mais la somme s'effectue sur $\mathcal{P}$.

Preuve de (4)

$$\begin{aligned}
\mathrm{Var}(V) &= \mathbb{E}(V^2) - \mathbb{E}(V)^2 = \mathbb{E}\left(\left(\sum_{i\in\mathcal{W}} I(M_i > 0)\right)^2\right) - \left(w - \sum_{i\in\mathcal{W}} \pi_i\right)^2 \\
&= \mathbb{E}\left(\sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} I(M_i > 0) I\big(M_j > 0\big)\right) - w^2 + 2w\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \pi_i\,\pi_j \\
&= \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \mathbb{E}\Big(I\big(M_i > 0 \cap M_j > 0\big)\Big) - w^2 + 2w\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \pi_i\,\pi_j \\
&= \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \Big(1 - \mathbb{P}(M_i = 0) - \mathbb{P}\big(M_j = 0\big) + \mathbb{P}(M_i = 0 \cap M_i = 0)\Big) \\
&-w^2 + 2w\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \pi_i\,\pi_j \\
&= w^2 - 2w\sum_{i\in\mathcal{W}} \pi_i + \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \pi_{ij} - w^2 + 2w\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \pi_i\,\pi_j \\
&= \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{W}} \big(\pi_{ij} - \pi_i\pi_j\big)
\end{aligned}$$

Pour $\mathrm{Var}\big(V_p\big)$, le développement est similaire mais les sommes s'effectuent sur $\mathcal{P}$.

Preuve de (5)

$$\begin{aligned}
\mathrm{Cov}(V, V_p) &= \mathbb{E}(VV_p) - \mathbb{E}(V)\mathbb{E}(V_p) \\
&= \mathbb{E}\left(\sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} I\,(M_i > 0) I(M_j > 0)\right) \\
&-ww_p + w\sum_{j\in\mathcal{P}} \pi_i + w_p\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \pi_i\,\pi_j \\
&= \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \big(1 - \mathbb{P}(M_i = 0) - \mathbb{P}(M_j = 0) + \mathbb{P}(M_i = 0 \cap M_i = 0)\big) \\
&-ww_p + w\sum_{j\in\mathcal{P}} \pi_i + w_p\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \pi_i\,\pi_j \\
&= ww_p - w\sum_{j\in\mathcal{P}} \pi_i - w_p\sum_{i\in\mathcal{W}} \pi_i + \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \pi_{ij} \\
&-ww_p + w\sum_{j\in\mathcal{P}} \pi_i + w_p\sum_{i\in\mathcal{W}} \pi_i - \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \pi_i\,\pi_j = \sum_{i\in\mathcal{W}}\sum_{j\in\mathcal{P}} \big(\pi_{ij} - \pi_i\pi_j\big)
\end{aligned}$$

Annexe B : loi d'un sous-échantillon sous l'hypothèse « bag-of-words »

On fait ici l'hypothèse (dite « bag-of-words ») d'un processus de génération textuelle obéissant à une loi multinomiale : chaque mot du texte est tiré dans un vocabulaire $\mathcal{W}$ de taille w, de

manière indépendante, avec remise, selon une probabilité p_i (avec $\sum_{i \in \mathcal{W}} p_i = 1$). Le processus de génération d'un texte est représenté par un vecteur aléatoire $T = (N_1, \dots, N_w)$, où N_i représente le nombre de tokens tirés de type i, avec $\sum_{i \in \mathcal{W}} N_i = n$. Dans ce contexte, un échantillon textuel donné est une réalisation $\mathbf{n} = (n_1, \dots, n_w)$, dont la probabilité est donnée par :

$$\mathbb{P}(T = \mathbf{n}) = \begin{cases} \dfrac{n!}{\prod_{i=1}^{w} n_i\,!} \prod_{i=1}^{w} p_i^{n_i} & \text{si } \sum_i n_i = n \text{ et } \mathbf{n} \in \mathbb{N}^w \\ 0 & \text{sinon} \end{cases}$$

Le processus de tirage (sans remise) d'un sous-échantillon de taille $m \leq n$ dans T peut à son tour être représenté par un vecteur aléatoire $S = (M_1, \dots, M_w)$ avec $\sum_{i \in \mathcal{W}} M_i = m$. Sachant $\mathbf{n} = (n_1, \dots, n_w)$, pour un sous-échantillon $\mathbf{m} = (m_1, \dots, m_w)$ donné (soit une réalisation de S), on a :

$$\mathbb{P}(S = \mathbf{m} | T = \mathbf{n}) = \begin{cases} \dfrac{\prod_{i=1}^{w} \binom{n_i}{m_i}}{\binom{n}{m}} & \text{si } n_i \geq m_i \text{ et } n - m \geq n_i - m_i \ \forall i \\ 0 & \text{sinon} \end{cases}$$

Pour un vecteur $\mathbf{m} = (m_1, \dots, m_w) \in \mathbb{N}^w$ donné, qui vérifie $\sum_i m_i = m$, posons l'ensemble $\mathcal{N}_{\mathbf{m}}^w$ des vecteurs possibles $\mathbf{n} \in \mathbb{N}^w$ qui permettent la réalisation de $\mathbf{m}$. Cet ensemble se définit comme :

$$\mathcal{N}_{\mathbf{m}}^w := \left\{ \mathbf{n} \in \mathbb{N}^w \,\middle|\, n_i \geq m_i, n - m \geq n_i - m_i \ \forall i \ ; \ \sum_i n_i = n \right\}$$

On peut alors décomposer la probabilité $\mathbb{P}(S = \mathbf{m})$ avec :

$$\begin{aligned} \mathbb{P}(S = \mathbf{m}) &= \sum_{\mathbf{n} \in \mathcal{N}_{\mathbf{m}}^w} \mathbb{P}\left(S = \mathbf{m} | T = \mathbf{n}\right) \mathbb{P}(T = \mathbf{n}) \\ &= \sum_{\mathbf{n} \in \mathcal{N}_{\mathbf{m}}^w} \frac{\prod_{i=1}^{w} \binom{n_i}{m_i}}{\binom{n}{m}} \frac{n!}{\prod_{i=1}^{w} n_i\,!} \prod_{i=1}^{w} p_i^{n_i} \\ &= \sum_{\mathbf{n} \in \mathcal{N}_{\mathbf{m}}^w} \frac{m!\,(n-m)!}{\prod_{i=1}^{w} m_i\,!\,(n_i - m_i)!} \prod_{i=1}^{w} p_i^{n_i} \\ &= \frac{m!}{\prod_{i=1}^{w} m_i\,!} \prod_{i=1}^{w} p_i^{m_i} \sum_{\mathbf{n} \in \mathcal{N}_{\mathbf{m}}^w} \frac{(n-m)!}{\prod_{i=1}^{w} (n_i - m_i)!} \prod_{i=1}^{w} p_i^{n_i - m_i} \end{aligned}$$

En posant $\mathbf{d} = \mathbf{n} - \mathbf{m}$, on voit que l'ensemble $\mathcal{N}_{\mathbf{m}}^w$ peut se réécrire comme :

$$\begin{aligned} \mathcal{D}_{\mathbf{m}}^w &:= \left\{ \mathbf{d} \in \mathbb{Z}^w \,\middle|\, d_i \geq 0, n - m \geq d_i \ \forall i \ ; \ \sum_i d_i = n - m \right\} \\ &= \left\{ \mathbf{d} \in \mathbb{N}^w \,\middle|\, \sum_i d_i = n - m \right\} \end{aligned}$$

car $\sum_i d_i = n - m$ avec $d_i \geq 0 \ \forall i$ implique que $n - m \geq d_i \ \forall i$. On a donc :

$$\begin{aligned}\mathbb{P}(S=\mathbf{m}) &= \frac{m!}{\prod_{i=1}^{w} m_i\,!}\prod_{i=1}^{w} p_i^{m_i} \underbrace{\sum_{\mathbf{d}\in\mathcal{D}_{\mathbf{m}}^{w}} \frac{(n-m)!}{\prod_{i=1}^{w} d_i\,!}\prod_{i=1}^{w} p_i^{d_i}}_{=1} \\ &= \frac{m!}{\prod_{i=1}^{w} m_i\,!}\prod_{i=1}^{w} p_i^{m_i}.\end{aligned}$$

Autrement dit, le tirage d'un sous-échantillon S suit une loi multinomiale avec les probabilités $p_1, \ldots, p_w$ et une taille d'échantillon de m. On peut donc considérer qu'il est le résultat du même processus de génération T, obéissant à la même loi multinomiale, mais avec une taille réduite $m \leq n$.

Références

Alcaraz Mármol G. (2011). Vocabulary input in classroom materials: Two EFL coursebooks used in Spanish schools. *Revista Espanola de Linguistica Aplicada*, 24: 9-28.

Benaroya H., Han S. M. et Nagurka M. (2005). *Probability models in engineering and science*. Boca Raton: CRC press.

Brugidou M. (2011). Le Grenelle de l'environnement : corpus et dispositif d'écriture. *Corpus*, 10: 155-178. https://journals.openedition.org/corpus/2069

Bybee J. L. (1985). *Morphology: A study of the relation between meaning and form.* Philadelphia: Benjamins.

Dubrocard M. (1988). Évaluation de l'étendue du lexique, quelques essais de simulation. In Labbé D., Thoiron P. et Serant D. (éd.), *Etudes sur la richesse et la structure lexicale*. Paris-Genève: Slatkine–Champion, pp. 43-66.

Gilles F. (2017). *Valorisation des analogies lexicales entre l'anglais et les langues romanes: étude prospective pour un dispositif plurilingue d'apprentissage du FLE dans le domaine de la santé*. (thèse de doctorat, Université Grenoble Alpes, France). https://tel.archives-ouvertes.fr/tel-01719853/document

Leblanc J.-M. (2010). Nouvelles fonctionnalités pour la visualisation des données textuelles et de résultats : Pour une approche plus ergonomique des dispositifs lexicométriques. In *Actes des 10èmes Journées d'Analyse statistique des Données Textuelles (JADT 2010)*, pp. 1057-1068.

Malvern D. D. et Richards B. J. (1997). A new measure of lexical diversity. In Ryan A. et Wray A. (éd.), *Evolving models of language*. Clevedon: Multilingual Matters, pp. 58-71.

McCarthy P. M. et Jarvis S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4): 459-488.

Rodriguez, L. (2012). Evolution lexicométrique des régionalismes dans un corpus franco-canadien (1993–2006). In *Actes des 10èmes Journées internationales d'analyse statistique des données textuelles (JADT 2012)*, pp. 871-882.

Serant D. (1988). A propos des modèles de raccourcissement de textes. In Labbé D., Thoiron P. et Serant D. (éd.), *Etudes sur la richesse et la structure lexicale*. Paris-Genève: Slatkine–Champion, pp. 77-91.

Tweedie F. J. et Baayen R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32: 323-352.

Xanthos, A. (2014). Textable: programmation visuelle pour l'analyse de données textuelles. In *Actes des 12èmes Journées internationales d'analyse statistique des données textuelles (JADT 2014)*, pp. 691-703.