Y a-t-il une mathématique des textes ? Le monde non-euclidien du discours

Florian Cafiero

Y a-t-il une mathématique des textes ? Le monde non-euclidien du discours

Florian Cafiero
Université Paris-Diderot

16 mai 2019

Séminaire de l'Observatoire de la Vie Littéraire (OBVIL) - Université Paris Sorbonne

Résumé

Cette communication traite de l'importance de l'utilisation de géométries adaptées dans l'étude des données textuelles et des phénomènes linguistiques. Le géométrie euclidienne, d'un usage commun et d'une compréhension intuitive, a parfois été utilisée pour l'étude des phénomènes linguistiques, et est encore présupposée, y compris inconsciemment, dans nombre de calculs simples effectués dans l'étude de textes. Bien qu'efficace pour décrire les phénomènes physiques, nous rappelons ici l'inadéquation de la géométrie euclidienne au monde du langage naturel, et les travers liés à l'usage de certains outils mathématiques aussi courants que la "moyenne" ou la notion de "distance". Des géométries plus complexes fournissent un cadre plus approprié pour comprendre les relations et les structures présentes dans les données linguistiques. Les avantages de l'utilisation de ces géométries adaptées sont illustrés par l'analyse de divers phénomènes linguistiques, notamment en stylométrie.

Mots clés : Analyse du discours ; Stylométrie ; Statistique textuelle ; Distance intertextuelle

La plupart des outils mathématiques que nous apprenons dès notre plus jeune âge, et que nous manipulons le plus souvent, sont destinés à appréhender la réalité physique visible. Ces mathématiques sont celles dont nous avons besoin pour notre quotidien, un quotidien qu'elles nous permettent en général de bien décrire, au moins en première approximation. Mais restent-elles adaptées lorsque l'on étudie des données textuelles ?

1 Au nom de la loi

1.1 La normalité des sciences de la nature

La loi Normale, dite aussi de Laplace-Gauss, est la plus célèbre, et sert de référence dans les sciences de la nature. Elle est représentée par la fonction de densité suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Malgré ses allures complexes, c'est en fait une loi à la fois simple et fondamentale. Elle représente ce qui arrive lors de la répétitions à grande échelle d'expériences aléatoires similaires 1. Un des exemples canoniques est la fameuse *planche de Galton*, une planche percée de clous régulièrement espacés, du haut de laquelle on fait tomber une série de billes.

Résumés statistiques

Si on veut condenser l'information en un ou quelques chiffre(s), on utilise des *résumés statistiques*. Le plus connu et le plus usité d'entre eux est bien sûr la *moyenne arithmétique* :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

En fonction des cas, toutes les valeurs entrant en jeu dans le calcul de la moyenne peuvent ne pas avoir la même importance. Dans ce cas, on attribue un poids différent aux diverses réalisation pour calculer la *moyenne arithmétique pondérée* par des poids w_i :

$$\bar{x} = \frac{\sum_{i=n}^{n} w_i x_i}{\sum_{i=n}^{n} x_i}$$

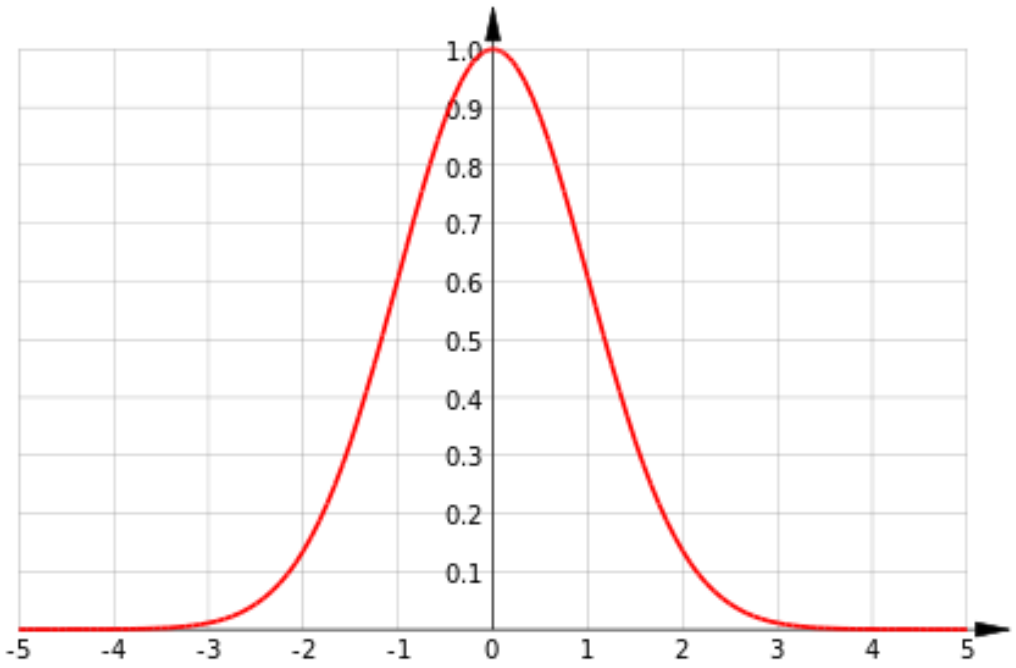
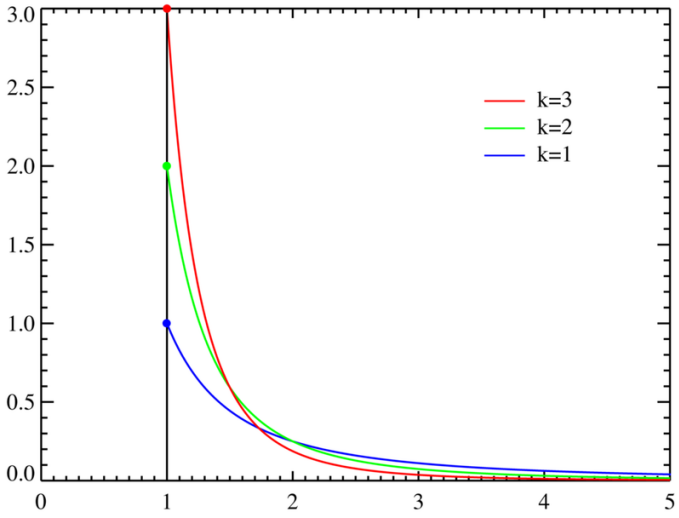


FIGURE 1 – Loi de Laplace-Gauss

FIGURE 2 – Loi de Pareto pour différents k



La moyenne arithmétique est particulièrement utile lorsque la distribution est « équilibrée », typiquement, lorsqu'elle suit une loi de Gauss.

1.2 L'anormalité des sciences humaines

Le monde est-il normal ?

De nombreux phénomènes naturels sont effectivement distribués selon la loi normale. Mais dans les sciences humaines en général, et dans les analyses textuelles en particulier, il est bien moins fréquent de rencontrer ce type de distribution !

Hors sciences de la nature : mondes sociaux et loi de Pareto

En sciences humaines et sociales, on rencontre couramment des phénomènes suivant une loi de Pareto, dont la densité s'écrit :

$$f(x; k, x_m) = k\frac{x_m^k}{x^{k+1}}$$

où k est un paramètre appelé « indice de Pareto »

Héritée du sociologue et économiste italien Vilfredo Pareto (1848-1923), cette loi visait à décrire des situations d'inégalités économiques : la majorité des richesses est accaparée par peu d'individus, et la majorité des individus a une forte probabilité d'être pauvre.

De nombreuses distributions rencontrées en sciences humaines et sociales ont en effet cette forme. On peut montrer qu'il existe des raisons structurelles pour que les distributions des revenus des ménages suivent toujours des lois de Pareto, qui servent encore à modéliser ce type de phénomène aujourd'hui [1]. On retrouve des lois de Pareto également dans la taille des villes [2]

2 Mot à mot

Les lois décrivant les propriétés fondamentales autour de l'apparition des mots dans un texte ne sont pas sans rappeler les lois de Pareto. Elles dépendent marginalement de la langue sur laquelle on travaille, tout comme la répartition des revenus peut marginalement dépendre des sociétés que l'on étudie. Elles sont cependant souvent valables en première approximation pour l'immense majorité des langues naturelles.

La loi d'Estoup-Zipf

Une loi prédisant l'ordonnancement des fréquence d'apparition des mots d'un vocabulaire dans un texte est d'abord formulée par Jean-Baptiste Estoup et développée et mise en valeur par son contemporain George Kingsley Zipf (1902-1950) [3].

— En comptant les occurrences de mots dans le roman *Ulysses* de James Joyce, George Zipf avait remarqué qu'à quelque chose près :

1. le mot le plus courant revenait 8 000 fois ;
2. le dixième mot 800 fois ;
3. le centième, 80 fois ;
4. le millième, 8 fois.

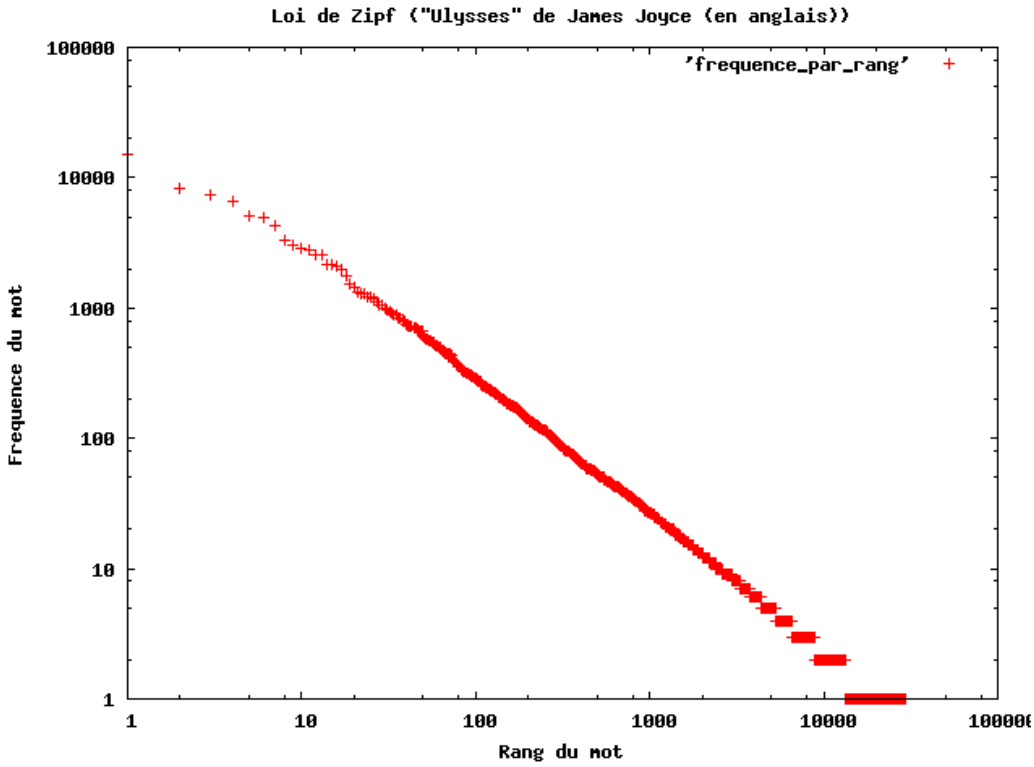
De cette observation est née l'idée que, dans un texte défini, le nombre d'occurrences $f(n)$ d'un mot est déterminé par son rang n dans l'ordre des fréquences, selon la formule :

$$f(n) = \frac{K}{n}, K \in \mathbb{R}, n \in \mathbb{N}*$$

Où K est une constante quelconque propre au texte

Pour le dire de manière plus simple : la fréquence d'apparition d'un mot multipliée par son rang serait une constante.

FIGURE 3 – Ulysses : apparition des mots



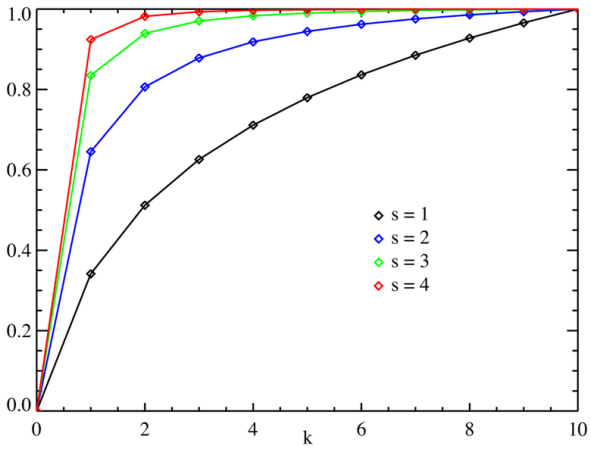


FIGURE 4 – Loi de Zipf-Mandelbrot : fonction de répartition

Loi de Zipf-Mandelbrot

On a depuis raffiné cette observation, sans l'avoir infirmé pour autant. Une généralisation a ainsi été proposée par Benoît Mandelbrot (1924-2010), qui introduit simplement des paramètres a, b, c, à ajuster selon le type de texte et la langue étudiés :

$$f(n) = \frac{K}{(a+bn)^c}, (a, b, c, K) \in \mathbb{R}^{\not 4}, n \in \mathbb{N}^*$$

Dans les faits, l'exposant c est, étonnamment, souvent proche de 1 - légèrement supérieur à 1 en français et en anglais contemporains. Cette généralisation permet notamment de mieux rendre compte des valeurs observées pour les mots les plus fréquents.

Première conséquence : une autre moyenne

Comment résumer l'usage du vocabulaire dans un texte ? En moyenne, un mot est utilisé "k fois par texte" n'a pas beaucoup de sens, si on le calcule avec le moyen usuel de la moyenne arithmétique. Les quelques mots outils utilisés à tour de bras vont drastiquement tirer cette moyenne vers le haut de manière ; pourtant, l'immense majorité des mots utilisés dans un texte n'apparaît qu'une poignée de fois, voir une seule fois. Repensons à nos exemples de loi de Pareto et de partage des richesses. Dans un pays où un dictateur et sa famille possèdent 90% des richesses, et où le reste de la population meurt de faim, calculer une "moyenne des revenus" n'a pas grand sens. Cette moyenne conduirait à considérer que la population se porte globalement bien, ce qui est un très mauvais résumé statistique de la situation.

Dans le cas où la distribution a une longue traîne, comme c'est le cas avec la loi de Zipf, on doit privilégier un résumé comme la *moyenne géométrique* :

$$\bar{x}^G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Cette moyenne est en effet moins sensible aux valeurs les plus élevées d'une série.

Seconde conséquence : calcul de spécificités

Ceci n'est pas sans conséquence sur certains calculs qu'on aurait pu croire simples, comme les calculs du vocabulaire *spécifique* à un auteur ou à un texte. Intuitivement, ce qui rend un mot / groupe de mots / phénomène linguistique spécifique à un locuteur ou à un texte etc., c'est qu'il est plus fréquemment utilisé par lui que par les autres. Le calcul à réaliser pour objectiver ces spécificités individuelles pourrait donc simplement consister à relever les fréquences de la forme qui nous intéresse, et la comparer à la fréquence de la même forme chez d'autres locuteurs / dans d'autres textes. Il est bien sûr nécessaire de travailler sur des fréquences relatives et non absolues -on doit rapporter le nombre d'occurrences à la longueur du texte. C'est souvent le calcul proposé, comme dans la base de données https ://www.frantext.fr/, qui calcule des fréquences relatives. Ce calcul simple est potentiellement une approximation convanable, mais est cependant encore insuffisant.

Considérer que la fréquence d'un mot divisée par la taille du texte est un bonne mesure des fréquences d'origine, c'est un réalité faire une hypothèse à laquelle on ne prend pas garde : la "normalité". La moyenne arithmétique est bien le maximum de vraisemblance dans le cas d'une loi normale.Mais l'apparition d'un mot ne suit pas une loi normale... Pierre Lafon [4] choisit donc de développer pour le Français un outil spécifique, calibré sur une loi hypergéométrique, qui collait bien aux données linguistiques pour notre langue. Cette loi ressemble à une gaussienne dissymétrique, avec une queue s'affaissant petit à petit vers les hautes fréquences

Dans ce cadre, on calcule la probabilité qu'une forme A apparaisse f fois dans une partie p de longueur t, la forme apparaissant F fois en tout dans l'ensemble du corpus dont la longueur totale est de T occurrences :

$$Pr(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}$$

Où le vocabulaire V est l'ensemble des mots recensés.

Indice de spécificités :

$$Pr(card\{A \in V | A \in p\} \geq f_{obs}) = \sum_{f_{obs}}^{card\{A \in V | A \in p\}} Prob(card\{A \in V | A \in p\} = f)$$

C'est par exemple le calcul que permet de réaliser par défaut le logiciel de textométrie TXM [5].

Troisième conséquence : le calcul de corrélations

Pour étudier la relation potentielle entre deux variables, ou, pour le dire plus précisément, pour vérifier si ces deux variables ont des variations concomitantes, opposées ou indépendantes, on réalise couramment des calculs de corrélation, et classiquement de *corrélation linéaire* ou *corrélation de Pearson*. Celle-ci va nous indiquer si la relation entre les deux variables se rapproche ou non d'une droite, telle que représentée sur ce type de graphique, en donnant cette fois l'intensité de la relation.

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Par construction, $Corr(X,Y) \in [-1;1]$

- Si Corr (X,Y)= 0, aucune corrélation entre les variations de X et Y
- Si Corr(X,Y)= 1, corrélation positive parfaite : quand X augmente de k%, Y augmente également de k%
- Si Corr(X,Y)= -1, corrélation négative parfaite : quand X augmente de k%, Y diminue de k%

Cet outil bien connu est régulièrement utilisé dans l'analyse de discours. Pourtant, comme on vient de le voir, les non-linéarités sont fréquentes dans ce contexte. Une loi comme la loi de Zipf-Mandelbrot nous indique par exemple qu'il est souvent plus intéressant de s'intéresser aux rangs des mots dans la distribution qu'à leur fréquence d'utilisation

La *Corrélation de Spearman* ou ρ *de Spearman* étudie non pas les valeurs numériques, mais le rang des valeurs. On calcule une corrélation de Pearson étudiant le lien pouvant unir les rangs des valeurs prises par les variables étudiées.

Une variation sur cette idée de corrélation de rangs est la *corrélation de Fechner-Kendall*, ou τ *de Kendall*. On s'intéresse alors au nombre de fois où les valeurs des deux variables évoluent dans le même sens. Encore une fois, par construction, $\tau \in [-1;1]$ Si elles varient tout le temps dans le même sens, $\tau = 1$, si elles varient tout le temps en sens opposé $\tau = -1$, si elles ne sont pas corrélées, $\tau = 0$ Mathématiquement :

$$\tau = \frac{(PC) - (PNC)}{\frac{1}{2}n(n-1)}$$

où PC= nombre de paires concordantes et PNC : nombre de paires non concordantes

3 Toujours plus : ces nouveaux mots qui n'en finissent pas d'apapraître

Lorsque je feuillette un texte, quelle est la probabilité que je vois un nouveau mot apparaître à un certain point de ma lecture ? C'est à ce cette question que la loi de Heaps (aussi appelée loi de Herdan) permet de répondre. Cette loi est en réalité très connectée aux propriétés décrites par loi de Zipf-Mandelbrot. Au prix de quelques hypothèses relativement acceptables, on peut montrer que les deux sont en fait (asymptotiquement) semblables.

Loi de Heaps (ou loi de Herdan, ou Heaps-Herdan)

Souvent attribuée à Harold S. Heaps [6], mais originellement évoquée par Gustav Herdan [7], la loi de Heaps-Herdan explicite et précise mathématiquement une observation relativement intuitive : la probabilité de découvrir, lors de la lecture cursive d'un texte, un mot qui n'était pas apparu jusque là est plus faible en fin de texte qu'en début de texte. Un phénomène est cependant relativement inattendu : aussi loin que l'on aille dans la lecture des textes, les nouveaux mots continuent à apparaître sans cesse !

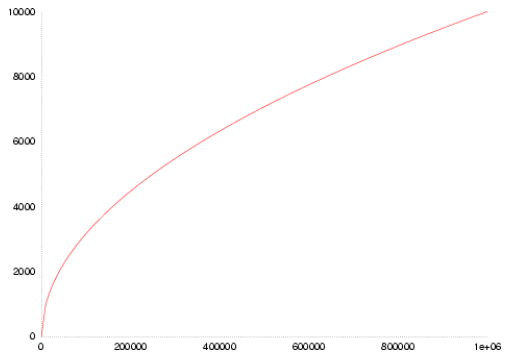
Mathématiquement, on écrit cette loi sous la forme :

$$V_R(n) = Kn^\beta$$

Où :

— $V_R(n)$ nombre de mots dans les n premiers mots d'un texte R

— K et β : paramètres. En pratique, K entre 10 et 100, β entre 0.4 et 0.6.



Conséquence : échantillonage

Il est parfois nécessaire de ne travailler que sur des échantillons de texte :

— Parce que les phénomènes que l'on veut analyser sont trop longs à encoder sur tout le texte

— Parce que l'on veut s'éviter des biais liés à la longueur de textes à comparer.

Quand et comment peut-on légitimement échantillonner ? Grâce à la loi de Heaps-Herdan, on sait qu'avec un échantillon (relativement court) d'un texte, on aura rapidement l'essentiel du vocabulaire d'un auteur. On sait par contre que l'on n'a aucune chance d'avoir l'intégralité du vocabulaire dans notre échantillon. Ce constat est, selon les cas, plutôt rassurant, car on peut travailler de manière assez fiable avec des échantillons courts ; ou plutôt inquiétant, car il est impossible d'avoir un échantillon parfaitement représentatif du vocabulaire d'un auteur.

4 Quand dire, c'est redire

Le phénomène de contagion (*clumping*)

L'apparition d'un mot particulier dans un texte est un phénomène statistiquement rare. La probabilité de voir apparaître le mot "galère" dans un texte est très faible. Mais dès qu'un mot w est apparu une fois, la probabilité qu'il réapparaisse augmente drastiquement :

$$Pr(w_2|w_1) >> Pr(w_1)$$

C'est ce qu'on appelle le phénomène de **contagion**.

Conséquence : binarisation ou ternarisation

Qu'un mot revienne plusieurs fois n'est donc pas aussi significatif que le fait qu'il apparaisse tout court. S'intéresser à la fréquence peut parfois aiguiller vers une fausse route. Si le mot "galère" revient autant de fois dans *l'Avare*, ce n'est pas par exemple parce qu'il est très distinctif du vocabulaire de Molière. Ce n'est pas non plus une marque d'une thématique de la pièce, qui n'est pas consacrée à la mer ou à l'esclavage... Pour détecter que l'auteur parle de la mer, on pourrait préférer voir qu'aux côtés du mot "galère", on trouve dans le vocabulaire les mots "navire", "rames", "flots" etc. Pour ne pas fausser certains calculs à cause de la contagion, on peut donc, selon

la problématique envisagée, travailler une uniquement sur des booléens. On ne compte plus la fréquence d'apparition d'un mot, mais on se demande seulement s'il apparaît ou pas. Des évaluations récentes ont proposé dans le même esprit de réaliser une ternarisation (*ternary quantization*)[8] des fréquences d'apparition d'un mot, avec des performances remarquables dans des tâches d'attribution d'autorité.

5 Un très long texte à pieds, ça use les souliers... ?

Calculer une similarité ou une dissemblance entre deux textes est une opération souvent nécessaire. On sélectionne tout ou partie de textes pour comprendre s'il y a eu plagiat, si deux textes ont le même auteur, sont de la même époque, s'ils évoquent les mêmes thématiques etc. Pour classer des textes de la sorte, les regrouper par points communs stylistiques ou sémantiques, il faut d'abord définir comment on mesurera la "distance" entre deux textes.

Tout comme dans le reste de nos exemples, un réflexe a d'abord été d'appliquer les mêmes recettes au monde du texte qu'au monde physique. Pour calculer une "distance intertextuelle", on pourrait se contenter de calculer une distance euclidienne :

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

C'est la mesure habituelle de la distance, celle du monde physique. Après ce que nous avons vu sur les lois par lesquelles ont peut approximer efficacement les phénomènes linguistiques de base, vous comprenez aisément qu'encore une fois, cette idée de calquer le monde de la physique sur celui de la langue sera voué à l'échec. On ne va pas à pied entre deux textes... Il n'y a donc aucune raison de penser *a priori* que la distance entre deux textes doive se calculer comme la distance qu'un humain parcourt quand il marche en ligne droite entre deux points. Il existe ainsi une très grande variété de métriques employables et employées pour calculer la dissemblance de la langue de textes différents. Parmi elles, la distance euclidienne fait par-

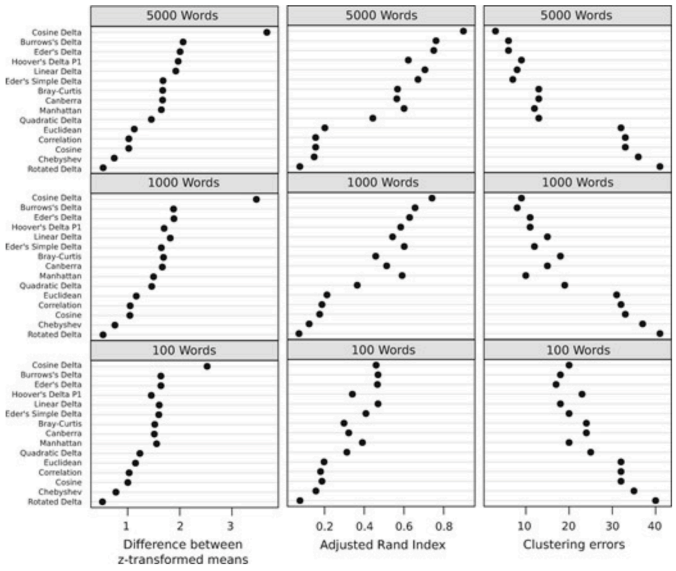
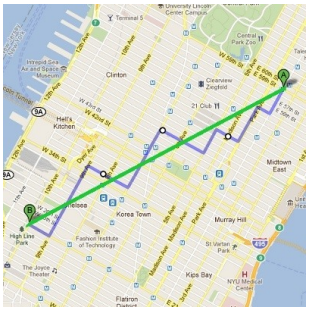


FIGURE 5 – Comparaison de l'efficacité de différentes distances pour des tâches d'attribution d'autorité [11]

tie des moins efficaces dans les tâches comme l'attribution d'un texte à son auteur 5... C'est ce qui explique l'absence de cette mesure de distance dans nos travaux d'attribution d'autorité par exemple [9, 10]. Nous présentons ici quelques alternatives souvent utilisées.

Métrique (2) : distance de Manhattan



Elle est la omme des valeurs absolues des différences entre coordonnées. On parle de géométrie du taxi [12], ou de Manhattan, car cette manière

de calculer correspond à un espace qui ressemblerait à celui que constitue Manhattan pour un chauffeur de voiture : il existe un grand nombre de trajets de distance strictement égale, selon comment on choisit de conduire à travers le plan en quadrillage qui caractérise la grande majorité de ce quartier de New-York

$$\sum_{i=1}^{n} |x_i - y_i|$$

Métrique (3) : distance de Canberra

Beaucoup de mesures de distance utilisées dérivent de la distance de Manahattan. La **distance de Canberra** [13] n'est rien d'autre qu'une distance de Manhattan pondérée.

$$D_{Canb}(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Cette distance permet d'atténuer l'importance de certaines différences majeures d'un point de vue numérique. On s'intéresse plus à l'existence de différences qu'à l'importance quantitative de chacune de ces différences.

Métrique(4) : delta de Burrows

Le **Delta de Burrows** [14] a longtemps été la distance la plus usitée en stylométrie. Considérés comme un outil aussi performant pour la prose que pour la poésie [15], Sa suprématie a cependant été récemment remise en cause [11].

Le Delta de Burrows combine la standardisation des comptes d'occurrences (z-transformation) et d'un changement de métrique (distance de Manhattan) :

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - B_i}{\sigma_i} \right|$$

Il est cependant difficile d'expliquer théoriquement le succès de cette mesure. Il est général difficile de se faire une intuition sur pourquoi une mesure de distance serait plus efficace qu'une autre. De manière générale, on peut dire qu'on attend d'une distance et des éventuelles transformations qu'elle

permette d'étudier des textes de tailles différentes. Pour savoir quelle distance est la plus "efficace", il est nécessité d'avoir recours à l'expérience, sur des corpus donnés, donnant parfois (souvent ?) des résultats contradictoires. Mais dans le cas du delta de Burrows, d'autres questions supplémentaires se posent. Argamon [16] remarque que le delta de Burrows est incohérent mathématiquement : la z-transformation a du sens dans le cas d'un monde euclidien. Mais elle est combinée à une distance de Manhattan... Propose alors deux alternatives pour rendre plus cohérent ce calcul :

— Delta linéaire d'Argamon : distance de Manhattan, mais normalisation non plus selon la moyenne arithmétique et l'écart-type, mais selon la médiane et la dispersion.

— Delta quadratique d'Argamon : distance Euclidienne et z-transformation.

Manque de chance : ces deux mesures plus cohérentes mathématiquement sont moins performantes que le Delta de Burrows...

Métrique (5) Similarité cosinus

Pour mesurer la similarité entre deux textes, on peut représenter chacun des textes comme un vecteur. On calcule alors le cosinus de l'angle formé par les deux vecteurs.

$$\cos(\theta) = \frac{A \times B}{\|A\| \times \|B\|}$$

Cette valeur est théoriquement comprise dans $[-1; 1]$. Mais pour des valeurs toujours positives ou nulles, comme c'est le cas en textométrie, la valeur de cette mesure sera toujours comprise dans $[0; 1]$.

Cette métrique est régulièrement considérée comme la plus performante. Intuitivement, on peut comprendre l'utilité de cette mesure : en se basant sur un angle, sur un rapport entre des valeurs, cette distance permet de s'affranchir plus facilement des questions de normalisation, des problèmes de taille de texte etc.

Références

[1] W. J. Reed, "The pareto law of incomes—an explanation and an extension," *Physica A : Statistical Mechanics and its Applications*, vol. 319, pp. 469–486, 2003.

[2] K. T. Rosen and M. Resnick, "The size distribution of cities : an examination of the pareto law and primacy," *Journal of urban economics*, vol. 8, no. 2, pp. 165–186, 1980.

[3] M. Petruszewycz, "L'histoire de la loi d'estoup-zipf : documents," *Mathématiques et sciences humaines*, vol. 44, pp. 41–56, 1973.

[4] P. Lafon, *Dépouillements et statistiques en lexicométrie*, vol. 24. Slatkine, 1984.

[5] S. Heiden, J.-P. Magué, and B. Pincemin, "Txm : Une plateforme logicielle open-source pour la textométrie-conception et développement," in *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, vol. 2, pp. 1021–1032, Edizioni Universitarie di Lettere Economia Diritto, 2010.

[6] H. S. Heaps, *Information retrieval, computational and theoretical aspects*. Academic Press, 1978.

[7] G. Herdan, *Type-token mathematics*, vol. 4. Mouton, 1960.

[8] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, and T. Vitt, "Understanding and explaining delta measures for authorship attribution," *Digital Scholarship in the Humanities*, vol. 32, no. suppl_2, pp. ii4–ii16, 2017.

[9] J.-B. Camps and F. Cafiero, "Setting bounds in a homogeneous corpus : A methodological study applied to medieval literature," *Revue des Nouvelles Technologies de l'Information*, no. MASHS 2011/2012. Modèles et Apprentissages en Sciences Humaines et Sociales Rédacteurs invités : Mar, pp. 55–84, 2013.

[10] F. Cafiero and J.-B. Camps, "Why molière most likely did write his plays," *Science advances*, vol. 5, no. 11, p. eaax5489, 2019.

[11] F. Jannidis, S. Pielström, C. Schöch, and T. Vitt, "Improving burrows' delta. an empirical evaluation of text distance measures," in *Digital Humanities Conference*, vol. 11, 2015.

[12] E. F. Krause, *Taxicab geometry : An adventure in non-Euclidean geometry*. Courier Corporation, 1986.

[13] G. N. Lance and W. T. Williams, "Computer programs for hierarchical polythetic classification ("similarity analyses")," *The Computer Journal*, vol. 9, no. 1, pp. 60–64, 1966.

[14] J. Burrows, "'delta' : a measure of stylistic difference and a guide to likely authorship," *Literary and linguistic computing*, vol. 17, no. 3, pp. 267–287, 2002.

[15] D. L. Hoover, "Testing burrows's delta," *Literary and linguistic computing*, vol. 19, no. 4, pp. 453–475, 2004.

[16] S. Argamon, "Interpreting burrows's delta : Geometric and probabilistic foundations," *Literary and Linguistic Computing*, vol. 23, no. 2, pp. 131–147, 2008.