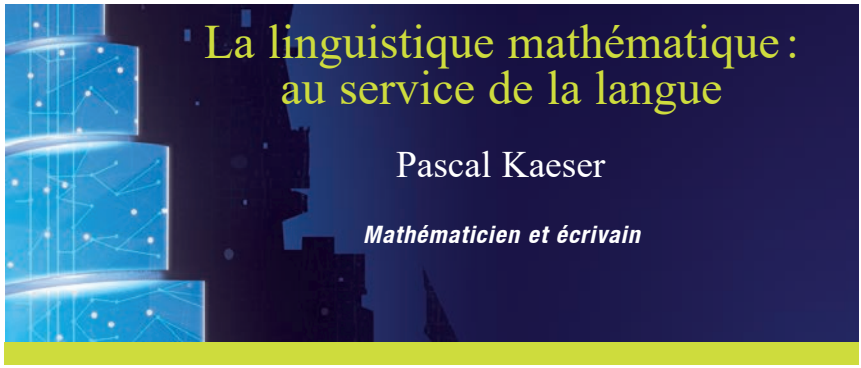


La linguistique mathématique : au service de la langue

Pascal Kaeser

Mathématicien et écrivain

Les mathématiques peuvent nous aider à étudier la langue. La première idée, naïve, consiste simplement à compter les mots : un texte est plus difficile à lire quand le nombre moyen de mots par phrase et le nombre moyen de syllabes par mot sont élevés… Essayons cependant de mettre en évidence, sous la forme d'un dialogue imaginaire, des liens plus profonds !

L'odyssée du nombre : des probabilités à la statistique

« L'idée de compter les mots d'un texte n'est après tout pas si mauvaise. Y a-t-il un domaine qui s'y intéresse sérieusement ?

– Oui, la *lexicométrie* consiste à appliquer la statistique à des textes. Au niveau le plus bas, on cherche à déterminer la richesse d'un vocabulaire, la fréquence d'un mot, d'une fonction grammaticale… À un niveau plus élevé, on se risque à définir des types de distance intertextuelle. Les unes se focalisent sur les mots, d'autres sur les fonctions grammaticales, d'autres encore sur les phonèmes, la ponctuation, *etc*. L'interprétation soulève des problèmes. Une distance peut rendre deux textes "suffisamment proches" pour justifier le soupçon qu'ils aient le même auteur, tandis qu'une autre distance tendrait plutôt à invalider pareille hypothèse.

– Alors, est-ce Molière ou Corneille qui écrivit le *Misanthrope* ?

– C'est un génie, peu m'importe son nom !

– Où nous entraîne encore l'odyssée du nombre sur les flots du *logos* ?

– À l'*Ulysse* de James Joyce. En classant par ordre décroissant les fréquences f_1, f_2… des mots présents dans cette œuvre, le statisticien et linguiste américain George Zipf découvrit une loi curieuse, qui sera publiée en 1949 : f_n vaut environ f_1/n. C'est la fameuse *loi de Zipf* :

George Kingsley Zipf
(1902–1950)
à l'âge de 15 ans.

© Freeport High School, Freeport (Illinois, États-Unis)

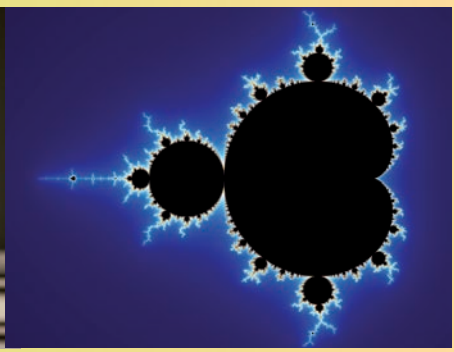
la fréquence d'occurrence d'un mot dans un texte est inversement proportionnelle à son rang dans l'ordre des fréquences.

– C'est donc une observation, une loi empirique. A-t-elle un fondement ?

– Il semble : dans les années 1950, Benoît Mandelbrot démontre, grâce à la théorie de l'information de Shannon, que la loi de Zipf est en fait un cas particulier d'une loi plus générale, où le choix des paramètres permet de mieux coller à la réalité. Il utilise pour ce faire la *loi statique de Shannon*, qui énonce que le coût de représentation d'une information augmente comme le logarithme du nombre des informations à considérer, et la *loi dynamique de Shannon*, selon laquelle les symboles les moins coûteux sont à exploiter en priorité. Tu sais, comme dans le code Morse : la lettre "e", la plus fréquente, est matérialisée par un simple point alors que le "x", plus rare, est codé par "trait point point trait".

– Une créature littéraire accouche de théorèmes mathématiques. C'est beau…

– Connais-tu l'*Eugène Onéguine* de Pouchkine ?



Le mathématicien franco-américain Benoît Mandelbrot (1924–2010) est le père des fractales.

© Rama, 14 mars 2007 *© Wolfgang Beyer, 2006*

– Oui, je l'ai lu dans une prodigieuse traduction en octosyllabes, due à André Markowicz.

– Amusante proximité ! Je veux en effet te parler d'Andreï Markov. En 1913, ce mathématicien russe dégagea une notion, nommée plus tard *chaîne de Markov*, d'une étude statistique portant sur les vingt mille premières lettres d'*Eugène Onéguine*, en ne retenant que l'aspect voyelle–consonne. Il compta les paires de voyelles adjacentes, les paires de consonnes adjacentes, et les autres paires, de manière à pouvoir évaluer par exemple la probabilité qu'une voyelle succède à une voyelle.

– Cette idée a-t-elle débouché sur un outil mathématique nouveau et efficace ?

– Au-delà de ce que tu peux imaginer : les chaînes de Markov sont omniprésentes en sciences et incontournables pour nombre d'ingénieurs ! »

Andreï Andreïevitch Markov (1856–1922) fut le premier à décrire les processus stochastiques.

© Domaine public

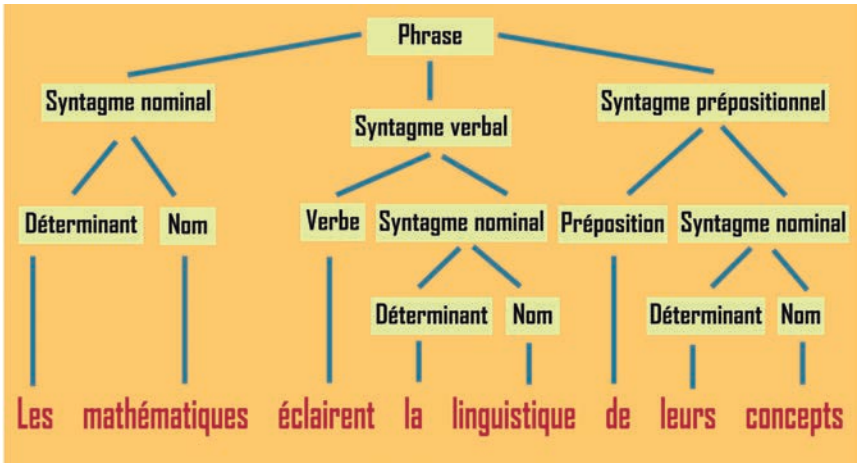
Le choix des termes : dites-le avec des mots… imagés

« Selon le philosophe allemand Arthur Schopenhauer, le monde doit être considéré comme volonté et comme représentation. Quelles figures me proposes-tu pour représenter le verbe ?

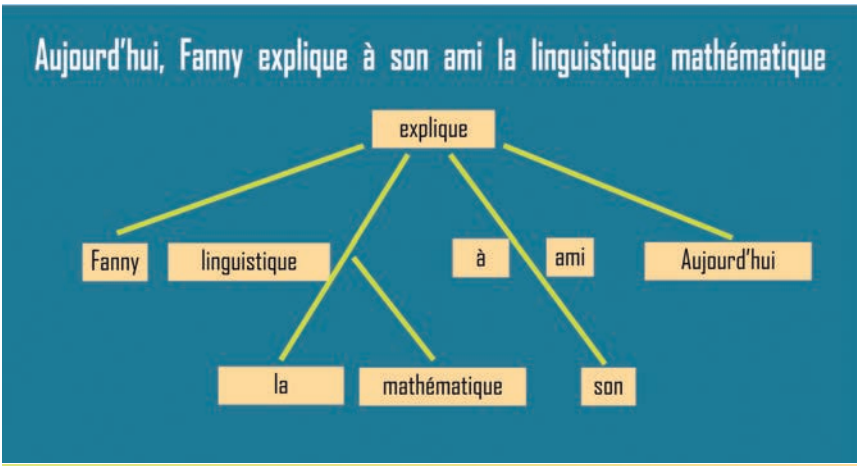
– Dans le jardin des mathématiques, l'homme a planté plusieurs arbres de la connaissance. L'*arbre syntagmatique* de Noam Chomsky (1957) décompose une phrase en groupes, puis les groupes en sous-groupes, *etc.*, jusqu'à obtenir au bout de chaque branche un mot de la phrase. L'*arbre de dépendance* de Lucien Tesnière (1959) hiérarchise les mots selon des relations de subordination : dans une phrase, à part le verbe, chaque mot est le dépendant d'un autre. L'*arbre à bulles* de Sylvain Kahane (1997) s'en inspire pour étudier plus finement les phrases complexes.

– Le choix des termes semble très suggestif et visuel…

– Un petit dessin vaut souvent mieux qu'un long discours !



Un arbre syntagmatique.



Un arbre de dépendances.

Un autre outil graphique est la clique. En 1998, Sabine Ploux et Bernard Victorri ont mené une étude sur la polysémie de "sec". Ils ont construit un graphe dont les soixante-trois sommets sont le mot "sec" et ses synonymes trouvés dans sept dictionnaires. Chaque fois que deux de ces soixante-trois mots sont synonymes, une arête les relie. Ce graphe n'est pas *complet* : en effet, les synonymes de "sec" ne sont pas tous synonymes entre eux (par exemple, "brusque" et "maigre" ne sont pas synonymes), et donc il existe des paires de sommets non reliés. Une

clique est un ensemble maximal de sommets tous reliés deux à deux. Le graphe des synonymes de “sec” comporte quatre-vingt-quatorze cliques, comme {sec ; fauché ; pauvre}. Une méthode classique permet d’associer à chaque clique un point dans un espace à soixante-trois dimensions. Avec des outils judicieux, on peut alors découper le nuage de points en plusieurs zones qui correspondent chacune à un sens de “sec”. Six acceptions principales se dégagent : qui manque d’eau ; maigre, décharné ; stérile, improductif ; qui manque de sensibilité ; bref, abrupt ; seul.

– Je serais probablement parvenu au même résultat en sondant ma mémoire…

– À l’ère de l’ordinateur, on justifie souvent la mathématisation d’une tâche de l’esprit par la possibilité d’automatisation qui en résulte.

– Est-ce une bonne chose ?

– Cette question nous entraînerait trop loin. Le rêve d’algébriser la grammaire, qui s’inscrit lui aussi dans le courant du traitement automatique des langues, a suivi plusieurs voies. Yehoshua Bar-Hillel propose dès 1953 des lois de simplification qui permettent de savoir si une phrase abstraite, c’est-à-dire une succession de fonctions grammaticales, appartient ou non à une catégorie donnée de grammaire formelle. Dans les années 1960, Chomsky remporte un grand succès, tant auprès des linguistes que des informaticiens, avec sa théorie des grammaires génératives. »

Des aventures dont certaines restent à imaginer

« Les grammaires génératives ? De quoi s’agit-il ?

– On part d’un symbole initial, ou d’une chaîne de symboles. Selon des règles fixées, on opère des substitutions successives jusqu’à la construction d’une phrase. Les règles qui définissent une grammaire générative sont de la forme suivante : telle chaîne C de symboles peut être remplacée par telle autre chaîne C’ de symboles ou tel symbole S peut être remplacé par tel mot M. Dans chaque cas, l’unicité n’est pas requise.

– Que de chemin parcouru depuis l’introduction d’une forme de structuralisme dans la linguistique par Ferdinand de Saussure au début du XX^{e} siècle ! Ou même depuis Leibniz, qui se demandait quel est le nombre d’énoncés, de taille variant de 1 à un entier fixé, que l’on peut formuler avec un alphabet de vingt-quatre lettres. Ces nouvelles perspectives ouvrent des horizons étonnants…

Portrait de Gottfried Wilhelm Leibniz (1646–1716) peint par Christoph Bernhard Francke vers 1700.

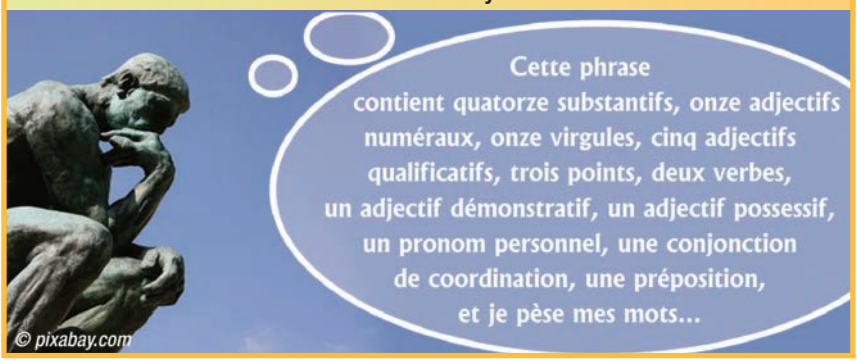
© Herzog Anton Ulrich Museum

– Eh oui, la langue nous permet d'accomplir de beaux voyages. Il y a ceux dont nous avons parlé, il y en a beaucoup d'autres. Traverser des espaces topologiques pour cueillir des fleurs de la sémantique, développer des méthodes combinatoires pour identifier des suites de lettres, ou engager monoïdes libres et demi-groupes pour mettre en concert la phonologie, combien d'aventures, dont certaines restent à imaginer, attendent l'amateur de mots !

Le pouvoir de la langue, c'est d'amener le poète, le philosophe, le mathématicien et le linguiste à s'embrasser. »

P. K.

Exercice de Style.



Pour en savoir (un peu) plus :

« Œuvres complexes. Rire, frémir, rêver, rimer, férir. »
Site Internet administré par Pascal Kaeser, ***www.pascalkaeser.ch*** .

« Fatrazie, jeux de l'esprit. »
Site Internet administré par Alain Zalmanski, ***www.fatrazie.com*** .