Dialogue sur la linguistique mathématique par Pascal Kaeser

[Ce texte est la première version d'un article (très modifié) paru dans la brochure
Maths Langages express, éditée par le Comité International des Jeux Mathématiques en 2017
(voir : www.cijm.org/accueil/productions-cijm/90-maths-express)]

– Comment draguer la langue avec les maths ?
– En jouant 4 atouts : la force du nombre, l'érotisme de la combinaison, le sourire de la figure, le tango de l'algèbre.
– Jusqu'à quel point peut-on compter sur le nombre ?
– La langue a son palais que le nombre mesure.
– Exemples ?
– Il y a la plaisanterie un peu lourde imaginée par ceux qui pèsent les mots. Partant de l'idée qu'un texte est probablement plus difficile à lire quand le nombre moyen de mots par phrase et le nombre moyen de syllabes par mot sont élevés, ils bricolent des indices qui gagnent du galon en confirmant des évidences comme : Proust est moins accessible que le Reader's Digest. Des institutions américaines jugent ces formules grossières suffisamment polies pour calibrer des contrats d'assurance et des formulaires de consentement éclairé.
– Éclaire-moi sur la lexicométrie ! De quoi s'agit-il ?
– D'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'un groupe de mots, d'une fonction grammaticale, etc. À un niveau plus élevé, on se risque à définir plusieurs 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 pareil soupçon.
– 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, f_3, etc. des mots présents dans cette œuvre, Zipf découvrit une loi curieuse (publiée en 1949) : f_n vaut environ f_1/n. Dans les années 50, Mandelbrot démontra grâce à la théorie de l'information que la loi de Zipf était un cas particulier d'une loi plus générale, où le choix des paramètres permet de mieux coller à la réalité.
– Une créature littéraire accouche d'une créature mathématique... C'est beau.
– Connais-tu l'*Eugène Onéguine* de Pouchkine ?
– Oui, je l'ai lu dans une prodigieuse traduction en octosyllabes, due au talent d'André Markowicz.
– Amusante proximité ! Je veux 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. Markov avait un caractère combatif. Quand l'église orthodoxe russe excommunia Tolstoï, Markov adressa une requête pour en être lui aussi exclu. En signe de protestation contre le veto exercé par le tsar en vue d'empêcher l'entrée de Maxime Gorki à l'Académie, Markov annonça qu'il refuserait dorénavant tout honneur venant de Nicholas II. Une polémique assez violente opposa Markov à Nekrassov, un mathématicien qui prétendait démontrer l'existence du libre arbitre en s'appuyant sur la théorie des probabilités. L'erreur de Nekrassov était de supposer que la loi des grands nombres ne s'appliquait qu'à des variables indépendantes.

– À propos de grands nombres, il paraît que des moines tibétains ont déterminé un système de contraintes qui amène à tester 9 milliards de possibilités pour écrire le véritable nom de Dieu. Un ordinateur est chargé d'imprimer la liste. Quand l'authentique nom sortira , Dieu fera disparaître l'univers.

– Tu es un farceur ! Ce que tu me racontes là, c'est l'argument d'une célèbre nouvelle de Clarke, publiée en 1953. Le pouvoir de l'ineffable nom de Dieu est une vieille croyance juive, que l'on rencontre aussi dans le mythe du Golem. La Kabbale médiévale tourne autour de l'idée qu'en explorant la combinatoire du langage, l'homme approfondira sa connaissance de Dieu, du Monde en soi.

– « Le jeu combinatoire semble être le trait essentiel de la pensée productive », disait Einstein.

– Oui. Un outil combinatoire par excellence s'obtient en disposant des mots ou des symboles sur des cercles concentriques mobiles. Au 13e siècle, Lulle exploite cet instrument pour élaborer des questions philosophiques (par exemple : la bonté est-elle grande en ce qu'elle contient des choses différentes ?) ; au 16e siècle, Bruno s'en sert pour composer des images littéraires (par exemple : une femme à cheval sur un taureau peigne ses cheveux en tenant un miroir dans sa main gauche, tandis qu'un adolescent avec un oiseau vert sur la main assiste à la scène) ; au 17e siècle, Harsdörffer tire de ce système 99 millions et demi de mots potentiels pour la langue allemande (en plaçant 48 syllabes préfixes sur le 1er cercle, 60 chaînes de caractères sur le 2e, 12 caractères sur le 3e, 120 chaînes de caractères sur le 4e et 24 syllabes suffixes sur le 5e).

– Leibniz n'a-t-il pas marché sur leurs traces ?

– Oui, de plusieurs façons. Il se posa la question suivante : quel est le nombre maximal d'énoncés (vrais, faux, insensés) formulables avec un alphabet de 24 lettres? Si M est la taille maximale d'un énoncé, la réponse est la somme d'une suite géométrique de raison 24 et de premier terme 24, soit: (24^(M+1) – 24)/23. Leibniz prend pour M le nombre de caractères, à raison de 100'000 par jour, qu'un homme peut lire en 1000 ans, ce qui donne environ: 3.65E10. Un petit calcul de logarithme nous permet alors de voir que le nombre maximal d'énoncés comporte plus de 50 milliards de chiffres. Dans une œuvre de jeunesse, Leibniz envisage de

hiérarchiser les idées. En combinant par 2 les idées simples et primitives d'une 1re classe, il obtient une 2e classe; en les combinant par 3, une 3e classe; etc. Il représente les termes de la 1re classe par des nombres et les combinaisons par des groupes de nombres. Quelques années plus tard, il améliore cette notation en associant des nombres premiers aux termes de la 1re classe et des produits aux combinaisons. Ainsi, les termes de la 2e classe correspondent aux nombres qui ont 2 facteurs premiers, etc. La divisibilité devient alors un critère pour savoir si une idée de classe N entre dans la composition d'une idée de classe supérieure à N.

– Ce que beaucoup de gens retiennent de Leibniz, c'est que Voltaire s'est foutu de sa gueule dans *Candide*.

– À tort, de mon point de vue. Si on accepte le postulat suivant : « le monde est l'œuvre d'un Dieu qui représente le plus haut degré d'intelligence et de bonté », alors il me paraît tout à fait raisonnable d'en conclure que notre monde est le meilleur des mondes possibles. Naturellement, nous pouvons aussi rejeter ce postulat...

– Selon 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 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 Tesnière (1959) hiérarchise les mots selon des relations de subordination. L'arbre à bulles de Kahane (1997) s'en inspire pour étudier plus finement les phrases complexes. Un autre outil graphique est la clique. En 1998, Ploux et Victorri ont mené une étude sur la polysémie de « sec ». Ils ont construit un graphe dont les 63 sommets sont le mot « sec » et ses synonymes trouvés dans 7 dictionnaires. Chaque fois que deux de ces 63 mots sont synonymes, une arête les relie. Ce graphe n'est pas complet, car les synonymes de « sec » ne sont pas tous synonymes entre eux, par exemple « brusque » et « maigre » ne sont pas synonymes. Une clique est un ensemble maximal de sommets tous reliés deux à deux. Le graphe des synonymes de « sec » comporte 94 cliques, par exemple {sec; fauché; pauvre} ou {sec; bref; brusque; tranchant}. Une méthode classique permet d'associer à chaque clique un point dans un espace à 63 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 : 1. qui manque d'eau ; 2. maigre, décharné ; 3. stérile, improductif ; 4. qui manque de sensibilité ; 5. bref, abrupt ; 6. 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. Bar-Hillel (1953) propose 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 60, Chomsky remporte un grand succès, tant auprès des linguistes que des informaticiens, avec sa théorie des grammaires génératives.

– De quoi s'agit-il ?

– L'idée est de produire une infinité de phrases à l'aide d'un nombre fini de règles de réécriture. Plus précisément, on part d'un symbole initial (ou 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 : telle chaîne de symboles peut être remplacée par telle autre chaîne de symboles ou tel symbole peut être remplacé par tel mot. Dans chaque cas, l'unicité n'est pas requise.

– Est-ce l'étude mathématique du langage qui a fait de Chomsky un champion d'une liberté maximale d'expression ?

– Peut-être… qui sait ?

– En France, la loi Gayssot réprime les propos négationnistes. Chomsky a signé une pétition réclamant l'abrogation de cette loi.

– User de l'arme juridique pour restreindre la liberté d'expression, n'est-ce pas illusoire ? Supposons que la loi m'interdise d'énoncer : « Il faut exterminer les Suisses ». Ai-je le droit, comme je viens de le faire, de citer cette phrase ? Dans le cas contraire, le juge, les avocats et les journalistes ont-ils, eux, le droit de citer cette phrase lors d'un procès contre une personne accusée de l'avoir dite ? Ai-je le droit de faire dire cette phrase a un personnage de fiction ? Un cinéaste a-t-il le droit de tourner un film qui met en scène un procès dans lequel juge, avocats et journalistes citent des propos interdits ? Ai-je le droit de dire : « Faut-il exterminer les Suisses ? Certaines personnes le pensent. » ? Ai-je le droit de dire : « Il ne faut pas exterminer les Suisses et je n'ai pas le droit de dire le contraire, hélas. » ? Ai-je le droit de dire : « Il faut terexminer les Suisses. » ? Comme l'avance Zinoviev dans *Les hauteurs béantes*, une loi limitant la liberté d'expression ne peut prévoir toutes les astuces qui permettront de la contourner.

– Il y a une parade : opter pour des lois vagues qui laissent au juge une grande marge d'appréciation.

– En Occident, la tendance est plutôt à la précision ; souvent, la lettre l'emporte sur l'esprit.

– La langue est probablement trop agile pour que la loi puisse en contrôler tous les mouvements, et pour que les mathématiques puissent en modéliser toutes les potentialités.

– Gödel, qui trouva des failles dans les mathématiques, semble en avoir aussi découvert une dans la constitution des USA. En 1947, il étudia soigneusement ce texte avant l'audition requise pour être naturalisé américain. Gödel, accompagné de ses deux témoins Einstein et Morgenstern, dit à l'examinateur, le juge Forman, que la constitution américaine, comme celle de l'Autriche, comportait une faille. Un moyen existait d'instaurer légalement une dictature. « Je peux le prouver », déclara-t-il. Un mémorandum de Morgenstern relate cette histoire. Hélas, on ne sait rien du

raisonnement de Gödel.

– Par contre, après sa mort, on a retrouvé dans ses papiers une preuve de l'existence de Dieu, qui formalise des idées de Saint-Anselme, Descartes et Leibniz.

– Toute preuve ontologique mise sur un idéalisme. La puissance déductive du langage ne permet d'aboutir qu'à l'existence d'un Dieu abstrait, d'une Idée de Dieu, d'un objet mathématique, en somme.

– Certains physiciens soutiennent que l'univers, tel que la science peut le modéliser, est en quelque sorte un objet mathématique. Ce point de vue nous offre une opportunité de conférer à un Dieu abstrait un statut ontologique comparable à celui de l'univers.

– Je ne crois pas. L'univers se modélise à travers un dialogue compliqué entre expérience et langage, tandis que le Dieu des preuves ontologiques sort d'un processus linguistique simple, beaucoup trop simple. C'est un retour à la Kabbale, avec d'autres outils. Je ne pense pas que la langue, même avec l'appui des maths, puisse nous conduire à Dieu, mais elle 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, engager les demi-groupes pour mettre en concert la phonologie, etc.

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