Des chif

Des physiciens et des linguistes font appel à l'algèbre linéaire pour analyser la structure d'œuvres littéraires comme «Moby Dick» ou «Hamlet». Petit voyage lettré dans les espaces vectoriels en compagnie du professeur Jean-Pierre Eckmann

Comment un texte peut-il traduire une pensée? Autrement dit, n'est-il pas curieux que l'on puisse réduire à une séquence de mots unidimensionnelle les idées complexes, profondes et intriquées que peut générer le cerveau humain? La difficulté de cet exercice se reconnaît par le fait que celui qui lit un texte ne comprend pas toujours correctement ce que l'auteur a voulu dire. Il est même considéré comme un art majeur que de savoir coucher par écrit ses idées de manière intelligible par tout le monde.

Mémorisation facilitée

Dans un article paru dans la revue *Proceedings of the National Academy of Sciences* du 23 mai, une brochette de physiciens et de linguistes a fait appel aux mathématiques pour tenter d'y voir plus clair dans cette faculté mystérieuse du langage. Jean-Pierre Eckmann, professeur au Département de physique théorique et à la Section de mathématiques, et ses collègues israéliens, espagnols et allemands ont développé un algorithme capable d'analyser objectivement la composition d'un texte afin d'en discerner les ficelles sous-jacentes. Il en ressort que le découpage d'un ouvrage en sections, chapitres et paragraphes distincts favorise des corrélations de longue durée entre des ensembles de mots, eux-mêmes définissant une idée ou un concept précis. Une telle disposition faciliterait la mémorisation de certains passages et leur rappel plus loin dans la lecture lorsque c'est nécessaire. Ce résultat confirme les hypothèses posées bien avant eux par les linguistes et les philosophes. Ils leur offrent du même coup une base formelle et quantitative qui leur manquait jusque-là. «*Des penseurs comme le Polonais Roman Ingarden et le Tchèque Bernard Bolzano ont réfléchi à ces problèmes il y a longtemps déjà,* explique Jean-Pierre Eckmann. *Bolzano a écrit en 1837 qu'un texte scientifique, pour être intelligible, doit avoir une structure hiérarchique en paragraphes, sections, chapitres, etc. Ingarden évoque lui aussi des "unités structurelles" et des "couches de compréhension". Selon le point de vue du philosophe, le cerveau parviendrait ainsi à comprimer des parties du texte pour en faciliter la mémorisation et les restituer plus loin dans la lecture. Une manière de rendre une profondeur à un discours forcément rectiligne.*»

Jean-Pierre Eckmann et ses collègues ont approché le problème de manière moins empirique, comme le feraient des physiciens devant un phénomène naturel. Pour eux, un texte représente un ensemble de mots, extraits de la totalité des mots existants dans une langue donnée (anglaise en l'occurrence), agencés selon une logique qu'il convient de découvrir. La méthode d'investigation qu'ils ont développée fait appel aux notions de l'algèbre linéaire (espace à plusieurs dimensions, vecteurs, matrices, valeurs propres, etc.). Les œuvres étudiées, elles, ont été choisies pour leur capacité reconnue à transmettre des idées de manière claire. Il y en a douze, dont *Moby Dick* de Herman Melville, *Les Aventures de Tom Sawyer* de Mark Twain, *Hamlet* de William Shakespeare ou encore *La Théorie de la relativité restreinte et générale* d'Albert Einstein.

Les chercheurs ont travaillé à l'intérieur d'un espace mathématique dans lequel chaque mot utilisé représente une coordonnée. Cet espace a donc autant de dimensions qu'il y a de termes différents dans un livre (sont exclus les conjonctions, prépositions et autres petits mots sans signification intrinsèque), c'est-à-dire plusieurs centaines de milliers parfois. Pour simplifier un peu, seuls les mots apparaissant à une fréquence considérée comme minimale ont été conservés, réduisant ainsi drastiquement le nombre de dimensions.

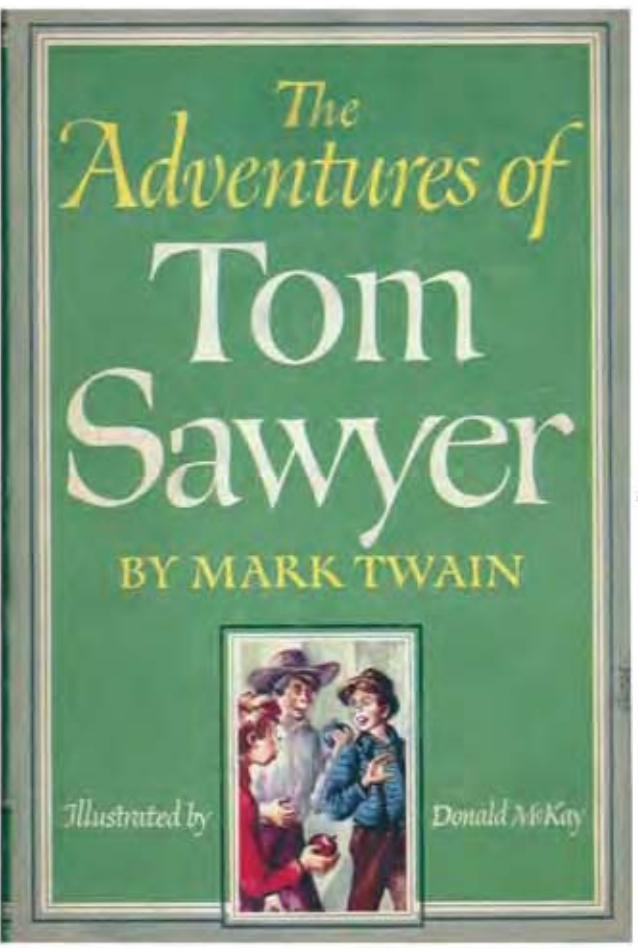
«Fenêtre d'attention»

L'étape suivante a consisté à définir une «fenêtre d'attention» longue de 200 mots qui correspond à la longueur de texte que le cerveau garde en «mémoire vive» au cours de la lecture. Avec les termes contenus dans cet intervalle, les chercheurs ont défini un vecteur désignant un point dans l'espace décrit plus haut. Au cours de la lecture, les composants de la «fenêtre d'attention» se modifient au fur et à mesure qu'elle glisse le long du texte. Du coup, le vecteur correspondant entame une promenade dans l'espace des mots pour décrire une trajectoire propre à l'œuvre étudiée.

Il est désormais possible de manipuler les textes comme des objets géométriques

Possédant un espace bien défini et un vecteur évoluant au cours du temps en fonction de la progression du texte, les chercheurs ont dès lors tout loisir d'appliquer la panoplie des opérations four-

fres et des lettres



«Les Aventures de Tom Sawyer» figurent parmi les douze œuvres décortiquées par les physiciens, tout comme «Moby Dick» ou «Hamlet».

nie par l'algèbre linéaire. En d'autres termes, il est désormais possible de manipuler les textes comme des objets géométriques. Pour les auteurs, le vecteur représente à chaque moment un concept, une idée contenue dans le texte. Ils ont donc analysé les directions principales que ce vecteur visite lors de sa promenade à travers le texte – certaines orientations s'avèrent en effet plus importantes que d'autres – et en ont déterminé les «valeurs propres». Pour le roman *Moby Dick,* par exemple, les principales composantes du vecteur le «plus important» fournissent immédiatement la trame de l'histoire: baleine, Achab, Starbuck (premier compagnon d'Achab), sperm (nom anglais pour cachalot), bateau, cri, aye (forme anglaise pour dire oui), Stubb (second compagnon d'Achab), sir et Léviathan. Les trois vecteurs suivants sont similaires avec l'apparition de quelques nouveaux mots comme blanc, requin, capitaine et navire. Le cinquième, lui, indique brusquement une scène précise, qui se déroule à l'intérieur d'une chambre: lit, chambre, Queequeg, dat (forme populaire pour that), aye, porte, Moby, Dick, propriétaire, Achab.

«Quantifier des intuitions philosophiques»

L'étape suivante a consisté en une analyse dynamique du vecteur. Lorsque ce dernier pointe dans une région de l'espace des mots, par exemple, combien de temps lui faut-il pour la quitter? Autrement dit, il s'agit de mesurer la corrélation entre les mots, une grandeur essentielle dans la compréhension du fonctionnement du langage. Dans les ouvrages étudiés, les chercheurs ont remarqué que cette corrélation diminue très lentement. Ils ont également observé que les résultats restaient identiques même en changeant de manière aléatoire la place des mots à l'intérieur des paragraphes, tout en laissant ces derniers dans le même ordre. D'où la conclusion de l'importance du découpage d'un texte pour améliorer sa compréhension.

«Nous commençons à comprendre de manière quantitative à quel point la structure d'un texte est utile pour le rendre intelligible, estime Jean-Pierre Eckmann. *Et il nous semble qu'un bon texte fait justement un usage efficient des techniques permettant de mémoriser ses différentes parties. Certes, ces découvertes ont été réalisées avant nous par les linguistes, mais nous avons trouvé une méthode pour quantifier leurs intuitions philosophiques et pour les connecter avec le peu que nous savons sur le fonctionnement du cerveau. C'est cette tension entre les mathématiques et des disciplines qui n'ont à première vue aucun lien avec elles qui me passionne. J'espère que d'autres recherches comme la nôtre aideront à clarifier la nature intrigante du cerveau et de la communication humaine.»* ■

Anton Vos