

Le sens des mots

La statistique de la distance entre deux apparitions du même mot permet de distinguer deux catégories de mots.

Tous les mots de cette phrase n'ont pas la même importance pour sa compréhension. Pour s'en persuader, il suffit de la relire en supprimant soit les huit mots *Tous, les, de, cette, n, ont, la, sa*, soit un seul des trois mots *phrase, importance* ou *compréhension*. Dans le premier cas, la phrase est encore compréhensible, dans le second, plus du tout. Certains mots organisent la syntaxe de la phrase : ce sont les mots que nous nommons «mots grammaticaux». D'autres servent plutôt à exprimer le sens, ce sont les «mots pleins». Nous présentons ici une méthode permettant de distinguer ces populations de mots selon la distance en mots séparant deux occurrences d'un même mot. La base de données utilisée est composée de deux années d'articles du quotidien *Le Monde*, soit environ 40 millions de mots.

Imaginons un texte où un mot *M* est réparti au hasard : en chaque position, la probabilité de trouver le mot *M* est indépendante de la nature des mots qui le précèdent. La loi de probabilité de la distance séparant deux apparitions de *M* est appelée, dans ce qui suit, loi de distribution aléatoire. Réciproquement, si la loi de distribution d'un mot est la loi de distribution aléatoire, la probabilité de le rencontrer en une position ne dépend pas des mots qui l'entourent.

En examinant la loi de distribution du mot *de*, on constate qu'excepté pour des petites distances (jusqu'à six mots), cette loi est très proche de la loi aléatoire : on rencontre le mot *de* dans un texte comme s'il y était réparti aléatoirement. En se fixant un seuil arbitraire de «ressemblance» par rapport à une distribution totalement aléatoire, nous pouvons établir une liste de mots qui ont ce comportement : par exemple, les prépositions, *de, par*, les pronoms, *en, il*, les déterminants, *la, les*, les formes conjuguées des verbes *mettre, faire*, qui sont des verbes support de la langue (verbes auxiliaires utilisés dans la nominalisation), les adverbes *probablement* et

relativement. Autrement dit, les mots grammaticaux ont une distribution quasi aléatoire. Le choix du seuil de vraisemblance est arbitraire, mais un mot ayant généralement plusieurs significations, la loi calculée est une moyenne sur ses différentes significations.

Pour les mots que nous ne savons classer dans l'une ou l'autre des catégories, la méthode fera un choix qui dépendra du seuil. Plus la tolérance sera faible, plus les mots trouvés seront «purement grammaticaux». Par exemple, le verbe *mettre* est un mot plein dans les phrases où il exprime le déplacement d'un objet, mais c'est un mot grammatical lorsqu'il est utilisé comme verbe support (comme dans *mettre dans l'embarras*).

Ce critère, qui permet de distinguer les mots grammaticaux n'est pas un critère de fréquence : certains termes fréquents comme *France* n'ont pas une distribution aléatoire. L'écart de ces distributions à la loi aléatoire pour les petites distances est de nature syntaxique : trouver deux occurrences successives du mot *de* est très peu probable, alors que, pour une distribution aléatoire, cette probabilité serait supérieure à celle de trouver deux occurrences séparées par un autre mot. De surcroît, la signature d'un mot (c'est-à-dire l'allure de la loi de probabilité) est relativement stable par

changement de corpus : dans des corpus aussi variés qu'une traduction de la Bible et un quotidien, la signature du mot *de* est semblable. La signature des mots grammaticaux ne dépend pas du contenu sémantique des textes, mais du style de l'auteur.

LES MOTS PLEINS

En revanche, on constate que la distribution des mots pleins (qui intuitivement semblent être «porteurs de sens») n'est pas aléatoire : la présence de certains mots dans une partie du texte dépend des autres mots de cette partie. Cette dépendance entre les différents termes est associée au contexte. Dans un article traitant de politique, il est plus probable de trouver des mots comme *ministre* ou *gouvernement*, que dans un article traitant de sport. Réciproquement, un article contenant le mot *gouvernement* traite plus probablement de politique que de sport. Ces deux remarques entraînent que, dans un article contenant le mot *gouvernement*, il est plus probable qu'ailleurs de trouver le mot *ministre*. On dit alors que le mot *gouvernement* a un effet de contexte positif sur le mot *ministre*, et le mot *match* a un effet de contexte négatif sur le mot *ministre*. Nous avons quantifié cet effet de contexte, en comparant la loi de distribution d'un mot dans un contexte donné, avec la loi de distribution aléatoire qu'il aurait s'il était à comportement grammatical.

Les perspectives ouvertes par ces études sont au moins triples. La première est l'étude de méthodes de compression de texte (ou d'index) performantes, fondées sur la finesse de la description de la loi de distribution des mots. La deuxième est l'élaboration de catégories sémantiques autour d'un mot, par l'examen des effets de contexte. À terme, on pourrait lever l'ambiguïté sur le sens des mots en étudiant leurs contextes, ce qui pourrait être appliqué à la traduction automatique. Par exemple, pour le mot *jeu*, nous avons, d'une manière totalement automatisée, obtenu trois ensembles de mots appartenant à trois champs sémantiques distincts : le contexte sportif, le contexte théâtral, le contexte musical et nous avons aussi obtenu une liste d'expressions contenant le mot *jeu* employé dans n'importe quel contexte. Enfin, on pourrait caractériser un auteur par les signatures, dans son œuvre, des mots à comportements grammaticaux.

Jean SENELLART – LADL, Paris

	PC	PS	RPR	UDF	écologie	spectacle
Barre	0,65	1,29	3,78	4,13	1,89	-1,55
Chirac	-0,08	0,86	5,97	3,64	1,49	-0,65
Cresson	-0,37	1,49	0,32	0,19	1,12	-1,30
Lalonde	0,28	2,77	1,19	1,66	9,43	-0,19
Marchais	2,51	2,09	2,34	1,23	5,02	-4,32
Mitterrand	-1,18	0,64	1,09	0,73	0,56	-1,22

Ces valeurs mesurent la probabilité de trouver une mention d'un personnage politique (*colonne*) dans un bloc de 100 mots autour de la mention d'un parti politique ou d'un autre mot (*ligne*). Plus la valeur est élevée, plus l'effet de contexte est important. La valeur maximale pour un personnage politique s'obtient généralement pour son propre parti, sauf pour le mot *Mitterrand* qui présente un comportement plus grammatical. Le mot *spectacle* à un effet de contexte négatif pour toutes les mentions de personnages politiques.

Les membranes de nos ancêtres

On trouve dans le pétrole la composition chimique des premières cellules vivantes.

Comment les premières cellules vivantes se sont-elles formées? La réponse que proposent Guy Ourisson, Yoichi Nakatani et leurs collaborateurs du Centre de neurochimie de Strasbourg se fonde sur des études *a priori* sans relation avec l'étude de la chimie prébiotique : l'analyse chimique du pétrole.

Il y a une trentaine d'années, G. Ourisson et Pierre Albrecht étudiaient les molécules constituant la matière organique des sédiments : pétroles, charbons, argiles, schistes, calcaires, etc. Ils observaient que la présence d'une molécule nommée cholestane, dans de nombreux sédiments, est un indice de la présence initiale de cholestérol (ce dernier, que l'on considère comme un ennemi de nos artères, est un constituant des membranes cellulaires de tous les êtres vivants, à l'exception des bactéries). Le cholestane est fréquemment accompagné de ses dérivés portant un ou deux atomes de carbone supplémentaires, structures que l'on retrouve dans les stérols des plantes ou de certains micro-organismes. Les «fossiles moléculaires» de stérols sont souvent accompagnés de produits nommés diastéranes, dont on peut obtenir le squelette par chauffage modéré des dérivés du cholestane avec des argiles : cette «simulation» renforce l'hypothèse d'une présence de cholestérol d'êtres vivants ayant conduit, par chauffage sous pression, dans les entrailles de la Terre, à la formation des pétroles.

Lors de cette étude de la matière organique des sédiments, les chimistes trouvèrent des composés qui furent nommés «géohopanoïdes», dérivés d'une molécule nommée hopane, à 30 atomes de carbone. Cette molécule est un triterpène, c'est-à-dire qu'elle est composée par l'enchaînement de motifs isoprène (voir la

figure), dont le squelette est formé de cinq atomes de carbone.

Les terpènes sont des molécules abondantes dans le monde naturel : on les trouve dans l'essence de térébenthine, dans l'écorce des bouleaux ou dans le parfum des roses ; ce sont des hormones de plantes, des constituants de membranes cellulaires d'archéobactéries, des hormones sexuelles d'animaux, des pigments de légumes... Leur structure fortement ramifiée rend leur biodégradation difficile, ce qui explique leur pérennité dans les sédiments.

Quelle était l'origine de ces géohopanoïdes? Les chimistes étaient perplexes, car aucun produit naturel connu n'avait une structure moléculaire ayant pu, même par maturation dans les sédiments, se transformer en une molécule ayant cinq atomes de carbone de plus que le hopane,

au demeurant rare. Les géohopanoïdes étaient des structures «orphelines».

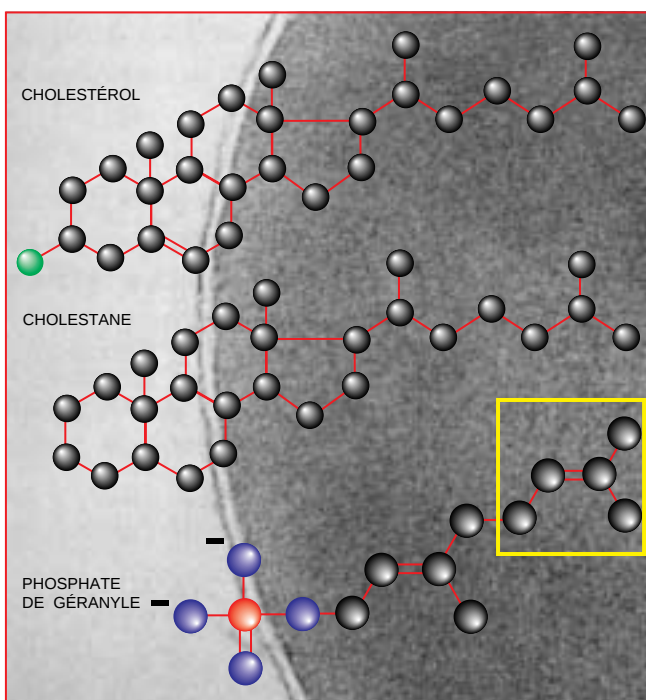
Nouveau détournement de l'histoire : les précurseurs des géohopanoïdes furent bientôt trouvés... dans des bactéries : ces «biohopanoïdes» étudiés par Michel Rohmer constituent une classe nouvelle de lipides bactériens ; tous sont amphiphiles, avec une partie composée seulement d'atomes de carbone et d'atomes d'hydrogène, qui n'a pas d'affinité pour l'eau, et une queue hydrophile. On sait aujourd'hui, grâce à des résultats obtenus en Allemagne et à Strasbourg, que les biohopanoïdes sont localisés dans les membranes bactériennes, où ils jouent un rôle comparable à celui du cholestérol cellulaire : ils rigidifient les membranes.

DU SÉDIMENT À LA MEMBRANE

Si les biohopanoïdes sont amphiphiles, comme les phospholipides qui composent les membranes de nos cellules, ils ont la même rigidité et les mêmes dimensions que le cholestérol, qui sert de renforteur de nos membranes cellulaires. D'autre part, les géohopanoïdes ne sont pas les seuls fossiles moléculaires dont les précurseurs soient des terpènes : l'équipe de P. Albrecht a trouvé un grand nombre de familles moléculaires différentes, dont certaines semblent dériver de lipides membranaires bactériens connus, tandis que d'autres sont encore orphelines, comme l'étaient les géohopanoïdes au moment de leur découverte.

Connaissant la structure de ces fossiles moléculaires, les réactions susceptibles de se produire au cours de la maturation dans les sédiments et les mécanismes généraux utilisés par les êtres vivants pour synthétiser leurs constituants, on en a déduit les structures des précurseurs des fossiles moléculaires isolés.

Les chimistes ont alors déterminé que ces précurseurs, connus ou postulés, étaient tous des dérivés terpéniques, qu'ils étaient tous amphiphiles et que leur partie hydrophobe avait soit les dimensions du cholestérol, soit une longueur double, avec deux têtes polaires, hydrophiles. Ils en ont déduit – et vérifié expérimentalement – que toutes ces molécules avaient un rôle de renforcement comparable à celui du cholestérol dans les membranes modèles. Ainsi le rôle des dérivés terpéniques serait membranaire, dans tous



Trois molécules associées aux membranes (liseré sur l'image du fond) : le cholestérol (en haut), le cholestane (au milieu) et le phosphate de géranyle (en bas). Les atomes de carbone sont en noir et les atomes d'hydrogène ne sont pas représentés ; ils complètent à quatre la valence de chaque atome de carbone. En vert figure un groupe hydroxyle OH, en violet des atomes d'oxygène et en rouge un atome de phosphore. Le cadre jaune indique un motif isoprène.

les êtres vivants ; dans les archéobactéries, ces molécules seraient les constituants de base des membranes, et, dans les organismes supérieurs, elles serviraient seulement de renforçateurs.

Pourrait-on alors reconnaître, parmi les substances isolées, quelles sont les plus anciennes ? Une molécule semblait ancienne quand sa synthèse pouvait s'effectuer à l'aide de peu d'enzymes et quand les réactions de synthèse étaient analogues à des réactions supposées prébiotiques. Les chimistes ont établi un arbre phylogénétique de tous les terpénoïdes connus ou postulés comme constituants membranaires, et ils ont observé que les constituants membranaires les plus anciens devaient avoir été les phosphates d'alcools polyterpéniques.

MEMBRANES PRIMITIVES

Ces derniers étaient connus, puisqu'ils sont les précurseurs biosynthétiques de tous les terpénoïdes, mais on ne les avait jamais préparés en quantités suffisantes pour vérifier s'ils formaient des vésicules. À Strasbourg, N. Plobeck s'est livré à cette synthèse, et il a alors obtenu des phosphates de bis-polyterpényle : ceux qui portaient deux queues hydrophobes courtes, à cinq atomes de carbone, étaient solubles dans l'eau, mais ceux qui portaient deux queues longues, à 15 ou à 20 atomes de carbone, formaient des vésicules. De surcroît, même le phosphate de bis-géranyle, avec deux queues à seulement dix atomes de carbone, formait également des vésicules.

Une vésicule n'est pas une cellule, mais elle en possède plusieurs propriétés caractéristiques. Par exemple, l'intérieur de la membrane elle-même constitue une phase lipidique : l'ajout d'une trace d'un hydrocarbure fluorescent, insoluble dans l'eau, mais soluble dans la membrane, rend fluorescente la mince pellicule qui entoure les vésicules. En outre, au moins sur de petites vésicules, l'environnement diffère des couches interne et externe conduit à des différences d'acidité et à des possibilités de transport qui sont à l'étude.

D'où viennent les phosphates d'alcools terpéniques ? On suppose aujourd'hui qu'ils résultent de réactions entre des alcools et des polyphosphates, présents dans les eaux volcaniques. Une simple catalyse acide due à des argiles, par exemple, devrait permettre d'obtenir les alcools polyterpéniques à partir de constituants plus simples, tels l'isobutène et le formaldéhyde. Ainsi la recherche des membranes de nos ancêtres les plus lointains passe, aujourd'hui, par l'exploration chimique de classes de molécules universelles, retrouvées grâce à l'étude du pétrole. □

Vaccin anti-cocaïne

*Chez la souris,
un vaccin diminue
les effets de la drogue.*

La consommation de cocaïne entraîne parfois des perturbations neurologiques, des anomalies comportementales et un besoin quasi irrépissable de consommer la drogue. Comment lutter contre cette dépendance ? George Koob et ses collègues de l'Institut de recherches Scripps ont considéré la cocaïne comme un agent étranger et «vacciné» des animaux contre cette substance. Ils ont publié dans la revue *Nature* que, de la même façon qu'un vaccin classique protège contre un micro-organisme pathogène, leur vaccin contre la cocaïne rend les souris vaccinées insensibles à la drogue.

Depuis plus de 1 000 ans, les feuilles de coca sont consommées pour leurs effets psychotropes. Selon les Incas, «les anges de Dieu avaient fait cadeau à l'homme de la feuille de coca pour rassasier ceux qui avaient faim, donner une vigueur nouvelle à ceux qui éprouvaient fatigue et défaillance, et faire oublier leurs misères aux malheureux.» Les feuilles de coca sont apparues en Europe au milieu du XIX^e siècle. En 1863, un pharmacien corse, Angelo Mariani, a déposé le brevet du vin Mariani, une boisson aux extraits de coca, vite populaire en Europe. La frontière entre boisson plaisante et médication s'est encore estompée quand, en 1886, le pharmacien américain, John Pemberton, a inventé le Coca Cola.

Les feuilles de coca produisent sensiblement les mêmes effets que la substance pure, mais la cocaïne se déverse plus progressivement dans l'estomac, puis dans le sang, et ses effets sont moins intenses que ceux de la poudre administrée par inhalation ou par injection.

La cocaïne favorise la libération de la dopamine et bloque sa recapture dans les zones «de la récompense», le noyau accumbens et le cortex préfrontal : la concentration en dopamine y augmente, procurant les effets mentionnés.

Pour dresser le système immunitaire contre un antigène étranger, celui-ci doit

être stable ; or la molécule de cocaïne se décompose spontanément par hydrolyse. Ainsi, il fallait trouver un moyen d'en empêcher la dégradation. L'équipe de G. Koob y est parvenue en associant un analogue stable de la cocaïne à une protéine-porteur étrangère : le composé stabilisé était alors susceptible d'activer le système immunitaire.

Afin de tester l'efficacité de l'immunisation, les chercheurs ont d'abord évalué l'effet de la cocaïne sur des rats : leur activité locomotrice augmente, et leur comportement devient stéréotypé. Puis, ils ont administré à quelques animaux le «vaccin» contre la cocaïne. Après une injection et deux rappels, l'équipe américaine a injecté de la cocaïne aux rats vaccinés et observé leurs réactions : la drogue n'avait plus d'effet.

Après la vaccination, le système immunitaire produit des anticorps. Lors d'une injection ultérieure de la drogue, les anticorps se fixent sur la cocaïne, l'empêchant de traverser la barrière hématoencéphalique, de pénétrer dans le cerveau et d'y exercer ses effets. L'immunisation contre la cocaïne abaisse la concentration en cocaïne dans le cerveau.

Si, devant le vide thérapeutique auquel sont confrontés les médecins qui veulent traiter les toxicomanes, ce résultat est encourageant, de nombreuses questions restent ouvertes : lors de l'expérience menée sur les rats vaccinés, la drogue n'était pas injectée par voie intraveineuse, mais par voie sous-cutanée, de sorte que les anticorps avaient le temps de se fixer sur la cocaïne et de l'empêcher d'agir. Lorsqu'un toxicomane s'administre la drogue par voie intraveineuse, l'action est immédiate et les anticorps neutralisants n'auraient sans doute pas le temps d'interdire à la cocaïne de pénétrer dans le cerveau. De surcroît, si les anticorps avaient le temps d'agir, les toxicomanes, n'éprouvant plus les effets qu'ils recherchent, risqueraient de s'administrer des doses croissantes, et de mourir d'overdose.

Enfin, le système immunitaire est très précis, et même s'il reconnaissait la cocaïne, il serait insensible aux analogues que les trafiquants ne manqueraient pas de synthétiser : ces molécules seraient suffisamment proches de la cocaïne pour en reproduire les effets, mais assez différentes pour échapper au système immunitaire. □



Publicité pour le vin Mariani.