

Pour ce faire, il est nécessaire de mesurer avec une grande précision l'état du système. Cela implique de pouvoir intégrer presque en temps réel les données des satellites et des stations locales aux simulations numériques de grande échelle, afin de déterminer la position des barrières de transport.

L'analyse de la marée noire de *Deepwater Horizon* montre qu'une barrière de transport a poussé la nappe de pétrole vers la côte Ouest de la Floride pendant deux semaines en juin 2010. Si cette information avait été disponible à l'époque, la décision des autorités pour contenir ou disperser le pétrole aux endroits critiques aurait pu être prise avec plus d'assurance.

La théorie des structures lagrangiennes cohérentes est aujourd'hui mathématiquement bien établie. Le défi est d'optimiser les temps de calcul pour des applications concrètes. Parmi les pistes possibles, de nouveaux travaux montrent que un pour cent des structures lagrangiennes parmi les plus attractives sont à l'origine des phénomènes

les plus marqués dans un écoulement. Le développement d'une méthode pour repérer ces éléments permettrait de gagner du temps. Des outils généraux d'extraction de structures attractives et répulsives, mais aussi d'autres configurations telles que des tourbillons ou des jets, sont en cours de développement.

Des applications diverses

Le suivi des débris dérivants et, à plus long terme, la surveillance des pollutions ne sont pas les seules applications possibles pour les structures lagrangiennes cohérentes. Cette approche pourrait être utile aux opérations de sauvetage en mer après un naufrage, en permettant de localiser plus rapidement dans quelle direction les victimes vont dériver. David Schmale et Shane Ross, de l'Institut polytechnique de Virginie, aux États-Unis, ont étudié le transport par le vent de spores du champignon *Fusarium*, un phytopathogène res-

pensable de maladies dans les récoltes. Ils ont montré que des barrières de transport attractives et les régions encadrées de deux barrières répulsives piègent les spores et peuvent les transporter sur des centaines de kilomètres. Le mathématicien Shawn Shadden, de l'Institut de technologie de l'Illinois, a étudié la circulation sanguine dans les ventricules du cœur. Il a montré que chez certains patients dont le cœur est volumineux, le sang reste plus longtemps dans le ventricule, ce qui peut entraîner une thrombose avec formation de caillots.

Ces exemples montrent que l'approche des structures lagrangiennes cohérentes peut être utile dans des contextes variés. Cette méthode a aussi révélé une vision plus pertinente du transport de débris dans des écoulements complexes. Les barrières de transport révélées par les structures lagrangiennes cohérentes constituent un squelette autour duquel s'organisent les flux et, à ce titre, sont un aspect fondamental de la compréhension des mouvements au sein d'un fluide. ■

france culture

LA MARCHÉ DES SCIENCES

Découvertes, inventions, aventures savantes au fil de l'Histoire

Aurélie Luneau
14h-15h / chaque jeudi

franceculture.fr

SCIENCE

Photo: M. D. / Getty Images

La structure en réseaux du langage

Ricard Solé, Bernat Corominas-Murtra et Jordi Fortuny

Des techniques d'analyse issues des mathématiques et de la physique révèlent les structures du langage et renseignent ainsi sur son origine et son évolution.

Seule l'espèce humaine emploie un système de communication symbolique capable de transmettre une information complexe. Certains animaux semblent avoir maîtrisé quelques éléments rudimentaires d'un tel système, mais leur mode de communication est beaucoup moins développé que notre langage. De même, aucune machine ne parvient à l'imiter pour l'instant. Comment avons-nous développé ce système de communication ?

Comme l'a signalé le linguiste américain Derek Bickerton, le langage ne laisse pas de fossiles. On doit donc recourir à des données indirectes. On étudie, par exemple, les changements culturels liés à l'art ou à la fabrication d'outils, qui traduisent des améliorations cognitives et, suppose-t-on, des facultés de communication. Ces dernières années, la génétique a aussi apporté quelques éléments. On a ainsi découvert un gène noté *FoxP2*, dont la mutation provoque des troubles de la parole. Sa séquence est identique chez l'homme moderne et chez les Néandertaliens, qui maîtrisaient peut-être certains aspects d'une communication orale.

Ces indices corroborent une autre affirmation de D. Bickerton : « Rien ne survient dans le monde sans laisser une trace, aussi ténue ou indirecte soit-elle. » En linguistique, l'analyse de ces traces a donné lieu à de grands débats, souvent très théoriques, concernant leur nature et leur origine (génétique, culturelle ou les deux).

Certaines recherches en linguistique ont récemment fait appel à des outils associés d'habitude aux mathématiques et à la physique statistique, tels que l'analyse de systèmes complexes. Celle-ci porte sur des systèmes composés d'un grand nombre d'éléments et qui présentent des propriétés dites émergentes : les propriétés du système global sont plus riches que la simple sommation des propriétés de chaque composante.

Le langage humain est un système de ce type. De même qu'une fourmière est plus qu'un simple agrégat d'insectes, le langage ne se limite pas à la juxtaposition de mots.

Dans une série de travaux récents, nous avons abordé l'étude quantitative du langage du point

« Naturellement destiné à l'exploitation de la pension bourgeoise, le rez-de-chaussée se compose d'une première pièce éclairée par les deux croisées de la rue, et où l'on entre par une porte-fenêtre. »

la rue

le père

on entre

il avait

monument

droite

toi

sur

donne

rude

deux

homme

jaune

pouvoir

maison

douze

pa

fa

no

si

es

bi

do

de

ti

ma

bu

lo

m

o

a

l

e

f

r

d

c

« La façade de la pension donne sur un jardinet, en sorte que la maison tombe à angle droit sur la rue Neuve-Sainte-Geneviève, où vous la voyez coupée dans sa profondeur. »

« La maison où s'exploite la pension bourgeoise appartient à madame Vauquer. Elle est située dans le bas de la rue Neuve-Sainte-Geneviève, à l'endroit où le terrain s'abaisse vers la rue de l'Arbalète par une pente si brusque et si rude que les chevaux la montent ou la descendent rarement. Cette circonstance est favorable au silence qui règne dans ces rues serrées entre le dôme du Val-de-Grâce et le dôme du Panthéon, deux monuments qui changent les conditions de l'atmosphère en y jetant des tons jaunes, en y assombrissant tout par les teintes sévères que projettent leurs coupoles. »

« Le père Goriot, vieillard de soixante-neuf ans environ, s'était retiré chez madame Vauquer, en 1813, après avoir quitté les affaires. Il y avait d'abord pris l'appartement occupé par madame Couture, et donnait alors douze cents francs de pension, en homme pour qui cinq louis de plus ou de moins étaient une bagatelle. »

Extrait du Père Goriot,
Honoré de Balzac, 1835



L'ESSENTIEL

- Les diverses langues du monde ont en commun certains modèles d'organisation et des lois statistiques.
- L'étude de la connectivité des réseaux linguistiques permet d'expliquer de nombreux aspects, tels que la polysémie (l'existence de mots ayant plusieurs sens) ou l'organisation en catégories sémantiques.
- La façon dont la syntaxe émerge chez l'enfant et le développement de formes de communication rudimentaires par des systèmes artificiels renseignent sur l'origine du langage.

D'après Ricard Solé et Josepilh Santiago

1. LE LANGAGE HUMAIN EST STRUCTURÉ
à de multiples niveaux. À partir de quelques règles de composition simples, les phonèmes se combinent en syllabes qui, à leur tour, s'associent en mots. Grâce à la syntaxe, les mots forment des phrases. Le nombre d'éléments possibles croît de façon exponentielle.

de vue systémique. Bien que les langues du monde arborent des différences lexicales et grammaticales notables, elles semblent avoir des structures et des lois statistiques communes. Leur analyse renseigne sur l'origine de la complexité du langage. En particulier, nous avons étudié les mécanismes qui régissent l'acquisition de la syntaxe par l'enfant et l'évolution du langage humain vers un système de communication efficace, souple et au pouvoir informatif illimité.

Pour analyser l'architecture du langage, une approche nouvelle et prometteuse s'appuie sur la théorie des réseaux, qui a connu un développement spectaculaire au cours des dix dernières années. L'idée principale consiste à traiter les mots comme des éléments de base (des nœuds du réseau), à déterminer certaines propriétés qui permettent de les relier et à étudier le maillage résultant.

À titre d'exemple, prenons l'ensemble des noms d'animaux et, en un temps limité à quelques secondes, prononçons tous ceux qui nous viennent à l'esprit. Nous obtiendrons

une liste du type: chien, chat, cheval, dauphin, baleine, cachalot... Des termes contigus ont presque toujours un lien. Ils peuvent tous deux désigner des animaux de compagnie, des représentants de la faune marine, des animaux rares dans la vie quotidienne (tels des reptiles ou des insectes)... Pour élaborer une liste, le cerveau recourt à des associations d'idées liées sémantiquement: le choix de l'animal suivant est déterminé par sa similarité avec le précédent selon des propriétés diverses (telles que la taille, l'habitat ou la familiarité).

Des catégories et des « animaux ponts »

Cette expérience a été menée en 2010 par une équipe de l'Université de Navarre, en Espagne, dirigée par Joaquín Goñi et Pablo Villoslada, avec la collaboration de l'un d'entre nous (B. Corominas-Murtra). En la répétant maintes fois avec de nombreux participants, nous avons établi une liste de mots de plus en plus longue et tracé un réseau qui prenait en compte leurs liens sémantiques (voir l'encadré page ci-contre). Ce réseau était organisé en groupes d'éléments ayant davantage de connexions entre eux qu'avec le reste du réseau. Ces groupes sont, par exemple, les insectes, les reptiles, les animaux rencontrés dans les zoos...

Cette structure reflète une organisation fondée sur des catégories. Elle révèle aussi l'existence, d'un point de vue sémantique, d'« animaux ponts », qui permettent de passer d'un groupe à un autre. Ainsi, le crocodile est perçu à la fois comme un reptile et comme un animal de zoo. Le maillage résultant est très interconnecté et facile à parcourir.

De tels réseaux renseignent aussi sur l'origine de certains types d'ambiguïtés sémantiques, apparemment universels. Ainsi, dans toutes les langues, il existe de nombreux termes ayant la même signification et, à l'inverse, des mots qui ont plusieurs sens. Les premiers sont les synonymes et les seconds sont qualifiés de polysémiques.

Le phénomène de la polysémie est surprenant. Le langage tend à favoriser une communication efficace et donc, en principe, à éviter l'ambiguïté. Pourquoi alors la polysémie est-elle omniprésente ? Si les mots polysémiques ne faisaient qu'entraver la communication, la constante évolution des langues (bien plus rapide que l'évolution génétique) les aurait éliminés.

En 2002, les physiciens argentins Mariano Sigman et Guillermo Cecchi ont analysé l'un des réseaux d'association sémantique les plus complets au monde: *WordNet*. Son édification a démarré en 1985 à l'Université de Princeton, aux États-Unis, sous l'impulsion du linguiste George Miller. M. Sigman et G. Cecchi ont utilisé les outils de la théorie des réseaux complexes pour analyser l'organisation globale de *WordNet*.

Confrontés aux mots *lion* et *rayures*, nous pensons souvent au mot *tigre*, qui apparaît comme une façon d'associer ces deux termes *a priori* sans rapport. Ainsi, en choisissant deux mots au hasard, nous pouvons les relier par une chaîne de vocables intermédiaires. Des chemins de mots, de longueur mesurable, apparaissent alors. De même, les liens triangulaires, où deux éléments liés à un troisième sont aussi associés l'un à l'autre, fournissent une idée du degré d'association locale entre concepts: par exemple, les mots *lumière* et *NUIT* sont reliés à *lampe*, et aussi entre eux. Les connexions entre les divers nœuds sont de nature distincte, mais on peut les considérer simultanément pour se représenter le réseau comme un tout organisé en différents niveaux.

Le langage humain, un petit monde

M. Sigman et G. Cecchi ont montré que *WordNet* est un réseau de type petit monde, c'est-à-dire où, d'une part, deux nœuds quelconques peuvent être reliés par un chemin assez court et, d'autre part, certaines zones comprennent un grand nombre de connexions par rapport au nombre de nœuds. Le phénomène du petit monde a été analysé en détail en 1998 par Duncan Watts et Steven Strogatz, de l'Université Cornell, aux États-Unis. Sa découverte remonte à quelques dizaines d'années plus tôt, lorsque des sociologues, tel l'Américain Stanley Milgram, ont étudié la proximité sociale entre individus. Imaginons que chacune des personnes qui constituent une société corresponde à un nœud d'un réseau. Si deux d'entre elles se connaissent, elles sont considérées comme connectées. Le degré de séparation sociale entre deux individus quelconques est défini comme le nombre minimum de « sauts » à effectuer entre les nœuds pour les relier.

À partir d'une série d'expériences fondées sur l'échange de cartes postales,

Milgram a déduit que la distance sociale moyenne entre les habitants des États-Unis s'élève à environ six degrés. Ainsi, même dans un pays comptant plusieurs dizaines de millions d'habitants, nous sommes bien plus proches les uns des autres que nous l'imaginons. Le phénomène du petit monde se manifeste dans de nombreux autres réseaux: gènes, circuits électroniques, etc. Il favorise le transport d'information ou d'énergie dans l'ensemble du réseau.

Bien que le lexique *WordNet* analysé par M. Sigman et G. Cecchi soit composé de 66 000 mots, la distance moyenne entre deux d'entre eux ne dépasse pas sept degrés. On peut donc naviguer aisément dans le réseau. En outre, ce dernier présente de nombreux liens triangulaires (une autre caractéristique des réseaux de type petit monde), ce qui facilite la libre association de concepts.

WordNet est également un réseau dit sans échelle. Ce type de réseau se caractérise par l'association d'un grand nombre d'éléments ayant peu de liens entre eux et de quelques nœuds fortement connectés. En 1999, Albert-László Barabási et Réka Albert, alors à l'Université Notre-Dame, aux États-Unis, ont montré que c'est la présence de tels superconnecteurs qui donne lieu au phénomène de petit monde. Bien que la plupart des nœuds de ces réseaux ne permettent pas de se déplacer avec rapidité, les chemins possibles se multiplient dès qu'un superconnecteur est atteint. Et il est facile, grâce à certains d'entre eux, de parvenir rapidement à destination. En outre, la perte de quelques superconnecteurs provoque une baisse de connectivité importante du réseau.

Dans le réseau sémantique, les superconnecteurs sont les mots polysémiques. M. Sigman et G. Cecchi ont analysé les conséquences de leur élimination: le degré de séparation moyen entre les mots est monté jusqu'à 11 et le nombre de liens triangulaires a été divisé par 300.

Une navigation facilitée

La polysémie confère donc une grande cohésion au réseau sémantique, car elle facilite aussi bien la navigation que l'association locale entre concepts (corrélée au nombre de liens triangulaires). L'ambiguïté favorise les liens sémantiques: c'est peut-être pourquoi toutes les langues du monde présentent de la polysémie. L'exécution de

QUAND LES RÉSEAUX SÉMANTIQUES EXPLIQUENT LA POLYSÉMIE

L'étude quantitative des propriétés émergentes du langage se fonde sur la théorie des réseaux complexes. On relie des mots au sein d'un réseau, dont on analyse les propriétés statistiques.

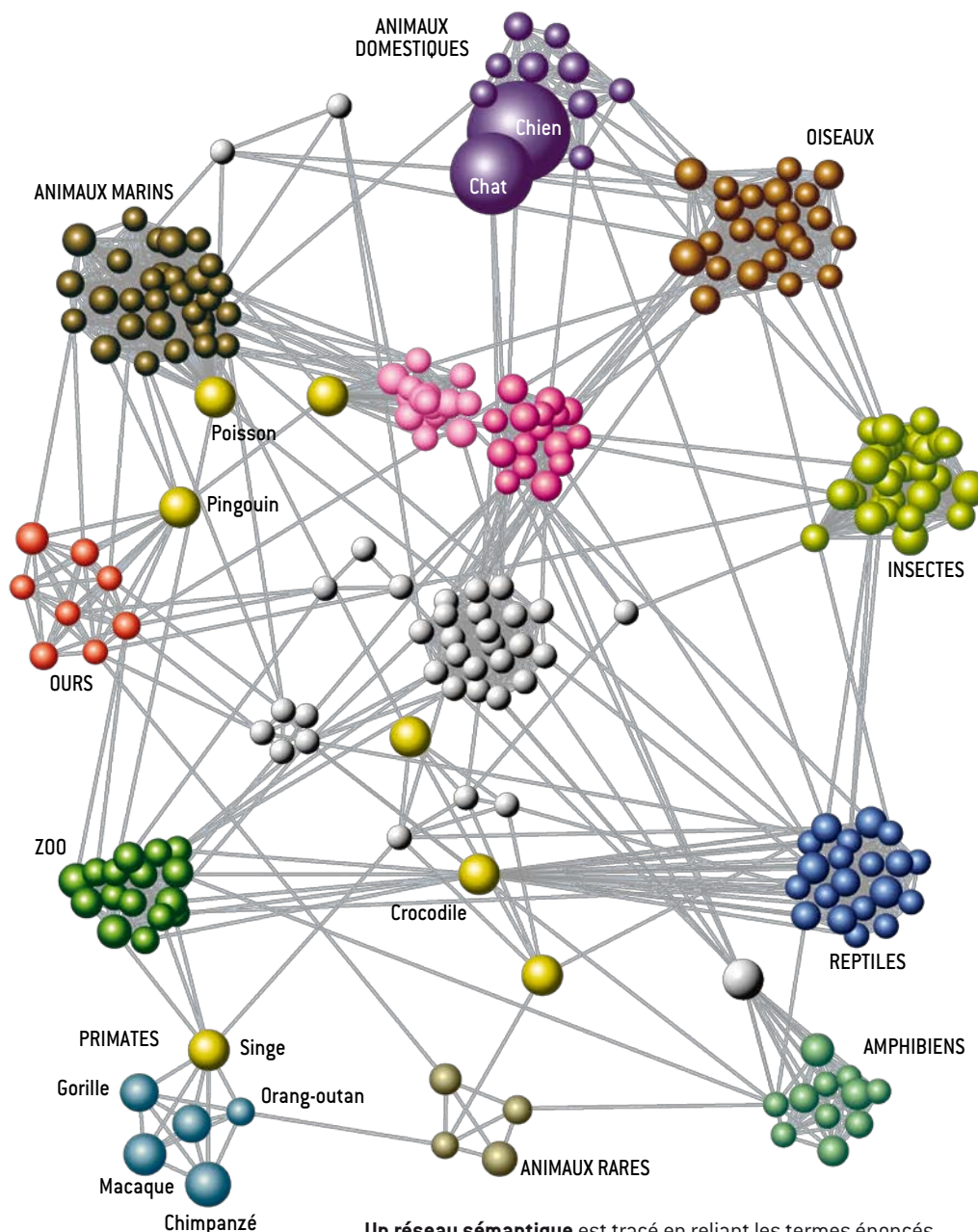
Le schéma ci-dessous illustre une expérience menée en 2010. Les sujets devaient énoncer en un temps court les noms de tous les animaux

qui leur venaient à l'esprit. Deux mots consécutifs avaient presque toujours un lien sémantique. On a alors tracé un réseau de relations sémantiques entre ces termes. On y distingue des groupes de mots (*figurés en couleur*) ayant plus de relations entre eux qu'avec le reste du réseau, et des mots ponts (*nœuds jaunes*), qui mettent en relation di-

vers groupes. Par exemple, le mot *singe* fait penser à la fois aux animaux de zoo et aux primates. Ces caractéristiques facilitent notablement la navigation au sein du réseau, ce qui constitue un avantage pour de multiples tâches cognitives.

Plusieurs études ont montré que les réseaux linguistiques sont sans échelle. Ils se caractérisent par un

petit nombre de nœuds très connectés, ce qui permet d'aller rapidement d'un point à un autre. Dans les réseaux sémantiques, ces « superconnecteurs » seraient les mots polysémiques, ce qui expliquerait leur existence dans toutes les langues (en apparence contre-productive, puisqu'ils introduisent de l'ambiguïté).



Un réseau sémantique est tracé en reliant les termes énoncés successivement, de façon spontanée, par les sujets. La dimension relative des nœuds symbolise la fréquence à laquelle apparaissent les différents mots dans les listes fournies.

tâches cognitives demande de naviguer à travers un vaste univers de mots et de significations, d'où l'utilité du phénomène de petit monde. Des études ultérieures sur d'autres réseaux linguistiques, tels ceux obtenus à partir des dictionnaires de synonymes et d'antonymes, ont confirmé ces propriétés.

Outre ses propriétés sémantiques, l'un des attributs essentiels du langage humain réside dans son pouvoir combinatoire. À partir d'éléments simples et de quelques règles de base permettant de les associer, nous créons d'autres éléments, plus complexes : nous unissons des phonèmes pour construire des syllabes, que nous combinons pour former des mots, avec lesquels nous constituons des phrases. Le nombre de possibilités croît alors de façon exponentielle.

Pour étudier ce potentiel combinatoire, le processus d'acquisition de la langue maternelle offre un champ d'investigation unique. Un bébé n'apprend pas les règles d'une langue comme le ferait un adulte, mais il les construit à partir de données dispersées, fournies par d'autres membres de sa communauté.

L'enfant subit une immersion linguistique nécessairement incomplète, car les mots et les phrases auxquels il est confronté ne peuvent couvrir toutes les éventualités du langage. Il ne reçoit ni une liste d'exemples exhaustive ni des instructions explicites quant à l'usage des règles linguistiques. Le locuteur se forge une connaissance tacite de la langue. Il emploie diverses règles, notamment grammaticales, mais n'est pas conscient de la structure sous-jacente à ses propres productions linguistiques.

Des premiers mots aux premières phrases

Au cours des deux premières années de sa vie, l'enfant entend souvent certains mots et quelques séquences courantes composées de deux vocables, ainsi qu'un faible nombre de phrases (comparé à l'univers des combinaisons possibles). Il passe alors par une phase durant laquelle il n'utilise que des mots isolés, puis par l'étape dite des deux mots, où il associe deux termes de façon assez primitive. Vers l'âge de deux ans, une transition se produit : l'enfant commence à construire des phrases complexes, en faisant peu d'erreurs, voire aucune. Ainsi, l'étape des trois mots n'existe pas. Une fois dépassée celle des deux mots, la

LES AUTEURS



Ricard SOLÉ est chercheur à l'Université Pompeu Fabra de Barcelone, en Espagne, et membre de l'Institut de Santa Fe, aux États-Unis.



Bernat COROMINAS MURTRA est chercheur au Département de science des systèmes complexes de l'Université médicale de Vienne, en Autriche.



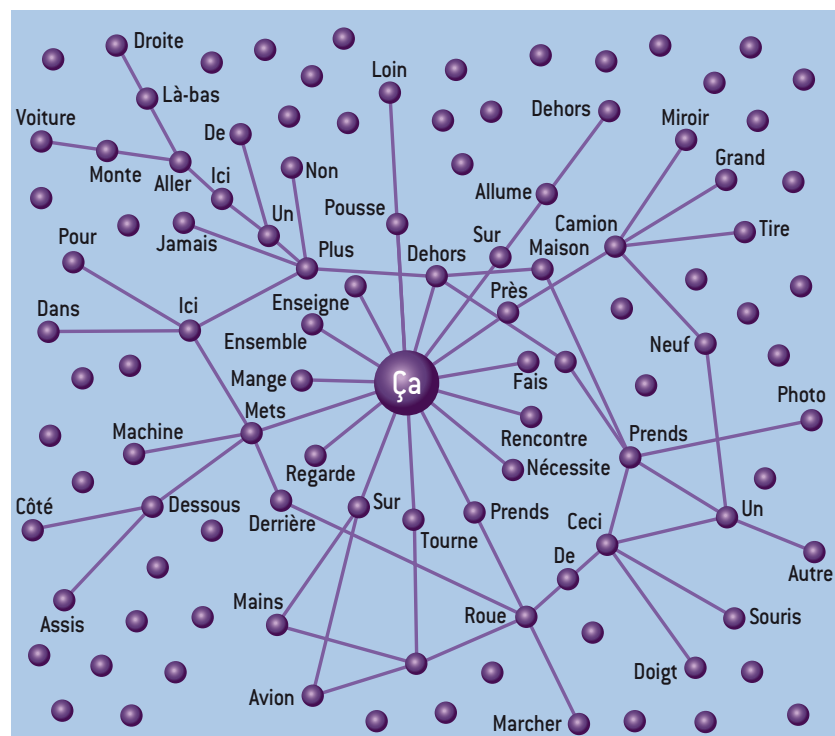
Jordi FORTUNY est chercheur à l'Université de Barcelone.

syntaxe infantile contient déjà une grande partie de la complexité de la syntaxe adulte. Comment l'enfant peut-il développer une connaissance aussi systématique et élaborée à partir d'une stimulation aussi partielle (comprenant très peu d'exemples) et souvent truffée d'ambiguïtés ?

Le processus d'acquisition de la langue a donné lieu à de vifs débats. Dépend-il d'un système cognitif prédisposé à cette acquisition ? En d'autres termes, existe-t-il un « organe du langage » qui facilite le passage d'une communication fragmentée à un discours général, comme le soutient Noam Chomsky ? Pour ce linguiste américain et d'autres, l'enfant dispose d'une grammaire universelle innée. Celle-ci le guiderait au cours de l'acquisition de la langue, en restreignant l'ensemble des combinaisons possibles et en lui procurant les éléments fondamentaux d'analyse. Pour apporter quelques éléments au débat, nous avons exploré la façon dont

L'ÉMERGENCE DE LA SYNTAXE

Vers l'âge de deux ans, les enfants passent directement de la phase dite des deux mots (au cours de laquelle ils n'associent que deux vocables de façon assez primitive) à une étape où ils construisent des phrases complètes. Il n'existe pas d'étape à trois mots. Comment cette transition se produit-elle ?



Au départ, le réseau d'associations syntaxiques a une structure essentiellement arborescente. La plupart des mots sont déconnectés ou ne sont reliés que par paires.

Ricard Solé, Bernat Corominas-Murtra et Jordi Fortuny

l'enfant relie les mots les uns aux autres à mesure de son apprentissage.

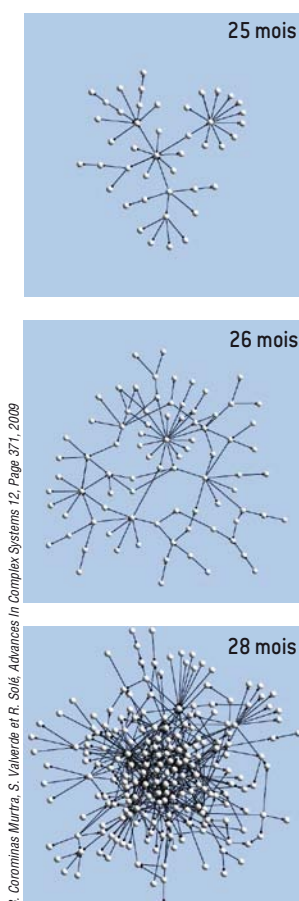
En 2009, nous avons mené une étude fondée sur la base de données CHILDES, qui contient des dialogues entre des parents et leurs enfants, régulièrement enregistrés pendant plusieurs mois. À partir de ces conversations – caractérisées par beaucoup d'imitation et de redondance, surtout au début –, on peut extraire les structures syntaxiques qui apparaissent peu à peu. On forme un réseau en liant les mots connectés dans les « phrases » de l'enfant. Il peut être nécessaire d'extraire d'abord les constituants fondamentaux de ces « phrases » *via* une analyse complexe, incluant l'examen détaillé du contexte de la conversation (les productions enfantines sont souvent peu claires, avec notamment des mots déformés).

Comment ce réseau évolue-t-il à mesure que l'enfant acquiert la syntaxe ? Au début, il est peu étendu et constitué de sous-réseaux

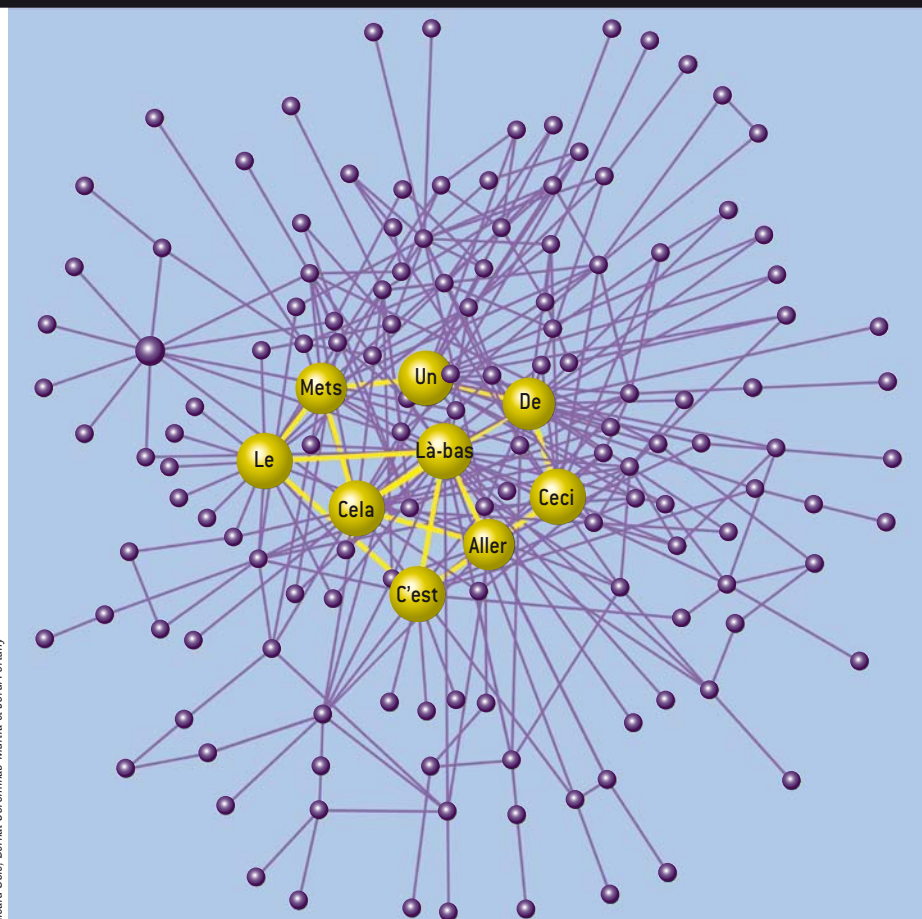
non connectés entre eux : chacun comprend un faible nombre de mots, présente peu de boucles et ressemble à un arbre (*voir l'encadré ci-dessous*). Un tel réseau indique que les relations syntaxiques entre les termes sont encore rares et illustre le fait que les productions enfantines se composent d'à peine plus de deux mots.

Vers l'âge de deux ans survient une transition entre un réseau peu connecté et un réseau caractérisé par un grand nombre de liens. L'organisation des relations syntaxiques se modifie de façon brutale. Comme dans les réseaux sémantiques, des réseaux sans échelle apparaissent : tous les mots sont désormais connectés et, si la plupart d'entre eux ne le sont qu'avec un ou deux autres, certains ont de très nombreux liens. Là encore, les réseaux sont de type petit monde, les mots pouvant être rapidement reliés par une relation syntaxique grâce aux nœuds très connectés, tels *un* ou *ça*.

Au cours de la transition, un changement qualitatif se produit dans la façon de créer des phrases. Avant cette transition, l'enfant ne prononce que des termes isolés ou des combinaisons de deux vocables complémentaires d'un point de vue sémantique, tels qu'un nom et un adjectif s'y rapportant ou un verbe et un nom. Les particules fonctionnelles (principalement les articles et les prépositions) sont quasi absentes. Plus tard, avec l'incorporation de ces dernières, de nouvelles constructions syntaxiques apparaissent. Elles se fondent sur l'emboîtement successif de sous-structures, qui permet de créer des structures de dimension illimitée. Par exemple, la phrase « L'enfant joue à la balle » peut être utilisée comme subordonnée dans la phrase « J'ai vu que l'enfant joue à la balle. » Ce mécanisme, qualifié de récursivité, est considéré par l'école chomskienne comme la propriété essentielle du langage humain.



B. Corominas Murtra, S. Valverde et R. Solé, *Advances in Complex Systems* 12, Page 371, 2009



Ricard Solé, Bernat Corominas-Murtra et Jordi Fontuy

Quand l'enfant atteint l'âge de deux ans, le réseau syntaxique se métamorphose : il passe d'un réseau arborescent à un réseau sans échelle. Les liens se multiplient et plusieurs superconnecteurs apparaissent (*nœuds jaunes sur l'image de droite*), permettant à l'enfant de produire des phrases complètes et bien organisées.

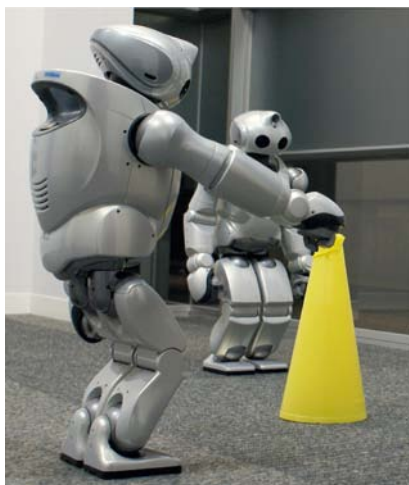
Il n'est efficace pour la communication que grâce à notre capacité d'interpréter les structures produites. Cette interprétation est compositionnelle, c'est-à-dire qu'elle prend en compte les éléments lexicaux d'un énoncé, mais aussi leurs relations dans le contexte de la phrase: ce n'est pas une simple « addition » de significations. La capacité d'interpréter de façon compositionnelle les structures récursives est la clé du potentiel expressif du langage. Pour de nombreux experts, son acquisition au cours de l'évolution est l'une des principales énigmes posées par l'étude des origines du langage.

Le moindre effort

De telles observations restant assez qualitatives, nous avons tenté d'analyser de façon plus quantitative les propriétés des structures syntaxiques. Nous nous sommes intéressés à la fréquence d'usage des mots dans une langue à partir de l'analyse de textes écrits. Elle suit une loi dite de Zipf, en référence au linguiste américain George Zipf qui l'a décrite pour la première fois dans les années 1930. Cette loi stipule que la fréquence d'occurrence $f(r)$ d'un mot est inversement proportionnelle à son rang r (qui désigne sa place dans une liste où les mots sont classés par fréquence d'utilisation décroissante). Autrement dit, si le mot le plus employé (de rang $r = 1$) apparaît un certain nombre de fois dans un texte, le deuxième plus utilisé ($r = 2$) apparaît deux fois moins souvent, le troisième trois fois moins souvent, etc. Il en découle que peu de mots sont très fréquents.

En 2011, nous avons examiné une hypothèse formulée par Zipf au milieu du siècle dernier. Selon cette hypothèse, le langage humain suivrait le principe du moindre effort, qui concernerait aussi bien l'émetteur que le récepteur d'un message. Moins un locuteur emploie de mots, plus il économise son effort, mais plus il augmente celui que son interlocuteur doit faire pour déchiffrer le message. D'après Zipf, le nécessaire compromis qui en résulte pourrait régir l'évolution des langues. Cela expliquerait en particulier qu'aucun lexique n'est composé d'un seul mot (moindre effort pour l'émetteur, qui ne doit apprendre qu'un mot), ni d'innombrables vocables ayant chacun un sens unique (moindre effort pour le récepteur, puisqu'aucun mot n'est ambigu).

Pour vérifier la pertinence de cette hypothèse, nous avons utilisé des résultats



Luc Steels, Laboratoire Sony, Paris

2. LES AUTOMATES de Luc Steels, de la Société Sony CSL, à Paris, apprennent à partir de leur environnement et inventent des mots pour se référer aux objets. En faisant de multiples essais et en communiquant, plusieurs robots peuvent développer un lexique commun et une grammaire simple.

■ BIBLIOGRAPHIE

B. Corominas-Murtra et al., **Emergence of Zipf's law in the evolution of communication**, *Physical Review E*, vol. 83, art. 036115, 2011.

R. V. Solé et al., **Diversity, competition, extinction : The ecophysics of language change**, *Journal of the Royal Society Interface*, vol. 7, pp. 1647-1664, 2010.

A.-L. Barabási et É. Bonabeau, **Les réseaux invariants d'échelle**, *Dossier Pour la Science*, n°66, 2010.

B. Corominas-Murtra et al., **The ontogeny of scale-free syntax networks : phase transitions in early language acquisition**, *Advances in Complex Systems*, vol. 12, pp. 371-392, 2009.

R. V. Solé et al., **Language networks: their structure, function and evolution**, *Complexity*, vol. 15, pp. 20-26, 2009.

D. Bassano, **Premiers pas dans le langage**, *Cerveau&Psycho*, n°3, 2003.

antérieurs qui quantifiaient les efforts du locuteur et du récepteur grâce à la théorie de l'information. En outre, nous avons supposé que le langage se complexifie en incorporant des éléments nouveaux et qu'il est doté d'une capacité informative illimitée. Nous avons alors montré que le système de communication le moins coûteux en termes d'efforts se conforme à la loi de Zipf. On retrouve donc bien une caractéristique des langues naturelles à partir de l'hypothèse du moindre effort.

Enfin, nous avons établi qu'il est possible de construire un système formel d'emboîtement de contenus où le passage de l'étape à deux éléments à celle à trois éléments implique une transition vers un potentiel de création illimité. Ce système fait intervenir la loi de Zipf, qui associe la précision (de multiples mots peu courants et spécifiques) et l'ambiguïté (peu de termes très fréquents, mais dont l'usage est générique). Si ce cadre théorique est encore abstrait, il suggère que la transition observée dans le processus d'acquisition de la langue obéit en partie à un principe organisationnel fondamental acquis par le cerveau au cours de l'évolution.

Pour étudier les origines de la complexité du langage, on doit aussi prendre en compte le rôle joué par l'évolution culturelle. Les langues subissent des changements notables en un temps bien plus court que celui de l'évolution génétique. Un texte écrit dans notre langue maternelle il y a à peine quelques centaines d'années contient de nombreux archaïsmes et peut être ardu à comprendre. L'entrelacement des évolutions linguistiques et culturelles est difficile à tester expérimentalement, mais la simulation informatique a permis de l'explorer.

Des langages artificiels

Dans le sillage du linguiste britannique Jim Hurford, divers chercheurs, tels Simon Kirby, de l'Université d'Édimbourg, ou Morten Christiansen, de l'Université Cornell, ont mené des expériences virtuelles au cours desquelles une population d'« agents », capables d'échanger et de décoder des mots, ont développé un certain nombre de propriétés caractéristiques du langage. Dans plusieurs d'entre elles, certaines règles d'ordonnement des mots sont apparues de façon spontanée. Selon plusieurs scientifiques, tels S. Kirby et Terrence Deacon, de l'Université de Californie, le langage

aurait évolué pour survivre dans un environnement défini par la nature du cerveau humain. Dans leurs modèles, le langage est considéré comme un virus qui infecte l'esprit de chaque individu et qui agit comme une entité symbiotique : les êtres humains en auraient besoin pour prospérer, voire pour survivre, tandis que le langage dépendrait d'eux pour se reproduire. Bien que cette hypothèse paraisse séduisante, elle renseigne peu sur les processus régissant l'évolution du langage.

Découvrons-nous un jour l'origine de ce dernier ? Le langage ne laissant pas de traces fossiles, nous ne retrouverons jamais la voie précise qu'il a empruntée. Néanmoins, nous pourrions peut-être reconstituer son développement, au moins à grands traits, en nous fondant sur l'existence de règles universelles qui semblent guider l'évolution vers la complexité. Bien que les langues du monde arborent des différences lexicales et grammaticales notables, la présence de modèles organisationnels généraux (tels que la loi de Zipf ou les réseaux de types petit monde et sans échelle)

suggère que le langage humain constitue la seule solution pour une communication à la fois sûre et flexible.

Une autre approche novatrice consiste à faire évoluer des formes de langage au sein de systèmes robotiques. Au lieu de construire des modèles de langage humain, on emploie des robots équipés de capteurs et d'un cerveau simplifié (une sorte de réseau neuronal). C'est l'approche adoptée par Luc Steels, du laboratoire de l'entreprise Sony (Sony CSL, pour *Computer Science Laboratory*), à Paris, et de l'Institut de biologie évolutive de Barcelone. Il a employé deux systèmes robotiques simples, dotés de vision et capables d'observer un ensemble d'objets et de les nommer. Ces systèmes pouvaient communiquer entre eux et échanger des informations sur les objets environnants.

L. Steels a ainsi tenté de recréer les premiers pas du langage. Ses robots devaient segmenter l'information relative à chaque objet, telles sa couleur ou sa localisation, et choisir un mot pour la nommer. Les termes employés étaient inventés, tels *bozopite* ou *malewina*. Durant la phase de communication,

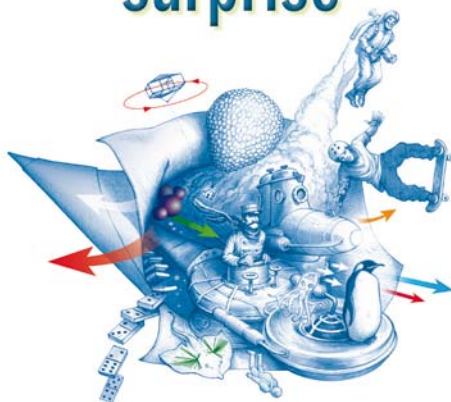
chaque robot tentait de déduire à quel objet ou attribut son compagnon faisait référence. De temps à autre, un accord surgissait et les deux machines commençaient à employer le même terme. À la fin, les robots utilisaient un grand nombre de mots pour indiquer l'orientation spatiale (en haut, en bas, etc.) et d'autres attributs tels que la couleur et la taille. D'autres études avec des robots plus évolués ont aussi montré l'émergence d'une grammaire rudimentaire.

De tels travaux renseignent sur l'importance de l'interaction physique du robot avec le milieu et de la perception qu'il a de lui-même et des autres dans l'espace. Ils permettent aussi d'étudier si la structure du langage humain constitue l'unique solution au problème de la communication complexe, ou seulement une parmi d'autres. Quels types de langages apparaîtront le jour où l'on utilisera des robots plus complexes ? Verra-t-on surgir un langage semblable au nôtre – ce qui confirmerait l'existence de lois universelles –, ou découvrira-t-on une forme de communication totalement nouvelle ? ■

La physique surprise

Jean-Michel Courty • Édouard Kierlik

La physique surprise



Jean-Michel Courty et Édouard Kierlik

Que nous réserve cette Physique surprise ?

Trente-neuf pièces à déguster sans modération, mitonnées par les auteurs. Le style et les croquis, l'actualité et la variété des thèmes vous séduiront, vous amuseront, vous surprendront. Oui, on peut franchir le mur du son en chute libre.

Non, les veines ne sont pas bleues. Oui, on peut séparer le chaud et le froid dans un courant d'air. Alors... Bon appétit !

Éditions Belin/Pour la Science 2013

184 pages – 24 euros – ISBN 978-2-8425-5121-9

Disponible en librairie et sur www.pourlascience.fr

Sonder les **Volcans** avec des rayons cosmiques

N. Lesparre, D. Gibert et J. Marteau

