Comment les robots construisent leur monde : Expériences sur la convergence des catégories sensorielles

Frédéric Kaplan et Luc Steels

Sony CSL Paris
6 rue Amyot 75005 Paris
kaplan@csl.sony.fr, steels@arti.vub.ac.be - http://www.csl.sony.fr

Prenez un nouveau-né. Laissez-le découvrir le monde à sa guise mais veillez à ce qu'il n'ait aucun contact linguistique pendant au moins plusieurs années. A quoi ressemblera sa vision du monde ? Aura-t-il inventé la notion de rouge, de vert, de grand ou de petit ? Ou en aura-t-il inventé d'autres ?

Si nos catégories sont basées principalement sur notre physiologie, en d'autres termes si elles sont avant tout innées, cet enfant aura sans doute une vision du monde assez proche de la nôtre. Ses catégories correspondraient à celles que nous utilisons d'ordinaires. Il ne connaîtrait simplement pas les mots pour les nommer.

Si par contre il y a une part importante d'arbitraire dans la manière dont nous organisons notre perception de l'environnement, cet enfant aurait très bien pu se forger un système de catégorisation efficace mais fort différent du nôtre.

Lorsqu'un Français dit "rouge", nous le comprenons. Pourtant nous ne serons jamais capables de voir pas ses yeux. Nous ne serons jamais capable de savoir ce que ce mot veut vraiment dire pour lui. Un mot peut être apparemment compris par tous sans que, pour autant, nous ne puissions le définir exactement. Nous nous bornons à constater que derrière le mot "rouge" se cache une notion apparemment commune, suffisamment partagée pour qu'en pratique nous nous comprenions.

Comment une catégorie sensorielle comme rouge, qui par essence est indéfinissable, peut-elle être partagée ? Si elle est innée, cela ne pose, bien sûr pas de problème. Si par contre elle est d'une manière ou d'une autre construite, il nous faut expliquer par quel processus des catégories potentiellement différentes entre les personnes deviennent des notions partagées.

Ce sont ces mécanismes que nous avons explorés au travers d'une expérience dans laquelle plusieurs robots ont dû s'accorder sur le sens des mots qu'ils utilisaient. Les "Têtes Parlantes", ainsi nommées en référence aux automates construits par l'Abbé Mical en 1783, sont des couples de robots programmés pour essayer de se comprendre. Les deux robots, dotés d'une caméra mobile, sont placés en face d'un grand tableau blanc magnétique. Sur le tableau sont posées des formes géométriques de formes et de couleurs variées.

Chaque robot est capable d'analyser les images venant de sa propre caméra. Une image est divisée en différents segments, correspondant aux objets présents dans l'image, selon un algorithme classique basé sur la couleur. Pour chacun de ces segments, les valeurs correspondant à une dizaine de canaux sensoriels sont calculées : Les positions X et Y du centre de gravité du segment dans l'image, la surface S du segment, la hauteur H et la largeur L du quadrilatère dans lequel le segment est inscrit, sa rectangularité (Rect) et six valeurs caractérisant la couleur moyenne du segment correspondant aux canaux opposés Rouge et Vert (R/V), Jaune et Bleu (J/B) et Blanc-noir (B/N). Chaque objet peut donc être analysé comme un point dans l'espace sensoriel défini par ces canaux.

Les robots construisent des catégories sensorielles en tentant de discriminer les objets les uns par rapport aux autres. Ces catégories correspondent à des domaines dans l'espace sensoriel des agents. Le robot choisit un objet dans l'image et le compare avec les autres. Une catégorie est dite discriminante si elle permet de caractériser cet objet sans ambiguïté par rapport aux autres. Un triangle rouge en haut à gauche pourra ainsi être différencié d'un rond vert en bas au centre et d'un carré bleu à gauche par des catégories correspondant à des valeurs élevées de rouge (R>v1) ou à une forme triangulaire (v2<Rect<v3) ou à une position élevée (Y>v4).

Si aucune catégorie discriminante n'existe, le robot en construit une en choisissant arbitrairement certains critères. Au fur et à mesure que le robot pratique ce jeu de discrimination, il se construit progressivement un répertoire de catégories. Au bout d'un certain temps, cet ensemble est suffisamment grand pour couvrir l'essentiel des situations qui peuvent se présenter sur le tableau blanc.

Si on laisse deux robots développer leur système de catégories, indépendamment l'un de l'autre, ils construisent la plupart du temps deux systèmes différents. Le premier robot développera par exemple un système essentiellement basé sur les positions. Il sera capable de catégoriser de manière fine les positions respectives des différents objets et ainsi de les discriminer sans ambiguïté. L'autre, par contre, aura pu de la même manière développer un répertoire spécialisé dans la discrimination des couleurs. Aucun système n'est a priori meilleur. Avec suffisamment de précision, chacune des dimensions de l'espace sensoriel peut être utilisée pour catégoriser les différents objets du tableau.

Que se passe-t-il maintenant si les deux doivent communiquer sur ce qu'ils voient ? Imaginons le jeu de langage suivant. Un robot joue le rôle de locuteur et l'autre, celui d'interlocuteur. Les robots alternent dans ces deux rôles. Le locuteur doit désigner à l'interlocuteur un objet en utilisant une forme verbale. L'interlocuteur interprète cette forme verbale et pointe vers l'objet désigné par le locuteur. Le jeu est un succès si le locuteur considère que l'interlocuteur a deviné juste. C'est un échec si le locuteur considère que l'interlocuteur pointe vers un autre objet du contexte ou s'il n'a pas réussi à interpréter le mot du locuteur. Dans le cas d'un échec, le locuteur indique à l'interlocuteur, de façon non verbale, le sujet qu'il voulait désigner.

Nous jouons à ce genre de jeu souvent lorsque nous parlons. Lorsqu' à déjeuner je demande à un ami de me passer le «vin», je m'attends à ce qu'il me tende la bouteille située de l'autre côté de la table. S'il me passe le pain ou s'il me regarde avec un visage interrogatif (peut-être ne comprend-il pas le Français ?), je lui indiquerai sans doute, par un geste de la main, l'objet désiré.

Pour jouer à ce jeu, les robots ont une mémoire associative qui leur permet d'associer à une catégorie un mot donné. Les robots ne peuvent que communiquer en utilisant ces mots et en pointant vers le tableau (ils utilisent pour cela la direction de leur caméra). Ils ne peuvent en aucun cas "inspecter" directement les structures internes de leurs partenaires.

Les robots devinent le sens des mots utilisés par leurs congénères en faisant des hypothèses. Par exemple si le premier robot a dit "Wapaku" en pointant vers le triangle rouge, le second pourra faire l'hypothèse que "Wapaku" correspond à sa propre catégorie désignant les segments dont la composante R est supérieure à v1. Il testera cette hypothèse dans les interactions futures et découvrira, par exemple, que "Wapaku" ne s'applique en fait que pour des segments triangulaires c'est-à-dire correspondant à un certain domaine de valeurs de la composante Rect. Ou bien, peut-être s'apercevra-t-il que "Wapaku" s'applique bien à une catégorie spécifiant la composante R mais de manière plus précise R>v2>v1. Au fur et à mesure des interactions, les robots construisent de nouvelles catégories et de nouvelles hypothèses qui viennent s'ajouter aux anciennes sans les remplacer.

Chaque hypothèse est évaluée par un score. Ce score dépend des succès et des échecs de l'hypothèse lors des communications précédentes. Les hypothèses, ayant des scores élevés, sont choisies préférentiellement pour caractériser une scène et pour exprimer le mot correspondant. Il s'agit pour les robots de se faire comprendre le plus efficacement en utilisant les notions correspondant aux mots qui, de leurs points de vue, semblent les mieux compris. C'est ce mécanisme qui progressivement conduit les robots à utiliser des catégories proches les unes des autres.

Si l'on se contente de faire cette expérience avec deux robots en face d'un unique tableau blanc, la convergence ne sera que partielle. En effet si le seul objet rouge est un triangle et si le seul triangle est rouge, il n'y aura alors aucun moyen de différencier les notions de rouge et de triangularité. Même si les deux agents se comprennent parfaitement en utilisant le mot "Bopako" pour désigner le triangle rouge, il est possible que pour l'un, ce mot soit associé à la couleur, et pour l'autre à la forme.

Pour l'expérience des Têtes Parlantes, nous avons installé des couples de robots dans des laboratoires, des musées et à l'occasion de conférences scientifiques. Ces installations étaient connectées les unes aux autres par Internet. Des agents logiciels transportant les connaissances des différents robots pouvaient se "téléporter" entre chaque plate-forme. Avec ce système, nous pouvions simuler une large population de robots interagissant les uns avec les autres. En multipliant les plates-formes, nous nous assurions aussi que les robots seraient confrontés à des environnements diversifiés et changeants. Pendant plus d'un an, à Paris, Bruxelles, Anvers, Tokyo, Amsterdam,

Londres, Cambridge et Lausanne, ces robots ont tenté de s'accorder sur les mots qu'ils utilisaient. Ils créaient les premiers mots d'une nouvelle langue.

Au bout de quelques mois, un lexique commun était apparu. Avec une vingtaine de mots les robots réussissaient à désigner, avec un minimum d'ambiguïté, chaque élément des scènes qui leur étaient présentées. En analysant les systèmes de catégories construits, on s'apercevait qu'à chacun de ces mots correspondaient des notions de couleurs, de formes ou de positions, relativement uniformes entre les robots. Alors que sans communication, les robots auraient construit des systèmes tous différents, la nécessité de communiquer avait provoquer indirectement une convergence sémantique des catégories sensorielles.

Mais les systèmes de catégories ne restèrent pas figés pour autant. Au fur et à mesure que de nouveaux robots rejoignaient la population, une forme de ***sélection culturelle*** a opéré. En effet, certaines catégories, parce qu'elles étaient plus simples à construire ou parce qu'elles se révélaient plus efficaces pour désigner sans ambiguïté certains groupes d'objets, furent implicitement favorisées dans la transmission culturelle vers les nouvelles générations de robots. Progressivement les systèmes de catégories se sont régularisés et simplifiés pour devenir plus faciles à apprendre, plus faciles à transmettre, plus efficaces pour décrire sans ambiguïté les environnements auxquels les robots étaient confrontés.

Cependant, même au terme de l'expérience, les robots n'ont pas convergé vers des systèmes uniformément partagés. Une polysémie résiduelle est inévitable. Le mot "Bozopite", par exemple, est un des mots les plus utilisés dans le lexique final. Les robots s'en servent pour désigner certains objets et semblent presque toujours compris. Pourtant pour un premier ensemble de robots, ce mot s'applique aux objets de grande surface (S>v1) et pour d'autres, il est utilisé pour les objets de grande largeur (L>v2). Il est difficile de rencontrer des situations pour lesquelles cette confusion pourrait apparaître car les objets "grands" sont souvent "larges". Il faudrait pour cela montrer aux robots un tableau avec des objets hauts et étroits. Notons que l'on retrouve en partie ce genre d'ambiguïtés dans les langues naturelles. En Anglais le sens premier du mot "large" est [de grande surface] alors qu'en Français "large" désigne plutôt [de grande largeur].

Nous pourrions remarquer qu'un autre type de polysémie résiduelle minimale pourrait également se présenter pour des mots qui caractérisent des sens si différents qu'ils ne sont jamais utilisés dans le même contexte. Les exemples dans la langue naturelle sont nombreux. Nous ne confondons jamais le mot "table" dans "table à repasser" et "table des matières" car les contextes d'apparition de ces deux sens n'ont aucun espace d'intersection. De même il serait possible d'imaginer ce type de polysémie chez nos robots. Nous ne l'avons cependant pas observé dans l'expérience.

Des expériences comme les Têtes Parlantes permettent d'explorer les dynamiques complexes qui assurent la construction du sens des mots et des systèmes de catégories qui leur sont associés. Résumons ici les principaux résultats observés pour l'évolution des systèmes de catégories sensorielles dans une population de robots.

1. Des robots qui ne communiquent pas construisent des systèmes de catégories efficaces mais non cohérents les uns avec les autres. Le fait de communiquer force la convergence de catégories associées aux mots échangés.

2. Pour un environnement et un mode d'observation donnés, les robots peuvent construire un lexique et un système de catégories où une certaine polysémie subsiste. Les mots sont ainsi associés à des catégories proches mais non identiques. Le degré de similarité dépend de l'environnement et du mode d'observation.

3. Cette polysémie résiduelle peut être réduite si l'environnement est ouvert et de nouveaux objets deviennent sujets d'interaction. Ces nouveaux objets révèlent certaines ambiguïtés sémantiques et amènent les agents à converger vers des catégories plus similaires.

4. Si de nouveaux robots se joignent régulièrement à la population, une forme de sélection culturelle des systèmes de catégories se met en place favorisant les catégories les plus faciles à apprendre, les plus générales et les plus efficaces.

5. Une polysémie résiduelle minimale subsiste cependant pour des catégories si similaires qu'aucune observation dans l'environnement ne permet d'en révéler la différence.

- - -

Kaplan, F. (à paraître) La naissance d'une langue chez les robots, Hermès, Paris

Kaplan, F. (2000) L'émergence d'un lexique dans une population d'agents autonomes, Thèse de l'université Paris VI.

Kaplan, F. (1999) Dynamiques de l'auto-organisation lexicale: simulations multi-agents et Têtes parlantes. In Cognito , 15:3-23.

Steels, L. (1999) The Talking Heads Experiment. Volume 1. Words and Meanings. Special Pre-edition.

Steels, L. (1997) Perceptually Grounded Meaning Creation. In Tokoro, M., editor, Proceedings of the International Conference on Multi-Agent Systems, Cambridge, MA, 1996. The MIT Press.

Steels, L. et Kaplan, F. (1999) Situated grounded word semantics. In Dean, T., editor, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI'99, pages 862-867, San Francisco, CA., 1999. Morgan Kaufmann Publishers.