

[Version préliminaire de la contribution parue en 2011 dans C. Garbay et D. Kayser (eds.) : *Informatique et sciences cognitives : influences ou confluences ?*, Paris : éd. de la MSH, pp. 287-294.]

Pertinence de l'informatique pour les sciences cognitives vue par une linguiste : informatique et langage

Catherine FUCHS (LATTICE : CNRS/ENS)

Le langage, caractéristique de l'espèce humaine, est au cœur de la cognition. Différentes disciplines des sciences cognitives (notamment la linguistique, la psychologie, la philosophie, la neurobiologie, et l'informatique) contribuent, chacune dans son ordre et selon ses problématiques propres, à éclairer cet objet.

Entre la linguistique et l'informatique, les liens sont anciens, notamment dans le domaine du traitement de la parole et du signal (que je n'aborderai pas ici) et dans celui du traitement automatique des langues (qui s'est principalement intéressé au traitement de l'écrit) : ces liens remontent à la fin des années 1940. Contrairement à ce que pourraient croire des non spécialistes, le traitement automatique des langues n'est pas un domaine homogène : "depuis la conception de modèles théoriques jusqu'à la fabrication d'outils opérationnels, s'étend une longue chaîne de travaux, dont l'hétérogénéité dans les objectifs, les méthodes et les démarches est manifeste" (C. Fuchs & B. Habert, 2004, p.1). L'intitulé même de ce domaine prête d'ailleurs à débats ; que cherche-t-on, en effet, à traiter de façon automatique : 'les langues' (dans leur diversité), 'la langue' (laquelle ? une langue particulière ? toute langue ?), ou bien 'le langage' (entendu comme faculté de langage ? ou assimilé à 'la langue', à l'instar de l'anglais *language* qui indistingue les deux ?) ? — sans compter que l'on parle souvent de traitement automatique du langage 'naturel', ainsi mis en rapport avec les langages 'artificiels' créés de toutes pièces, tels les langages formels.

La linguistique computationnelle et la validation de modèles théoriques

C'est précisément en se fondant sur le rapprochement entre langues naturelles et langages formels que s'est développée, depuis le milieu des années 1950, une branche de la linguistique connue sous le nom de 'linguistique computationnelle' (en d'autres termes, 'calculatoire'). L'objectif commun aux linguistes, mathématiciens et logiciens de ce courant était de décrire le fonctionnement des langues à la manière d'une machine (et donc grâce à une machine), en termes de calculs correspondant au traitement d'informations diverses — essentiellement syntaxiques au départ. D'où la recherche des 'structures mathématiques du langage' (Harris) et l'élaboration de différents types de 'grammaires formelles' (grammaire 'universelle' de Chomsky, mais aussi divers autres formalismes syntaxiques plus implémentables) ; puis un certain nombre de tentatives visant à prolonger cette démarche formelle au plan de la sémantique (Montague, ...).

Cette branche de la linguistique, qui s'est beaucoup diversifiée au fil des années, entretient avec l'informatique des liens se situant essentiellement au plan théorique et épistémologique. Elle s'efforce en effet de construire, à l'aide de formalismes logico-algébriques, des représentations métalinguistiques censées correspondre aux connaissances intériorisées par les sujets parlants (humains). Elle travaille donc au niveau de l'architecture structurale des connaissances sur les langues, et s'inscrit *de facto* dans le paradigme cognitiviste classique, dit 'computo-représentationnel-symbolique' : il s'agit de calculer sur des symboles pour construire des représentations. Mais les 'calculs' en question sont assez rarement implémentés de façon effective, en sorte que l'on en reste souvent à une pure linguistique théorique et formelle. Et lorsqu'il y a recours effectif à l'informatique, c'est pour ainsi dire "de l'extérieur", afin de disposer d'un outil de validation de modèles théoriques (car rien ne remplace une implémentation effective pour déceler d'éventuelles incohérences dans les règles de calcul). Mais les linguistes de cette obédience se sont assez peu engagés au sein d'équipes pluridisciplinaires motivées par les enjeux opérationnels et applicatifs du traitement automatique des langues.

Le traitement automatique des langues et l'échec (relatif) des grands projets

Les premiers travaux en matière de traitement automatique des langues s'originaient assez largement du courant cybernétique. Nés à la fin des années 1940 dans le contexte politique de la 'guerre froide', ils ont tout d'abord porté sur la traduction automatique de textes (cf. C. Fuchs, 1993). Dans ce domaine, tout comme (ultérieurement) dans ceux de la compréhension automatique de textes, puis de la génération automatique de textes, l'ampleur des ambitions proclamées était considérable : méconnaissant à l'évidence la complexité de l'objet et de la tâche, les chercheurs se sont, durant plusieurs décennies, lancés dans de grands projets censés offrir une solution globale à tous les problèmes posés par le traitement d'objets textuels en langue 'naturelle'. (On notera au passage que le traitement automatique, contrairement à la linguistique computationnelle, ne vise pas à formaliser, pour elles-mêmes, les règles de la langue, mais à opérer certains types de traitements sur des textes, c.à.d. sur des produits du système de la langue).

Le reflux actuel des grands projets, consécutif aux déceptions et aux échecs des premiers travaux, ne saurait toutefois faire oublier les enseignements et les acquis de cette période pionnière, qui ont grandement bénéficié aux recherches ultérieures. Très vite, en effet, les chercheurs ont pris conscience des problèmes posés par la recherche de modèles globaux de traitement, c'est-à-dire d'une architecture non seulement structurale mais aussi fonctionnelle des connaissances sur la langue. Car faire traduire, comprendre ou produire un texte par la machine, c'est lui donner les moyens de faire intervenir au bon moment et à bon escient les divers types de connaissances nécessaires. Dès lors, faut-il hiérarchiser ces connaissances et les mobiliser séquentiellement ou bien en parallèle, ou encore les faire interagir ? Une telle problématique inscrivait à l'évidence le traitement automatique des langues sur le terrain de la cognition, tant artificielle qu'humaine. A ce titre, elle appelait en droit une collaboration entre l'informatique, la linguistique et la psychologie : c'est précisément ce qui devait se produire, autour des années 1980, période marquée par l'essor de l'intelligence artificielle (cf. G. Sabah, 1988/89).

Les principales questions auxquelles se sont trouvés confrontés les chercheurs, et sur lesquelles la réflexion des informaticiens a, depuis lors, rejoint — au moins partiellement — celle des linguistes et des psycholinguistes (cf. M. Fayol ed., 2002, et J-F. Le Ny, 2005) concernent, d'une part, les formalismes de représentation des connaissances, et d'autre part

les niveaux de connaissances. Sur ces deux points, les limites des approches de la langue en termes de grammaires formelles sont apparues. C'est pourquoi le traitement automatique des langues s'est progressivement tourné vers d'autres options théoriques, dont beaucoup se trouvaient, par ailleurs, partagées par divers courants de la linguistique dite 'cognitive' (cf. C. Fuchs, 2004). Concernant le premier point, la remise en question de l'isomorphie postulée entre langue 'naturelle' et langages formels logico-algébriques a conduit à l'élaboration de formalismes réputés plus adéquats au traitement du langage : logiques non classiques (cf. D. Kayser, 1990), d'un côté, formalismes d'inspiration topologique, de l'autre. Sur le second point, le besoin de représenter le sens des phrases et des textes a conduit à se préoccuper davantage des connaissances sémantiques et pragmatiques. D'où un intérêt croissant pour des phénomènes comme l'ambiguïté, les glissements de sens, la référence, l'implicite, l'ellipse, ou la typicité, ainsi que pour la dimension du contexte et pour l'idée d'une pluralité des niveaux de sens (cf. D. Kayser, 1991).

Les grands projets ambitieux des débuts se fondaient, clairement, sur l'analogie entre l'esprit(-cerveau ?) et la machine. Nombre de travaux conduits par la suite dans le sillage de l'intelligence artificielle semblent avoir filé la métaphore au point de chercher à fabriquer des programmes de traitement automatique des langues dont, non seulement les résultats (les produits de sortie) mais également les processus de traitement auraient vocation à reproduire ceux de l'humain. En d'autres termes, des outils visant non seulement à 'émuler' mais plus fondamentalement à 'simuler' ce que nous pouvons savoir du comportement langagier des êtres humains. Objectif plus qu'ambitieux, au demeurant : comment, par exemple, programmer une machine pour qu'elle soit en mesure de reproduire la variabilité des modes de calcul du sens par l'humain (cf. les notions de grammaires subjectives, de pondérations et focalisations variables, de cheminements interprétatifs diversifiés, etc.) ?

Du traitement automatique à l'ingénierie linguistique : la fabrication d'outils opérationnels

Depuis quelques décennies, les recherches en matière de traitement du langage à l'aide de l'informatique semblent, à première vue, poursuivre des objectifs moins ambitieux. Il s'agit désormais de fabriquer des outils limités mais opérationnels, plutôt que de vouloir construire des artefacts quasi indiscernables de l'humain : on est passé du global au local, du traitement 'automatique' (censé remplacer l'humain) au traitement 'assisté par ordinateur' (réputé aider l'humain en le déchargeant de certaines tâches fastidieuses et longues à exécuter pour lui).

Cette nouvelle phase présente un certain nombre de caractéristiques qui la différencient des précédentes. Les approches se proclament volontiers empiriques : le bricolage et l'éclectisme théorique ne semblent plus tabous. Plutôt que de vouloir explorer les textes en profondeur pour aboutir à une compréhension complète, on revendique une analyse "light", qui en écume quantitativement et statistiquement la surface afin d'en extraire une compréhension limitée à des objectifs ponctuels. Les nouveaux besoins en matière d'accès rapide à l'information à partir de documents électroniques, par exemple, semblent justifier un tel choix. A l'instar de nombre de linguistes actuellement (cf. B. Habert & al., 1997), la recherche informatique travaille sur corpus — sur de très gros corpus, qui se comptent en centaines de milliers, voire en millions de mots. Enfin, elle recourt très largement à des techniques d'apprentissage, en particulier dans le but de mettre en évidence de façon automatique certaines régularités présentes dans des échantillons significatifs des données

textuelles à traiter : c'est ce que l'on appelle 'l'acquisition de connaissances à partir de données'.

Le linguiste voit, à juste titre, dans ces nouveaux types d'approches, la promesse d'outils opérationnels, susceptibles de lui faciliter le travail sur de grandes masses de données en langue 'naturelle'. Sous réserve, toutefois, que sa langue d'étude soit une langue de grande diffusion, susceptible d'être jugée "rentable" (!) par les 'industries' (ou 'ingénieries') de la langue : cf. J-M. Pierrel (ed.), 2000. La recherche informatique se situe ainsi, par certains côtés, en amont de la bureautique : elle confectionne des outils permettant au simple utilisateur qu'est alors le linguiste de se livrer sur ses propres données de langue à une pratique empirique assistée par ordinateur. Il est de fait que l'on dispose déjà, à l'heure actuelle, d'une palette appréciable d'instruments et de ressources électroniques facilitant ainsi l'accès à de vastes données langagières et leur gestion (cf. B. Habert, 2005). Ce qui est visé ici, c'est donc une symbiose entre l'artefact et l'humain, basée sur leur complémentarité : capacité de stockage, vitesse de travail, fiabilité du traitement, du côté de la machine, souplesse, adaptabilité, diversité des parcours, du côté de l'humain.

Informatique et linguistique : influences ou confluence ?

En amont, c'ad. au niveau d'une recherche plus fondamentale, certains informaticiens concepteurs de nouveaux outils d'accès au sens revendiquent explicitement l'ancrage cognitif de leur entreprise (cf. T. Poibeau, 2006). A l'instar de la linguistique cognitive, ils soulignent l'importance du texte, la prédominance de la sémantique (ils parlent de 'compréhension partielle', et postulent la non autonomie de la syntaxe), le rôle du contexte et invoquent l'existence de mécanismes cognitifs généraux (catégorisation, perception).

Toutefois, à certains égards, le fossé semble s'être creusé entre la recherche informatique et la linguistique. Depuis une bonne dizaine d'années en effet, les approches en termes d'apprentissage artificiel ont assez largement pris le pas sur celles qui optent pour le transfert de l'expertise humaine. Il n'est certes jamais plaisant, pour un linguiste, de penser que l'informaticien peut se passer des systèmes de règles patiemment élaborés par les linguistes et que, à ce compte, la collaboration interdisciplinaire risque de devenir caduque. Mais, par-delà cet antagonisme, se pose une série de questions. De quelle nature sont les 'connaissances' que l'on prétend ainsi 'acquérir' plutôt que 'représenter' ? Sont-ce vraiment des 'connaissances', ou bien de simples régularités observées, toujours susceptibles d'être invalidées ? Ne nous cachons pas que nous traversons actuellement une phase aiguë de fantasme technologique (on rivalise sur la taille des corpus, le quantitatif, les statistiques, la puissance des machines, ...) : course folle ? Or, malgré les efforts en ce sens, il ne semble pas que la question de l'évaluation des nouveaux outils ait encore reçu de réponse satisfaisante. Que signifie, en effet, d'arriver à des scores de réussite avoisinant les 90% (pour des cas triviaux, dans des domaines restreints !), sachant que ce sont souvent les quelques pourcents restants qui, qualitativement, font toute la différence — précisément parce qu'ils correspondent aux secteurs de la langue qui échappent aux régularités statistiques superficielles. Loin de recouvrir des "exceptions" inexplicables ou des "irrégularités" préjudiciables, de tels secteurs constituent au contraire de véritables révélateurs : ils témoignent de l'existence de certaines propriétés essentielles, et encore inexplorées, de la langue. Ces propriétés sur lesquelles, justement, se penche le linguiste ...

Mais, si l'approche des faits de langue par les informaticiens paraît bien souvent réductrice aux linguistes, reconnaissons, en retour, que l'approche des linguistes est

facilement jugée parcellaire (et donc peu utilisable) par les informaticiens : à une couverture large mais superficielle des faits semble donc s'opposer une multitude de descriptions et de théories locales "pointues" mais hétérogènes, consacrées chacune à un type particulier de phénomène. A ce compte, le dialogue serait-il impossible ? Sans doute pas, mais à quelles conditions ce dialogue peut-il devenir constructif, au plan de la recherche fondamentale ? L'informaticien tend à privilégier un modèle unifié, au sein duquel il espère pouvoir raffiner au fil du temps et prendre en compte tel ou tel phénomène spécifique analysé par le linguiste. Pour l'informaticien, cela nécessite que les analyses linguistiques des différents phénomènes soient théoriquement compatibles et cumulables. Pour le linguiste, cela suppose qu'un modèle unifié et adéquat de la langue soit d'ores et déjà disponible. Dans un cas comme dans l'autre, ces conditions ne sont peut-être pas (encore) réunies à l'heure actuelle ...

Je remercie Daniel Kayser pour sa relecture attentive et ses remarques pertinentes, qui ont permis d'améliorer la version finale de cette contribution ; il va de soi, toutefois, que les points de vue ici exprimés n'engagent que leur auteur.

Références

- FAYOL, Michel (ed.) (2002) : *Production du langage*, Paris : Hermès.
- FUCHS, Catherine (2004) : "Pour introduire à la linguistique cognitive", in C. Fuchs (ed.) : *La linguistique cognitive*, Paris : Ophrys/MSH, pp. 1-24.
- FUCHS, Catherine (1993) : "Traduction automatique", in C. Fuchs & al. : *Linguistique et traitements automatiques des langues*, Paris : Hachette, pp. 193-222.
- FUCHS, Catherine & Benoît HABERT (eds.) (2004) : "Traitement automatique et ressources numérisées pour le français", *Le Français Moderne*, LXXII : 1, Paris : CILF.
- HABERT, Benoît (2005) : *Instruments et ressources électroniques pour le français*, Paris : Ophrys.
- HABERT, Benoît, NAZARENKO, Adeline & SALEM, André (1997) : *Les linguistiques de corpus*, Paris : Colin/Masson.
- KAYSER, Daniel (1990) : "Adéquation et inadéquation de la logique au traitement sémantique des langues", *Modèles Linguistiques*, XII : 1, Lille, pp. 119-136.
- KAYSER, Daniel (1991) : "Meaning representation versus knowledge representation", in N. Cooper & P. Engel (eds.) : *New inquiries into meaning and truth*, New-York : St Martins Press, pp. 163-186.
- LE NY, Jean-François (2005) : *Comment l'esprit produit du sens*, Paris : O. Jacob.
- PIERREL, Jean-Marie (ed.) (2000) : *Ingénierie des langues*, Paris : Hermès.
- POIBEAU, Thierry (2006) : "Quelques applications d'accès au sens vues au travers de la sémantique cognitive", Communication lors de la journée des 20 ans du LIPN (17 octobre 2006, Université Paris-Nord).

SABAH, Gérard (1988/1989) : *L'intelligence artificielle et le langage*, 2 vol., Paris : Hermès.