

# Linguistique computationnelle : pourquoi, comment ?

Bruno Mery

13 septembre 2006

La *linguistique computationnelle*<sup>1</sup> est une discipline transversale à de multiples domaines, dont en premier lieu l'informatique et la linguistique, mais également la psychologie, la neurologie ou la philosophie. Elle peut rebuter au premier abord, car, étant très spécialisée, elle n'est véritablement abordable pour aucun chercheur généraliste de ces disciplines, et encore moins pour un lecteur extérieur désireux de s'informer.

Ce court article n'a pas une véritable vocation scientifique. Sa seule ambition est d'exposer les tenants et aboutissants de ce champ de recherche à toute personne intéressée, avant qu'elle ne s'attaque au vif du sujet. Il est composé d'un historique court et approximatif, suivi des motivations philosophique, d'une typologie des spécialisations, de quelques exemples d'applications et enfin d'un aperçu des champs liés. D'autres articles plus détaillés sont à la disposition du curieux aux adresses internet données en notes, ou dans les liens présents sur mon propre site<sup>2</sup>, entre autres.

---

<sup>1</sup>[http://en.wikipedia.org/wiki/Computational\\_linguistics](http://en.wikipedia.org/wiki/Computational_linguistics)

<sup>2</sup><http://labri.fr/~mery>

# Philosophie et évolution

## Origines

Des études formelles sur les mécanismes utilisés pour la communications, qui forment l'abstraction que l'on nomme "langue", ont été menés depuis l'antiquité – la plus ancienne grammaire générative connue est celle du Sanskrit<sup>3</sup>. Il fallut attendre le *XIX<sup>eme</sup>* siècle et Ferdinand de Saussure<sup>4</sup> pour que la théorie linguistique s'intéresse de près à la structure de la phrase, mais ce n'est que relativement récemment que la nécessité à fait de ces études formelles un domaine particulièrement actif, avec l'avènement de l'ère de l'information.

Dans les années 1950, les informaticiens, soucieux d'élargir encore un champ de recherche en plein effort, imaginèrent la réalisation de traducteurs automatiques, comme on avait pu réaliser des calculateurs. Encouragés par les succès rencontrés durant la Première Guerre Mondiale par Turing<sup>5</sup>, portant sur le décryptage, ces chercheurs comptaient rapidement intégrer la compétence langagière à leurs systèmes.

Quelques années plus tard, on dut bien se rendre à l'évidence et admettre qu'il est extrêmement facile de sous-estimer l'incommensurable complexité de la langue. En dehors de résultats positifs pour des domaines extrêmement spécialisés, les applications pratiques n'étaient pas à la hauteur des espérances de départ. La traduction automatique, tout comme beaucoup de domaines relevant de l'intelligence artificielle, tomba légèrement en désuétude.

La linguistique computationnelle est née à cette période, du besoin de l'informatique, pour des applications telles que la traduction, de disposer d'une formalisation complète des mécanismes de la langue. Les recherches effectuées depuis ont permis beaucoup de progrès en ce sens, mais on sait que le chemin est encore très long...

---

<sup>3</sup><http://en.wikipedia.org/wiki/Vyakarana>

<sup>4</sup>[http://en.wikipedia.org/wiki/Ferdinand\\_de\\_Saussure](http://en.wikipedia.org/wiki/Ferdinand_de_Saussure)

<sup>5</sup>[http://en.wikipedia.org/wiki/Alan\\_Turing](http://en.wikipedia.org/wiki/Alan_Turing)

## Buts et spécialisations

Le but de la discipline est donc de formaliser *l'ensemble* des mécanismes de la langue afin, en pratique, d'arriver à la *compréhension* intégrale, par un système informatique, de tout énoncé compréhensible par un humain. Des études donnent à penser que ce problème était *IA-complet*<sup>6</sup>, ce qui impliquerait la création d'un système disposant d'une intelligence comparable ou supérieure à l'humain pour le résoudre entièrement. Par contre, divers objectifs à plus court terme sont plus facilement atteignables, et ont donné lieu à plusieurs spécialisations de la discipline. Il s'agit de :

**La phonologie :** vise à transcrire un flot sonore ininterrompu en ses composantes élémentaires, afin de les retranscrire en mots.

**La morphologie :** Étudie la manière dont se forment les mots, de façon à déterminer si un terme est le féminin d'un autre, par exemple. Accords, flexion et composition forment les concepts liés à ce champ, associé également à l'*étymologie*.

**Le lexique :** Se rapporte à toutes les études portant sur les mots. De multiples approches sont possibles à ce niveau, donnant lieu à des interrogations sémantiques ou syntaxiques.

**La syntaxe :** L'analyse de la structure, de la construction de la phrase. Permet de déterminer qu'un mot est sujet ou complément d'un autre, et de connaître l'organisation de la phrase, afin de pouvoir en déduire le sens (ou, plus prosaïquement, les accords à effectuer).

**La sémantique :** L'analyse de la structure logique des phrases, en vue de dégager le sens des mots et des constructions. Champ extrêmement vaste.

**La pragmatique :** Ensemble de cas particuliers, ou problématique visant à associer au sens d'une phrase toutes les indications et le contexte extérieur qui est supposé connu par le locuteur. Par exemple, l'association d'un sobriquet avec la personnalité connue de l'époque correspondante.

---

<sup>6</sup><http://catb.org/jargon/html/A/AI-complete.html>

## **Théorisation industrielle de formalismes variés**

L'une des principale caractéristiques de la recherche en linguistique computationnelle est la création prolifique par les scientifiques de *formalismes* visant à décrire certains mécanismes. Il peut s'agir de systèmes capables de produire ou de reconnaître un langage, en donnant sa syntaxe : les grammaires génératives de Noam Chomsky<sup>7</sup> et toutes leurs dérivées, par exemple ; ou ce peuvent être des mécanismes chargés de représenter, sous forme d'une formule logique, le sens d'une phrase, comme les grammaires catégorielles logiques.

Quoi qu'il en soit, ces formalismes sont extrêmement nombreux de par la nécessité de théoriser un ou plusieurs systèmes capables de représenter toutes les langues. De plus, certains sont plus utiles pour des langages artificiels (langages codés, langages de programmations), certains sont destinés à la morphologie... Le chercheur nouvellement arrivé dans ce domaine doit prendre le temps de choisir le ou les formalismes les plus adaptés à son problème particulier, et peut-être le créer s'il n'existe pas déjà.

---

<sup>7</sup><http://www.chomsky.info/>

# Le Traitement Automatique des Langues

## Un champ plus pratique

Le Traitement Automatique des Langues est l'appellation d'un pan entier d'applications pratiques de principes de linguistique computationnelle. Avec la démocratisation de l'informatique, les besoins dans ce domaines se font de plus en plus grands, et la motivation financière également ; les investissements en Traitement Automatique des Langues sont très élevés, de par leur application immédiate, mais ceux en recherche fondamentale et en théorie de la linguistique computationnelle sont également soutenus, car un progrès théorique dans ce domaine peut donner un avantage certain dans l'industrie de l'information.

## Applications

Les applications déjà réalisés du Traitement Automatique des Langues sont excessivement nombreuses et lucratives. Citons :

- Traduction. Ces outils vont bien au delà de la translittération de Babelfish ou Google, et sont généralement peu connus, spécialisés dans un domaine précis et utilisés par les entreprises intéressées.
- Aide à la traduction. Plus utile, pour de la traduction généraliste ou littéraire, qu'un processus automatique, ces outils donnent différentes options pour chaque mot, avec des informations contextuelles.
- Synthèse vocale. Permet d'exprimer un texte écrit à l'oral avec une voix synthétique. Peut être d'un certain confort pour l'utilisateur moyen, et bien plus utile pour les mal- ou non-voyants, ainsi que pour suppléer à l'aphasie ou au mutisme.
- Reconnaissance vocale. Permet de dicter des textes qui seront reconnus, et donc de se passer de sténo-dactylo.
- Génération de textes suivant des patrons. Peut avoir des applications multilingues, telle la génération des notices de médicaments pour

l'international (qui dispose d'un formatage très précis).

- Génération de textes aléatoires, ou "littérature combinatoire", expérimentation chère, par exemple, à Raymond Queneau, qui peut également servir à donner des exemples de phrases en cycle Primaire pour illustrer quelques règles d'orthographe.
- Correction de fautes de frappes, à la manière de ispell, par exemple.
- Correction orthographique, utilisant une analyse syntaxique pour vérifier les accords et déclinaisons, suivant la langue (particulièrement difficile, donc demandé, en français).
- Interprétation de questions, d'ordres ou de commandes. Un certain type d'interface pouvant être plus intuitif que des commandes formatées shell ou des actions contextuelles restreintes par l'interface graphique, parfois couplée à la reconnaissance vocale.
- Expression de résultats et messages en langage clair.
- Psychanalyse virtuelle.
- ...

## Domaines connexes

Quelques unes de ces applications ont eu des répercussions dans d'autres domaines, et certaines techniques extérieures sont venues enrichir le Traitement Automatique des Langues par des interactions prévisibles ou parfois totalement fortuites. Voici ces domaines plus ou moins connexes :

- Compilation. Utilise des méthodes proches pour analyser un texte efficacement, mais ce texte est du code source (en général algébrique).
- Recherche d'informations (*data mining*).
- *Web Sémantique* : tentative de standardiser un certain formatage pour faciliter la recherche de données.
- Ontologies : modèles représentant les objets et leurs relations, par exemple dans leurs définitions. Également : modélisation, méta-données...
- Théories des langages formels. Porte sur les langages (ou ensembles de mots) en tant qu'objets mathématiques.
- Logique : classique, intuitionnelle... Modèles divers de déduction de preuves mathématiques, ayant une application directe dans le calcul de prédicats, qui forme une part importante de la théorie de la linguistique computationnelle.
- Théorie des types,  $\lambda$ -calcul : principes de représentation des aspects logiques de la sémantique.
- Théorie des graphes ou des arbres : bases pour certaines constructions utilisées pour la syntaxe ou la sémantique.
- Bien entendu : linguistique, mathématique, psychologie, neurosciences...