Traitement automatique du langage naturel

http://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/linguistique003.html

1 Introduction

Toutes les sociétés humaines découvertes de par le monde pratiquent au moins une langue. On en dénombre actuellement environ 5 000 différentes, dont beaucoup sont en voie de disparition faute de locuteurs. Même si l'acquisition du vocabulaire se poursuit tout au long de la vie, tout être humain normalement constitué et inséré depuis sa naissance dans un groupe social est capable, vers l'âge de 5 ans (donc bien avant qu'il ne maîtrise le "raisonnement"), de tenir une conversation courante dans sa langue maternelle. Aucun singe - et aucun ordinateur ! - ne peut en faire autant. Parler est bien encore, à l'heure actuelle, "le propre de l'homme".

Pour désigner les langues humaines, on parle maintenant des "langues naturelles", parce que ce sont en quelque sorte des créations collectives spontanées auxquelles on ne peut pas attribuer de date de naissance précise. Les langues naturelles s'opposent ainsi principalement aux "langues artificielles" ou "formelles" que sont notamment les langages de programmation informatiques ou la logique mathématique. On peut très bien aussi classer parmi les "langues naturelles" les créations linguistiques intentionnelles comme l'Esperanto ou le Volapük, qui ne sont les langues maternelles de personne, ou encore certaines "langues des signes" inventées spécialement pour répondre à un besoin : leur point commun est qu'elles sont toutes sont destinées à être *utilisées par des humains pour communiquer*.

L'objet de la linguistique ou, comme on le dit plutôt désormais, des *sciences du langage*, c'est l'étude scientifique des langues naturelles et, à travers elles, du langage en tant que "faculté de langue" universellement distribuée dans l'espèce humaine. Derrière l'apparente diversité des langues humaines, les linguistes essaient de traquer des fonctionnements communs, des structures partagées, des *universaux*. Les linguistes ne sont pas nécessairement polyglottes ; ils cherchent plus à comprendre les *principes* qui régissent les langues qu'à multiplier les connaissances qu'ils ont de certaines d'entre elles (même si les deux ne sont bien sûr pas incompatibles).

Contrairement à une idée courante, la linguistique n'est pas prescriptive : elle ne dit pas comment "bien parler" ou "bien écrire". Les langues naturelles sont des systèmes vivants qui changent, interagissent, se transforment. Les linguistes se contentent de les observer *telles qu'elles se parlent et s'écrivent*, sans chercher à contrôler ou à limiter leurs évolutions naturelles. Pour cette étude, l'informatique joue un rôle de plus en plus considérable, via le domaine du *traitement automatique du langage naturel* (TALN). L'informatique est une discipline scientifique

récente, qu'il ne faut pourtant pas réduire à la simple utilisation d'ordinateurs et de programmes. Son nom la désigne comme la science du "traitement automatique de l'information". Elle est en fait l'héritière d'une longue tradition mathématique et logique de *modélisation du calcul*. Plus précisément, on peut dire que les fondements de l'informatique sont doubles :

- le codage des données à l'aide d'*éléments discrets* (les fameux 0/1)
- le codage effectif des traitements à l'aide d'*algorithmes*

C'est par ce biais qu'on va aborder le traitement automatique du langage. Introduire une démarche informatique dans un domaine revient en effet toujours à se poser les mêmes questions :

- quelles sont les données pertinentes de ce domaine, comment les coder ?
- quels sont les traitements pertinents de ce domaine, comment les coder ?

Maîtriser une langue requiert la manipulation de nombreuses données, et la mise en œuvre de nombreux traitements. Les linguistes les ont progressivement mis à jour et caractérisé, les informaticiens ont progressivement contribué à les modéliser. Le TALN est né de leur interaction. Ce chapitre est destiné à faire un tour d'horizon (forcément simplificateur) de ce domaine.

2 Histoires croisées de la linguistique et de l'informatique

Les dates qui suivent ne constituent absolument pas une histoire exhaustive ni de la linguistique, ni de l'informatique. Ce sont plutôt des points de repère marquants qui concernent soit l'une de ces disciplines exclusivement (ce que l'on marquera par (l) pour "linguistique" et (i) pour "informatique" respectivement) soit les deux (ce que l'on marquera par (il) pour "informatique linguistique"). Souvent, les auteurs cités sont ceux qui ont introduit une distinction entre deux concepts, objets ou approches, féconde pour leur discipline. Ces distinctions sont en général destinées à favoriser une des deux branches de l'alternative, pour préciser la démarche suivie et les choix qu'elle entraîne. C'est à ce titre que ces auteurs sont évoqués.

Avant le XVIIIème siècle

Les précurseurs de la linguistique sont les (très nombreux) auteurs de grammaires descriptives d'une langue donnée (l), les précurseurs de l'informatique sont les mathématiciens qui décrivent des méthodes générales de calcul (par exemple pour résoudre des équations) et les inventeurs de "machines à calculer" mécaniques comme Pascal et Leibniz (i).

1660

Publication de la "Grammaire générale et raisonnée" (connue sous le titre "Grammaire de Port-Royal") d'Arnaud et Lancelot. Son ambition était de décrire les règles du langage en termes de principes rationnels universels (l).

XVIII et XIXème siècles

C'est le règne de la linguistique comparative et historique. On compare les langues entre elles et on cherche à en déduire des lois d'évolution générales. Du rapprochement entre diverses langues (latines, grecques, perses, germaniques, celtes, slaves, etc.), émerge l'hypothèse que toutes ont un "ancêtre commun" qui sera appelé plus tard "indo-européen" (l). Le XIXème siècle connaît aussi de grands progrès en mathématiques, et voit naître la logique "booléenne" (ou "propositionnelle") par Boole (1815-1864) puis la "logique des prédicats du 1er ordre" par Frege

(1848-1925). Les débuts de l'automatisation du travail (métiers à tisser activés par des cartes perforées) inspirent aussi les premiers projets de calculateurs mécaniques. Les plus novateurs et visionnaires sont dûs à l'ingénieur et mathématicien anglais Babbage (1791-1871), qui a conçu les plans de machines ayant les mêmes capacités de calcul que les ordinateurs actuels. Elles n'ont malheureusement pas pu être construites de son vivant (i).

1916

(l) Publication posthume (ce sont des notes de cours publiées par deux de ses étudiants) du "Cours de linguistique générale" du linguiste suisse Ferdinand de Saussure (1857-1913). Saussure introduit plusieurs distinctions et concepts importants :

- il caractérise le langage comme la construction sociale d'un *système de signes*. Un signe est l'association arbitraire entre un *signifiant* (défini comme l"image acoustique" d'un mot) et un *signifié* (un concept, la représentation mentale d'une chose). Les signes font sens par les *rapports qu'ils entretiennent les uns avec les autres dans le système*.
- il considère que le *langage*, en tant que faculté générale de s'exprimer au moyen de signes, se distingue de la *parole*, qui serait plutôt l'utilisation concrète de signes linguistiques particuliers (et qu'il n'étudiera pas plus avant).
- il distingue les dimensions *diachronique* (évolution au cours du temps) et *synchronique* (rapports entre les signes à une époque donnée) du langage. Les études historiques et comparatistes se sont focalisées sur la première de ces dimensions, lui entend privilégier la seconde (nous aussi par la suite).
- il distingue deux axes d'analyse d'un discours, en tant que suite de signes : l'axe *syntagmatique* est celui de la succession linéaire des unités qui constituent le discours (un syntagme est une suite d'unités adjacentes) ; l'axe *associatif ou paradigmatique* provient des liens que les signes présents dans le discours entretiennent avec d'autres signes non présents dans le discours mais en rapport avec eux dans le système. Suivant l'axe syntagmatique "un petit chat" est un syntagme dont le sens provient de la combinaison des signifiés de "petit" et "chat", mais ces sens eux-mêmes sont associés suivant l'axe paradigmatique avec d'autres ("petit" par opposition à "grand", etc.).

Années 30-40

(l) Le "cercle de Prague" prolonge les analyses de Saussure et promeut une "linguistique structurale". Ses membres les plus connus sont Roman Jakobson (1896-1982) et Nicolas Troubetzkoy. On leur doit notamment l'invention de la *phonologie* : étude des sons élémentaires (les *phonèmes*) qui jouent le rôle d'unités distinctives dans une langue donnée. C'est aussi à Jakobson qu'on doit d'avoir identifié six *fonctions* permises par le langage dans un contexte de communication :

- la fonction *expressive* permet au locuteur d'exprimer ses sentiments ;
- la fonction *conative* permet d'agir sur le destinataire (donner un ordre...) ;
- la fonction *référentielle* permet d'informer sur le monde extérieur : il faut bien reconnaître que les modèles informatiques ont souvent tendance à limiter le langage à cette fonction ;
- la fonction *phatique* permet juste de s'assurer du bon fonctionnement de la "ligne" de communication ("allo"...) ;
- la fonction *poétique* met l'accent sur la forme du message plus que sur son contenu informationnel ;
- la fonction *métalinguistique* permet de parler du langage grâce au langage (comme le fait ce document !) ;

Alan Turing (1912-1954)

Mathématicien anglais. En 1936, il propose le dispositif plus tard appelé "machine de Turing" qui donne une caractérisation mathématique précise à la notion d'algorithme. Cette proposition peut être considérée comme la date de naissance de l'informatique (i). En 1950, il

publie un nouvel article fondamental dans lequel il décrit un test pour juger de la capacité des machines à penser : ce test, appelé depuis "test de Turing" est fondé sur un jeu de dialogue. En quelque sorte, il énonce qu'une machine peut être dite intelligente si elle est indiscernable d'un humain dans une situation de dialogue courant. Il prédit qu'en l'an 2000, des machines réussiront ce test (ça ne s'est pas vraiment réalisé). (il)

1945

Von Neumann (1903-1957), mathématicien et physicien américain d'origine hongroise, définit dans un rapport le plan de construction des ordinateurs, tels qu'ils sont encore conçus de nos jours (des prototypes plus rudimentaires ont été construits avant). (i)

1952

Première conférence sur la traduction automatique, organisée au MIT (Massachussets Institute of Technology) par Yehoshua Bar-Hillel (1915-1975). C'est l'époque de la guerre froide, la compétition russes/américains bat son plein. L'informatique, elle, n'en est qu'à ses balbutiements et les premiers programmes de traduction doivent se contenter de dictionnaires bilingues et de quelques règles de restructuration élémentaires. La légende propage qu'en partant de *"The spirit is willing but the flesh is weak"* (l'esprit est fort mais la chair est faible), après un aller-retour russe-anglais, on obtint alors *"The vodka is strong but the meat is rotten"* (la vodka est forte mais la viande est pourrie)... Le terme "Intelligence artificielle" (IA), lui, est inventé lors d'une autre conférence, en **1956** (il).

André Martinet (1908-1999)

(l) Linguiste français, connu notamment pour avoir caractérisé les langues naturelles par la propriété de la *double articulation*. Toutes les langues humaines sont doublement articulées parce qu'elles combinent des éléments (discrets) à deux niveaux différents :

- la "première articulation" est celle qui permet la combinaison "d'unités douées chacune d'une forme vocale et d'un sens" qu'il appelle des "monèmes" (le terme n'est plus vraiment employé : nous utiliserons plutôt celui de "morphème" mais, en première approximation, disons que ce sont les "mots"). On peut dire que ce niveau est celui de la *syntaxe*.
- la "deuxième articulation" décrit comment chaque "monème" est lui-même décomposable en une succession d'unités phoniques élémentaires dépourvues de sens, les *phonèmes*.

Il semble bien que toutes les langues humaines, y compris les langues des signes utilisées dans les populations ayant une déficience auditive et/ou articulatoire (même si, dans ce cas, les unités employées ne sont pas de nature phonique), soient doublement articulées, alors que ce n'est le cas d'aucun autre système de transmission d'information, notamment de ceux en usage dans les autres espèces animales (nous détaillons cette affirmation dans la section suivante).

1966

Deux programmes célèbres datent de cette époque : "Eliza" de Weizenbaum (simulation d'un dialogue avec un psychothérapeute, voir par ex : Eliza) et "Student" de Bobrow (résolution de problèmes mathématiques simples). Les deux étaient fondés sur la recherche de "mots-clés" dans les données qu'on leur fournissait, mots-clés qui servaient à remplir les "trous" de formulaires définis *a priori*, sans prise en compte du contexte d'énonciation. Par exemple, dans un dialogue avec Eliza, dès que l'utilisateur mentionnait un lien de parenté ("father"/ "mother"/"brother"...), le programme enchaînait en demandant "Tell me about your [father/mother/brother...]". Bar-Hillel critique cette approche en citant l'exemple suivant, où le sens des mots dépend de leur contexte : "The pen is in the box" (le crayon est dans la boîte)/

"The box is in the pen" (la boîte est dans le parc). Cette même année, paraît le raport ALPAC (Automatic Language Processing Advisory Comitte), commandé 2 ans avant par l'Académie des sciences américaines. Le rapport est très critique vis-à-vis des recherches menées dans le domaine du traitement de la langue à cette époque, et conclut qu'elles mènent à une impasse. Suite au rapport, les financements publics se tarissent en Amérique (il).

Noam Chomsky (né en 1928) :

Linguiste et activiste politique américain d'une productivité exceptionnelle, professeur de linguistique au MIT depuis 1961, qu'une enquête de 2005 désigne comme "le plus grand intellectuel vivant". A travers les différentes théories qu'il a contribué à élaborer, il a toujours argumenté en faveur de la *primauté de la syntaxe* sur tous les autres niveaux d'analyse du langage. Il a toujours adopté aussi une position innéïste et rationnaliste : selon lui, les hommes disposent à la naissance d'un "organe du langage" de nature mentale. Dans cette perspective, apprendre une langue particulière revient à instancier une "grammaire universelle" innée. Il a élaboré de nouveaux concepts et promu plusieurs distinctions inédites :

- il distingue la *compétence* linguistique (connaissance des règles de fonctionnement d'une langue) de la *performance* (mise en oeuvre effective de ces règles en compréhension ou en production). On peut rapprocher cette distinction de celle entre *langage* et *parole* introduite par Saussure. C'est seulement la *compétence* qui sera l'objet de son attention, la *performance* relevant plutôt de la psychologie. La linguistique se doit en effet d'étudier le langage indépendamment de son substrat biologique (gènes, zones du cerveau, etc.) : on considère avoir affaire à un locuteur abstrait idéal.
- il considère que le but de toute théorie linguistique est l'explication des *jugements de grammaticalité* dont sont capables tous les locuteurs d'une langue (surtout quand c'est leur langue maternelle). Pour lui, la *structure de surface* (syntaxe) d'un énoncé détermine sa *structure profonde* (les relations sémantiques qu'il exprime).

La chronologie des travaux de Chomsky a marqué l'histoire de la syntaxe des 50 dernières années :

- 1957 : publication de *Syntactic Structures* (structures syntaxiques), ouvrage fondateur. Dans les années 60, l'approche se raffine dans la théorie des "grammaires génératives et transformationnelles" qui deviendra ensuite dans les années 70 la "théorie standard" de la syntaxe.
- années 80 : l'approche "Principle and Parameters" (principes et paramètres) précise la nature de la "grammaire universelle" innée postulée par Chomsky : elle est constituée d'une liste de propriétés et de paramètres ne pouvant prendre qu'un nombre fini de valeurs. Ainsi, apprendre une langue particulière revient à acquérir son vocabulaire spécifique et à identifir la valeur des paramètres qu'elle instancie. Cette théorie sera reprise ensuite sous les termes "Government and Binding" (théorie du gouvernement et du liage, souvent abrégée en "GB").
- depuis 1995, Chomsky promeut un "programme minimaliste", qui est une reformulation de ses théories précédentes mais orientée par un principe d'économie.

Chomsky est aussi généralement bien connu des informaticiens via la *hiérarchie de Chomsky* qui caractérise des familles de langages de complexité croissante. Cette hiérarchie est issue de la mathématisation de la notion de "grammaire générative", qui remonte aux années 60. C'est le fondement de la *théorie des langages formels*, branche dans laquelle sont étudiées les propriétés des langages artificiels comme les langages de programmation informatiques. On évoquera cette hiérarchie au chapitre 5 dans ce document (il).

1972

L'informaticien Terry Winograd présente son programme intitulé SHRDLU (ce nom proviendrait de l'ordre décroissant de la fréquence des lettres en anglais : ETAOIN*SHRDLU*...), permettant des interactions langagières avec un ordinateur sur un domaine restreint à un *monde de blocs*. Ce monde est constitué d'un nombre fini d'objets de forme géométrique simple (cubes, boules, cylindres, pyramides, etc.), disposés dans un environnement limité (l'équivalent d'une table). Les interactions se limitent à la possibilité de poser des questions sur l'état de ce monde simplifié ("Combien y a-t-il de cubes verts à droite de la boule rouge ?") et de donner des ordres permettant de le modifier ("Mettre le cylindre sur le cube bleu."). L'originalité de ce programme est qu'il ne se contente plus de mots-clés ou de "traitements de surface" rudimentaires : le monde étant parfaitement circonscrit, il pouvait être entièrement modélisé.

Années 70-80

Elles sont marquées en TALN par l'effervescence de la *sémantique formelle* pour représenter des connaissances et formaliser des raisonnements : nouvelles logiques (logique floue, logiques modales, etc.), "scripts" (Roger Schank), "frames" (Marvin Minsky), "réseaux sémantiques", "graphes conceptuels" (John Sowa) naissent à cette époque. La pragmatique, c'est-à-dire l'étude de l'utilisation du langage en contexte, dans des situations concrètes (déjeuner au restaurant ou réservation de billets de trains, par exemple), est aussi prise en compte dans ces modélisations. Les "systèmes experts", basés sur des modèles symboliques du même genre, constituent alors la vitrine de l'IA (il).

Depuis les années 90

L'augmentation considérable de la capacité de stockage et de calcul des ordinateurs, ainsi que le développement exponentiel d'Internet, permettent l'émergence d'une *linguistique de corpus* fondée sur l'exploitation de textes au format numérique. Cette linguistique plus empirique, *fondée sur les données* plutôt que sur des modèles formels abstraits, fait un usage intensif de calculs numériques et statistiques. Cette évolution va de pair avec les progrès de "l'apprentissage automatique", branche de l'IA consacrée à l'écriture de programmes qui s'améliorent avec l'expérience, grâce à des exemples. L'idée est d'utiliser des corpus pour apprendre automatiquement à fixer les paramètres de modèles (souvent statistiques) utilisables sur de nouvelles données. Cette démarche, prédominante dans la recherche actuelle, est abordée dans le chapitre 8 de ce document (il).

3 Les niveaux d'analyse du langage

La chronologie précédente a montré que ce n'est que progressivement que les sciences du langage ont précisé leur objet d'étude. Pour caractériser les *données* et les *traitements* pertinents de ce domaine, nous proposons le schéma de la figure 2.1.

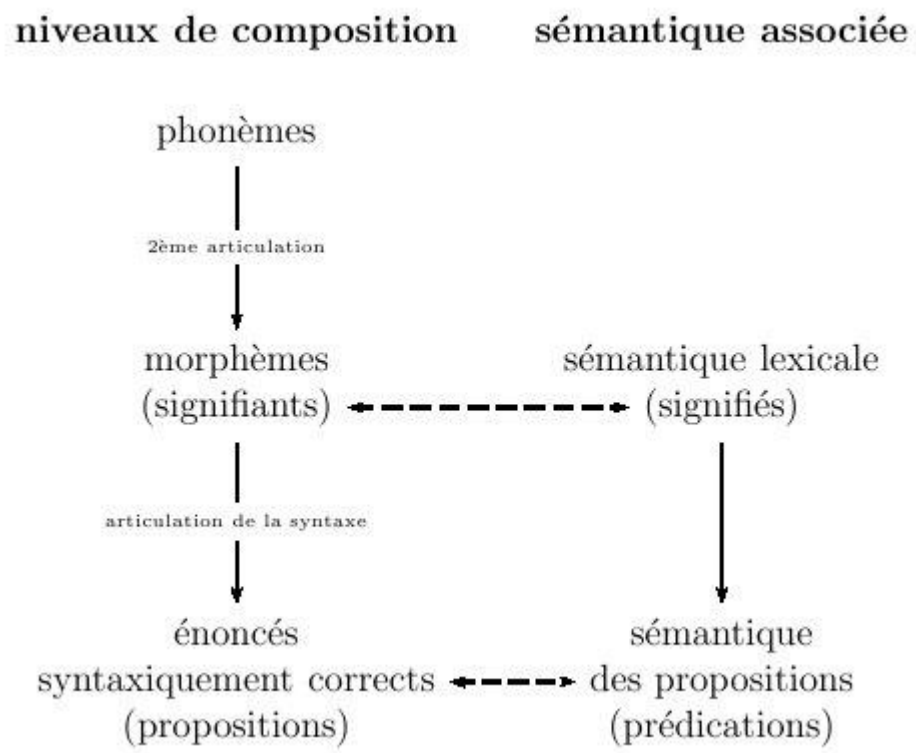

Figure 2.1 : hiérarchie des niveaux d'analyse des langues naturelles

Dans ce schéma, les principales unités d'analyse figurent de haut en bas, des plus simples aux plus complexes. Chacun des "nœuds" du schéma est constitué d'un ensemble de *données discrètes*. Les "flèches verticales descendantes" symbolisent des *règles de composition* opérant des traitements combinatoires sur les nœuds. Les "flèches bidirectionnelles horizontales en pointillés" traduisent, elles, des relations d'association qui ne sont pas biunivoques (ou pas bijectives) entre données. La non-univocité de ces flèches reflète les phénomènes d'*ambiguïté*, présents dans toutes les langues naturelles (sur lesquels nous reviendrons bien sûr par la suite).

Ce schéma permet surtout de mettre en évidence plusieurs des spécificités des langues naturelles évoquées précédemment :

- la dimension "verticale" du schéma est l'axe *syntagmatique*, tandis que sa dimension "horizontale" correspondrait plutôt à l'axe *paradigmatique*.
- la "double articulation" du langage se retrouve dans les deux niveaux de traitements combinatoires successifs qui occupent l'axe syntagmatique.
- ce schéma opère une claire distinction entre deux niveaux souvent confondus : celui de la *sémantique lexicale*, qui étudie le sens d'unités individuelles, et la *sémantique propositionnelle* qui étudie le sens d'énoncés complets, auxquels on peut attribuer une *valeur de vérité*.

Nous prétendons que ce qui caractérise les langues naturelles, c'est *l'ensemble des niveaux de description et relations présents dans ce schéma*. Précisons pourquoi.

Chacun des deux niveaux de l'axe syntagmatique traduit une combinatoire *ouverte* (au sens où la liste des éléments qu'elle produit est potentiellement infinie) d'*éléments discrets*. Ce

dispositif original diffère fondamentalement de codages de type *analogique*. On cite souvent, pour illustrer la communication animale, le cas des abeilles, qui informent leurs congénères de l'emplacement d'une source de nourriture en modulant de façon *continue* l'amplitude et l'orientation de leur vol : leur danse fait un "huit", incliné suivant un angle qui indique la direction à suivre et dont la hauteur est proportionnelle à la distance à parcourir. C'est un exemple typique de codage *analogique* (mais tous les modes de communication animale ne sont pas analogiques). Comme le remarque le psycholinguiste canadien Stephen Pinker (né en 1954), les deux systèmes les plus sophistiqués de transmission d'information qui ont été sélectionnés par la nature, à savoir les langues naturelles et le code génétique, reposent tous les deux sur des unités *discrètes*. La nature a inventé le codage "numérique" bien avant les informaticiens. Sans doute est-ce dû à la fiabilité de la transmission ainsi permise...

La maîtrise d'associations arbitraires "signifiant/signifié" au niveau lexical semble accessible à certaines espèces animales (principalement des singes) à qui on a pu enseigner l'usage d'un répertoire non négligeable de symboles : gestes empruntés à une langue des signes ou dessins abstraits arbitraires. Mais aucune espèce autre que l'homme n'a développé cette capacité dans la nature, sans enseignement explicite. Et (même si ce point est encore discuté) aucune non plus n'a semblé être capable de maîtriser complètement un niveau de combinaison syntaxico-sémantique plus complexe : la simple juxtaposition de symboles ne suffit pas à faire un langage doublement articulé.

Le niveau de l'association entre "unités signifiantes" et "sémantique lexicale" est un jalon important de notre schéma. Il constitue une étape fondamentale de l'acquisition de leur langue maternelle par les enfants, aux alentours de leur première année de vie. Certains auteurs argumentent aussi qu'il était peut-être la base d'un "protolangage" que nos lointains ancêtres auraient inventé avant d'avoir recours aux langues naturelles proprement dîtes. Evidemment, ce genre d'hypothèses est difficile à valider mais elle a le mérite de mettre l'accent sur la complexité considérable des langues humaines, dont on a du mal à imaginer comment elles ont pu émerger "d'un seul coup" dans une espèce particulière.

Pourtant, le critère de la "double articulation" est probablement insuffisant pour distinguer à lui tout seul les langues humaines de langages d'un autre genre apparus récemment : les langages de programmation informatiques (qui, il est vrai, étaient encore peu connus et diffusés en dehors des spécialises à l'époque de Martinet). On peut en effet argumenter que ces langages sont, eux aussi, doublement articulés :

- la première articulation est celle des règles à suivre impérativement pour écrire un programme syntaxiquement correct ;
- la deuxième articulation caractérise la construction des unités lexicales de ce programme (soit mots clés du langage, soit identifiants de variables ou noms de fonctions) à partir des unités distinctives élémentaires que sont les caractères alphanumériques autorisés (pendant écrit des phonèmes).

Il est possible aussi d'associer une "sémantique" aux programmes informatiques, aux deux niveaux évoqués par le schéma. Sa nature n'a pourtant rien à voir avec celle véhiculée par les langues naturelles : son domaine se réduit à l'arithmétique, son "monde" est celui du calcul opérationnel, et elle exclut toute ambiguïté. Les langages de programmation radicalisent en

quelque sorte les propriétés formelles des langues naturelles. Pour donner aux ordinateurs l'incroyable pouvoir expressif de ces langues, à savoir leur capacité à référer à ce qui leur est extérieur, à dire des choses sur le monde, il reste à réduire cette description du monde à un calcul. C'est paradoxalement leur caractère flou et *ambigu* qui rend finalement les langues naturelles plus aptes à réaliser ce programme.

Notre schéma permet ainsi de bien situer les unes par rapport aux autres les distinctions évoquées précédemment, qui se focalisent chacune sur une portion de sa structure. Il est toutefois loin d'être exhaustif. Un schéma plus complet devrait aussi faire figurer, à titre de niveau intermédiaire, la combinatoire propre à la *morphologie*, à laquelle nous consacrerons bien sûr un chapitre. Ne sont pas évoqués ici non plus les niveaux d'analyse qui vont au-delà des propositions (analyse des discours, des textes ou des dialogues, pragmatique...). Pourtant, les hommes se racontent en permanence des anecdotes et des histoires, et c'est au bout du compte ce *comportement narratif* qui caractérise le mieux l'espèce humaine. Mais la figure 2.1, et tout particulièrement le "rectangle" qu'elle fait apparaître à sa base, circonscrit en quelque sorte le périmètre de ce document.

Le principal intérêt pour nous de ce schéma, malgré son caractère réducteur, est qu'il met bien en évidence les données (les nœuds du schéma) et les traitements (les flèches) qui vont pouvoir donner lieu à une modélisation informatique. Il reste à voir comment traduire en code informatique les informations qui y figurent. La prise en compte des données discrètes de la colonne "niveaux de composition" ne pose pas de grosses difficultés, puisqu'il s'agit de coder des données discrètes par d'autres données discrètes. La traduction sous forme d'algorithmes des modes de combinaisons de ces données est un problème nettement plus intéressant. Dans ce domaine, théories linguistiques et modèles informatiques doivent collaborer. Quant au codage de la dimension "sémantique" du schéma, la plus cruciale et la plus problématique, elle est l'objet de nombreuses théories dont nous allons aussi essayer de rendre compte. Mais avant cela, voyons comment concevoir l'architecture "classique" d'un système complet de "compréhension du langage".

4 La chaîne de traitements "standard"

Pour comprendre comment les humains utilisent une langue naturelle, il ne suffit pas d'avoir identifié les différents niveaux de connaissances impliqués dans la compréhension de cette langue : il faut aussi savoir comment ces connaissances sont exploitées. Ce champ d'étude ne relève plus à proprement parler de la linguistique, mais de la *psychologie*, voire de la *psycholinguistique*. La figure 2.2 propose une chaîne de traitements qui a une certaine plausibilité. C'est avec ce genre de schémas que les psychologues cognitivistes tentent de modéliser le fonctionnement de l'esprit humain.

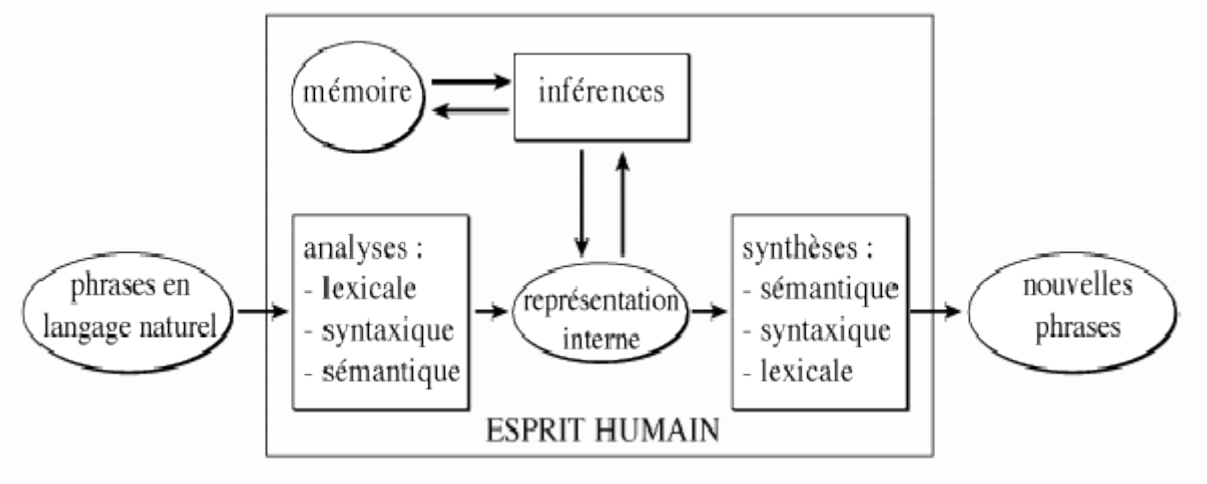

Figure 2.2 : chaîne de traitements classique de compréhension du langage

Dans ce schéma, ici encore très simplificateur (il n'intègre pas, par exemple, la composante orale du langage), les données figurent dans des ovales, tandis que les traitements sont représentés dans des rectangles. Les deux principaux rectangles, intitulés respectivement "analyses" et "synthèses", correspondent aux deux tâches principales accomplies par les locuteurs d'une langue. Le fait de bien les séparer provient de l'observation de patients souffrant de lésions cérébrales qui affectent en particulier une de ces compétences et pas l'autre. Il n'est pas nécessaire pourtant de faire l'hypothèse que chacun des traitements évoqués dans ce schéma soit réalisé par une aire cérébrale spécifique. Il suffit de considérer qu'il met en évidence un enchaînement des *fonctions*, indépendamment de leur "implantation" dans un substrat biologique concret.

Suivant ce schéma, *comprendre un énoncé* revient donc à le transformer, via une "analyse", en une représentation interne, tandis que pour en *produire* ou en *générer* un, il faut traduire linguistiquement une telle représentation via une "synthèse". Chacune de ces tâches nécessite de prendre en compte l'ensemble des niveaux d'analyse identifiés précédemment, mais dans un ordre différent et en faisant chaque fois des hypothèses différentes sur ce qui est connu et ce qui doit être accompli.
Maintenant, pour construire un système artificiel complet avec lequel des humains pourraient interagir via une langue naturelle, une première approche possible consiste à tenter de reproduire une architecture de ce genre, en traduisant les "fonctions" en programmes. C'est en quelque sorte le projet du "traitement automatique du langage naturel" dans sa forme originelle. Il a donné lieu à de très nombreux travaux ces 50 dernières années, et nous allons prendre le temps dans les chapitres qui suivent de présenter ses principaux résultats. Nous verrons notamment que beaucoup des outils formels imaginés pour modéliser certains des traitements cités ici sont en fait exploitables à la fois en analyse et en synthèse (c'est le cas, par exemple, des automates et grammaires formelles). D'autres sont plus spécifiques.

Pourtant, le schéma de la figure 2.2 n'est instancié dans sa totalité dans aucun système informatique existant. C'est un cadre idéal théorique, très influencé par une conception

cognitiviste de l'esprit humain, où les "représentations internes" sont en général de nature symbolique. Pour une application particulière, on se limite la plupart du temps à implémenter une toute petite portion de cette architecture.

Mais il est aussi possible d'envisager une toute autre approche, qui ne se soucie pas de crédibilité psychologique, et se concentre sur l'efficacité pragmatique de ses programmes. Cette mutation est représentative de l'évolution qu'a connue l'"intelligence artificielle" ces dernières années. Après avoir longtemps tenté d'imiter le fonctionnement de l'esprit humain, les chercheurs du domaine essaient plutôt désormais d'exploiter au mieux les capacités de mémoire et de calculs de leurs machines. Des *modèles symboliques formels*, on est passé aux *modèles statistiques fondés sur l'analyse des données*. Nous aborderons cette mutation dans le dernier chapitre de ce document, centré sur l'ingénierie linguistique, telle qu'elle est exploitée en "fouille de textes".

Dans la suite de ce document, nous allons en quelque sorte passer en revue les niveaux d'analyse du schéma de la figure 2.1, en présentant chaque fois les théories linguistiques qui les décrivent et quelques-uns des modèles informatiques auxquels ils ont donné lieu. Les données et les traitements seront "parcourus" de haut en bas et de gauche à droite, en passant des unités les plus simples aux unités les plus complexes, sans cacher les difficultés propres à l'informatisation de chacun d'eux. Seul le dernier chapitre sera consacré aux applications qui utilisent certains des outils ainsi décrits.

5 Sites Web

Avant de commencer l'exploration des niveaux d'analyse du langage, faisons un détour par le "test de Turing". Ce test est un jeu au cours duquel un juge humain, en situation de dialogue médiatisé par une machine avec une entité distante, doit décider au bout d'un temps fixé à l'avance si son interlocuteur est un humain ou un programme. Son origine remonte à un article philosophique que Turing, l'inventeur de l'informatique, a fait publier en 1950. Selon lui, un programme qui "passerait ce test" -autrement dit qui ne serait pas identifié comme tel par le juge- devrait se voir reconnu les mêmes capacités que celles attribuées "naturellement" aux humains -en particulier celle de penser... Ce test fait donc des capacités linguistiques en situation de dialogue le critère principal de l'"intelligence".

Il existe à l'heure actuelle une compétition qui propose d'instancier le "test de Turing" dans un cadre contrôlé : elle s'appelle le "Loebner Price", du nom du généreux donateur qui offre un prix et une médaille au vainqueur. Les sites suivants offrent la possibilité d'une petite conversation à distance avec un programme ayant participé à cette compétition. A vous de voir si vous leur accorderiez le prix...

- www-ai.ijs.si/eliza/eliza.html
- www.ellaz.com/EllaASP8/Direct.aspx
- www.bearbot.co.uk/
- www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1