Une petite introduction au Traitement Automatique des Langues Naturelles

François Yvon

0.1 Introduction

0.1.1 Préambule

On regroupe sous le vocable de *traitement automatique du langage naturel* (TALN) l'ensemble des recherches et développements visant à modéliser et reproduire, à *l'aide de machines*, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication. Il sera donc question ici de langage humain, d'où l'adjectif *naturel*, et non pas de langage formel, tel que C ou encore ADA. Ce *naturel* fait d'ailleurs tout le sel de l'affaire : les langages formels sont précisément conçus et optimisés dans l'optique de manipulations algorithmiques. Il en va tout autrement pour le langage naturel, dont le traitement automatique pose des difficultés majeures (voir la section 0.1.3). Précision importante, nous nous limiterons quasiment exclusivement au traitement du langage sous forme écrite, le traitement de la parole étant encore, en dépit de convergences de plus en plus marquées avec le traitement de l'écrit, une question de traitement du signal.

Pourquoi s'intéresser à l'automatisation du traitement du langage naturel ? Comme pour l'essentiel des champs de connaissance ressortissant de l'IA, on peut identifier deux sources principales de motivation à l'étude du TALN : d'une part la volonté de modéliser une compétence fascinante (le langage), afin de tester des hypothèses sur les mécanismes de la communication humaine, ou plus généralement sur la nature de la cognition humaine ; d'autre part le besoin de disposer d'applications capables de traiter efficacement les monceaux d'informations « naturelles » (documents écrits ou sonores) aujourd'hui disponibles sous forme électronique (mels, pages HTML, documents hypermédias, etc). Cette double motivation transparaît au long de l'histoire du TALN, histoire dont les grandes lignes sont esquissées à la section 0.1.2.

Le TALN est ainsi un champ de savoir et de techniques élaborées autour de problématiques diverses. Les concepts et techniques qu'il utilise se trouvent à la croisée de multiples champs disciplinaires : l'IA « traditionnelle », l'informatique théorique, la logique, la linguistique, mais aussi les neuro-sciences, les statistiques, etc.

Loin de présenter un panorama complet du domaine, notre objectif dans ce cours d'introduction est d'une part d'introduire les principaux concepts et les problèmes posés par le TALN, et d'autre part de présenter les formalismes utilisés pour modéliser certains de ces problèmes, en particulier les problèmes liés à l'analyse de la syntaxe des phrases. Ceci reduit de fait énormément de notre champ d'investigation, puisque nous n'aborderons donc que marginalement les questions liés à la comprhénsion et à l'interprétation, et pratiquement pas les question liées à la production (génération) automatique de documents.

Guide de lecture Après les deux sous-sections qui complètent cette introduction, nous commencerons par clarifier quelques concepts linguistiques, en étudiant les différents niveaux de représentation et de traitement des énoncés linguistiques (voir la section 0.2). La section suivante est consacrée à un large tour d'horizon des applications actuelles des outils de traitement du langage naturel.

0.1.2 Brève histoire du traitement automatique du langage naturel

Historiquement, les premiers travaux importants dans le domaine du TALN ont porté sur la traduction automatique, avec, dès 1954, la mise au point du premier traducteur automatique (très rudimentaire). Quelques phrases russes, sélectionnées à l'avance, furent traduites automatiquement en anglais. Bien que le vocabulaire ne comptât que 250 mots et la grammaire 6 règles, cette expérience a déclenché de nombreux travaux dans ce domaine. C'est en effet l'époque où l'URSS remporte succès après succès dans la course à l'espace et où les militaires américains sont très désireux de suivre les publications techniques soviétiques, sans pour autant faire apprendre le russe à tous leurs ingénieurs.

En 1962, la première conférence sur la traduction automatique est organisée au MIT par Y. Bar-Hillel. Depuis 1954, de lourds financements ont été investis et nombre de recherches ont été lancées, avec un optimisme que l'on peut considérer aujourd'hui comme exagéré. Les principaux travaux présentés concernent alors la fabrication et la manipulation de dictionnaires électroniques, car les techniques de traduction consistent essentiellement à traduire mot à mot, avec ensuite un éventuel réarrangement de l'ordre des mots. Cette conception simpliste de la traduction a conduit à l'exemple célèbre suivant : la phrase *The spirit is willing but the flesh is weak* (l'esprit est fort mais la chair est faible) fut traduite en russe puis retraduite en anglais. Cela donna quelque chose comme : *The vodka is strong but the meat is rotten* (la vodka est forte mais la viande est pourrie) !

Ce qui ressort de cet exemple, c'est que de nombreuses connaissances contextuelles (i.e. portant sur la situation décrite) et encyclopédiques (i.e. portant sur le monde en général) sont nécessaires pour trouver la traduction correcte d'un mot (par exemple ici *spirit*, qui, suivant les contextes peut se traduire comme *esprit* ou comme *alcool*). Autre exemple fameux, celui de la traduction du mot *pen* dans les deux phrases suivantes :

(1) (a) *the box is in the pen* (la boîte est dans l'enclos)

(b) *the pen is in the box* (le stylo est dans la boîte).

Le problème fondamental de la représentation des connaissances et de leur utilisation est donc posé, après moins de dix ans de recherches sur la traduction automatique. Ce problème est alors considéré comme insoluble, ce qui fait dire à Bar-Hillel que le problème de la traduction automatique est probablement insoluble.

Un groupe d'experts (l'Automatic Language Processing Advisory Council : ALPAC) rédige alors un rapport dans lequel il apparaît que la traduction automatique, en l'état des connaissances de l'époque, coûte environ deux fois plus cher que la traduction humaine et donne des résultats nettement moins bons. Cette considération purement économique amène un arrêt de la plus grande part des financements publics aux Etats-Unis puis en Europe.

Malgré l'échec des tentatives de traduction automatique, les années 50 voient néanmoins l'apparition d'idées fondamentales à la naissance desquelles les financements en traduction automatique n'ont certainement pas été étrangers. Zellig Harris publie ses travaux les plus importants de linguistique (linguistique distributionnaliste) entre 1951 et 1954. Il est suivi par N. Chomsky, qui publie en 1957 ses premiers travaux importants sur la syntaxe des langues naturelles, et sur les relations entre grammaires formelles et grammaires naturelles. Très schématiquement, la démarche de Chomsky est axée sur la volonté de formuler, à travers l'étude du langage, des hypothèses intéressantes sur la cognition. Le langage est une faculté à la fois universelle (tous les humains développent spontanément, pour peu que l'environnement s'y prête, un langage), et spécifique à l'espèce humaine (aucune espèce animal ne possède de système de communication comparable, dans sa richesse et dans sa complexitè, au langage humain). En conséquence, la mise à jour des propriétés que possèdent tous les langages humains est aussi un moyen de mettre en évidence certaines propriétés de l'appareillage cognitif universellement utilisé pour traiter le langage (*la grammaire universelle*).

On peut également situer en 1956, à l'école d'été de Dartmouth, la naissance de l'intelligence artificielle. Posant comme conjecture que tout aspect de l'intelligence humaine peut être décrit de façon suffisamment précise pour qu'une machine le simule, les figures les plus marquantes de l'époque (John Mc Carthy, Marvin Minsky, Allan Newell, Herbert Simon) y discutent des possibilités de créer des programmes d'ordinateurs qui se comportent intelligemment, et en particulier qui soient capables d'utiliser le langage.

Les élèves de Marvin Minsky, au MIT, développent divers systèmes (BASEBALL (1961), SIR (1964), STUDENT (1964), ELIZA (1966) ...) mettant en œuvre des mécanismes de traitement simples, à base de mots-clés. Leurs résultats, en particulier le comportement assez spectaculaire d'ELIZA, qui simule un dialogue entre un psychiatre et son patient, relancent les recherches sur la compréhension automatique du langage. La plupart de ces systèmes ne fonctionnent toutefois que dans des contextes de communication extrêmement restreints, et, s'ils utilisent quelques formes grammaticales prédéfinies dans le traitement

des phrases, se passent pratiquement de *syntaxe* et totalement de *sémantique* ou de *pragmatique* (tous ces concepts sont définis à la partie 0.2).

Des réflexions importantes sur la représentation des connaissances voient aussi le jour, principalement à l'initiative de Ross Quillian, qui préconise l'utilisation de réseaux sémantiques pour représenter le sens des mots et des phrases en explicitant les relations des divers concepts entre eux grâce à des liens qui indiquent le sens des relations.

Terry Winograd, en réalisant en 1972 SHRDLU, le premier logiciel capable de dialoguer en anglais avec un robot, dans le cadre d'un micro-monde (quelques blocs de couleurs et de formes variées, posés sur une table), montre que les diverses sources de connaissances (à propos de la structure des phrases, de leur sens et de ce qu'elles désignent dans le monde) doivent et peuvent interagir avec les modules d'analyse et de raisonnement.

Les années 70 voient ensuite le développement d'approches surtout sémantiques (Roger Schank, Yorick Wilks, ...), le rôle de la syntaxe étant pratiquement omis ou, tout du moins considéré comme secondaire. L'importance du contexte et le rôle essentiel d'une bonne connaissance du domaine traité pour comprendre un texte est ainsi mis en avant. On ne se limite plus au seul sens objectif et on remarque que la signification subjective dépend très étroitement d'informations implicites qui font partie des connaissances générales communes aux interlocuteurs. M. Minsky tente alors d'élaborer un cadre général de représentation des connaissances, les *frames*, alors que R. Schank s'efforce d'identifier clairement les diverses connaissances nécessaires dans un système interprétant le langage naturel. Les recherches ont alors cessé de se limiter à l'interprétation de phrases seules pour aborder le traitement d'unités plus importantes comme les récits et les dialogues.

Parallèlement, les modèles syntaxiques connaissent en informatique des développement et des raffinements continus, et des algorithmes de plus en plus performants sont proposés pour analyser les grammaires les plus simples (grammaires régulières et algébriques). Depuis Chomsky, ces formalismes grammaticaux sont toutefois considérés comme trop simples pour modéliser correctement les phénomènes observés dans les langues naturelles. Ces développements des gramaires formelles sont donc largement sous-estimés, jusqu'à ce qu'au milieu des années 70, divers travaux théoriques, en particulier ceux de Ronald Kaplan et de Martin Kay, réhabilitent ces formalismes dans le cadre du traitement de la *morphologie* et de la *phonologie* des langues naturelles. Ces années voient également une recrudescence des travaux en syntaxe des langues naturelles, et à l'émergence de nouveaux formalismes de description grammaticale, qui étendent de manière informatiquement gérable les grammaires algébriques. Ce sont tout d'abord les réseaux de transition augmentés (abbréviation anglaise ATN), puis les *grammaires d'unification*, que nous étudierons plus en détail pendant les cours de syntaxe.

Bien évidemment, et quelle que soit leur élégance, les propositions issues de l'intelligence artificielle jusqu'au début des années 80 ne permettent pas d'échapper à l'obligation d'affronter la complexité de la tâche de description préalable des connaissances sur la langue et sur le monde. C'est pourquoi une partie importante des travaux actuels vise à analyser et à formaliser *des mécanismes d'acquisition automatique des connaissances*, qui permettent d'extraire directement de lexiques ou de corpus de documents, des règles de grammaire, ou encore des connaissances sémantiques.

Aujourd'hui, le champ du traitement du langage naturel est un champ de recherche très actif. De nombreuses applications industrielles (traduction automatique, recherche documentaire, interfaces en language naturel), qui commencent à atteindre le grand public, sont là pour témoigner de l'importance des avancées accomplies... mais également des progrès qu'il reste encore à accomplir.

0.1.3 Les difficultés du TALN : ambiguïté et implicite

Les difficultés que l'on rencontre en TALN sont principalement de deux ordres, et ressortent soit de l'ambiguïté du langage, soit de la quantité d'implicite contenue dans les communications naturelles.

Ambiguïté Le langage naturel est ambigu, et ce à quelque niveau qu'on l'appréhende. Cette ambiguïté, loin d'être marginale, est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc). Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, comme en témoignent les exemples suivants :

- ambiguïté des graphèmes (lettres) dans le processus d'encodage orthographique : comparez la prononciation du *i* dans *lit*, *poire*, *maison* ;
- ambiguïté des terminaisons dans les processus de conjugaison et d'inflection : ainsi un /s/ final marque à la fois le pluriel des noms, des adjectifs, et la deuxième (parfois également la première) personne du singulier des formes verbales ;
- ambiguïté dans les propriétés grammaticales et sémantiques (i.e. associées à son sens) d'une forme graphique donnée : ainsi *manges* est ambigu à la fois morpho-syntaxiquement, puisqu'il correspond aux formes indicative et subjonctive du verbe *manger*), mais aussi sémantiquement. En effet, cette forme peut aussi bien référer (dans un style familier) à un ensemble d'actions conventionnelles (comme de s'assoir à une table, mettre une serviette, utiliser divers ustensiles, ceci éventuellement en maintenant une interaction avec un autre humain) avec pour visée finale d'ingérer de la nourriture (auquel cas il ne requière pas de complément d'objet direct) ; et à l'action consistant à effectivement ingérer un type particulier de nourriture (auquel cas il requiert un complément d'objet direct), etc. Comparez en effet :

  (2) (a) *Demain, Paul mange avec ma sœur.*

  (b) *Paul mange son pain au chocolat.*

  ainsi que les déductions que l'on peut faire à partir de ces deux énoncés : de (a), on peut raisonnablement conclure que Paul sera assis à une table, disposera de couverts,... ; tout ceci n'est pas nécessairement vrai dans le cas de l'énoncé (b).
- ambiguïté de la fonction grammaticale des groupes de mots, illustrée par la phrase :

  (3) *il poursuit la jeune fille à vélo.*

  Dans cet exemple *à vélo* est soit un complément de manière de poursuivre (et c'est *il* qui pédale), soit un complément de nom de *fille* (et c'est *elle* qui mouline) ;
- ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions. Ainsi, dans

  (4) *Tous mes amis ont pris un verre*

  on peut supposer que chacun avait un verre différent, mais dans

  (5) *Tous les témoins ont entendu un cri*

  il est probable que c'était le même cri pour tous les témoins. De même, lorsque l'on évoque *les chiens et les chats de Paul*, l'interprétation la plus naturelle consiste à comprendre *de Paul* comme le complément de nom du groupe *les chats et les chiens* ; cette lecture est beaucoup moins naturelle dans *les chiens de race et les chats de Paul* ;
- ambiguïté sur l'interprétation à donner en contexte à un énoncé. Comparez ainsi la « signification » de *non*, dans les deux échanges suivants :

  (6) (a) *Si je vais en cours demain ? Non* (négation)

  (a) *Tu vas en cours demain ! Non !* (j'y crois pas)

Conformément au parti pris de ce cours d'introduction, nous avons surtout insisté sur les ambiguïtés de reconnaissance (compréhension), mais les problèmes se posent naturellement de manière symétrique pour ce qui est de la génération : comment choisir les phrases produites de manière à limiter les ambigïtés pour le recepteur ? Comment sélectionner parmi un ensemble de synonymes ? parmi un ensemble de paraphrases ?

Implicite L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites. La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible, si l'on ne dispose pas de bases de connaissance additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique). Un exemple typique est la désambiguïsation du référent du pronom personnel *il* dans les trois énoncés suivants : *Le professeur envoya l'élève chez le censeur parce qu' ...* :

(7) a. *... il lançait des boulettes* (*il* réfère probablement à l'élève)

b. *... il voulait avoir la paix* (*il* réfère probablement au professeur)

c. *... il voulait le voir* (*il* réfère probablement au censeur)

En l'absence de telles connaissances, bien d'autres problèmes de compréhension deviennent pratiquement insurmontables : pensez par exemple aux ellipses, aux métaphores, et, plus généralement, aux figures de style.

Fort heureusement, il existe de nombreuses applications pour lesquelles ces difficultés peuvent être, dans une large mesure, circonscrites. Dès lors, en effet, que l'on restreint le cadre des textes analysés à un sous-domaine particulier (textes juridiques, textes scientifiques, serveur d'information spécialisé dans les informations sportives...), il devient possible d'une part d'ignorer un grand nombre d'ambiguïtés, en particulier sémantiques (par exemple dans le contexte de textes juridiques, on pourra probablement négliger la possibilité qu'un *avocat marron* désigne un fruit un peu trop mûr) ; et d'autre part de représenter formellement un grand nombre des connaissances nécessaires à la compréhension des énoncés du domaine considéré. En fait, certains domaines d'activité ou contextes d'interactions spécifiques semblent restreindre de manière drastique l'ensemble des énoncés possibles (ou acceptables), simplifiant de manière considérable le traitement de ces véritables *sous-langages* par une machine.

0.2 Les niveaux de traitement

Nous introduisons dans cette section les différents niveaux de traitement nécessaires pour parvenir à une compréhension complète d'un énoncé en langage naturel. Du point de vue de l'ingénieur, ces niveaux correspondent à des modules qu'il faudrait développer et faire coopérer dans le cadre d'une application complète de traitement de la langue. Mais il n'est pas absurde de voir également dans ces nivaux, tant ils semblent demander des connaissances et des mécanismes différents, un modèle des différents composants de la machinerie cognitive mobilisée dans la production et la compréhension du langage.

Vous le verrez, au fur et à mesure que l'on progresse dans cette hiérarchie des niveaux, les difficultés s'accumulent, et les outils aujourd'hui disponibles se font moins performants, ou ne sont opérants que pour des sous-domaines particuliers. Toutefois, bien des applications ne nécessitent pas une compréhension complète des énoncés, et ne mettent en œuvre que des traitements correspondant aux niveaux les mieux compris et automatisés.

0.2.1 Introduction

Considérons à titre d'exemple l'énoncé :

(8) *Le président des antialcooliques mangeait une pomme avec un couteau*

et envisageons les traitements successifs qu'il convient d'appliquer à cet énoncé pour parvenir automatiquement à sa compréhension la plus complète. Il nous faudra successivement :
– segmenter ce texte en unités lexicales (mots) ;
– identifier les composants lexicaux, et leurs propriétés : c'est l'étape de traitement **lexical** ;
– identifier des constituants (groupe) de plus haut niveau, et les relations (de dominance) qu'ils entretiennent entre eux : c'est l'étape de traitement **syntaxique** ;
– construire une représentation du sens de cet énoncé, en associant à chaque concept évoqué un objet ou une action dans un monde de référence (réel ou imaginaire) : c'est l'étape de traitement **sémantique** ;
– identifier enfin la fonction de l'énoncé dans le contexte particulier de la situation dans lequel il a été produit : c'est l'étape de traitement **pragmatique** ;
La séquentialité de ces traitements est une idéalisation. Dans la pratique, il est préférable de concevoir ces niveaux de traitement comme des processus coopératifs, qui échangent de l'information dans les deux sens (à la fois des niveaux « bas » vers les niveaux « hauts », et en sens inverse) : il est ainsi souvent nécessaire de faire appel à des informations sémantiques pour trouver la « bonne » structure syntaxique d'une phrase, etc.

Ces niveaux conceptuels, correspondant ou non à des modules distincts de traitement, se retrouvent dans d'autres applications du TALN. Ainsi une application de génération de texte impliquera la production d'un argumentaire (pragmatique), la construction de représentations des significations à engendrer (sémantique), la transformation de ces représentations sémantiques en une suite bien formée de mots (morpho-syntaxe), etc.

0.2.2 Traitements de « bas niveau »

Même si on n'insistera pas dans la suite sur la partie segmentation, il faut bien reconnaître que la tâche, quoique grandement facilitée en français[1] par la présence de séparateurs explicites (les espaces et autres signes de ponctuations), n'a rien de trivial, à cause de l'ambiguïté (à nouveau) desdits séparateurs. On notera également que la sémantique de ces séparateurs varie suivant les langues. Quelques exemples typiques :
, sépare des propositions, la partie décimale et numérique des nombres réels ;
. marque les fins de phrase, mais apparaît aussi dans les sigles (*S.N.C.F*), les abréviations (*M. Jacques*) ;
- apparaît comme séparateur de mots composés, mais également pour désigner l'opérateur arithmétique, dans les scores (*victoire 3-0 des ...*) ;
' signale une élision, mais apparaît également dans certains noms propres (*O'hara*, *Guiwarc'h*), dans les notations du temps (*il a couru le cent mètre en 12'3.*), etc.
Un bon[2] segmenteur se doit donc de bien recenser tous les usages possibles dans la langue écrite des signes de ponctuation.

Signalons également pour mémoire un second problème, correspondant à un traitement de « bas niveau » des textes électroniques, et qui est relatif à leur format. Alors qu'il y seulement 10 ans, pour l'essentiel les documents informatiques étaient stockés dans un format relativement pauvre (texte ASCII non-accentué en général), les documents d'aujourd'hui sont disponibles dans des formats enrichis (HTML, RTF, etc) qui contiennent des multitudes d'informations telles que les marques de fin de paragraphe, les changements de fontes, etc. Ces informations renseignent manifestement sur la structure et le contenu du document, de

[1] Ce n'est pas le cas dans toutes les langues : en thaïlandais, par exemple, les ruptures de mots de sont pas orthographiées.
[2] Pour information, on considère que les meilleurs segmenteurs pour le français ont un taux d'erreur pour l'indentification des fins de phrase compris entre 5 pour mille et 1 pour cent (suivant la nature des textes à segmenter).

la même manière que l'intonation d'une phrase renseigne sur son contenu. L'absence de normalisation de l'usage de ces marqueurs rend leur exploitation difficile, mais il est certain que la tendance actuelle va vers une intégration des règles ou conventions d'utilisation de ces marqueurs dans les systèmes de traitement (cf. par exemple l'initiative internationale de normalisation des systèmes de balisage de corpus écrits connue sous le nom de *Text Encoding Initiative*[3]). Dans le même ordre d'idées, la banalisation de l'usage du mel est en train de faire émerger de nouvelles conventions typographiques, qui introduisent elles aussi comme une forme de marquage partiel de la prosodie dans l'écrit (usage de *smileys*, des oppositions majuscule/minuscule, etc).

0.2.3 Le niveau lexical

Objectifs du traitement lexical

Le but de cette étape de traitement est de passer des *formes atomiques* (tokens) identifiées par le segmenteur aux *mots*, c'est-à-dire de *reconnaître* dans chaque chaîne de caractère une (ou plusieurs) unité(s) linguistique(s), dotée(s) de caractéristiques propres (son sens, sa prononciation, ses propriétés syntaxiques, etc).

En poursuivant sur l'exemple (8), l'étape d'identification lexicale devrait conduire à un résultat voisin de celui donné ci-dessous, dans lequel on peut constater en particulier l'ambiguïté d'une forme telle que *président* : cette chaîne correspond à deux formes du verbe *présider* (indicatif et subjonctif), ainsi à une forme nominale, et sa prononciation diffère selon qu'elle représente un nom ou un verbe.

le - det. masc. sing., /lə/ ; pron. pers. masc. sing., /lə/
président - vrb 3pers. plur. prés. ind./ subjonctif [présid+ent], <présider(X), présider(X,Y)>, /pKezidɐ/ ; nom masc. sing., ←présider : action de X, <president(X)>, /pKezidɑ̃/
des - det. masc./fem. plur., /dɛ+z/ ; prep. contr. *de les*. ...
antialcooliques - adj. masc./fem. plur. [anti+alcool+ique+s], ← alcoolique : s'opposer à X, antialcoolique(X), /ɑtialkolikə+z/ ; nom. masc. sing. [anti+alcool+ique+s], ← antialcoolique (adj) : être X, antialcoolique(X), /ɑ̃tialkolikə/
mangeait - vrb (1,3) pers. sing. imp. ind., [mang+e+ait], <manger(X),manger(X,Y)>, /mɜ̃ɛ/
pomme - nom fem. sing., [pomme], <pomme(X),fruit(X),golden(X)...>, /pɔmə/

Accès lexical direct

Comment cette identification est-elle réalisée ? On conçoit aisément que pour les mots les plus fréquents, comme *le*, la solution la plus simple soit de rechercher la forme dans un lexique pré-compilé. Dans les faits, c'est effectivement ce qui se passe, y compris pour des formes plus rares, dans la mesure où l'utilisation de formalismes de représentation compacts permettant un accès optimisé (par exemple sous la forme d'automates d'états finis), et l'augmentation de la taille des mémoires rend possible la manipulation de vastes lexiques (de l'ordre de centaines de milliers de formes).

Pour autant, cette solution ne résoud pas tous les problèmes. Le langage est *création*, et de nouvelles formes surgissent tous les jours, que ce soit par emprunt à d'autres langues (il n'y a qu'a écouter parler les enseignants des autres modules de la dominante informatique !), ou, plus fréquemment, par l'application de procédés réguliers de créations de mots, qui nous permettent de composer pratiquement à volonté de nouvelles formes *immédiatement compréhensibles par tous les locuteurs de notre langue* : si j'aime lire *Proust*, ne peut-on pas dire que je *m'emproustise*, que de *proustien* je deviens *proustiste*, voire *proustophile*, puis que, lassé, je me *désemproustise*... Ce phénomène n'a rien de marginal, puisqu'il est admis que, même si l'on dispose d'un lexique complet du français, environ 5 à 10 % des mots d'un article de journal pris au

[3] http://www.uic.edu/orgs/tei/

hasard ne figureront pas dans ce lexique. La solution purement lexicale atteint là ses limites, et il faut donc mettre en œuvre d'autres approches, de manière à traiter aussi les formes hors-lexique.

Introduction à la morphologie

Si vous pouvez comprendre une forme telle que *proustien*, c'est que vous avez été capables de reconnaître dans cette forme des composants plus petits, nommément une *racine*, *Proust*, qui réfère au nom d'un écrivain, et un *suffixe*, *ien*, qui permet de manière régulière de construire des adjectifs à partir de noms propres. La linguistique traditionnelle appelle ces composants plus petits les *morphèmes*, et l'étude de leurs combinaisons la *morphologie*. La morphologie s'attache à décrire deux types de phénomènes relativement distincts :

– les processus d'ajustement de forme imposés par les conditions syntaxiques d'utilisation du mot : ainsi le pluriel se construit en français le plus souvent par ajout d'un *s*, le féminin d'un adjectif par suffixation d'un *e*, le futur se marque par adjonction d'un *r* (et d'une conjugaison spécifique dépendant du nombre et de la personne), certaines langues ont des sytèmes de cas qui permettent de différencier une forme en position sujet (cas nominatif) d'une forme en position objet (cas accusatif) ... Ces processus sont les processus *flexionnels*, qui se caractérisent par leur systématicité[4] (tous les noms ont une forme pluriel...), et leur régularité relative (sauf accident historique, la forme plurielle se construit en français ajoutant un *s*).
– les processus de créations de nouvelles formes à partir de formes existantes, qui sont les processus qu'étudie la morphologie *dérivationnelle*. Les processus dérivationnels entraînent le plus souvent un changement de la catégorie morpho-syntaxique : un nom se transforme en verbe, un verbe en adjectif... Ces processus se caractérisent par leur moindre prédictibilité, aussi bien en termes des mots qui y sont soumis (certains verbes du premier groupe n'ont pas de dérivé nominal pour désigner l'action liée au verbe), qu'en termes de signification du dérivé construit. L'histoire de la langue a en effet gommé des filiations, figé des dérivés, dont le sens ne s'analyse plus aujourd'hui directement à partir de ces composants : la *cassure* est bien le résultat de l'action de *casser*, comme la *morsure* celui de l'action de mordre, mais une *gageure* n'est pas toujours relier à une action de *gager*, et la *mâchure* est une pièce de velours dont les bords sont écrasés (mâchés). Les constructions dérivationnelles sont donc soumises à une pression vers la *lexicalisation*, c'est à dire vers un figement de l'association globale entre forme et sens, qui peu à peu, devient indécomposable.

Les phénomènes morphologiques sont plus (les flexions) ou moins (les dérivations) bien décrits, et ces descriptions ont permis aux compulinguistes de développer des analyseurs morphologiques, qui mettent en œuvre deux types de connaissance :

– les règles de combinaison morphématiques, qui expriment les conditions sous lesquelles l'association de deux morphèmes est possible, ainsi que son résultat (i.e. la nature syntaxique du dérivé, et si possible, sa signification). Ainsi, une règle telle que :

  (9) $N/\text{singulier} \rightarrow N + s/\text{pluriel}$

  exprime la construction du pluriel d'un nom ; une règle telle que :

  (10) $X + er/\text{verbe infinitif} \rightarrow X + ure/\text{nom féminin singulier}$

  la nominalisation d'un verbe du premier groupe... Pour l'essentiel, les formalismes de représentation et les mécanismes de traitement associés à ces règles de combinaisons morphématiques sont les mêmes que ceux utilisés pour le traitement syntaxique. Dans la mesure où ces règles sont susceptibles de s'enchaîner, elles permettent de fournir une analyse arborescente des mots, telles que celle représentée à la figure 1.
  Notez que si, en français, de telles règles prennent souvent la forme de l'adjonction d'un *affixe* (suffixe ou préfixe), ce n'est pas le cas de toutes les langues. La morphologie des verbes arabes, par exemple, repose

[4] Enfin, quasi systématicité : il existe bien quelques verbes défectifs (*clore*, *gésir*), quelques noms sans singulier (*us*)...

sur des alternances de voyelles autour et *à l'intérieur* d'un squelette de consonnes, qui correspond à la racine.
– des règles d'ajustement ortographiques, souvent imposées par des nécessités phonologiques, et qui permettent de préserver l'intégrité de la forme orale d'un mot, ainsi que sa conformité aux règles décrivant les successions possibles des sons pour une langue donnée. Reprenons l'exemple des dérivés en *-ure*. En première approche, ces dérivés sont le plus souvent construits en remplaçant la terminaison *er* du verbe correspondant. *gageure* fait exception à cette règle, puisque seul le *r* de la terminaison verbale est remplacé, ce qui permet de conserver l'identitié phonique de la racine (i.e. le fait qu'il se prononce comme *gage*). On retrouve ce type d'ajustement sous des formes diverses (alternance de *c* et *ç*, alternance de *é* et de *è*...) à de nombreux endroits des systèmes flexionnels et dérivationnels du français.

Vous noterez finalement que la connaissance de la structure morphologique se révèle dans bien des cas indispensable par exemple pour prononcer correctement une forme : ainsi *tia* se prononce différemment dans *antialcoolique* et dans *martial*, le *s* de *asocial* ne se prononce pas *z*, etc.

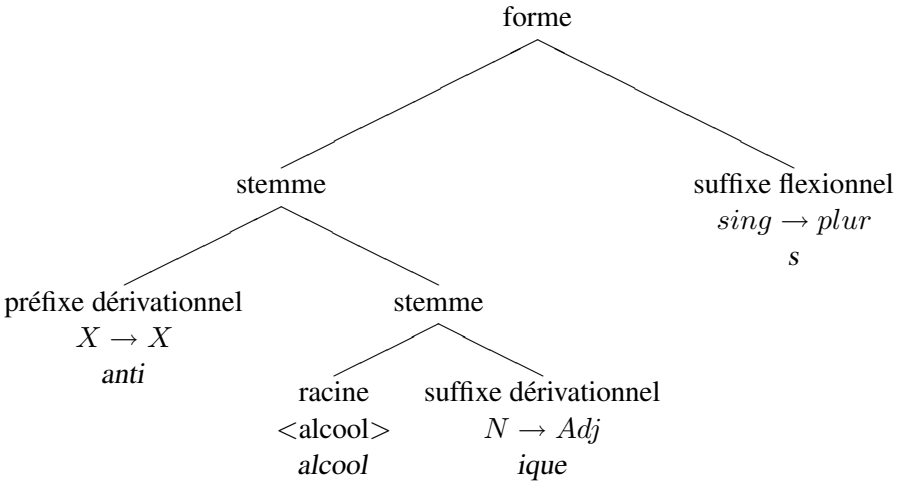

FIG. 1 – Une décomposition arborescente de la forme *antialcooliques*

Une dernière source majeure de création lexicale, particulièrement importante dans les domaines techniques, est la composition. De nombreux *mots composés* ont en effet acquis des sens particuliers tout à fait précis qui ne sont plus directement déductibles de leurs coposants. Quelques exemples :
– noms composés : *coupe-papier*, *compte courant*, *pomme de terre*, ...
– adverbes composés : *en effet*, *de temps à autre*, ...
– conjonctions composées : *parce que*, *si bien que*, ...
– composés verbaux : *mettre de l'eau dans son vin*, *prendre le taureau par les cornes*, ...
–

Cette tendance est exarcerbée dans les domaines techniques, entraînant la nécessité de normalisation de ces composés au sein de *terminologies*) : pensez à *système expert*, *réseau de neurones*, *système distribué*, *langage objet* ou encore à *base de données*. La catégorie des mots composés est toutefois difficile à appréhender, car ses frontières sont floues et mouvantes. Si, en effet, tout le monde s'accorde sur la nature d'un certain nombre d'expressions figées, qui fonctionnent syntaxiquement comme des unités indissociables (essayez pour voir d'insérer un déterminant ou un adjectif dans le composé *poule au pot*, et demandez-vous ce que le grand Henri en aurait dit !), le statut d'un certain nombre de syntagmes nominaux complexes qui apparaissent de manière récurrente, surtout dans les domaines techniques, est beaucoup moins clair, et en constante évolution, au fur et à mesure que les vocabulaires spécialisés se stabilisent. Un terme tel que *réseau de neurones* admet des variantes (*réseau neuronal*, *réseau de neurones formels*...), son sens est quasiment directement déductible de ses composants, mais il n'a plus aujourd'hui toute la « flexibilité » syntaxique d'un syntagme quelconque. Pour ces composés, en particulier pour ceux dont la signification ne se déduit plus

simplement de leurs composants atomiques, il n'y a bien souvent pas d'autre solution que de les recenser dans des lexiques. La construction automatique ou semi-automatique de tels lexiques constitue un champ de recherche actif, avec de nombreux débouchés industriels.

Ces formes composées, il va sans dire, contribuent à augmenter de manière sensible le nombre d'ambiguïté (ici des ambiguïtés de segmentation), puisque dans bien des cas, pour un même groupe, il existe une interprétation figée et une interprétation « littérale ». Un groupe tel que *avant que* fonctionne ainsi comme une locution dans (11-a), et comme deux mots isolés dans (11-b) :

(11) (a) *il est parti avant que la soirée fût finie ;*

(b) *il aurait pu dire avant que la soirée serait mondaine ;*

Les lexiques électroniques et les analyseurs morphologiques constituent un maillon important de la chaîne de traitement. Pour garantir une bonne couverture de la langue, un lexique électronique pour le français devra ainsi typiquement contenir 80 000 formes canoniques correspondant à 500 00 formes fléchies. Si l'on tient compte des formes composées, il faudra prévoir plusieurs millions d'entrées ! Ainsi, pour les 40 000 substantifs simples du dictionnaire DELAS du LADL (Université Paris VI), on estime qu'il existe au moins 200 000 formes nominales composées. Pour les 2 000 adverbes simples, plus de 6 000 adverbes composés... Ce dictionnaire électronique, comme beaucoup d'autres aujourd'hui, peut être consulté sur la toile à l'adresse : http://www-ceril.univ-mlv.fr/Dictionnaires/. Vous pouvez également jeter un œil aux démonstrations d'outils de traitement morphologiques développés par Rank Xerox : http://www.rxrc.xerox.com/research/mltt/toolhome.html.

0.2.4 Le niveau syntaxique

Syntaxe et grammaires

La syntaxe est l'étude des contraintes portant sur les successions licites de formes qui doivent être prises en compte lorsque l'on cherche à décrire les séquences constituant des phrases grammaticalement correctes (voir ci-dessous pour la notion de grammaticalité) : toutes les suites de mots ne forment pas des phrases acceptables. Les contraintes envisagées sont de nature variée et correspondent à des propriétés *sélectionnelles* (telles que les règles d'accord en genre, en nombre, en cas, ...) ou *positionnelles* (telles que celles qui contrôlent les positions relatives des mots dans la phrase, ..).

La description des contraintes caractéristiques d'une langue donnée se fait par le biais d'une *grammaire*. Les modèles et formalismes grammaticaux proposés dans le cadre du traitement automatique du langage sont particulièrement nombreux et variés. On aura l'occasion de revenir en détail sur une classe particulière de modèles, à savoir la classe des grammaires d'unification.

Le niveau syntaxique est donc le niveau conceptuel concerné par le calcul de la validité de certaines séquences de mots, les séquences grammaticales ou bien-formées. On conçoit bien l'importance d'un tel traitement dans une application de génération, pour laquelle il est essentiel que la machine engendre des énoncés corrects. Dans une application de compréhension, la machine analyse des textes qui lui sont fournis, et dont on peut supposer qu'ils sont grammaticaux. Pourquoi donc, dans ce cas, mettre en œuvre des connaissances syntaxiques ?

Une première motivation provient de ce que justement, les textes ne sont pas toujours grammaticaux, par exemple à cause de coquilles et fautes d'orthographes. Une analyse syntaxique peut donc permettre de choisir entre plusieurs corrections à apporter à une phrase incorrecte, mais également se révéler bien utile pour améliorer les sorties d'un système de reconnaissance optique de caractère ou encore un système de reconnaissance de la parole.

Une seconde raison est que l'entrée du module syntaxique est une série de formes étiquetées morpho-

syntaxiquement, une forme pouvant avoir plusieurs étiquettes différentes. Une première fonction du module syntaxique consiste donc à désambiguïser la suite d'étiquettes, en éliminant les séquences qui correspondent à des énoncés grammaticalement invalides. Poursuivant sur notre exemple jouet, le module syntaxique devrait éliminer la suite d'étiquettes `det vrb` pour les deux premiers mots de la phrase (8), un verbe n'étant jamais, en français, précédé d'un article.

Les constituants syntaxiques

La troisième raison est que les énoncés naturels ne sont pas simplement des suites de mots, mais sont organisés en constituants de taille supérieure au mot (les *syntagmes*), qui entretiennent entre eux des relations de dominance et de contrôle. Le second but de l'analyse syntaxique est donc d'associer, à chaque énoncé, sa structure de constituants. L'organisation syntagmatique des énoncés est marquée de multiples manières dans le langage parlé par le biais de la prosodie (pause, accentuation, montée ou descente mélodique marquée, allongement de la syllabe finale, etc). Sa retranscription au niveau graphique est moins systématique, au travers des signes de ponctuation.

Pourtant, l'existence d'une telle organisation hiérarchique ne fait pas de doute, comme le montre l'examen du syntagme *le chien de ma voisine*. L'entité désignée par ce groupe est un type de chien, pas une sorte particulière de voisine. On peut donc dire, d'une certaine manière, que dans ce syntagme, *chien* domine *voisine*. Le constituant dominant d'un syntagme est appelé la *tête* du syntagme.

L'existence de composants dans cette structure hiérarchique est attestée par un certain nombre de faits syntaxiques :

– la possibilité de variations *paradigmatiques* entre composants de tailles différentes. Les deux énoncés suivants :

(12) (a) *Le chien de la voisine mange*

(b) *Le chien mange*

ont manifestement la même structure, ce qui impose de considérer que *chien de ma voisine* dans (a) joue le même rôle que (i.e. est un constituant syntaxiquement équivalent à) *chien* dans (b).

– les contraintes qui portent sur les déplacements de constituants dans des transformation telles que la formation du passif à partir de l'actif, ou la construction d'interrogatives. Prenons d'abord l'exemple du passif : l'observation d'un exemple tel que :

(13) *Jean casse la boîte - La boîte est cassée par Jean*

suggère une règle simple pour la passivisation : déplacer les deux mots qui suivent le verbe en position sujet, mettre le verbe au passif, et déplacer les deux mots qui précèdent le verbe en position post-verbale, en les faisant précéder de *par*. Cette règle est toutefois mise en échec par un exemple tel que :

(14) *Jean a cassé la boîte de Paul*

qui donne au passif, parmi les deux possibilités suivantes :

(15) (a) *La boite de Paul a été cassée par Jean*

(b) **La boîte à été cassée par Jean de Paul*

l'énoncé (a) et non le (b) : la passivation nécessite le déplacement du groupe *la boîte de Paul* dans son intégralité, manifestant l'existence d'un constitutant structuré, qui joue dans (14) le même rôle que *la boîte* dans (13).
De même, la question correspondant à

(16) *L'homme qui est grand est dans la chambre*

est, parmi les deux possibilités suivantes :

(17) (a)**L'homme qui est-il grand est dans la maison*

(b) *L'homme qui est grand est-il dans la maison*

l'énoncé (b) : la règle de construction de la forme interrogative prend en compte le fait que le premier *est* appartient à un constituant (*l'homme qui est grand*) dont le constituant dominant (la tête) est *homme*.

– le test de la conjonction : dans une phrase quelconque, il est possible de rajouter des éléments par conjonction ; cependant cette possibilité est fortement contrainte, et l'on ne peut pas effectuer cette opération pour tous les groupes de mots. Ainsi, à partir de : *La fille de Minos se repose dans son île*, (a), (b), (c) et (d) sont possibles, mais pas (e) ni (f) :

(18) (a) *Le fils et la fille de Minos se reposent dans leur île*

(b) *La fille de Minos et de Pasiphaé se repose dans son île*

(c) *La fille de Minos se repose et tisse dans sa chambre*

(d) *La fille de Minos se repose dans son île ou dans son bateau*

(e)**La fille de et la nièce de Minos se reposent dans leur île*

(f)**Le fils de Jean se repose ou lamente dans son île*

Un but important de l'analyse syntaxique est donc d'identifier les différents constituants et sous-constituants, ainsi que de repérer les relations que ces groupes entretiennent entre eux, et les fonctions syntaxiques qu'il remplissent (sujet, objet direct, object indirect, circonstant...). En d'autres termes, il s'agit d'associer à une séquence linéaire monodimensionnelle d'unités lexicales une structure hiérarchique rendant compte des relations entre ces unités.

Paraphrase et réduction syntaxique

Il existe finalement une quatrième motivation à la mise en œuvre d'une analyse syntaxique, qui découle de la multiplicité des paraphrases possibles d'un même énoncé. Reprenons l'exemple du passif : on peut s'accorder en première approximation sur le fait que (a) et (b) dans :

(19) (a) *Jean casse la boîte*

(b) *La boîte est cassée par Jean*

ont le même sens. Il est donc souhaitable, avant de passer la main aux modules suivants, d'essayer de rendre en compte de cette identité en *normalisant* l'énoncé (b), de manière que le traitement sémantique ne soit confronté quà un nombre limité de structures différentes. Le même sens (à quelques nuances près) est aussi véhiculé par[5] :

(20) (a) *Jean, il casse la boîte*

(b) *La boîte, c'est Jean qui la casse*

(c) *La boîte, Jean, il la casse*

(d) ...

Dans le cas particulier du passif, une telle normalisation implique, d'identifier le sujet réel de la phrase au passif (*Jean*), ainsi que l'objet (ce qui est cassé, à savoir la *la boîte*), de manière à pouvoir reconstruire la structure syntaxique canonique.

Les arbres syntaxiques

Traditionnellement, le résultat de l'analyse syntaxique est représenté sous la forme d'un arbre, ce qui permet d'identifier simultanément les frontières de constituants, ainsi que les relations de dominance qu'ils entretiennent. On obtiendrait ainsi, dans le cas de l'exemple 8, une analyse similaire à celle figurée sur la figure 2.

[5] Molière a aussi donné de très bon exemples...

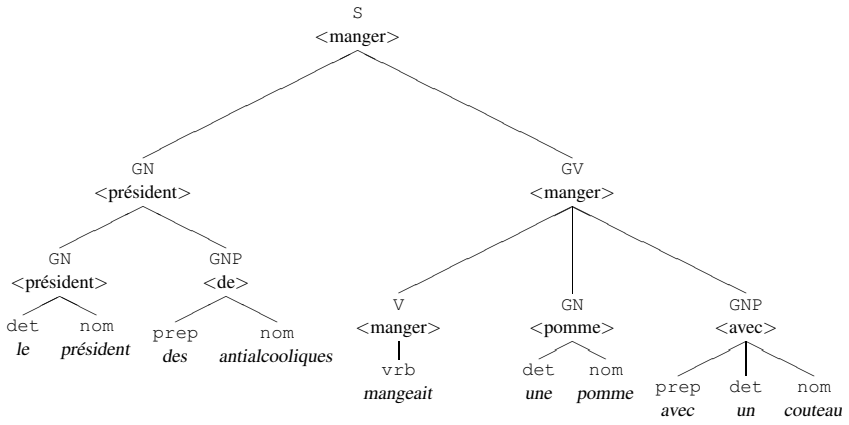

FIG. 2 – Représentation arborée de l'énoncé (8)

Au niveau le plus haut de l'arbre, on trouve un nœud étiqueté `S`, associé au concept `<manger>`. Ce nœud couvre toute la phrase, traduisant le fait que cet énoncé parle de l'action de manger. Au niveau juste inférieur, on trouve deux constituants principaux, l'un étiquetté `GN` (groupe nominal), correspondant au constituant *Le président des antialcooliques*, et associé au concept `<président>`, l'autre étiqueté `GV`, associé à `<manger>`. C'est donc un président qui mange. La lecture se poursuit selon le même schéma, en descendant récursivement les branches de l'arbre.

Quelques difficultés du traitement syntaxique

La conception d'analyseurs syntaxiques fiables et rapides est un problème redoutablement ardu. Le syntacticien est en effet confronté à une double contrainte : lutter contre la prolifération des ambiguïtés, tout en décrivant des phénomènes extrêmement complexes et subtils. Or, dans la pratique, ces deux contraintes sont largement contradictoires.

La réalité avec laquelle tout syntacticien doit composer d'emblée, c'est l'ambiguïté lexicale, qui fait que de très nombreuses formes graphiques correspondent à plusieurs entrées lexicales différentes, comme :
– *souris* : formes verbales de *sourir*, nom féminin singulier et pluriel ;
– *petit* : adjectif ou nom masculin singulier ;
– *la* : déterminant ou pronom personnel féminin singulier, nom masculin ;
– *mousse* : formes verbales de *mousser*, nom masculin, nom féminin ;
Si l'on se limite aux simples catégories syntaxiques de base, environ 50% des mots d'un texte sont ambigus, c'est-à-dire qu'ils correspondent possiblement à plusieurs catégories morpho-syntaxiques. Conséquence directe : une phrase de 20 mots a en moyenne 2^{10} interprétations différentes au niveau des étiquettes des feuilles l'arbre syntaxique !

Cette ambiguïté n'est pas seulement statique (lexicale), mais également dynamique (liée au contexte) : les phénomènes syntaxiques de *translation* rendent ainsi ambigus tous les adjectifs (emploi nominal), tous les participes passés, etc. Quelques exemples de translations :
– *chien* (emploi adjectival) *il est vraiment chien !*
– *vert* dans *mangez du vert !*, *affreux* dans *Les affreux ont encore tout cassé*
– *blessé* : *il est blessé* (v.s. *il a blessé*), *le blessé*, etc.
De surcroît, la description des phénomènes syntaxiques requiert bien souvent des descriptions lexicales bien plus précises que les simples étiquettes morpho-syntaxiques. Prenons l'exemple du verbe *parler*. Ce verbe a en fait quatre variantes syntaxiques, correspondant à des schémas verbaux (schémas de sous-catégorisation) différents. Ces quatre variantes sont attestées par les phrases suivantes :

(21) (a) *Jean parle*

(b) *Jean parle à Marie*

(c) *Jean parle de Paul*

(d) *Jean parle de Paul à Marie*

Ce comportement particulier de *parler* diffère très fortement de celui d'un verbe comme *courir*, pour lequel parmi les constructions précédentes, seul (a) est possible[6]. Cette propriété du verbe *parler* doit figurer dans la sa description lexicale. Sinon, la grammaire ne sera pas capable à la fois d'accepter la construction (c) pour *parler*, et de la refuser pour *courir*. La conséquence, malheureusement, est que pour chaque occurrence du verbe *parler*, il y aura systématiquement quatre interprétations possibles, correspondant aux quatre schémas ci-dessus. Cette ambiguïté est d'ailleurs parfois parfaitement justifiée, comme dans :

(22) *Je parle à la maîtresse de Marie*,

énoncé pour pour lequel il est impossible, au niveau syntaxique, de choisir entre l'usage du schéma (b) et celui du schéma (d).

En résumé, l'ambiguïté lexicale est donc largement sous-évaluée par le chiffre d'une forme ambigüe sur deux, et pose un réel problème aux analyseurs syntaxiques, qui ont souvent à envisager des dizaines, voire des centaines de milliers de structures ou sous-structures concurrentes.

L'ambiguïté lexicale est aggravée par les ambiguïtés purement syntaxiques, en particulier par les ambiguïtés de *rattachement*, dont (22) donne quelques figures prototypiques. Le problème est par exemple le suivant : un groupe nominal introduit par une préposition peut jouer le rôle de complément du nom comme celui de complément du verbe, ou encore de complément circonstanciel. À une même phrase peuvent correspondre donc plusieurs structures arborées différentes. Ce phénomène est illustré par les quelques exemples qui suivent :

(23) (a) *Elle mange une glace à la fraise* vs *Elle mange une glace à la plage* ;

(b) *J'ai été voir un film avec Marilyn Monroe*

(c) *C'est la fille du cousin qui boit* ;

(d) *Il a parlé de déjeuner avec Paul* ;

Les phénomènes syntaxiques à décrire sont souvent complexes, et demandent des descriptions lexicales et syntaxiques très fines, qui aggravent plus qu'ils ne résolvent le problème de l'ambiguïté. Ceci explique peut-être pourquoi il n'existe, à l'heure actuelle, et en dépit de 30 années de recherches intensives dans ce domaine, aucun analyseur de syntaxe complet pour aucune des langues naturelles. Il existe, par contre de nombreux lemmatiseurs, capables de désambiguïser un énoncé au niveau morpho-syntaxique. Il existe également des parenthéseurs, capables d'identifier grossièrement la structure des constituants, ainsi que des analyseurs plus complets, fonctionnant toutefois dans des domaines restreints, capables de découvrir les relations syntaxiques entre constituants.

Quelques pointeurs qui permettent de se faire une idée des capacités des systèmes d'analyse du français : l'analyseur et l'étiqueteur du français de l'université de Genève à l'adresse : `http://latl.unige.ch/latl/` ; celui de l'université de Caen : `http://www.info.unicaen.fr/~giguet/java/textes/index.html`.

0.2.5 Le niveau sémantique

Intuitivement, la sémantique se préoccupe du sens des énoncés. Une phrase comme

(24) *Le jardin de la porte mange le ciel*,

[6] *courir* admet naturellement d'autres constructions que n'admet pas *parler*, comme dans *il court après les filles*.

bien que grammaticalement parfaitement correcte, n'a pas de sens dans la plupart des contextes. Mais qu'est-ce que le sens ? Pour une expression référentielle comme *la bouteille de droite* dans la phrase

(25) *Sers-toi du vin. Non, pas celui-là, prends la bouteille de droite*,

le sens correspond à l'objet (au concept) désigné. Dans cet exemple, le sens dépend étroitement du contexte[7] : il faut une représentation de la scène pour savoir de quelle bouteille, et donc de quel vin, il s'agit.

Pour une expression prédicative, comme

(26) *Il commande un Margaux 1982*,

le sens peut être représenté par un prédicat logique comme `<demander(paul,chateau_margaux_82)>`. L'identification d'un tel prédicat dépend encore une fois du contexte. Le verbe *commander* aurait en effet renvoyé à un autre prédicat s'il s'était agi de commander un navire.

Les représentations conceptuelles

La définition que nous donnons ici de la sémantique est assez proche de celle d'un modèle en logique formelle. La sémantique, en logique, repose sur le choix d'un ensemble appelé domaine. À chaque constante intervenant dans les formules logiques, on associe un élément du domaine, et à chaque prédicat d'arité n on associe une relation n-aire sur ce domaine. Comprendre le sens d'un énoncé linguistique revient, en première approximation, à constituer une expression de type logique[8] qui renvoie à une relation entre des objets de la situation considérée. La construction du sens correspondant à la phrase se fait de proche en proche, à partir du sens trouvé pour les constituants. L'exemple ci-dessous permet de voir comment la phrase (émise au restaurant à l'intention du serveur)

(27) *Je voudrais le même vin que celui qu'ils boivent à la table de droite*

peut être traduite, étape par étape, par le prédicat `<commander(paul, margaux 82)>`. Vous noterez que la construction obtenue résulte non seulement des entrées linguistiques, mais aussi de *l'observation de la scène où se produit l'interaction*.

Une telle représentation strictement logique du sens peut sembler inadéquate et réductrice. Par exemple, *Il met le vin sur la table* est très mal représenté par `<mettre(paul, vin) & sur(vin, table)>`. Une solution consiste à utiliser des foncteurs, c'est-à-dire des symboles fonctionnels intervenant comme arguments des prédicats : `<mettre(paul, vin, sur(vin, table))>`. Cette méthode, cependant, distingue mal la contribution respective des foncteurs et des prédicats dans la construction du sens, et se révèle peu opératoire (essayer avec *il court autour du lac* ou *il montre en direction du sud*). Une autre solution consiste à faire reposer la représentation du sens non plus sur le prédicat, mais sur une combinatoire fonctionnelle très détaillée, ce qui donnerait quelque chose comme : `<événement(action(paul, cause(vin, aller(vin, chemin(?, sur(table))))))>`.

Grâce à cette combinatoire fonctionnelle, le sens est construit par l'insertion de la description conceptuelle de chaque mot dans le schéma conceptuel en cours de construction pour la phrase. Prenons un exemple :

(28) (a) *Luc a avoué ce vol a Guy* ;

(b) *Luc a attribué ce vol à Guy* ;

(c) *Luc a décrit ce vol à Guy* ;

[7] Historiquement, les sémanticiens ont essayé de se limiter à l'étude d'un sens littéral indépendant du contexte, celui que l'on essaie de donner dans les dictionnaires. Cette restriction est de peu d'intérêt en TAL : l'absence de contexte revient souvent, en fait, à se référer à un contexte prototypique qui suffit rarement à la construction du sens porté par l'expression linguistique écrite ou orale.

[8] La logique n'est pas de seul formalisme possible de représentation de ces relations conceptuelles, et d'autres modes de représentation sont également utilisés, en particulier les *graphes conceptuels*.

Je	?(paul)
veux	commander(paul, ?)
le même vin que	commander(paul, ?Vin)
celui qu'	commander(paul, ?Vin) & ?(?Vin)
ils	commander(paul, ?Vin) & ?(?Personnes, ?Vin)
boivent	commander(paul, ?Vin) & boire(?Personnes, ?Vin)
à la table	commander(paul, ?Vin) & boire(?Personnes, ?Vin) & situé(?Personnes, ?Table)
de droite	commander(paul, ?Vin) & boire(?Personnes, ?Vin) & situé(?Personnes, ?Table) & situé(?Table, à_droite)
(repérage visuel de la table)	commander(paul, ?Vin) & boire(?Personnes, ?Vin) & situé(?Personnes, table_1)
(repérage des convives)	commander(paul, ?Vin) & boire(personnes_1, ?Vin)
(repérage du vin)	commander(paul, margaux_82)

TAB. 1 – Construction d'une représentation conceptuelle

Dans cette description simplifiée, les ? désignent des variables ; ?Vin désigne une variable de type vin ; table_1 dénote un objet particulier de la scène, de type table ; l'emploi du même nom de variable en deux positions astreint les deux positions à recevoir la même valeur.

Ces trois énoncés sont syntaxiquement identiques (même structure, même forme passive, même forme pronominale et interrogative, ...) et pourtant, les représentations conceptuelles qu'on leur associe sont assez différentes : dans (a), c'est Luc qui a commis le vol, dans (b), c'est Guy (en principe), et enfin dans (c) Luc et Guy ne sont pas nécessairement les auteurs du vol. C'est la représentation *conceptuelle* des trois verbes, stockée dans le lexique, qui permet ces interprétations. Par exemple, le verbe *avouer* renvoie à une action qui porte sur une autre action, avec la contrainte que les deux actions ont le même acteur. Une telle représentation fonctionnelle permet de limiter l'ensemble des concepts de base, en faisant appel à des entités et des relations supposées universelles : événement, objet, action, localisation, cause, déplacement, etc. Un mot comme *boire* sera ainsi présenté dans le lexique comme une action, l'acteur causant le déplacement d'un objet de type liquide dans sa gorge. Sur le plan technique, de telles représentations peuvent être stockées dans des *réseaux sémantiques* ou des *schémas* (frames) de manière à tirer partie des possibilités d'héritage au sein de hiérarchies conceptuelles (le concept de vin, s'il est correctement déclaré, peut ainsi hériter des propriétés de liquide, d'aliment, et indirectement de matière). Les hiérarchies conceptuelles, telles que Wordnet (consultable à la page :hhtp://www.infres.enst.fr/~yvon/Projets/Wapi), sont donc des outils importants pour le traitement sémantique.

La représentation conceptuelle permet de veiller au respect des compatibilités de type, au moment de l'insertion de la définition conceptuelle d'un mot dans le schéma de la phrase. Par exemple, le foncteur `<cause>` exige deux arguments : un patient de type objet, et un événement. Dans l'exemple du vin mis sur la table, le concept `<aller>` produit bien un événement qui peut jouer le rôle d'effet. Ces principes permettent de répondre à l'un des principaux défis posés à toute théorie sémantique : respecter les contraintes qui lient une expression linguistique avec son sens. Par exemple, la phrase (24) viole de nombreuses contraintes de nature conceptuelle : si "de" désigne une appartenance, comment un jardin peut-il appartenir à une porte ou en faire partie ? Comment un être inanimé comme un jardin peut-il manger ? etc. En revanche, une phrase comme *le président boit du vin* est sémantiquement admissible, car la construction `<événement(action(président, boire(président, vin)))>` respecte les contraintes de type, dans la mesure où l'on suppose que le concept de président hérite des propriétés d'un être humain, et que le concept de vin hérite des propriétés d'un liquide et d'un aliment. C'est également grâce à ces contraintes de type que l'on peut assigner correctement les référents du pronom *le* dans les deux continuations de la phrase suivante : *le bus a écrasé un passant...*

(28) (a) (...) *je l'ai entendu freiner*

(b) (...) *je l'ai entendu crier*

le réfère au bus dans (a) et au passant dans (b) car, dans les contextes typiques, les bus ne crient pas plus que les passants ne freinent. Grâce à l'utilisation d'un dictionnaire et du contexte d'énonciation, on peut ainsi reconstituer le sens de certains énoncés sous la forme d'une expression conceptuelle. Toutefois, les glissements de sens (ainsi cette dernière expression où l'on prête à une entité abstraite comme le sens la faculté de glisser), omniprésents dans l'usage linguistique, restent très difficiles à analyser avec les techniques actuelles.

Les limites des représentations fonctionnelles

La traduction fonctionnelle, même sous la forme conceptuelle que nous avons décrite, reste réductrice. Elle semble impuissante à représenter les nuances (*il mit furtivement le vin sur la table*) ou le temps (*juste avant l'arrivée de Paul*). Les extensions de la logique « classique » (logiques modales, logiques temporelles) qui ont été proposées pour remédier à ces limitations sont complexes et difficilement utilisables. Cet aspect réducteur de la représentation logique a conduit de nombreux sémanticiens à insister sur les aspects graduels du sens. Certains ont proposé de représenter le sens dans des espaces topologiques, le voisinage rendant compte de la proximité sémantique. D'autres ont montré que la compréhension des énoncés reposait pour une large part sur l'emploi de métaphores, souvent spatiales (comme l'emploi du verbe reposer dans la phrase précédente).

Ces deux points de vue sur la sémantique, le point de vue logico-conceptuel et le point de vue gradualiste, sont difficilement conciliables. En revanche, ils peuvent être combinés avec profit dans une démarche algorithmique. La phrase *il mit furtivement le vin sur la table* peut ainsi recevoir une représentation conceptuelle du type de celle que nous avons donnée, jointe à une représentation imagée qui la situe dans le temps et qui qualifie le déroulement de l'action décrite. Cette dernière représentation suppose que soient modélisés l'axe temporel et la scène à laquelle la phrase fait référence (la salle de restaurant, la position des tables, etc.).

L'interface syntaxe-sémantique

La détermination de la structure syntaxique est essentielle pour la construction du sens. Le principe de base est que les syntagmes se projettent sur les constituants conceptuels, que nous avons représentés par des foncteurs. Autrement dit, dans *le président boit du vin*, *président* ou *le président* portent un sens, alors que *président boit* ou *boit du* ne sont pas signifiants. Concrètement, la reconnaissance d'un syntagme peut déclencher son interprétation immédiate, même si cette interprétation est provisoire (par exemple si elle contient des variables liées ou des variables demandant une instanciation visuelle, comme c'était le cas pour "le même vin que ..."). Qu'en est-il des ambiguïtés, si fréquentes lorsqu'on réalise une analyse syntaxique ? Sont-elles seulement syntaxiques ou sont-elles en réalité d'origine sémantique ? En général, le lien syntaxe-sémantique est si étroit que deux sens différents correspondent à des structures syntaxiques différentes. L'ambiguïté de la phrase

(29) *il poursuit la jeune fille à vélo*

est sémantique (qui est sur le vélo ?), mais aussi syntaxique (c.f. la page 4). Cependant, dans un contexte concret, cette phrase a peu de chances d'être sémantiquement ambiguë. En effet, les contraintes sémantiques vont vraisemblablement bloquer toute ambiguïté. L'article défini de *la jeune fille* pose comme contrainte que l'objet désigné par cette expression puisse être identifié de manière univoque. Si l'on sait qu'il y a trois jeunes filles dans l'histoire, on s'attend à ce que *à vélo* serve à la détermination. C'est donc la deuxième interprétation qui est la bonne. En revanche, une entité ne peut pas être surdéterminée (i.e. déterminée de plusieurs manières). Si l'on remplace *la jeune fille* par un nom propre :

(30) *il poursuit Christelle à vélo*,

le complément *à vélo* ne peut plus être une détermination du groupe nominal *Christelle*, et doit donc être rattaché au verbe.

Dans certains cas également, l'ambiguïté potentielle n'apparaît pas au niveau syntaxique. Par exemple, la phrase

(31) *c'était un vin d'origine inconnue*

est ambiguë. Soit l'origine du vin n'était pas précisée (absence d'étiquette), soit elle était précisée (e.g. *Château Laforge du Plantier, 23750 Proumillac*) mais ne correspondait pas à une région connue. Nous sommes en présence d'une ambiguïté purement sémantique. Elle sera éventuellement levée par le contexte, au niveau sémantique ou pragmatique.

0.2.6 Le niveau pragmatique

Le niveau pragmatique est parfaitement dissociable du niveau sémantique. Alors que la sémantique se préoccupe du sens des énoncés, la pragmatique porte sur les attitudes (vérité, désirabilité, probabilité) que les locuteurs adoptent vis à vis des énoncés et sur les opérations logiques que ces attitudes déclenchent. Historiquement, certains linguistes ont appelé pragmatique tout traitement du langage faisant intervenir le contexte d'énonciation. Ce critère présente fort peu d'intérêt, dans la mesure où les processus sémantiques sont les mêmes, que le contexte intervienne ou non. En revanche, il existe une distinction très importante, basée sur la notion d'inférence logique. Considérons l'exemple suivant :

(32) (a) *Pierre : viendras-tu au bal ce soir ?*

(b) *Marie : j'ai entendu que Paul y sera !*

La seconde phrase sera interprétée comme une réponse négative si l'on sait que Marie n'aime pas Paul. Cette interprétation n'est pas de nature sémantique. À partir de la compréhension du sens de l'intervention de Marie, Pierre réalise une inférence logique en utilisant une connaissance contextuelle, l'inimitié entre Paul et Marie. Pierre conclut que Marie ne veut pas aller au bal, autrement dit il reconstruit l'attitude de Marie par rapport à son propre énoncé. Cette opération n'est pas une construction conceptuelle, c'est une opération logique. Elle appartient donc à la pragmatique.

Le niveau pragmatique possède ses propres lois. De même qu'au niveau sémantique, on s'attend à ce que tout énoncé ait un sens (i.e. permette une construction conceptuelle), on s'attend au niveau pragmatique à ce que tout énoncé soit *pertinent*. C'est cette contrainte de pertinence qui nous permet de prêter à Marie et à Pierre l'inférence conduisant au refus d'aller au bal. En d'autres termes, nous nous attendons à ce que la mention de la présence de Paul soit logiquement connectée à l'intention de Marie d'aller ou non au bal. Certaines ambiguïtés sont de nature purement pragmatique. Par exemple, le sens de "il pleut " n'est généralement pas ambigu. En revanche, l'attitude qu'on veut communiquer est ambiguë. Veut-on signifier qu'il faut se désoler (on ne pourra pas aller se promener) ou se rejouir (on n'aura pas besoin d'arroser la pelouse) ? Seul le contexte pragmatique et la capacité à faire des inférences permettront à l'auditeur de lever cette ambiguïté. La pragmatique correspond au niveau argumentatif du langage. Prenons l'exemple classique de la personne disant

(33) *il fait plutôt froid ici*

pour demander en fait que son interlocuteur se lève pour fermer la fenêtre. Supposons que la connaissance d'arrière-plan inclue la relation causale :

$$\text{froid\_dehors} \,\&\, \text{fenetre\_ouverte} \;\rightarrow\; \text{froid\_dedans}$$

ainsi que le caractère indésirable de ce dernier état de fait :

$$\text{froid\_dedans} \;\rightarrow\; \text{INDESIRABLE}$$

Selon certaines théories pragmatiques, la pertinence de l'intervention "il fait plutôt froid ici" est liée à son effet cognitif, qui réside ici dans le caractère indésirable de l'état évoqué. Mais l'esprit de l'auditeur n'en reste pas là, et le locuteur le sait. Il ne peut s'empêcher d'envisager ce qui peut rendre la situation moins indésirable. En l'occurrence, des deux termes qui peuvent empêcher l'enchaînement causal, seul le terme <fenêtre_ouverte> peut être rendu faux. La conclusion qui consiste à envisager de fermer la porte est donc logiquement inférée par l'auditeur. Ce dernier, qui prête au locuteur des capacités équivalentes aux siennes, comprend qu'il s'agissait là de l'intention ayant motivé l'intervention de son vis-à-vis[9].

Les techniques pour représenter le rôle des arguments font appel à la logique et à la planification. Elles sont utilisées principalement pour la gestion de dialogues « finalisés », c'est-à-dire orientés vers la résolution d'une tâche, par exemple la réservation d'un billet d'avion. Les programmes de gestion de dialogue fonctionnent à partir d'une représentation de la tâche en termes de buts, à la manière des techniques de résolution de problème ou de planification. A chaque instant, on tient à jour les connaissances et les intentions prêtées à l'interlocuteur. Chaque reconnaissance d'une entrée linguistique est interprétée pragmatiquement comme interférant avec le plan en cours.

Le niveau pragmatique est aussi invoqué à un niveau plus élevé, celui de l'organisation de larges tronçons de textes ou de discours. Il s'agit alors de repérer les relations rhétoriques et structurelles entre les passages. Les techniques correspondant à ce niveau de traitement sont encore très mal maîtrisées. Le niveau pragmatique, même si les techniques qui lui correspondent ne sont pas encore stabilisées, apparaît moins difficile à aborder que le niveau sémantique. Il semble en effet qu'il repose sur un ensemble de principes fixes, comme le principe de pertinence, qu'il s'agit de modéliser correctement. La détermination de l'intention argumentative de l'auteur ou du locuteur est essentielle dans bon nombre d'applications, notamment la gestion de dialogue, le résumé de texte, la traduction automatique, les systèmes d'aide contextuelle ou d'enseignement, etc. On attend donc des progrès significatifs à ce niveau dans les années qui viennent.

0.3 Les applications du TALN

Concernant les applications, la demande de TALN provient, pour dire vite, de deux tendances « lourdes » : d'une part la nécessité de concevoir des interfaces de plus en plus ergonomiques, d'autre part la nécessité de pouvoir traiter (produire, lire, rechercher, classer, analyser, traduire) de manière de plus en plus « intelligente » les informations disponibles sous forme textuelle, de manière à pouvoir résister à leur prolifération exponentielle. Les applications des techniques de TAL sont donc nombreuses et variées. Nous avons regroupé ces applications en trois grandes familles, qui correspondent aux aides à la lecture de documents, aux aides à la production de documents, et enfin aux interfaces homme-machines. Ces familles d'applications sont détaillées dans les paragraphes qui suivent.

0.3.1 Le traitement documentaire

Les applications les plus immédiates du TALN sont celles qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel, comme par exemple :

– La traduction automatique (ou l'aide à la traduction automatique). Cette application, qui a historiquement suscités les premiers efforts de recherche en TALN, reste un enjeu économique et politique[10] de première importance. Si de tels traducteurs existaient, il serait sans doute beaucoup moins crucial de recourir, pour

[9] Les contraintes liées à la pertinence permettent de nombreuses prévisions. Par exemple, le premier locuteur aurait obtenu le même effet en disant *il fait froid dehors*. De même, la théorie permet d'analyser la non-pertinence comique de la réplique de Pierre Dac : ce n'est pas parce que je fermerai la fenêtre qu'il fera moins froid dehors.

[10] Pensez au nombre de traductions, et de traducteurs, nécessaires au fonctionnement quotidien de l'Union Européenne, qui reconnaît aujourd'hui plus de 20 langues officielles !

assurer une large diffusion à des documents, à une langue véhiculaire « universelle » telle que l'anglais ; Il est important de noter que même si la traduction complète indépendante du domaine est encore hors d'atteinte, on peut obtenir de bons traducteurs spécialisés (domaines techniques), qui constituent un moyen efficace de préparer l'intervention manuelle d'un traducteur. Il existe également des environnements de travail fournissant des ressources lexicales (dictionnaires bilingues étendus) pour l'aide à la traduction. Enfin, il existe une application intéressante des systèmes complètement automatiques de traduction, qui se fonde sur le fait que la recherche documentaire, ou le filtrage manuel de document, qui peut nécessiter une lecture diagonale d'un grand nombre de documents, se fera toujours mieux dans la langue natale de l'utilisateur...

De nombreux systèmes commerciaux sont aujourd'hui disponibles (Systran, Logos, Metal, Aleth-Trad...), et certains moteurs de recherche proposent une traduction automatique des pages html. Une démonstration relativement récente des capacités actuelles des systèmes de traduction commerciaux est disponible à l'adresse : http://www.systransoft.com/trans.html.

– La recherche de documents « intéressants » dans des bases documentaires. La prolifération des outils de recherche documentaire sur la toile, qui traitent quotidiennement des millions de requêtes, montrent bien l'importance de la demande en la matière. Les performances de ces moteurs témoignent du chemin qu'il reste à parcourir dans ce domaine. Si Google semble aujourd'hui sortir du lot, d'autres moteurs de recherche valent certainement la peine d'être connus : ainsi http://www.metacrawler.com, http://www.search.com, ou encore www.exalead.com.

  De plus en plus d'outils fournissent également des outils de recherche spontanée d'adresses potentiellement intéressantes (à partir de profil utilisateurs), ou encore de surveillance automatique des publications dans des domaines donnés.
– Le routage, classement ou l'indexation automatique de documents électroniques sont des variantes applicatives du paradigme de la recherche documentaire.
– Plus complexe est la tâche de trouver (ou de produire à la demande) des réponses précises aux questions de l'utilisateur (tâche de "question-réponse"). Là encore, des solutions techniques existent : voir par exemple http://faqfinder.ics.uci.edu.
– La lecture automatisée de documents, par exemple pour les stocker dans des structures formelles de données, ou pour en extraire des résumés ;
– L'analyse d'un corpus de documents relatifs à un thème donné (histoire, stylométrie, veille technologique, etc). Une application typique de ce domaine consiste à fournir des outils de visualisation et d'expolration dynamique de champs disciplinaires (scientifiques, par exemple). Vous trouverez une démonstration d'un tel système à l'adresse http://websom.hut.fi/websom/.

Pour ce qui concerne les outils d'indexation et de recherche documentaire tels que ceux qui sont disponibles sur la toile, les techniques actuellement mises en œuvre sont essentiellement statistiques, et ne font appel, pour des raisons évidentes de coût de traitement, qu'aux outils de « bas-niveau » du TAL : segmentation et lemmatisation. La lemmatisation consiste à retrouver le lemme d'une forme fléchie, c'est-à-dire à lui retirer son ou ses suffixes flexionnels, c'est donc une forme très simplifiée d'analyse morphologique. Notons toutefois que les systèmes d'indexation les plus récents intègrent des mécanismes plus complexes, principalement :

– des thésauri décrivant des relations entre concepts, comme l'hyponymie (le concept X est une instance particulière de Y), l'hyperonymie (le concept X est une généralisation de Y), de synonymie (le concept X est équivalent au concept Y), ou encore d'antonymie (le concept X est opposé au concept Y). À l'aide de ces réseaux de relation, il est possible d'augmenter la portée d'une requête d'information, en étendant par exemple les termes de la requête par les termes synonymes. Vous pouvez consulter un de ces thésaurus (Wordnet), à l'adresse : http://www.infres.enst.fr/~yvon/Projets/Wapi ;
– des mécanismes de reconnaissance des mots composés. Ces mécanismes mettent en œuvre des grammaires régulières locales, qui vont détecter toutes les occurences des patrons typiques de production des mots-composé. Ainsi en Français, ces grammaires détecteront les séquences *N Adj* (e.g. *langage naturel*), *N PREP N* (e.g. *réseau de neurones*, *machine à écrire*)...

0.3.2 La production de documents

Si autant de documents électroniques sont aujourd'hui disponibles, c'est bien que quelqu'un les a écrit. Dans le domaine de l'aide à la production de texte (la génération de textes), les applications du TALN sont également nombreuses :

– les claviers « auto-correcteurs » (par exemple pour les handicapés) ;
– la reconnaissance optique de caractères. De nombreux systèmes commerciaux sont aujourd'hui disponibles, avec des performances très satisfaisantes : Recognita, Omnipage, ScanWorX... ;
– les correcteurs d'orthographe ou de syntaxe. De tels correcteurs sont aujourd'hui disponibles dans la majorité des systèmes de traitement de texte commerciaux, avec des performances variables suivant les mécanismes de correction mis en œuvre, qui vont de la recherche lexicale tolérante à l'analyse syntaxique partielle ou complète de la phrase. Une démonstration des capacités de correction en ligne est accessible à l'adresse : http://www.spellonline.com.
– les correcteurs « stylistiques », ou les aides intelligentes à la rédaction intégrant des thésaurus, des connaissances sur les « bonnes » pratiques rédactionnelles, etc.
– l'apprentissage assisté par ordinateur des langues naturelles ;
– la génération automatique de documents à partir de spécifications formelles. En fait, de nombreux secteurs d'activité impliquent la production massive de textes très stéréotypés à partir de spécifications plus ou moins formelles (textes juridiques, compte-rendu d'exploration d'une base de donnée, rapports d'analyses statistiques, documentations techniques, etc). Pour cette classe de documents, il est parfaitement possible de générer automatiquement, sinon des textes complètement définitifs, du moins des versions préliminaires qui seront ensuite finalisés par des rédacteurs humains.

On retrouve dans ces applications la même dialectique que dans les applications destinées à faciliter la gestion de documents. D'un côté des applications à large couverture, qui utilisent essentiellement des ressources lexicales, avec des fonction d'accès tolérant (permettant la correction d'erreurs) au lexique : c'est le cas des applications qui tournent autour de la correction orthographique. De l'autre, des applications qui intègrent des mécanismes de traitement de plus haut niveau (typiquement la génération), mais qui ne fonctionnent que pour des domaines beaucoup plus restreints.

0.3.3 Les interfaces naturelles

Dernier domaine d'application, qui est sans doute celui dans lequel la demande de traitements linguistiques est la plus forte, le domaine des interfaces naturelles (i.e. en langage naturel) telles que :

– l'interrogation en langage naturel de bases de données (traduction langage naturel $\leftrightarrow$ SQL) ou de moteurs de recherche sur la toile. De multiples applications de ce type commencent à se mettre en place sur la toile. Consultez par exemple : http://infolab.cs.uchicago.edu/faqfinder/, qui, grâce à une recherche intelligente dans les « Foires Aux Questions » (FAQs) des forums de discussion, fournit un véritable petit Quid en ligne ;
– les interfaces vocales, qui mettent en œuvre de manièNre variable suivant les applications des modules de reconnaissance de parole, synthèse de parole, génération et gestion de dialogue, accès aux bases de connaissance,..., chacun de ces modules demandant des traitements spécifiques (désambiguïsation morpho-syntaxique et identification de syntagmes pour la synthèse, grammaires stochastiques pour la reconnaissance de la parole...).
  Les premiers systèmes « grand-public » de dictée vocale commencent à véritablement arriver sur le marché (Via Voice, le produit d'IBM, Dragon Dictate, et bien d'autres encore), et l'intégration dans Windows d'une API de traitement de la parole devrait, dans les années qui viennent, faire littéralement exploser le marché des technologies vocales. De nombreux services commerciaux existent déjà, ou sont proches de la commercialisation qui font appel à l'ensemble de ces techniques (reconnaissance-dialogue-synthèse), pour des applications diverses telles que les ordinateurs « main-libre », la lecture téléphonique de courriers

électroniques, la réservation de billets (train, avion)... la liste est pratiquement sans limite. Le marché des technologies vocales représentait en 1995 quelques centaines de millions de dollars pour les États-Unis, et est estimé, pour l'an 2 000, à près de 3 Milliards de dollars.
De nombreuses démonstration de la qualité de la parole de synthèse sont disponibles sur la toile, en particulier à l'adresse : http://www.ldc.upenn.edu/cgi-bin/ltts/list.
Pour ce qui concerne la reconnaissance, IBM par exemple propose un démonstrateur, téléchargeable à l'adresse http://www.software.ibm.com/is/voicetype/demo_98.html.
Pour enfin voir des agents dialoguants, faites donc un tour sur la page des "bots" à l'adresse suivante : http://bots.internet.com.

0.4 Conclusion

L'étude du langage naturel et des mécanismes nécessaires à la mise en œuvre à son traitement automatique par des machines est un domaine d'études foisonnant, et riche en applications potentielles ou émergentes. De nombreux progrès restent à accomplir pour mieux comprendre cette faculté et pour bâtir des systèmes capables de soutenir la comparaison avec l'humain, mais l'état des connaissances en permet aujourd'hui de proposer de nombreuses solutions efficaces à des problèmes et des demandes réels.

Une des limitations de pratiquement tous les systèmes de traitement un peu sophistiqués est qu'ils font appel à une somme importante de connaissance d'expert : lexiques, règles de grammaires, réseaux sémantiques... Ceci explique en partie pourquoi il n'existe pas de système de traitement qui soit à la fois complet (i.e. intégrant tous les niveaux de traitement) et indépendant du domaine (i.e. capable de traiter avec une même efficacité n'importe quel type de texte). Il existe une autre raison, moins visible, qui limite l'avancée des progrès en TALN, et qui est que, pour un bon nombre de phénomènes, l' état de la connaissance linguistique est insuffisament formalisée pour pouvoir être utilisée par les concepteurs de systèmes de TALN.

En réaction, de nombreux développement récents font appel à l'utilisation massive de corpus ou de lexiques (mono ou multilingues) pour extraire directement ces connaissances à l'aide de techniques statistiques. Ainsi, des analyseurs morphologiques, des étiqueteurs, des grammaires, des systèmes de traduction basés sur l'exploitation de corpus et bien d'autres outils encore ont été développés dans les années passées, avec des résultats plus qu'encourageants. Les avantages de ces outils sont (en théorie) multiples :
– ils sont indépendants (au moins partiellement) de la langue, c'est-à-dire qu'ils peuvent être utilisés pour extraire des connaissances relatives à des langues différentes (suffisamment voisines toutefois) ;
– ils sont facilement portables d'une application à l'autre, et l'adaptation à un nouveau domaine est facilitée ;
– d'un point de vue plus théorique, ils permettent, parce qu'ils sont fondés sur l'examen d'échantillons réels de langue, d'intégrer directement dans les modèles linguistiques tous les phénomènes liés à la performance.
L'utilisation de techniques d'apprentissage automatique et d'acquisition de connaissance (comme les modèles de Markov, les réseaux de neurones, les arbres de décision, les algorithmes génétiques, etc) est donc aujourd'hui une tendance importante en TALN, qui concentre les efforts de nombreuses équipes de recherche.

Bibliographie

Alfred Aho, Ravi Sethi, and Jeffey Ullman. *Compilateurs, principes et outils*. Traduction Française (1989), Editions Interalliés, Paris, 1972.

René Carré, Jean-Francois Dégremont, Maurice Gross, Jean-Marie Pierrel, and Gérard Sabah. *Langage Humain et Machine*. Presses du CNRS, 1991.

Ralph Grishman. *Computational Linguistics : An Introduction*. Cambridge Universtiy Press, 1986.

John E. Hopcroft and Jeffrey D. Ullman. *Introduction to automata theory, languages and computation*. Addison-Wesley, 1979.

Gérard Ligauzat. *Représentation des connaissances et linguistique*. Armand Colin, Paris, 1994.

Jean-Marie Pierrel. *Ingéniérie des Langues*. Hermès, Paris, 2000.

Gérard Sabah. *L'intelligence artificielle et le langage, vol. 1*. Hermès, Paris, 1988.

Gérard Sabah. *L'intelligence artificielle et le langage, vol. 2*. Hermès, Paris, 1989.