Initiation au TAL : Devoir 7

1. Testez le logiciel "Cordial" (accessible uniquement depuis les postes du LaboC) avec un corpus de votre choix. Quels sont les avantages et les désavantages ?

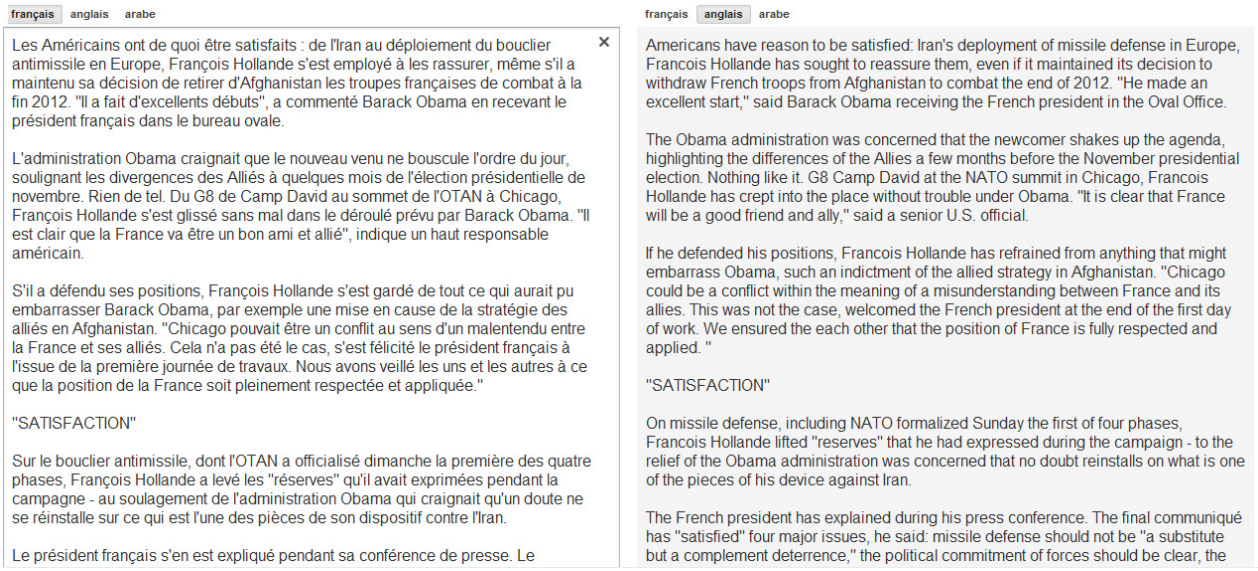
Avantage :

- Permet la correction orthographique, grammaticale et syntaxique, dispose d'un dictionnaire orthographique, et d'un dictionnaire de définitions de plus d'un million de mots.
- Il permet la correction grammaticale de 4 langues.
- Propose des outils de correction typographique et stylistique.

Inconvénient :

- Se limite a Windows.

2. Testez les deux outils de Traduction automatique avec un corpus de votre choix: Google Translat et Systran, expliquez de manière très brève pourquoi parle-t-on d' "un schéma général Fichier".



français anglais arabe

Les Américains ont de quoi être satisfaits : de l'Iran au déploiement du bouclier antimissile en Europe, François Hollande s'est employé à les rassurer, même s'il a maintenu sa décision de retirer d'Afghanistan les troupes françaises de combat à la fin 2012. "Il a fait d'excellents débuts", a commenté Barack Obama en recevant le président français dans le bureau ovale.

L'administration Obama craignait que le nouveau venu ne bouscule l'ordre du jour, soulignant les divergences des Alliés à quelques mois de l'élection présidentielle de novembre. Rien de tel. Du G8 de Camp David au sommet de l'OTAN à Chicago, François Hollande s'est glissé sans mal dans le déroulé prévu par Barack Obama. "Il est clair que la France va être un bon ami et allié", indique un haut responsable américain.

S'il a défendu ses positions, François Hollande s'est gardé de tout ce qui aurait pu embarrasser Barack Obama, par exemple une mise en cause de la stratégie des alliés en Afghanistan. "Chicago pouvait être un conflit au sens d'un malentendu entre la France et ses alliés. Cela n'a pas été le cas, s'est félicité le président français à l'issue de la première journée de travaux. Nous avons veillé les uns et les autres à ce que la position de la France soit pleinement respectée et appliquée."

"SATISFACTION"

Sur le bouclier antimissile, dont l'OTAN a officialisé dimanche la première des quatre phases, François Hollande a levé les "réserves" qu'il avait exprimées pendant la campagne - au soulagement de l'administration Obama qui craignait qu'un doute ne se réinstalle sur ce qui est l'une des pièces de son dispositif contre l'Iran.

Le président français s'en est expliqué pendant sa conférence de presse. Le

français anglais arabe

Americans have reason to be satisfied: Iran's deployment of missile defense in Europe, Francois Hollande has sought to reassure them, even if it maintained its decision to withdraw French troops from Afghanistan to combat the end of 2012. "He made an excellent start," said Barack Obama receiving the French president in the Oval Office.

The Obama administration was concerned that the newcomer shakes up the agenda, highlighting the differences of the Allies a few months before the November presidential election. Nothing like it. G8 Camp David at the NATO summit in Chicago, Francois Hollande has crept into the place without trouble under Obama. "It is clear that France will be a good friend and ally," said a senior U.S. official.

If he defended his positions, Francois Hollande has refrained from anything that might embarrass Obama, such an indictment of the allied strategy in Afghanistan. "Chicago could be a conflict within the meaning of a misunderstanding between France and its allies. This was not the case, welcomed the French president at the end of the first day of work. We ensured the each other that the position of France is fully respected and applied. "

"SATISFACTION"

On missile defense, including NATO formalized Sunday the first of four phases, Francois Hollande lifted "reserves" that he had expressed during the campaign - to the relief of the Obama administration was concerned that no doubt reinstalls on what is one of the pieces of his device against Iran.

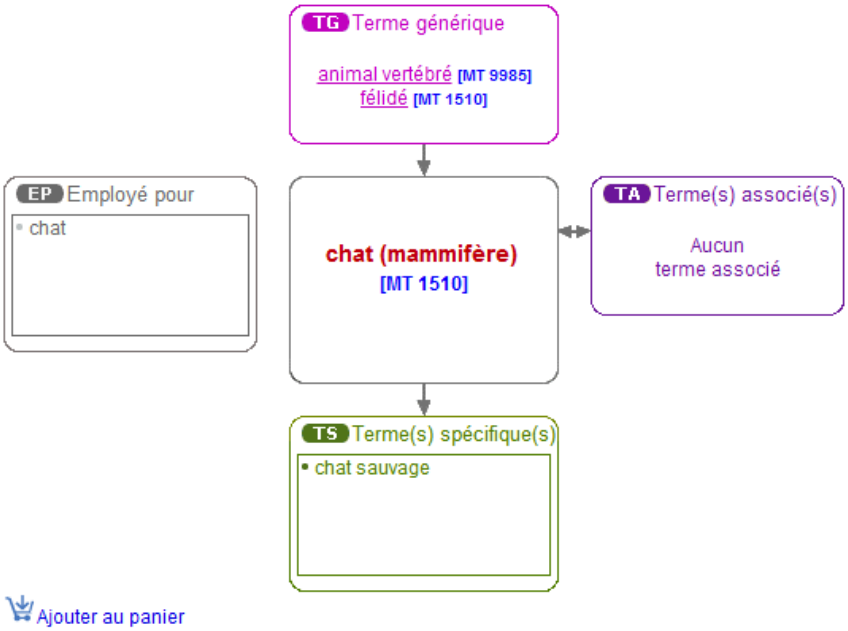
The French president has explained during his press conference. The final communiqué has "satisfied" four major issues, he said: missile defense should not be "a substitute but a complement deterrence," the political commitment of forces should be clear, the

On parle d'un un « schéma général Fichier » car il y a un enchaînement de modules traitant chacun un domaine reconnu de la linguistique :

Phrases a traiter → reconnaissance des mots → phrase étiquetée → analyse syntaxique → arbre syntaxique → analyse sémantique → représentation du sens → traitement du sens → action de la machine.

Pour ce qui est de la traduction, on représente le sens de la phrase ou du texte à traduire dans une sorte de métalangage, ou de langage-pivot, à partir duquel la génération d'une phrase ou d'un texte en n'importe quelle langue serait possible.

3. Testez MOTBIS, Qu'est-ce c'est un thésaurus? Comment ça marche ?



Un thésaurus, est une liste organisée de termes représentant les conceptsd'un domaine de la connaissance.

C'est un langage contrôlé utilisé pour l'indexation et la recherche de ressources documentaires dans des applications informatiques spécialisées. Les thésaurus sont donc une catégorie de langages documentaires parmi d'autres. Les termes (dans l'exemple ci-contre : *véhicule*, *navire*,…) sont reliés entre eux par des relations de synonymie (terme équivalent), de hiérarchie (terme générique et terme spécifique) et d'association (terme associé) ; chaque terme appartient à une catégorie ou domaine.

Le thésaurus est un outil linguistique qui permet de mettre en relation le langage naturel des utilisateurs et celui contenu dans les ressources. Cette technique pallie les limites du langage naturel, très riche mais aussi souvent ambigu. Le thésaurus évite ainsi les risques induits par les synonymies, les homonymies et les polysémies présentes dans le langage naturel.

Contrairement à undictionnaire auquel il est souvent rapproché, un thésaurus ne fournit qu'accessoirement des définitions, les relations des termes et leur sélection l'emportant sur la description des significations.

Il établit des relations entre les concepts: relation hiérarchique, relation d'association, d'appartenance à un groupe de concepts

4. Rédiger un compte rendu de lecture sur le document "Une petite introduction au Traitement Automatique des Langues Naturelles Fichier".

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 àproduire et à comprendre des énoncés linguistiques dans des buts de communication. Il sera donc question ici de langage humain.

Le langage naturel, dont le traitement automatique pose des difficultés majeures. Précision importante, nous nous limiterons au traitement du langage sous forme écrite,

Pourquoi s'intéresser à l'automatisation du traitement du langage naturel ? 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 ; 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.

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.

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 àl'analyse de la syntaxe des phrases. Nous n'aborderons que marginalement les questions liés à la comprehénsion et à l'interprétation, etpratiquement pas les question liées à la production (génération) automatique de documents.

- ➔ Clarifier quelques concepts linguistiques, en étudiant les différents niveaux de représentation et de traitementdes énoncés linguistiques
- ➔ 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). 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 concernentalors 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 fortemais la viande est pourrie) !

Ce qui ressort de cet exemple, c'est que de nombreuses connaissances contextuelles et encyclopédiques sont nécessaires pour trouverla traduction correcte d'un mot.

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. Un groupe d'experts 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 **le langage est une faculté à la fois universelle et spé-cifique à l'espèce humaine**. **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 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.**

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 avecun robot, dans le cadre d'un micro-monde montre que les diverses sources de connaissances doivent et peuvent interagir avec les modules d'analyse et de raisonnement.

Les années 70 voient ensuite le développement d'approches surtout sémantiques 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. **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 développements et des raffinements continus, et des algorithmes de plus en plus performants sont proposés pour analyser les grammaires les plus simples. Depuis Chomsky, ces formalismes grammaticaux sont toutefois considérés comme trop simples pour modéliser correctement les phénomènes observés dans les langues naturelles. Ces développements des grammaires formelles sont donc largement sous-estimés, jusqu'à ce qu'au milieu des années 70, divers travaux théoriquesréhabilitent ces formalismes dans le cadre du traitement de la morphologie et de la phonologie des langues naturelles. Ce sont tout d'abord les réseaux de transition augmentés 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, qui commencent à atteindre le grand public, sont là pour témoigner de l'importance des avancées accomplies mais également des progrès qu'il reste 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. Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, exemples :

- 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 marqueà 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 à un ensemble d'actions conventionnelles avec pour visée finale d'ingérer de la nourriture et à l'action consistant à effectivement ingérer un type particulier de nourriture
- ambiguïté de la fonction grammaticale des groupes de mots
- ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions.
- ambiguïté sur l'interprétation à donner en contexte à un énoncé.

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 en général et sur le contexte de l'énonciation (connaissance dynamique). Dès lors, en effet, que l'on restreint le cadre des textes analysés à un sous domaine particulier, il devient possible d'une part d'ignorer un grand nombre d'ambiguïtés, en particulier sémantiques 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, 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.

0.2.1 Introduction

Considérons à titre d'exemple l'énoncé :

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 ouune 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 étéproduit : **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 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 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, dans les notations du temps (il a couru le cent mètre en 12'3.), etc.

Un bon segmenteur se doit donc de bien recenser tous les usages possibles dans la langue écrite des signes de ponctuation. Ces informations renseignent manifestement sur la structure et le contenu du document, de 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. 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).

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. 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 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

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
- 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 qu'en termes de signification du dérivé construit.
- 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).
- 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.
Vous noterez finalement que la connaissance de la structure morphologique se révèle dans bien des casindispensable 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.
(voir décomposition arborescente)

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.

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 ».

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.

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. 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.

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.

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'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
- les contraintes qui portent sur les déplacements de constituants dans des transformations telles que la formation du passif à partir de l'actif, ou la construction d'interrogatives.
- 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.

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'ils remplissent.

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.

(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. (→ cf arbre syntaxique de : le président des antialcooliques mangeait une pomme avec un couteau)

Au niveau le plus haut de l'arbre, on trouve un nœud étiqueté S, associé au concept <manger>. 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>. La lecture se poursuit selon le même schéma, en descendant récursivement les branches de l'arbre.

Quelques difficultés du traitement syntaxique

Le syntacticien est 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. 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 : 4 variantes syntaxiques, correspondant à des schémas verbaux (schémas de sous-catégorisation) différents. Ces quatre variantes sont attestées par les phrases suivantes :

13(21) (a) Jean parle

(b) Jean parle à Marie

(c) Jean parle de Paul

(d) Jean parle de Paul à Marie

Pour chaque occurrence du verbe parler, il y aura systématiquement quatre interprétations possibles, correspondant aux quatre schémas ci-dessus. Cette ambiguïté est d'ailleurs

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, 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.

Les phénomènes syntaxiques à décrire sont souvent complexes, et demandent des descriptions lexicaleset 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.

0.2.5 Le niveau sémantique

Intuitivement, la sémantique se préoccupe du sens des énoncés.

14bien que grammaticalement parfaitement correcte, n'a pas de sens dans la plupart des contextes. Mais qu'estce 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 : 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 « 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.

<u>Les représentations conceptuelles</u>

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. Comprendre le sens d'un énoncé linguistique revient, en première approximation, à constituer une expression de type logique 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 (émiseau restaurant à l'intention du serveur)

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

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 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

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.

L'ambiguïté de la phrase : il poursuit la jeune fille à vélo

est sémantique (qui est sur le vélo ?), mais aussi syntaxique. 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.

Dans certains cas également, l'ambiguïté potentielle n'apparaît pas au niveau syntaxique. Par exemple, la phrase « 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 mais ne correspondait pas à une région connue. Nous ommes 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 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 uneopération logique. Elle appartient donc à la pragmatique.

La pragmatique correspond au niveau argumentatif du langage. Prenons l'exemple classique de la personne disant « 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 :

froid_dehors & fenetre_ouverte → froid_dedans

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

froid_dedans → 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

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. 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. 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. 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 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.

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 de première importance.

- 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.
- 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").
- 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).
- 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
- 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.

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.

– les correcteurs « stylistiques », ou les aides intelligentes à la rédaction intégrant des thésaurus, desconnaissances 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..

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 ↔ SQL) ou de moteurs de recherche sur la toile. De multiples applications de ce type commencent à se mettre en place sur la toile

– 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...).

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 une des limitations de pratiquement tous les systèmes de traitement un peu sophistiqués 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 et indépendant du domaine. 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.

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, ils peuvent être utilisés pour extraire des connaissances relatives à des langues différentes
- 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 est donc aujourd'hui une tendance importante en TALN, qui concentre les efforts de nombreuses équipes de recherche.