Compte rendu de lecture :

Introduction au TALN et à l'ingénierie linguistique, I.Tellier
(Chapitres 3, 4 et 5)

Les langues naturelles sont avant tout orales (de nombreuses langues ne sont pas écrites), et, de ce fait, elles se définissent par des propriétés acoustiques. Celles-ci ne seront abordées dans ce livre que brièvement, car ce dernier s'intéresse principalement au traitement des données écrites.

Les aspects oraux du langage peuvent être traités dans deux domaines distincts.
Le premier, la phonétique, donne lieu à une description physique des sons des langues naturelles. On peut les enregistrer notamment à l'aide d'oscillogramme et de spectrogramme pour mieux les étudier, ce que fait la sous branche de la phonétique acoustique.
La phonétique articulatoire, elle, étudie les sons du point de vue articulatoire. Les sons sont décrits à l'aide de traits articulatoires renvoyant aux organes de productions humains du langage, puis ils sont représentés à l'aide de symbole dans des alphabets phonétique (le plus utilisé étant l'Alphabet Phonétique International).
Le deuxième domaine à traiter des aspects oraux du langage est la phonologie. Elle regroupe les sons en classe et étudie leur fonctionnement dans la langue. Elle s'intéresse principalement aux phonèmes (sons distinctifs dans une langue donnée, qui peuvent former une paire minimale, c'est-à-dire deux mots de sens différents qui ne diffèrent que par un son, mais aussi aux sons de classe équivalente, c'est-à-dire les sons qui peuvent être remplacés l'un par l'autre dans un même mot sans que son sens ne change).

On peut également se pencher sur d'autres aspects acoustiques du langage, qui ne rentrent pas forcément dans les études de ces deux domaines. L'accent tonique, par exemple (« une augmentation de hauteur ou d'intensité de la voix lors de la prononciation d'une syllabe »), ou l'étude des langues à tons (où un même son peut changer de sens suivant le ton avec lequel il est prononcé), ou encore la prosodie (« règles de prononciation globales qui influent sur la mélodie d'un énoncé »).

Les applications de TAL concernant le signal sonore peuvent être diverses. Il peut s'agir de reconnaissance vocale (on part de l'oral pour aller vers l'écrit, en faisant une analyse du signal sonore) ou d'une production orale (une lecture) à partir d'un texte écrit (il s'agit alors davantage de synthèse).
En ce qui concerne la reconnaissance vocale, plusieurs difficultés se posent : la prise en compte des bruits environnants, l'adaptation aux voix des locuteurs malgré les variations interindividuelles (les systèmes mono locuteurs demandent une préparation préalable afin de saisir les spécificités de la voix du locuteur), et la difficulté de segmenter le flux continu de parole.
De plus, lorsque l'on passe à l'écrit, il faut ensuite choisir la bonne orthographe, mettre en place un traitement spécialisé pour ce qui relève du symbole (ponctuation, nombre, abréviations…), mais également tenir compte de la syntaxe pour que le texte soit cohérent.
Lorsqu'il s'agit de passer de l'écrit à l'oral, il faut arriver à gérer le problème des mots homographes hétérophones (ces mots qui s'écrivent pareil mais se prononcent différemment), ainsi que celui des règles de prononciation, de la ponctuation, mais également de la prosodie…

Différentes techniques ont été tentées pour établir des systèmes de reconnaissance vocale : un système travaillant par règles d'identification des phonèmes, un autre grâce à des bibliothèques d'exemples de prononciation de suite de phonèmes. Mais le plus efficace est celui qui utilise des modèles statistiques : après avoir inventorié des suites de phonèmes, on fait un calcul statistique sur leur fréquence d'apparition dans la langue, puis une probabilité de leur fréquence d'apparition suivant le contexte qui précède. Cependant, cela nécessite la présence de corpus importants.

A ce jour, des systèmes performants existent en reconnaissance vocale mais qui dépendent beaucoup de l'environnement et des conditions d'utilisation. Le chapitre finit par une proposition de quelques sites de synthèse vocale.

Nous passons maintenant à la morphologie (l'étude des sons combinés entre eux qui donnent un sens).

L'auteur commence d'abord par essayer de définir la notion de mot. A cause des contre exemple de l'apostrophe, du point dans les dates ou initiales, du tiret, ou même des mots composés séparés par des blancs, les séparateurs ne sont pas un critère fiable pour définir un mot.
Un mot ne peut pas non plus être défini comme la plus petite unité de sens possible : en effet, il existe des unités plus petites (comme la marque du pluriel). Ainsi, la notion de mot est peu utilisée par les linguistes, qui préfèrent celle de morphème : « unité linguistique minimale ayant une forme et un sens », objet d'étude de la morphologie.

Il y a différents types de morphèmes : les morphèmes lexicaux (qui ont généralement un sens, appartiennent à une liste ouverte et extensible), parfois appelés lexèmes, et renvoyant généralement aux noms, verbes, et adjectifs. Il y a ensuite les morphèmes grammaticaux (qui n'ont pas de sens, ont essentiellement un rôle grammatical, appartiennent à une liste fermée et non extensible). Il s'agit alors ici davantage de s déterminants, propositions, conjonctions et éventuellement auxiliaires. Cependant, cette classification est discutable (dans le cas des pronoms, ou du morphème « dans » classé comme grammaticale alors qu'ayant un sens bien précis), et insuffisante en ce qui concerne la gestion des entités nommées (catégorie importante pour le TAL), désignant les noms propres et variables numériques.

La morphologie est la discipline qui étudie la combinaison des morphèmes entre eux pour formes des « unités lexicales ». Il y a deux façon d'opérer cette combinaison : par la composition (simple concaténation de morphèmes lexicaux) et l'affixation (qui fait interagir des morphèmes lexicaux et grammaticaux entre eux pour créer des unités lexicales). Dans ce dernier phénomène, les morphèmes grammaticaux sont désignés comme des affixes et ne peuvent pas être autonomes. Ces affixes peuvent être soit dérivationnels (avec un contenu plus ou moins lexical) ou flexionnels (servant à exprimer des flexions d'un verbe, des variations de genre ou de nombre…). On peut avoir plusieurs affixations successives pour arriver au résultat final d'une unité lexicale. Les noms propres peuvent eux aussi être sujets à de telles variations.

Chaque unité lexicale peut avoir plusieurs propriétés : tout d'abord, elle possède une forme lemmatisée, ou lemme, qui est celle qui apparaît dans les dictionnaires. Chaque unité lexicale peut également être définie selon sa catégorie grammaticale (l'appartenance à une catégorie étant définie par des propriétés morphologiques et syntaxiques). Enfin, les unités lexicales portent des informations de flexion lorsqu'elles sont présentes dans un énoncé.

Même si, à l'heure actuelle, il est possible de stocker toutes les informations concernant les flexions de mots dans une mémoire d'ordinateur, il semble plus efficace de stocker celles-ci de manière synthétiques. Ceci est alors possible de deux manières : soit en faisant une « structure de données » (donc en encodant des informations sous forme de listes) soit grâce à des « automates finis » (qui établissent des règles morphologiques).

Pour stocker des listes de mots de façon économique, on utilise parfois des « arbres à lettres », c'est-à-dire une représentation de plusieurs mots sous forme d'arbre. Cela permet certes de stocker plusieurs mots ensembles de manière organisée, mais cela ne rend pas compte de la structuration des mots en morphèmes.
Les automates finis, quant à eux, sont des modèles informatiques qui se traduisent en programme et permettent de passer d'un état initial d'un mot à un état final (avec de possibles états intermédiaires) à l'aide de transitions modélisés en fonctions (le tout étant appelé chemin). On voit alors ici clairement apparaitre les différents morphèmes, ou plus exactement les mots et leurs différentes variantes morphologiques, en un nombre fini : les automates utilisent donc un langage fini. Ils sont de plus relativement bien adapté aux phénomènes d'affixations, et par extension aux phénomènes de conjugaison des verbes.

Parmi les langages reconnus par les automates, il y a en particulier ce qu'on appelle les « langages réguliers ou rationnels ». Ce sont les langages qui peuvent être exactement reconnus par un automate. Ces langages peuvent également être exprimés à l'aide d'expressions régulières.

Encore une fois à la fin de ce chapitre, on trouve quelques sites internet utiles, ainsi qu'une mention rapide d'Unitex (programme gratuit et libre qui fonctionne à l'aide d'automates).

Nous arrivons maintenant au niveau de la syntaxe. Il existe plusieurs analyses.

Commençons par l'analyse distributionnelle. A partir d'un corpus conséquent dans une langue donnée, on observe la distribution d'une unité (c'est-à-dire ses contextes d'apparition) et on définit des classes distributionnelles (des mots qui partagent les mêmes environnements). A partir de ce système, on peut alors définir une grammaire. Malheureusement, le critère de corpus n'était pas satisfaisant pour Chomsky (car en nombre fini, ce que ne sont pas les possibilités d'une langue).

Chomsky présente alors un nouveau critère pour classer des unités : la notion de grammaticalité. La grammaire est donc ce qui est capable de formuler des jugements de grammaticalité (si cela est bien formé ou non). Cette aptitude caractérise la compétence d'un locuteur.

Avant d'aller plus loin, il faut se poser la question de l'emploi du mot « phrase ». ce n'est pas la seule entité à être composée de plusieurs mots : prenons par exemple, les chunks (qui renvoient aux groupe de mots comme les groupes nominaux, verbaux…), les termes (des groupes nominaux éventuellement composés d'autres groupes nominaux) et les clauses (séquence de mots contenant au moins un sujet et un prédicat). La structure intermédiaire principale entre le mot et la phrase est en fait le syntagme (« mot ou suite de mot auquel on peut associer une catégorie syntaxique). Il ne faut pas confondre non plus phrase et énoncé, notion souvent utilisée par les linguistes et renvoyant à une unité de production textuelle ou de prise de parole. Ici, on va s'intéresser plus particulièrement aux propositions (suite d'unité disant des choses sur le monde et auquel on peut attribuer une valeur de vérité).

Une analyse syntaxique équivaut à « découper une suite d'unités en plusieurs sous groupes adjacents qui vont ensemble ». Lorsqu'on met en pratique ceci sur une phrase, on peut obtenir plusieurs catégories différentes : déterminant (précèdent et introduisent les noms communs), nom, verbe, préposition… A l'aide de substitution, on peut alors délimiter les différents syntagmes (en remplaçant un groupe de mots par un mot unique, comme un groupe nominal par un nom propre). On peut représenter l'analyse syntaxique obtenue de différentes manières : soit sous forme de phrase délimitée à l'aide de parenthèses, soit sous forme d'arbre.

Il peut être difficile d'associer une structure syntaxique à une proposition, comme c'est notamment le cas pour les phrases ambigües (par exemple « l'homme observe sa voisine avec des jumelles »). Dans ce cas, il est possible de faire deux représentations syntaxiques distinctes, selon le sens que l'on choisit. De plus, certains mots appartiennent à plusieurs catégories grammaticales, ce qui ne facilite pas leur classement.

La représentation sous forme arborescente n'explique pas à elle toute seule la structuration interne d'un syntagme (composée d'une tête, à laquelle sont rattachés un spécifieur et un complément). De plus, elle ne permet pas non plus de représenter certains phénomènes tels que l'ellipse, l'apposition ou la thématisation.

Le domaine de la modélisation de la syntaxe est celui qui a le plus évolué au cours des dernières années dans le domaine du TAL. Ici, c'est la théorie des langages qui va principalement être abordée, c'est-à-dire la théorie qui traite des grammaires et des langages en général.

La difficulté, lorsqu'on veut construire une grammaire en informatique, est que non seulement il n'existe pas de corpus de mots finis, mais qu'en plus même si cela était le cas, la possibilité de construire des phrases ou non grammaticales serait elle encore infinie. Le modèle des automates finis, déjà évoqué plus haut, pourrait correspondre à ce qu'on attend d'une grammaire.

Cependant, la compétence linguistique que possèdent les être humains est difficilement stockable dans des automates. De plus, ces derniers sont incapables de produire des structures arborescentes reflétant une structure syntaxique à partir d'une phrase. Enfin, les automates auraient des difficultés à traiter des répétitions telles que celles produites par des propositions relatives enchâssées successives.

Les réseaux de transition récursifs (RTRs) sont une généralisation des automates finis qui répond aux limites évoquées précédemment. Il s'agit d'un ensemble d'automates qui possèdent plus ou moins les mêmes propriétés que les automates finis (un état initial et final, ainsi que de transitions). Un automate s'occupe alors de traiter ce qui relève des GN, un autre des GV. Les différents automates ont la possibilité de s'appeler entre eux. Ils rendent également compte des structures arborescentes. Les RTRs sont donc plus puissants que les automates finis, même si aujourd'hui, ils ne sont plus l'outil principalement utilisé.

Une grammaire formelle sont des grammaires composées : d'un vocabulaire comprenant une suite finie de mots (comprenant l'ensemble des mots composant la suite dont on veut tester la grammaticalité), un vocabulaire non fini (pour les étapes intermédiaires de l'analyse), et d'un ensemble « règles de production » ou « règles de réécriture »).

On remarquera que n'importe quel automate peut être transformé « en une grammaire équivalente, c'est-à-dire reconnaissant exactement le même langage »).

Selon la hiérarchie de Chomsky, il existe plusieurs classes de grammaires formelles, qui sont de différentes nature et possèdent différentes propriétés. On a les grammaires rationnelles ou régulières, qui peuvent être transformées par un automate. On a également les grammaires hors-contexte, qui peuvent être transformées en RTR, ou encore les grammaires sensibles au contexte, en enfin, les grammaires formelles en général. Toutes ces grammaires sont emboitées les unes dans les autres (la grammaire formelle étant la plus large et la grammaire rationnelle la plus restreinte). On a donc une hiérarchie dans les grammaires qui permettent de classer les différents langages.

En ce qui concerne la hiérarchie de Chomsky appliquée aux langues naturelles, cela est assez difficile à déterminer : cependant, le consensus actuel s'attache aujourd'hui à l'idée que « les langues naturelles sont légèrement sensibles au contexte ».

Les grammaires formelles ont donné naissance à de nombreuses autres grammaires : les grammaires de constituants, les grammaires de dépendance, et les grammaires minimalistes.

Enfin, l'auteur propose de découvrir les grammaires formelles grâce à quelques sites et programmes.