INITIATION AU TAL

Devoir 4

I. Qu'est-ce *WordNet* ? Quels sont les avantages et les désavantages ?

WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour. WordNet est distribué sous une licence libre, permettant de l'utiliser commercialement ou à des fins de recherche.

Avantages :

- Disponible en téléchargement gratuit
- Offre une grande couverture lexicale
- Les noms sont classés en un système de catégories complet et précis comprenant plusieurs niveaux d'imbrication.
- Sa richesse et sa précision en font un outil de choix.

Désavantages :

- Distinctions de sens très (trop) fines, sans méthodologie précise pour les découper, sans repérer des processus lexicaux type métonymie, métaphores…
- Système de classification peu élaboré pour les verbes par rapport aux noms, qui sont organisés en un système hiérarchique « plat » avec moins de niveaux d'imbrication, où on passe très rapidement d'un concept spécialisé à un concept très général.
- Aucune catégorisation hiérarchique définie pour les embranchements des adjectifs et des adverbes.

II. Rédiger un compte rendu de lecture sur le livre de Mme Isabelle Tellier (Chapitre 3, 4 et 5)

Chapitre 3 : Aspects acoustiques du langage

Les langues naturelles sont avant tout orales, beaucoup n'ont pas de transcription écrite. Il est naturel que de nombreuses propriétés de ces langues découlent de considérations acoustiques. Mais certains documents écrits peuvent être la transcription de données orales. Nous aborderons les aspects oraux du langage, dans le but d'introduire les concepts fondamentaux du domaine, et d'illustrer certains modes de raisonnements linguistiques ou certains modèles informatiques qui seront repris dans la suite du texte.

I. Description linguistique

1.1 Phonétique

La phonétique est la branche de la linguistique qui étudie les sons des langues naturelles, indépendamment de leur sens. Pour caractériser les sons émis par deshumains qui parlent, on peut partir de leurs descriptions physiques, telles que les mesurent des dispositifs électro-acoustiques comme les oscillogrammes et les spectrogrammes.
La phonétique acoustique étudie les propriétés de diagrammes tels que l'oscillogramme et le spectrogramme.
La phonétique articulatoire va, elle, s'attacher à étudier les sons élémentaires d'une langue via la configuration physiologique nécessaire pour les produire. Ainsi, chaque émission vocale peut être décrite par un ensemble de traits articulatoires caractérisant la position des organes intervenant dans la prononciation (langue, gorge, glotte, nez et lèvres).
Exemple de description articulatoire :
[t] : consonne occlusive sourde dentale non nasale
Elle peut être paraphrasée comme suit :

- consonne : un son qui est produit grâce à un obstacle
- occlusive : l'obstacle doit être total
- sourde : ne pas générer de vibrations
- dentale : on l'obtient en mettant la langue sur les dents
- non nasale : aucun air ne passe dans le nez

Toutes les propriétés doivent être réalisées en même temps pour que le son soit correctement produit. Cette description ne tient pas compte des différences physiologiques pouvant exister entre deux individus dont les émissions varient en timbre, en hauteur et en intensité. Des alphabets phonétiquesont ainsi été définis, dans lesquels chaque symbole correspond à un ensemble de traits articulatoires. Le plus utilise est l'API (Alphabet Phonétique International), il existe depuis 1888 et est censé transcrire les productions orales de n'importe quelle langue naturelle.

1.2 Phonologie

Chaque langue opère une sélection dans la liste des sons que la physiologie rend possibles. Et, parmi les sons « adoptés », elle opère un regroupement en classes d'équivalences. Deux sons élémentaires appartiennent à des classes différentes s'il est possible de trouver deux unités lexicales différentes qui ne diffèrent d'un point de vue acoustique que par ces deux sons. On appelle cela une « paire minimale ». En revanche, ils sont équivalentss'il est impossible de trouver deux unités lexicales différentes qui ne diffèrent que par ces deux sons.
Exemples:

- « zona » et « sauna » sont des mots différents en français, et ils ne diffèrent quepar leur son initial. Ces deux sons ne sont donc pas équivalents en français.Ils le sont pourtant en espagnol, où ils ne permettent de faire la distinction d'aucun couple de mots.
- « cote » et « côte » sont des mots différents et ils ne diffèrent que par leur voyellecentrale : ces deux sons ne sont pas équivalents en français.
- en français, le fait de « rouler les r », de les « grasseyer » ou de les prononcer « normalement » ne permet de distinguer aucun mot. Cestrois sons sont donc équivalents en français. En espagnol, en revanche, on peutdistinguer « pero » (mais) et « perro » (chien) par le fait de rouler ou non le son« r ». Ces deux sons ne sont pas équivalents dans cette langue.

Deux sons sont ainsi équivalents si, en remplaçant l'un par l'autre, le mot prononcé reste le même. La relation d´équivalence considérée est celle de substituabilité en préservant l'identité de l'unité de niveau supérieur. Une classe d'équivalence de sons élémentaires distinctifs dans une langue donnée est appelée un phonème. Chaque phonème est donc spécifique d'une langue donnée.
La phonologie est la branche de la linguistique qui étudie les propriétés des phonèmes. La transcription phonologique d'un énonce peut se coder en utilisant certains des symboles de l'API mis entre barres obliques.
Exemple :
Le phonème correspondant à toutes les prononciations possibles de "r" en français est note /r/.
Le français comprend environ 33 phonèmes : 20 consonnes et 13 voyelles.

1.3 Autres aspects acoustiques

Dans certaines langues, comme l'anglais, chaque mot reçoit un accent tonique qui se traduit par une augmentation de l'intensitéde la voix lors de la prononciation de la syllabe sur laquelle il est localisé.
Dans les « langues à tons », comme le chinois et la plupart des langues du monde, le même son, suivant la tonalitéavec laquelle il est prononcé, peut changer le sens du mot dont il fait partie. Ces propriétés sont de nature discrète et peuvent être reproduites à l'écrit par des symboles spécifiques. On admet en principe qu'elles ne concernent pas la langue française.
On désigne par prosodie, les règles de prononciation globales qui influent sur la mélodie d'un énonce. En français, suivant l'intonation qu'on y met, « tu viensdemain » peut devenir une affirmation, un ordre ou une question sans que le sens des mots présents ne change pour autant. La prosodie est de nature continue.

II. Modélisation informatique

La modélisation de l'oral n'est pas l'objectif principal. Le domaine se rattache au « traitement du signal », branche de l'électronique et de l'automatique plus que de l'informatique, car elle traite de données continues.
Certaines des techniques utilisées pour manipuler des données orales sont employées à d'autres fins.

2.1 Domaines et problèmes

Il existe de nombreuses applications, y compris « grand public », au traitement de la langue en tant que « signal sonore ».

- En analyse : la reconnaissance vocale peut servir à identifier un locuteur par sa voix, à identifier la langue qu'il parle, à reconnaıtre l'ordre qu'il donne ou à transcrire automatiquement, sous forme écrite, ce qu'il dit. Dans ce dernier cas, on passe généralement par deux étapes successives : d'abord une transcription du son en une séquence de phonèmes, puis en un texte écrit.

- En synthèse : il s'agit de produire une lecture orale à partir d'un texte écrit. Cela revient à transformer un texte en une succession de phonèmes, puis en une émission acoustique.

En analyse, de nombreuses difficultés doivent être surmontées : tout d'abord, il faut distinguer le son de la voix des bruits environnants et adapter le système au timbre et à la hauteur de voix du locuteur. Les systèmes mono locuteurs, c'est à-dire destines à n'être utilisés que par une seule personne, nécessitent en général une phase d'apprentissage au cours de laquelle il est demandéà cette personne de lire un texte standard.
Ces derniers ne doivent pas pour autant être trop rigidement fixes, pour éviter que le système échoue en cas de modification momentanée de la voix du locuteur comme lorsqu'il est enrhumé par exemple. Quant aux systèmes multi locuteurs, ils doivent être capables de s'adapter aux variations interindividuelles.
La difficulté majeure à affronter est la segmentation du flux continu de paroles en unités discrètes. Un même phonème peut être prononce de façonstrèsdifférentes, suivant son voisinage avec les autres phonèmes (un /a/ en début ou en fin de mot ne se prononce pas du tout de la mêmefaçon), et certains phonèmes ont tendance à être « avalés » dans une prononciation courante.
Pour passer d'une séquence de phonèmesà un texte écrit, il faut opérer des regroupements, et retrouver parmi des homophones, celui qui doit est utilisé dans la transcription.

Pour que le texte écrit final soit correct, il faut tenir compte des accords, et donc de la syntaxe.
La synthèse vocale pose, moins de difficultés, mais elle doit surmonter quelques pièges. L'un d'eux provient des homographes hétérophonies (mots s'écrivant pareil mais qui ne se prononcent pas pareil).

2.2 Outils formels ou statistiques utilisés

Pour programmer un système de reconnaissance ou de synthèse vocale, de nombreuses techniques ont ététestées. A une certaine époque, on a tentéde traduire sous la forme de « bases de connaissances » façon « systèmes experts » lesrègles d'identification des phonèmes. Une règle typique d'un système de ce genre serait de la forme :
SI le spectrogramme montre une "barre d'explotion"ET est immédiatement suivi d'une « voyelle arrière-droite »ALORS la consonne correspondante est une « dentale »(c'est-à-dire /d/ ou /t/).
Les méthodes les plus performantes font appel à des techniques avancées en traitement du signal et à des modèles statistiques. Un exemple de modèlestatistique élémentaire est un « n-gramme » (succession de n éléments).
Exemple : si on prendn=2, on obtient des bi-grammes.
Certains couples comme /de/ ou /du/ seront trèsfréquents, d'autres très rares voire toujours absents (comme /pz/). Avec un n-gramme, on fait en général l'hypothèse que le n ième élément d'une suite ne dépend que des n-1 ièmes qui le précédent.
On trouve dans le commerce des systèmes de reconnaissance vocale trèsefficaces. Leurs performances dépendentbeaucoup de l'environnement (plus ou moins bruyant) et des conditions d'utilisation : un texte lu devant un micro, parlé aux téléphones ou capté lors d'un dialogue ne sera pas du tout reconnu de la mêmefaçon. Les meilleurs systèmes sont capables de reconnaitre un flux de parole continu en faisant moins de 5% de fautes.

2.3 Sites Web

Plusieurs sites réalisent des synthèses vocales en ligne de différentes langues, en laissant l'utilisateur fixer un certain nombre de paramètres (timbre, sexe, hauteur de la voix), et parfois en lui permettant de visualiser le résultat d'analyses intermédiaires réalisées par le programme.

Chapitre 4 Morphèmes, morphologie

En combinant des sons élémentaires qui ne veulent rien dire,on finit par réussirà « dire » quelque chose, à signifier. Il y a là un saut qualitatif considérable qui justifie qu'on lui associe un niveau d'analyse spécifique. Ce niveau correspond à la notion de « mot ». La linguistique, préfèrele terme de « morphème ».

1 Description linguistique

1.1 Problèmes avec la notion de « mot »

Un mot estselon lecritèreformel est ce qui, dans un texte, est compris entre deux séparateurs.

- l'apostrophe est, la plupart du temps, la marque d'une séparation entre deux mots : « j'ai », « l'arbre », « d'un »"... sont bien constitués de deux « mots ». Pourtant,« aujourd'hui » et « prud'homme », malgré l'apostrophe qu'ils contiennent,ne sont généralementconsidérés que comme un seul mot.
- le point est apparemment un séparateur pour isoler les phrases.
- le tiret est un séparateur plus problématique : rien ne nous autoriseàne trouver qu'un seul mot dans « porte-monnaie »ou « entre-déchirer »,et à en trouver plusieurs dans « (cet) homme-là », « est-ce-que », « voulez-vous » ?
- le caractère blanc n'est pas un séparateur fiable : on aurait envie defaire de « parce que » ou de « pomme de terre » un seul mot, tandis que « du » et« au », résultenteux d'un amalgame(« du » pour « de le », « au » pour « à le »),en font plutôt deux à eux tous seuls.
- faut-il considérer que « chat » et « chats » sont les mêmes mots ? Le "s" ne porte-t-il pas une nuance de sens spécifique ?
- le sens d'un mot unique comme « inimitables » semble décomposable enunités de sens élémentaires, au point que si le verbe « troufigner » avait un jour un sens, on en déduirait celui de « introufignable ».

Saussure préfère la notion de « signe » tandis que Martinet parle de « monème ». La linguistique contemporaine utilise le terme de « morphème » : unité linguistique minimale ayant une forme et un sens. L'étude des morphèmes et de leurs modes de combinaison est l'objet de la morphologie.

On peut considérer que « pomme de terre », « parce que » ou même « casser sa pipe » ne constituent chacun qu'un seul morphème. On peut même admettre l'existence de morphèmes discontinus à l'intérieur desquels peuvent s'insérer d'autres morphèmes :

Exemple : En français, la marque de la négation« je ne dort pas » ou du passécomposé « j'ai bien dormi ».

Le découpage en morphèmes d'un énonce peut s'avérerproblématique :

Exemple : les mots de « pomme de terre » forment, la plupart du temps, qu'un seul morphème, dans une phrase comme « Pour les protéger, il a recouvert les pommes de terre », ils sont à prendre comme morphèmes distincts.

1.2 Les différents types de morphèmes

On repartit les morphèmes en deux groupes : les **morphèmes lexicaux** et les **morphèmes grammaticaux**. Pour les distinguer, voici les critères suivants :

- **critère sémantique** : les morphèmes lexicaux ont la particularité « d'avoir un sens par eux-mêmes », de « référerà quelque chose dans le monde ». On range dans un premier groupe les noms communs, les verbes et les adjectifs. Le deuxième groupe englobe, les mots qui ont un rôle grammatical, syntaxique comme les déterminants, des prépositions, des conjonctions, ainsi que des auxiliaires (« être » et « avoir ») et des morphèmes qui caractérisent le genre et le nombre des noms, les temps de conjugaison et les personnes des verbes. Mais le critère sémantique est trop vague, il ne comprend pas par exemple les pronoms et les adverbes.
- **Critère énumératif** : les morphèmes grammaticaux appartiennent à une listefermée, tandis que les morphèmes lexicaux appartiennent àune liste ouverte, pouvant évoluer dans le temps. Quand on joue àinventer de nouveaux mots, ce sont toujours des morphèmes lexicaux (cf. le« troufigner »). Cela nous permetde classer les pronoms parmi les morphèmes grammaticaux, et les adverbes avec les morphèmes lexicaux.

Dans certaines traditions linguistiques, on utilise le terme « lexème » pour désigner« morphème lexical », et « morphème » pour désigner « morphèmegrammatical ».

1.3 Combinaisons de morphèmes

La morphologie étudie comment différents morphèmes se combinent entre eux pour former des unités plus complexes qu'on appelle unités lexicales, car elles vérifient les critères associés aux morphèmes lexicaux. On distingue deux façons d'opérer des combinaisons : la composition et l'affixation.

Exemple:

- bon + homme = bonhomme

On peut étendre ce mécanisme à certains groupes comme "pomme de terre" ou "laisser tomber". Ce qui justifie que ces derniers se comportent comme un seul morphème. L'affixation est un mécanisme qui fait interagir des morphèmes lexicaux, sous la forme de « racines », et des morphèmes grammaticaux, sous la forme des « affixes ».

Les affixes sont des morphèmes non autonomes, ils ne peuvent exister sans être rattachés à une racine. Ils sont de deux types distincts :

- les affixes **dérivationnels** ont un contenu quasi-lexical : ce sont soit des préfixescomme « de », « re », ... soit des suffixes comme « -eur», « -ement »…
- les affixes **flexionnels** servent à exprimer les variations en genreet en nombre des noms et des adjectifs, et les conjugaisons des verbes. Ils se positionnent en français après les affixes dérivationnels.

Une même unité lexicale peut être construite à l'aide de plusieurs affixations successives. Les suffixes ont la particularité de pouvoir changer la catégorie grammaticale de la racine à laquelle ils s'associent.

1.4 Les informations associées à une unité lexicale

Nous assimilons la notion d'unité lexicale à celle de« mot ». On associe à chaque unité lexicale présente dans un énoncé des propriétés.
Tout d'abord, on associe à chaque unité lexicale une forme lemmatisée: c'est la forme sous laquelle on va trouver cette unité dans un dictionnaire de la langue courante. Elle correspond à un choix arbitraire parmi les flexions possibles que peut subir l'unité en question. En français, la forme lemmatisée (ou lemme) des noms et des adjectifs est celle du masculin singulier, tandis que celle des verbes est l'infinitif. Les autres unités sont invariables et se confondent donc avec leur lemme.
On ne parle pas en général de forme lemmatisée pour les entités nommées (mais elles sont souvent aussi invariables).
Les unités appartiennent à des catégories grammaticales : certaines sont des noms communs, d'autres des verbes ou des adverbes... Les anglo-saxons parlent d'étiquettes « part-of-speech » (ou POS).
Depuis Chomsky, une grammaire est un dispositif capable de trier les suites de mots d'une langue donnée en « grammaticales »ou « non grammaticales », et cela indépendamment de leur sens. Cette capacitéà formuler des jugements de grammaticalité est ce qui constitue la compétence d'un locuteur. Si on admet l'existence d'une telle capacité, alors on dira que deux unités lexicales appartiennent à la même catégorie si on peut remplacer l'une par l'autre dans n'importe quel énoncé, sans modifier sa grammaticalité.
Exemple :

- les mots « livre » et « rhinocéros » appartiennent à la même catégorie car« le livre est sur l'étagère » et « le rhinocéros est sur l'étagère » sont tous les deux des énoncés grammaticaux.
- « livre » et « regarder » n'appartiennent pas à la même catégorie car « le regarde est sur l'étagère » n'est pas grammatical.

On peut dire que les catégories grammaticales sont des classes d'équivalence pour la relation de substituabilitéen préservant la grammaticalité de l'unité de niveau supérieur (l'énoncé).
Les accords en genre et en nombre imposés par la langue française font que, pour préserver la grammaticalité d'un énoncé, un nom commun féminin singulier ne peut être remplacé que par un autre nom commun féminin singulier.
La catégorie des « verbes » n'est pas homogène car il faudrait distinguer :

- les verbes **intransitifs** comme « dormir », qui caractérisent un individu sujetunique (on dit qu'ils ne « réclament pas de complément d'objet »)
- les verbes **transitifs** comme « aimer » qui caractérisent une relation entre deuxindividus : un sujet et un objet.
- les verbes **bi transitifs** comme « donner » qui caractérisent une relation à trois :un sujet et deux objets (quelqu'un donne quelque chose à quelqu'un d'autre).

Les substitutions entre verbes ne peuvent préserver la grammaticalité qu'en tenant compte de ces propriétés.

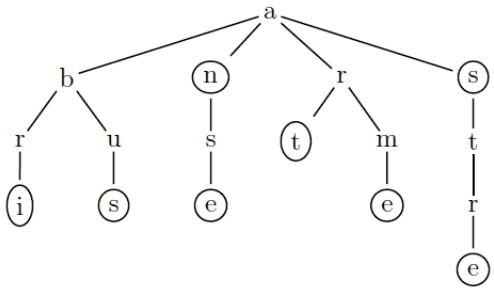
On associe aux unités lexicales présentes dans un énoncé des informations de flexion : le genre, le nombre et éventuellement le cas pour les noms, les pronoms et les adjectifs ; la personne, le nombre, le temps, le mode et la voix pour les verbes.

2 Modélisation informatique

La capacité de mémoire des ordinateurs actuels permet de stockerl'intégralité des formes fléchies d'une langue. On a pourtant intérêt à enregistrer ces différentes formes de façon synthétique. Nous évoquons deux approches possibles : la première est une structure de données, une manière efficace de coder des informations, la deuxième est un modèle issu de l'informatique théorique: les automates finis. Ils permettent de représenter les découpages morphologiques sous la forme de règles plutôt que sous la forme de listes.

2.1 Arbre à lettres

Pour stocker un dictionnaire de façon économique, il existe des organisations plus efficaces que les simples listes : **les arbres.**
Exemple :



Le point de départ des arbres, situé en haut de la figure, s'appelle racine, tandis que chaque point intermédiaire est un nœud.
Ces nœuds sont reliés les uns aux autres par des branches qui se développent de haut en bas jusqu'à des feuilles. Les fils d'un nœudintermédiaire (c'est-à-dire qui n'est pas une feuille) sont les nœudssitués au niveau immédiatementinferieur et reliés au premier par une branche. Un chemin est une succession de nœuds partant de la racine et suivant les branches en descendant.

Mais les arbres à lettres ne rendent absolument pas compte de l'organisation morphologique des mots : les morphèmes n'y sont pas du tout apparent. Pour remédier à ce problème, on va utiliserle modèle des **automates finis.**

2.2 Automates finis

Un automate fini est un dispositif constitué de :

- un vocabulaire fini V : le vocabulaire sera constitué de l'ensembledes morphèmesconsidérés.
- un ensemble fini Q d'états, où figurent : au moins un état initial, au moins un état final, une fonction de transition f : cette fonction énumère, pour chaque état possibleq $\in$Q et chaque élément possible du vocabulaire v $\in$V, l'état que l'on peut atteindre en partant de q et en utilisant v : f(q, v) $\subseteq$Q.

Un automate est un modèle informatique car il se traduit de façon quasi immédiate en un programme.
Exemple d'un automate élémentaire : Cela va permettre de rendre compte des régularités de découpages morphologiques:« inimitables », « imitée », « analysées », « inanalysable », « désirable », « indésiré », « innommable »... La liste des morphèmes est constituée du préfixe« in », et des racines « imit », « désir », « analys » et « nomm », des suffixes « able » et « é » et des flexions « e » et « s ». Ces morphèmesdéfinissent l'ensemble V de notre automate.
A tout automate, on peut associer une représentation graphique, que l'on nomme un graphe.
Un automate permet de caractériser un ensemble de chemins. Un chemin dans un automate commence à partir d'un état initial, suit des transitions et aboutit à un état final. Le langage de l'automate est l'ensemble de suites d'éléments du vocabulaireV qui correspondent à un tel chemin.
Le langage de notre automate contient tous les mots évoqués au début et leurs variantes morphologiques. C'est un langage fini car on peut énumérer tous ses éléments dans une liste finie.
Les automates sont bien adaptés à la modélisation des phénomènes d'affixation. On peut alors construire des automates qui explicitent les règles de conjugaison des verbes, mais il faut alors autant d'automates différents qu'il y a de familles de conjugaisons différentes possibles. Dans ce cas, on peut aussi faire en sorte que les états finaux distincts identifient les différentes personnes (1ère, 2ème ou 3ème) et le nombre (singulier ou pluriel) de la conjugaison. Le verbe « troufigner », pourrait avoir toutes les chances de se conjuguer comme « aligner » et tous les autres verbes réguliers du 1er groupe. Le fait de disposer d'un automate pour caractériser cette conjugaison évite d'en inventer une nouvelle, et économise son stockage en mémoire.

➔ Un des principaux intérêts des automates est qu'ils peuvent fonctionner aussi bien en analyse qu'en synthèse. En analyse, on dit que l'automate « reconnaît » un mot s'il existe un chemin correct étiqueté par ce mot. En synthèse, on dit qu'on « génère » ou « engendre » un mot en le produisant au fil des transitions de l'automate.

2.3 Expressions régulières

Les automates finis ont été étudiés par Chomsky et les pionniers de l'informatique théorique, à partir des années 60, et un très grand nombre de leurs propriétés formelles ont alors été explicitées. Les automates finis reconnaissent une classe de langages particulière appelée langages réguliers ou langages rationnels.
Un langage est régulier s'il existe un automate qui reconnaît exactement ce langage.
Une expression régulière est une suite de symboles prise parmi :

- un vocabulaire fini V
- le symbole ϵ qui représente la « chaîne vide »
- les symboles : « | » et «. », ainsi que les parenthèses « (« et ») »
- les symboles « + » et « * »utilisés en exposants

Elle doit êtrebâtie en respectant les règles de construction suivantes :

- tout élément de V ∪ {∈} est une expression régulière
- si U est une expression régulière, alors (U) est aussi une expression régulière

- si U et T sont toutes les deux des expressions régulières, alors U|T et U.T sont des expressions régulières : U|T = UUT = {w|w ∈U ou w ∈T} représente le choix entre un élément de U et un élément de T, tandis que UT = {ut|u ∈U et t ∈T} représente l'ensemble des concaténations d'un élément de U et d'un élément de T
- si U est une expression régulière, alors U+ et U* sont des expressions régulières :

U+ = {u1u2...un|∀i, ui ∈U} représente l'ensemble des concaténations un nombre quelconque de fois (au moins une) d'éléments de U, et U* = U+ ∪ {ϵ}est la mêmeconcaténation, y compris 0 fois (ce qui donne la chaîne vide).
Le symbole ϵjoue le rôle d'élément neutre pour la concaténation : pour toutsymbole w ∈V, w.ϵ= ϵ.w= w.

Chapitre 5 : Le niveau de la syntaxe

Le niveau de la syntaxe explique comment mettre bout à bout des unités lexicales afin de bâtir des énoncés dont le sens global est plus que la simple somme des sens de ces unités. Il constitue la « première articulation »de toute langue naturelle. Sous l'influence de Chomsky, ce niveau est celui qui a fait l'objet du plus grand nombre de travaux, d'études et de modèles ces 50 dernières années.

1 Description linguistique

1.1 De l'analyse distributionnelle à la notion de grammaticalité

Jusqu'aux années 50, le courant dominant en matière de description linguistique est l'analyse distributionnelle. Pour étudier une langue, il faut disposer d'un échantillon aussi représentatif que possible de cette langue : un corpus. Une langue n'a pas besoin d'être comprise pour êtreétudiée : toutes ses propriétés doivent pouvoir être extraites des régularités et redondances observées dans le corpus.
La distribution d'une unitéprésente dans un corpus est définie comme l'ensemble de ses environnements, c'est-à-dire des suites d'unités qui la précèdent et qui la suivent dans ce corpus, dans une fenêtre dont la taille est bornéeà l'avance.
L'ensemble des unités qui partagent un environnement constituent une classe distributionnelle.
Exemple :« bébé » et « marmot » appartiennent à la même classe car ils doivent apparaitre, dans tout bon corpus, dans les mêmes environnements (précédés de « le » et suivis de « pleure », par exemple...). On peut définir une grammaire, comme un ensemble de classes et de listes d'environnements associés. Une grammaire est l'usage distributionnel qui est fait de ses unités linguistiques.
Pour Chomsky, tout corpus est nécessairement incomplet parce fini, alors qu'une langue permet de construire un nombre potentiellement infini de phrases différentesà partir d'un nombre fini d'unités. Pour ranger les unités dans des classes, ce qui sera déterminant est le critère degrammaticalité.
Une grammaire est un dispositif capable d'opérer des jugements de grammaticalité, c'est-à-dire de trier les suites d'unités en « correctement formées » (grammaticales) ou non.

Pour bien comprendre les conséquences de cette conception voici les distinctions suivantes :
- la grammaticalité est différente de la fréquence d'apparition dans un corpus.

Exemple : « le petit_est mort » (le tiret remplace une unité lexicale), les unités « téléphone » et « chanterons » sont improbables. Pourtant, le statut des deux énoncés ainsi construits ne serait pas le même : « le petit téléphone est mort » est parfaitement grammatical, ce qui n'est pas le cas de « le petit chanterons est mort ».

- la grammaticalité n'est pas synonyme d' « interprétabilité ».

Exemple de Chomsky: « d'incolores idées vertes dorment furieusement » est un énoncé grammatical, mais n'a pas de sens. Et un énoncé comme « vous faire moi rigoler » est interprétablemais non grammatical.

Chomsky propose de remplacer le critère empirique observable de « présence dans un corpus » par un critère mental plus abstrait, dont les effets ne sont visibles qu'indirectement : on peut demander à un locuteur d'opérer autant de jugements de grammaticalité que l'on veut, même s'il est incapable d'expliquer comment il s'y prend. Sa « grammaire mentale » est pour lui une « boîte noir »" dont il ne perçoit que les entrées et sorties. Le rôle du linguiste est d'essayer d'ouvrir la boîte noire.

1.2 Des phrases aux propositions

Une phrase est difficile àcaractériser, aussi bien avec des critères formels qu'avec des critères sémantiques.

Plusieurs unités de descriptions sont constituées de plusieurs « mots » tout en restant généralement plus petites qu'une phrase :

- les chunks : plus petites séquences de mots auxquelles on peut associer une catégorie comme « groupe nominal » ou « groupe verbal ». Mais un tel groupe ne constitue un chunk que si lui-même ne contient pas un autre groupe de même nature.

Exemple : « le chat du voisin », n'apporte pas assez de propriétés nouvelles pour justifier de passer à un « niveau d'analyse » fondamentalement nouveau.

- les termes : noms communs, entitésnommées ou groupes nominaux éventuellementcomposés d'autres groupes nominaux. Les termes identifient un concept précis dans un domaine de spécialité et peuvent se servir de « mots clés » dans une indexation.
- les clauses : séquences de mots contenant au moins un sujet et un prédicat. Une phrase peut en général se découper en plusieurs clauses emboitées les unes dans les autres.

Exemple : Chaque couple de parenthèses marque les frontières d'une clause : « ((La dérégulation des compagnies de chemins de fer, qui a commencé en 1980)), a permis (aux affréteurs de marchandises de négocier leurs tarifs.)) ».

Un syntagme est un mot ou une suite de mots consécutifs auquel on peut associer une catégorie syntaxique.

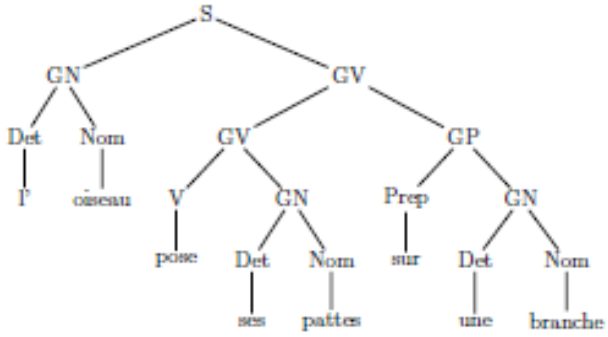
Les linguistes utilisent la notion d' « énoncé ». L'existence d'un niveau d'analyse spécifique, c'est l'existence de suites d'unités auquel on peut attribuer une valeur de vérité (vrai/faux). Cette suite s'appelle une proposition. Ce critère est de nature sémantique. Son inconvénient est qu'il écarte certaines « phrases » courantes comme les exclamations ou les questions.

1.3 Structures syntaxiques

Exemple d'une analyse syntaxique "syntagmatique" : Nous partons de la proposition « l'oiseau pose ses pattes sur une branche. » Repartons donc des affectations des mots présents dans cette proposition à ces catégories :

l'	oiseau	pose	ses	pattes	sur	une	branche
Dét	Nom	Vtr	Dét	Nom	Prép	Dét	Nom

On regarde quelles suites de mots consécutifs on peut remplacer par une autre suite, voire par un unique autre mot en préservant la grammaticalité :
« L'oiseau » peut être remplacé par un nom propre comme « Titi ». C'est aussi le cas de « une branche » qui peut être substitué par « Jean », ainsi que « ses pattes », qui peut être remplacé par « Médor ». Nous noterons l'ensemble de toutes les successions d'unités lexicales substituables à ces suites la classe des "groupes nominaux" : GN.
(L'oiseau)GN pose (ses pattes) GN sur (une branche) GN.
Il est possible de remplacer « sur une branche » par d'autres groupes nominaux introduits par une préposition comme « avec un soupir », « dans une heure »". Ce sont des « groupe prépositionnel » : GP. Son identification dans notre phrase initiale amène le nouveau parenthésage suivant (l'étiquette du groupe est attachée à la parenthèse fermante qui le délimite) : (l'oiseau)GN pose (ses pattes)GN (sur (une branche) GN) GP.
Le verbe « pose » qui est un verbe transitif peut être substitué par un verbe intransitif unique (comme « ronfle »), c'est la catégorie "groupe verbal" : GV. On peut également remplacer « pose ses pattes » par « ronfle » mais aussi « pose ses pattes sur une branche ». Le parenthésage se complexifie et donne : (l'oiseau)GN ((pose (ses pattes) GN) GV (sur (une branche) GN) GP) GV.
La proposition initiale est composée d'un groupe nominal suivi d'un groupe verbal, qui eux-mêmes se décomposent en sous-groupes de différentes natures. On va représenter cette structure hiérarchique avec une représentation arborescente :



Un syntagme est un groupe de mots qui correspond à un sous arbre d'unarbre d'analyse syntaxique complet. Exemple : « pose ses pattes » est un syntagme de catégorie GV.

1.4 Ambiguïtés

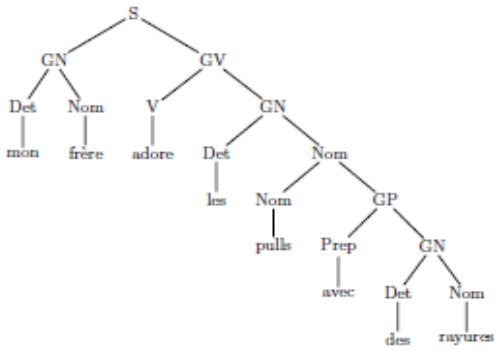
Exemple : « Mon frère adore les pulls avec des rayures ».

mon	frère	adore	les	pulls	avec	des	rayures
Dét	Nom	Vtr	Dét	Nom	Prép	Dét	Nom

On identifie « les pulls avec des rayures » comme un GN, alors que ce n'était pas possible pour « ses pattes sur une branche ». Le GP « avec des rayures » se rattache à une catégorie Nom plutôt qu'à une catégorie GV.
Certaines phrases autorisent plusieurs interprétations différentes correspondant à plusieurs arbres différents : on dit qu'elles sont ambigües.
Exemple : « l'homme observe sa voisine avec des jumelles ». Cette phrase peut se comprendre de deux manières différentes :

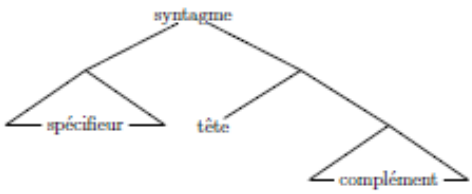
- soit elle signifie que les jumelles sont l'instrument grâce auquel l'homme réalise ses observations.
- soit elle signifie que la voisine observée, possède des jumelles.

Plusieurs phénomènes syntaxiques classiques peuvent donner lieu à une ambiguïté. Dans cette phrase on a un problème de rattachement prépositionnel, le groupe GP pouvant s'accrocher soit à un GV (ou un V), soit à un Nom (ou un GN) et donc la représentation serait la suivante :



1.5 Problèmes avec la structuration arborescente

Nous n'avonspas de raison syntaxique de traiter différemment « le chat mange la souris » et « le chat mange la nuit ». Pourtant, dans le premier cas, « la souris »est l'objet de l'action, tandis que, dans le deuxième cas, « la nuit »précise le moment de sa réalisation. On peut reconnaitre leur différence de statut au fait que certains GN peuvent changer de place et pas d'autres : Exemple : « la nuit le chat mange » ou même« la nuit mange le chat »préservent le sens initial de la phrase mais pas « la souris mange le chat ». Les relations de type « sujet », « complément d'objet direct »…sont appelésrôlessémantiques. L'identification de ces rôles, traitant desémantique propositionnelle se superpose en quelque sorte sur la structure arborescente de la proposition.
Chomsky précise la structuration interne générale des syntagmes.Il remarque que chaque syntagme contient une unité lexicale privilégiée qu'il désigne comme sa « tête » : la tête d'un syntagme nominal est son nom commun principal, celle d'un syntagme verbal son verbe principal... Les autres composants du syntagme s'organisent autour de sa tête de façonrégulière, et ce quelle que soit la nature de ce syntagme. Enfrançais, les têtes sont précédées d'un spécifieur et suivies de compléments :



Les structures arborescentes ne sont pas toujours suffisantes pour rendre compte de toutes les constructions linguistiques. Voici quelquesdifficultés :

- l'ellipse, qui autorise l'effacement de certains constituants pour éviter unerépétition, comme dans « elle est d'accord, moi non ».
- l'apposition, qui est la juxtaposition de syntagmes de même nature ayant le mêmeréférentsémantique : « Jean, mon voisin, un bon ami, m'a rendu visite ».
- la thématisation, où l'ordre des mots permet d'introduire des référents sur lesquels on met successivement l'accent : « moi, mon papa, sa voiture, elle est rouge ».

Dans ces situations, le rattachement entre les syntagmes qui est délicat. Certaines traditions d'analyse syntaxique se passent mêmecomplétement des arbres, et mettent en avant à la place la notion de dépendances (relation orientée entre deux mots) entre unités lexicales.
Un arbre rend visible la construction interne commune d'un nombre potentiellement infini de phrases différentes : c'est ce qui fait son expressivité.
On dispose maintenant de corpus arborés, c'est-à-dire de textes parenthèses et étiquetés syntaxiquement, rendant explicite la structure des phrases qu'il contient. Ces données, mises au service des informaticiens et des linguistes jouent un grand rôle dans la recherche actuelle.

2 Modélisation informatique

C'est sans doute dans le domaine de la modélisation de la syntaxe que le plus de travaux ont été produits ces 50 dernièresannées en TALN. Les recherches ont avancé en parallèle avec plusieurs autres branches de l'informatique : exemple : les « langages évolué » dans lesquels les informaticiens d'aujourd'hui écrivent leurs programmes (Java, C++, Python…) nécessitent pour être « compilés », une phase « d'analyse syntaxique ».
Nous allons voir la « théorie des langages », traitant des grammaires et des langages en général. C'est son adéquation aux langues humaines qui nous intéressera.

2.1 Le retour des automates finis

Une grammaire doit êtreconçuecomme un dispositif capable de « classer » une suite de mots quelconque en « grammaticale » ou « non grammatical ». Ladifficulté du problèmeréside dans le paradoxe suivant : même en supposant l'ensemble de tous les mots possibles d'une langue comme fini on peut construire avec cet ensemble fini un nombre potentiellement infini de phrases grammaticales, et un nombre potentiellement infini de « phrases » non grammaticales. La clé du problème tient dans une notion fondamentale en informatique : la récursivité.
Imaginons un automate dont le vocabulaire fini V contient l'ensemble de tous les mots possibles du français (flexions et conjugaisons comprises), et dont chaque « chemin » correspondrait à une phrase syntaxiquement correcte. Confronté à une suite de mots quelconque, l'automate n'aurait qu'àvérifier si cette suite correspond à un de ses chemins pour savoir si elle est grammaticale.
Par définition, il est possible d'utiliser un nombre quelconque de fois ces transitions, et donc de juger grammaticales un nombre infini de chaines possibles.

2.2 Limites des automates finis

La grammaire de la langue françaisene peut pas se représenter sous la forme d'un énorme automate fini où figurerait de la récursivité. Nous allons voir trois des principaux arguments cités à l'encontre de cette hypothèse. Ce sont trois façonsdifférentes de dire la même chose, mais en mettant l'accent sur une facette ou sur une autre de l'analyse linguistique.

L'argument est dû à Pinker, psycholinguiste canadien, et relève surtout de la psychologie cognitive. Il consiste à constater que stocker en mémoire ce qui constitue la « compétence » des locuteurs d'une langue sous la forme d'un unique automate ne serait ni économique ni efficace. En français, comme en anglais et dans beaucoup d'autres langues, on utilise la même construction pour les « GN » de la proposition dont ils font partie. S'il fallait bâtir un « automate complet du français », il faudrait donc répéter la portion d'automate qui décrit la construction des groupes nominaux à plusieurs endroits dans cet automate global : au début pour les GN sujets, après le verbe pour les GN objets directs…Ce n'est pas comme cela que la mémoire humaine fonctionne. On a d'ailleurs pu observer que lorsqu'un enfant a entendu une seule fois un nom (ou toute autre unité lexicale, ou toute portion de structure) dans une certaine position grammaticale, il est capable de le réemployer spontanément et instantanément dans une autre position grammaticale.

Le deuxième argument est proche du premier : il met en avant le fait que les automates sont incapables de produire les structures arborescentes. Ces structures sont le résultat direct de notre critère de substituabilité : s'il est possible d'étiqueter plusieurs portions d'arbres (ou sous arbres) par GN.

Le dernier argumentremonte auxpremières intuitions de Chomsky : il s'appuie sur un théorème mathématique. Le théorèmeénonce qu'il est impossible d'engendrer avec un automate fini certains langages comme le langage L = {anbn|n ≥ 1} (construit sur le vocabulaire V = {a, b}), c'est-à -dire le langage qui réunit toutes les suites constituées d'un nombre quelconque (non nul) de symboles a suivi du même nombre de symboles b. Ce langage comprend les chaines : ab, aabb, aaabb... Un automate fini ne peut pas reconnaitre ce langage.

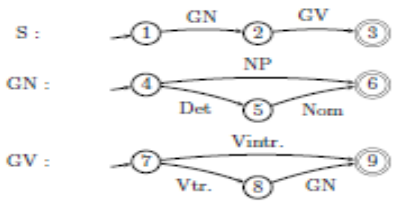
Exemple : Supposons qu'un tel automate existe, appelons-le A et cherchons à en déduire une contradiction. Le raisonnement suit les étapes suivantes :

- le langage L étant évidemment infini, A doit inclure au moins une boucle récursive (directe ou indirecte).
- choisissons parmi le langage engendré par A la suite de symboles correspondant à un chemin dans A qui emprunte exactement une fois cette boucle. Soit w cette suite de symboles. Par définition, on peut la décomposer en 3 morceaux : le morceau correspondant au chemin parcouru « de l'état initial jusqu'au début de la boucle », celui correspondant au chemin « dans la boucle » et celui du chemin « après la boucle jusqu'à un état final » (ce dernier et le premier pouvant éventuellementêtre vides) : w = u1u2u3 où u2≠ϵ est la suite des symboles dans la boucle.
- tous les chemins qui commencent et qui se terminent comme le précédent mais qui, au lieu d'emprunter une seule fois la boucle, l'empruntent un nombre quelconque de fois sont des chemins corrects de l'automate : donc toutes les suites de symboles qui appartiennent à l'ensemble u1u2*u3 font partie du langage de A.
- Essayons de choisir parmi les suites de la forme anbn celle qui pourra jouer le rôle de w, et en particulier dans cette suite la portion u2 qui pourra être répétée tout en restant dans le langage L : c'est impossible...

Ce théorème, appelé « lemme de pompage »donne une limite théorique rigoureuse à l'expressivité des automates finis.

2.3 Réseaux de Transitions Récursifs

Les Réseaux de Transition Récursifs (ou RTRs) sont une généralisation des automates finis qui répondent exactement aux arguments de la section précédente.



Un RTR se présente comme un ensemble d'automates qui ont les propriétés suivantes :

- chaque automate de l'ensemble a au moins un état initial et au moins un état final.
- chaque automate est associé à une étiquette, marquéeà sa gauche : on les appelle aussi les symboles non terminaux du RTR car ils servent de vocabulaire intermédiaire, et ne se retrouveront pas dans les suites d'unités lexicales dont on teste la grammaticalité. Dans les exemples linguistiques, ces symboles seront des catégories grammaticales. Parmi elle, figure le symbole Sdésignant la catégorie des propositions syntaxiquement correctes.
- les transitions des automates sont étiquetées soit avec des symboles non terminaux associes à un automate, soit avec des symboles « terminaux », qui sont les unités lexicales.

Les mots avec lesquels on va reconnaitre ou engendrer des phrases figurent à la fin dans de simples listes, alors que les automates qui les précédent n'ont que des transitions étiquetés avec des catégories. Il faut voir ces listes finales comme des automates élémentaires contenant simplement un état initial, un état final et autant de transitions de l'un à l'autre que de mots dans la liste.

Le critère de grammaticalitéassociéà ce nouvel objet est similaire à celui des automates finis: une suite de mots est reconnue si elle correspond à un chemin dans l'automate étiqueté par S du RTR. Comme précédemment, les transitions étiquetées par un mot peuvent être franchies par la reconnaissance de ce mot. Cependant pour franchir une transition qui porte un symbole non terminal, il est nécessaire de parcourir un chemin complet dans l'automate associé à ce symbole. Les automates du RTR ont donc la possibilité de s'appeler les uns les autres, à la manière des fonctions ou des procédures dans les langages de programmation impératifs. Exemple : « le chat mange la souris » peut être jugé grammatical:

- « le » est de catégorieDét et « chat » de catégorieNom, donc « le chat » est un chemin dans l'automate GN (passant par les états 4, 5 et 6). Ce chemin autorise à son tour à passer de l'état 1 à l'état 2 dans l'automate S.
- « la souris » est un chemin dans GN et« mange » étant de catégorieV intr, « mange la souris » est un chemin dans l'automateGV (passant par les états 7, 8 et 9). Donc "mange la souris" permet, dans l'automate S, de passer de l'´état 2 à 3. La phrase est donc acceptée par l'automate associé à S et donc par le RTR dans son ensemble.

Ce dispositif répond à l'argument de Pinker, car les GN en position sujet y sont traités par le même automate que ceux en position d'objet. Si on ajoute un élémentà la liste des noms, il pourra donc être utilise indifféremment dans l'une ou l'autre de ces positions.

Les RTRs rendent parfaitement compte des structures arborescentes.
Enfin, le langage L = anbn, impossible àgénérer avec un automate fini, peut-être reconnu par un RTR. L'unique automate a la propriété de pouvoir « s'appeler lui-même » via la transition portant sa propre étiquette.

2.4 Grammaires formelles

Les grammaires formelles sont au cœur de la « théorie des langages » des informaticiens.
On définit une grammaire formelle G comme un quadruplet d'éléments :
G = (V, N, P, S)où :

- V est le vocabulaire terminal de G : pour une application linguistique, V coïncidera avec l'ensemble fini des mots pouvant figurer dans les suites dont on veut tester la grammaticalité. Les éléments de V s'écriront avec des minuscules latines.
- N est le vocabulaire non terminal de G : comme dans les RTRs, ce vocabulaire servira lors d'étapesintermédiaires de calculs. Il est qualifié de « non terminal » car aucun de ces symboles ne doit se retrouver dans les productions finales reconnues ou engendrées par la grammaire. Il contient entre autres, le symbole S, qui identifie les suites grammaticales(4ème élément du quadruplet) : S ∈N. Traditionnellement, les symboles non terminaux sont écrits en majuscules latines.
- Le dernier élément, P, est un ensemble fini de « règles de production » ou « règle de ré écriture ». Chaque règle de P est de la forme : $\alpha \longrightarrow \beta$ où$\alpha \in (V \cup N)^+$ et $\beta \in (V \cup N)^*$. Ainsi, α et β sont des suites de symboles pris parmi les éléments de V et de N, avec la seule différence que β peut être une liste vide (β = ϵ est autorisé) mais pas α. Une telle règle doit être comprise comme : « α peut êtreremplacé par β ».

Etant donnée une grammaire G =(V, N, P, S)et une suite de symboles quelconque $u \in (V \cup N)^+$, on dit que G permet de dériverv à partir de u en une seule étape, et on note $u \longrightarrow v$ si les conditions suivantes sont réunies :

- u peut se décomposer en trois morceaux : u = xαy avec α ≠ϵ
- la règle : $\alpha \rightarrow \beta \in P$
- v = xβy.

Si une règle de la grammaire précise que « α peut êtreremplacé par β », cette règle peut s'appliquer à l'intérieur de n'importe quelle suite de symboles qui contient α. On peut appliquer successivement autant de règles que l'on souhaite à une suite de symboles. On notera $u \rightarrow^* v$ une dérivation en plusieurs étapes successives.
On appelle langage engendré (ou reconnu) par une grammaire G = (V, N, P, S)et on note L(G) l'ensemble des suites de symboles terminaux que l'on peut obtenirpar dérivations successives en partant de l'unique symbole S. On a ainsi :
$L(G) = \{w \in V^* \mid S \rightarrow^* w\}$.
Exemple : Soit la grammaire G = (V, N, P, S)définie par :

- V = {le, la, chat, souris, dort, mange}
- N = {S, GN, GV, Det, Nom, V tr, V intr}
- P = {S $\rightarrow_1$ GN GV, GN $\rightarrow_2$ Dét Nom, GV $\rightarrow_3$ V intr,

GV $\rightarrow_4$ V tr GN, Det $\rightarrow_5$ le, Det $\rightarrow_6$ la, Nom $\rightarrow_7$ chat,
Nom $\rightarrow_8$ souris, V intr $\rightarrow_9$ dort, V tr $\rightarrow_{10}$ mange}
➔ « Le chat mange la souris » fait partie du langage engendré par G. On a numéroté les règles et soulignéà chaque étape la portion de la suite qui est réécrite par la règle en question :

S $\rightarrow_1$ GN GV
GN GV $\rightarrow_2$ Dét Nom GV
Dét Nom GV $\rightarrow_5$ le Nom GV
le Nom GV $\rightarrow_7$ le chat GV
le chat GV $\rightarrow_4$ le chat V tr GN
le chat V tr GN $\rightarrow_{10}$ le chat mange GN
le chat mange GN $\rightarrow_2$ le chat mange Det Nom
le chat mange Det Nom $\rightarrow_6$ le chat mange la Nom
le chat mange la Nom $\rightarrow_8$ le chat mange la souris
On a donc bien : S $\rightarrow^*$ le chat mange la souris.
L'ordre d'application des règlesétant arbitraire, il existe d'autres séquences de dérivations qui produisent le mêmerésultat. D'autres suites grammaticalement correctes peuvent être obtenues par cette grammaire comme : « la souris dort », ou « le chat mange le chat ». Comme aucune contrainte d'accords en genre n'a été prise en compte dans ces règles, on peut aussi produire « la chat mange le souris ».
Il faudrait introduire des symboles non terminaux distincts pour les catégories Dét et Nom, suivant qu'ils sont masculins ou féminins, et adapter la règle 2 pour n'autoriser que les associations entre Dét et Nom du mêmegenre. Une phrase sera reconnue comme syntaxiquement correcte si, à partir de ses mots et en « remontant le sens » des règles de réécriture, on peut arriver au symbole S.

2.5 Transformation des automates et des RTRs en grammaires

N'importe quel automate ou RTR peut êtretransformé en une grammaire équivalente. Le langage L = la.Ferrari.passa.tres*.vite. La grammaire G = hV, N, P, Si qui lui correspond est définie par :

- V = {la, Ferrari, passa, tres, vite}
- N = {S = Q_1, Q_2, Q_3, Q_4} : les états de l'automate correspondent aux symboles non terminaux de la grammaire (la lettre Q est traditionnellement utilisée pour nommer ces symboles). L'état initial joue le rôle de S, tandis que l'état final ne sert que de point d'arrivée pour une transition.
- P = {S → la Q_2, Q_2 → Ferrari Q_3, Q_3 → passa Q_4, Q_4 → tres Q_4,Q_4 → vite}

Exemple :« la Ferrari passa trèstrès vite » :
S → la Q_2
la Q_2 → la Ferrari Q_3
la Ferrari Q_3 → la Ferrari passa Q_4
la Ferrari passa Q_4 → la Ferrari passa tres Q_4
la Ferrari passa tres Q_4 → la Ferrari passa tres tres Q_4
la Ferrari passa tres Q_4 → la Ferrari passa tres tres vite
Grâceà cette grammaire, on peut associer un arbre à chaque phrase du langage L. Mais les arbres obtenus ont une forme particulière : ils ne se développent toujours que « vers la droite ». On appelle ce type d'arbre des « peigne ».
Voici la méthodeà employer pour transformer un automate fini quelconque en une grammaire formelle :

- le vocabulaire terminal V de la grammaire est identique au vocabulaire de l'automate
- définir autant de symboles non terminaux Q_n qu'il y a d'états non terminaux n dans l'automate : l'état initial correspond au symbole S (Q_0 = S) ;

- pour tout état terminal m de l'automate à partir duquel part une transition, introduire un nouveau symbole Q_m non terminal dans la grammaire ;
- pour toute transition étiquetée par un symbole terminal quelconque a ∈V partant d'un étatn et aboutissantà un étatm dans l'automate :
  - si m est un état terminal, ajouter la règle : $Q_n \rightarrow a$
  - si m n'est pas un état terminal, ou bien si m est terminal mais il existe une transition qui en part, alors ajouter la règle : $Q_n \rightarrow a\ Q_m$

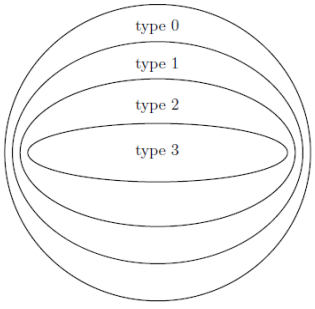
Les cas de récursivité indirecte seront eux aussi « lisibles » à partir des règles de la grammaire.

2.6 Hiérarchie de Chomsky

Dans la famille des grammaires formelles plusieurs « classes » se distinguent, précisément, par leur « expressivité ».La hiérarchie de Chomsky explicite la nature et les propriétésde ces classes. Comment savoir à quelle classe appartient une grammaire donnée ? Tout est dans la forme des règles de l'ensemble P. Soit une grammaire G = (V, N, P, S):

- si toutes les règles de P sont de la forme : A → a ou A → a B (ou bien si elles sont toutes de la forme : A → a ou A → B a) avec A, B∈N et a ∈V, alors n dit que G est une grammaire rationnelle ou régulière, ou encore que G est de type 3. Les grammaires qui proviennent de la transformation d'un automate, sont de ce type et inversement, toute grammaire de ce type peut être transformée en un automate.
- si toutes les règles de P ont une partie gauche réduite à un seul symbole non terminal, c'est-à-dire sont de la forme : A → ... avec A ∈N et n'importe quelle suite de (V ∪N) *à droite de la flèche, alors on dit que G est une grammaire hors-contexte (anglais « context-free ») ou algébrique, ou encore que G est de type 2. Les grammaires qui proviennent de la transformation d'un RTR sont de ce type et inversement, toute grammaire de ce type peut êtretransformée en un RTR.
- si toutes les règles de P sont telles que le nombre total de symboles de V ou de N à gauche de la flèche est toujours inferieur ou égal au nombre total de symboles à droite de la flèche, on dit que G est une grammaire sensible au contexte (traduction de l'anglais « context-sensitive ») ou encore que G est de type 1.
- toutes les grammaires formelles, quelle que soit la forme de leurs règles, sont de type 0.

Il y a 4 grandes familles de grammaires, emboitées les unes dans les autres. Toute grammaire qui vérifie celui d'une certaine classe vérifie aussi nécessairement ceux des classes de type inferieur.



Cette hiérarchie sur les grammaires permet de classer les langages.

Un langage L sur un vocabulaire V peut toujours êtreconsidéré comme une partie de V*(la partie des suites d'unités syntaxiquement correctes). Un langage L est de type n s'il existe une grammaire G de type n telle que L(G) = L et si aucune grammaire de type m > n ne satisfait cette propriété. Certains des noms de ces classes proviennent de propriétésmathématiques que nous n'évoquerons pas. Mais, pour les grammaires de type 2, l'explication est simple : elles sont appelées « hors-contexte » parce que chacune de leurs règles de réécriturespécifient comment remplacer un symbole non terminal indépendamment du contextedans lequel il se trouve, c'est-à-dire des symboles -terminaux ou non terminaux- qui sont ses voisins dans la chaîne de réécriture. Cette propriété nous assure que les structures produites par ces grammaires prennent toujours la forme d'arbres.
Au contraire, une grammaire de type 1 peut contenir des règles contextuelles.
Exemple : de la grammaire de type 1 définie par les ensembles V = {a, b, c}, N = {S, A, B} et les règles : S → aBSc, S → aBc, Ba → aB, Bc → bc et Cette hiérarchie sur les grammaires permet aussi, bien sûr, de classer les langages. Ces règles ne permettent jamais de raccourcir la chaîne en train d'êtreproduite. La plupart d'entre elles autorisent simplement à déplacer certains symboles ou à remplacer un symbole non terminal par un terminal, mais uniquement s'il se trouve dans un certain contexte, spécifié par ses symboles voisins. Cette grammaire permet d'engendrer le langage L = anbncn (n ≥ 1).
Pour synthétiser toutes ces propriétés, voici le tableau suivant :

type	nom des grammaires	forme des règles	structures produites	exemple typique	modèle équivalent
3	régulières ou rationnelles	$A \longrightarrow a$ ou $A \longrightarrow aB$	peignes	$a^* = a^n$	automates finis
2	hors-contextes ou algébriques	$A \longrightarrow ...$	arbres	$a^n b^n$	RTRs
1	sensibles au contexte	$\alpha \longrightarrow \beta$ $\lvert\alpha\rvert \leq \lvert\beta\rvert$	?	$a^n b^n c^n$	existe mais compliqué
0	quelconques	quelconque	quelconques	quelconques	machines de Turing !

2.7 Position des langues naturelles dans la hiérarchie de Chomsky

Les langues naturelles ne sont pas de type 3. La classe des grammaires de type 2, ou algébriques,est un candidat qui présente de sérieux arguments.
Nous avons déjàévoqué le langage anbncn, ne peut êtreengendré que par une grammaire de type 1. Voici un autre exemple de langage {ww|w ∈V*}, c'est-à-dire l'ensemble des suites de symboles (terminaux) quelconques répétées deux fois successivement à l'identique. Pour obtenir de telles chaines, il est nécessaire de générer en même temps les mêmes symboles dans chacune des deux suites : ce qui équivautà avoir une structure avec des branches « qui se croisent ».
Or, certains linguistes prétendent avoir reconnu des constructions relevant d'un des deux langages précédents dans le dialecte suisse alémanique, ou dans le génitifen géorgien ancien, ou encore dans la manière de compter en chinois...
L'argument, qui bien sûr suppose aussi l'adhésionpréalable à la distinction compétence/performance, est plus difficile àvérifier que celui qui portait sur les propositions relatives enchâssées, mais il est de même nature.
Depuis quelques années, une classe intermédiaire contenant toutes les grammaires de type 2 mais strictement plus petite que l'ensemble de celles de type 1 a étéidentifiée : on l'appelle la classe des grammaires légèrement sensibles au contexte (traduction anglaise de « middly context-sensitive »). Cette nouvelle classe présente toutes les « bonnes

propriétés »qu'on pouvait espérer : les grammaires qui en font partie permettent de produire les langages précédemmentévoqués, et l'analyse syntaxique y est réalisable par des algorithmes efficaces. Plusieurs formalismes ont étéproposés pour caractériser les grammaires « légèrement sensibles au contexte » : le plus célère d'entre eux est celui des « tree adjoining grammars » ou TAG.

2.8 Autres formalismes

Les accords en genre et en nombre des langues romanes obligent également la mise en place de mécanismes particuliers. Les grammaires formelles décrites précédemment sont donc en fait très difficilement exploitables telles quelles en traitement des langues. De manière générale, on appelle grammaire de constituant tout formalismequi réalise des analyses syntaxiques sous la forme de groupes de mots consécutifsrécursivementemboîtés les uns dans les autres. On peut rattacher à cette famille les grammaires LFG, GPSG, HPSG...
La principale alternative à ce choix est représentée par les grammaires de dépendances. Dans les formalismes de cette famille, les relations de dépendances entre couples de mots remplacent les constituants.
Les formalismes le plus en usage ces dernièresannéesont un point commun : ils sont presque tous lexicalisés. Un formalisme lexicaliséopère une distinction claire entre les informations syntaxiques rattachées aux éléments de son vocabulaire (terminal) d'une part, et les règles qui permettent de combiner entre elles ces informations d'autre part. L'idéalest de disposer d'un nombre restreint de règlesgénériques, communes à toutes les instances de grammaires du formalisme. Ce qui distingue une grammaire d'une autre se limite alors aux informations rattachéesà chacun de ses mots. Le principal intérêt de cette distinction est de faciliter les procédures de mises à jour : puisque les règlessont fixées une fois pour toute, seules les informations lexicales peuvent éventuellementêtremodifiées. Les formalismes lexicalisés les plus connus sont : les grammaires d'unification, les grammaires catégorielles et les LTAG (variante lexicalisée des TAG).
Enfin, un dernier type de formalisme : les grammaires minimalistes, qui essaient de formaliser les derniers écrits de Chomsky sur la syntaxe. C'est un modèlelexicalisé, qui ne fait usage que de deux règles génériques : une règle de « fusion » et une règle de « déplacement ». Sa particularité est de faire l'hypothèse que certaines constructions syntaxiques sont le résultat de déplacements de constituants qui laissent derrière eux des « traces ».
Une trace est en quelque sorte la « place vide »laissée par un constituant qui a étédéplacé : bien qu'invisible, elle est censée avoir des effets mesurables sur l'ensemble de la phrase dont elle fait partie.