Modèles de Langage et Analyse Syntaxique

Cours 1 - Introduction aux modèles syntaxiques

Antoine Rozenknop
antoine.rozenknop@lipn.univ-paris13.fr

30 septembre 2010

Plan

1 Introduction

La syntaxe est une partie très importante du TAL, et peut-être l'une des plus traitées. Nous rappelons dans cette section les différents niveaux de traitement nécessaires pour parvenir à une compréhension complète d'un énoncé en langage naturel. Au fur et à mesure que l'on progresse dans cette hiérarchie des niveaux, les difficultés augmentent, et les outils aujourd'hui disponibles se font moins performants. Toutefois, bien des applications ne nécessitent pas une compréhension complète des énoncés, et n'utilisent que les niveaux de traitement les mieux compris et automatisés.

Considérons à titre d'exemple l'énoncé suivant :

(1) *Le président des antialcooliques mangeait une orange 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 :

1. segmenter ce texte en unités lexicales (mots) ;

   « le », « président », « des »,...

2. identifier les composants lexicaux et leurs propriétés : c'est l'étape de **traitement lexical** ;

   le : article défini, masculin, singulier
   président : nom commun, masculin, singulier
   mangeait : verbe conjugué, troisième personne du singulier, imparfait
   ...

3. identifier des constituants (groupes) de plus haut niveau, et les relations (de dominance) qu'ils entretiennent entre eux : c'est l'étape de **traitement syntaxique** ;

   « le président », sujet de « mangeait une orange avec un couteau »
   « une orange », complément d'objet direct de « mangeait »
   ...

4. construire une représentation du sens de cet énoncé, en associant à chaque concept évoqué un objet ou une action dans un monde de référence (réel ou imaginaire) : c'est l'étape de **traitement sémantique** ;

   « il y a un président qui mange une orange avec un couteau »

5. Identifier enfin la fonction de l'énoncé dans le contexte particulier de la situation dans lequel il a été produit : c'est l'étape de **traitement pragmatique**.

   Selon le contexte : si l'on sait que quelqu'un a volé une orange chez le marchand, le but de l'énoncé est de dire que c'est peut-être le président.

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 des informations dans les deux sens (à la fois de niveau « bas » vers des niveaux « hauts », et en sens inverse) : il est ainsi souvent nécessaire d'obtenir 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 TAL. Ainsi, une application de génération de texte impliquera la production d'un argumentaire (pragmatique), la construction des représentations des significations à engendrer (sémantique), la transformation de ces représentations sémantiques en une suite bien formée de mots (morphosyntaxe), etc.

Le niveau lexical est relativement bien maîtrisé aujourd'hui : il suffit pratiquement d'identifier les mots, ce qui peut se faire à l'aide d'un bon dictionnaire informatique, éventuellement accompagné de règles de morphologies, qui permettent de « construire » les mots à partir de leur racine, en leur ajoutant des préfixes et/ou des suffixes (exemple : les règles de conjugaison).

Ce niveau n'est cependant pas trivial, car on se heurte à un certain nombre de problèmes, parmi lesquelles on peut signaler :

1. l'identification des noms propres, des sigles, des nombres ;
2. le repérage des « entités nommées » : Ex : centre équestre, chasse d'eau, fraise des bois...
3. les problèmes d'ambiguïté (homonymes) Ex : voler, brise
   La désambiguïsation des homonymes nécessite le recours aux traitements syntaxique et sémantique (voire pragmatique).

Les niveaux sémantique et pragmatique nécessitent des connaissances de plus en plus étendues du monde. Pour le TAL, nous sommes obligés de nous restreindre à des micro-mondes (pour le moment), à des domaines restreints.

Ex : compréhension d'articles scientifiques en génomique.

Le niveau syntaxique n'est pas aussi bien maîtrisé que le niveau lexical, mais comme il ne nécessite pas autant de connaissances sur le « monde » que les niveaux sémantique et pragmatique, les développements intensifs des 30 dernières années tentent de construire des représentations et des algorithmes suffisamment généraux pour traiter automatiquement une langue donnée, sans se restreindre à des sous-domaines. C'est le niveau qui nous intéresse dans ce cours.

2 La syntaxe

2.1 Notions de Grammaticalité et d'Interprétabilité

2.1.1 Pourquoi ?

Pourquoi, alors que les phrases (2a) et (3a) sont grammaticales, seule (3b) est grammaticale ?

Grammatical ?

(2) a) Cette valise est pleine à craquer.
    b) * A craquer, cette valise est pleine.
(3) a) Ce nom est difficile à prononcer.
    b) A prononcer, ce nom est difficile.

Pourquoi les phrases (4a) et (5a) sont ambiguës, alors que (4b) et (5b) ne le sont pas ?

Ambigu ?

(4) a) Tout le monde admire son enfant.
    b) Son enfant est admiré par tout le monde.
(5) a) Tous les enfants ont vu une sorcière.
    b) Une sorcière a été vue par tous les enfants.

2.1.2 Mystère !

Un locuteur natif a l'intuition des phrases de sa langue. Il est capable de reconnaître si un énoncé est bien formé (grammaticalement correct) ou pas. Mais un énoncé grammaticalement correct peut être interprétable ou non. Inversement, un énoncé interprétable peut très bien être grammaticalement incorrect.

Ex : n.i. = non interprétable, * = agrammatical

n.i. : D'incolores idées vertes dorment furieusement. (Chomsky)
* : Moi vouloir cigarettes acheter.
* n.i. : Libérer Jean l'interprétation.
n.i. : Les silences éclatent dans l'eau cérébrale.

Le fait que tout locuteur d'une langue puisse émettre ces types de jugements vis-à-vis des séquences qu'on lui soumet implique l'existence d'une compétence linguistique sous-jacente partagée par l'ensemble des locuteurs.

Cette compétence comporte, entre autres, des connaissances et des règles qui déterminent la bonne formation syntaxique, et qui associent une (des) interprétation(s) sémantique(s) appropriée(s) aux séquences bien formées.

Cet ensemble de connaissances et de règles permet aux locuteurs de produire et d'interpréter des énoncés.

Or, il est certain que ces règles ne leur ont jamais été enseignées de façon explicite : aucune grammaire ne fournit la liste exhaustive de ces règles.

Lorsqu'on interroge ces mêmes locuteurs, ils sont généralement incapables de formuler les raisons pour lesquelles ils acceptent ou rejettent une séquence.

De même, le choix d'une ou plusieurs interprétations possibles relève généralement d'un « mystère ».

Chaque locuteur possède en quelque sorte « une grammaire intériorisée », mais il est incapable d'y accéder.

L'objectif principal de la linguistique consiste à rendre compte de cette compétence, en expliciter le contenu et le fonctionnement.

Si cette entreprise est menée à bien, elle nous renseigne sur les propriétés universelles du langage.

Certains linguistes font l'hypothèse de l'existence d'une grammaire universelle, commune à toutes les langues du monde.

Note : Déterminer si un énoncé est grammatical ou interprétable n'est pas toujours aussi facile que dans les exemples précédents.

Exemples (du plus au moins interprétable/grammatical) :

Pierre est encore arrivé en retard.
Pierre est arrivé en retard encore.
Pierre encore est arrivé en retard.
Pierre est arrivé en encore retard.

2.2 Définitions :

Définitions

*La **syntaxe** est l'étude scientifique de la construction des phrases. Elle vise à formuler les régularités sous-jacentes à leur organisation, c'est-à-dire qu'elle vise à déterminer les règles (principes) qui gouvernent les relations de combinaison et de dépendance entre les mots et les groupes de mots au sein de la phrase.*

*La **grammaire** est la description des contraintes caractéristiques d'**une langue donnée** (ce que l'on voit à l'école avec le français). La grammaire est donc constituée des principes syntaxiques universels augmentés des principes spécifiques à une langue.*

Les modèles et formalismes grammaticaux proposés dans le cadre du TAL sont nombreux et variés. On reviendra par la suite sur certains des principaux formalismes, après avoir introduit les notions et les problèmes linguistiques associés à ce niveau syntaxique.

Le niveau syntaxique n'est pas concerné par l'interprétabilité des énoncés : c'est le niveau conceptuel concerné par le calcul de la validité de certaines séquences de mots, les séquences **grammaticales** ou **bien formées**.

Dans une application de génération de texte, on conçoit bien l'importance d'un tel traitement, car il est essentiel que la machine engendre des énoncés corrects, les énoncés agrammaticaux étant difficiles à comprendre pour les utilisateurs.

En revanche, dans un application de compréhension automatique, la machine analyse des textes qui lui sont fournis, et dont on peut supposer qu'ils sont grammaticaux. Pourquoi donc mettre en œuvre dans ce cas des connaissances syntaxiques ?

Une première raison est que l'entrée du module syntaxique est une série de « formes » étiquetées morpho-syntaxiquement, chaque forme pouvant avoir plusieurs étiquettes différentes. Une première fonction du module syntaxique consiste donc à désambiguïser la suite d'étiquettes, en éliminant les séquences qui sont grammaticalement invalides.

Exemple :	**La**	**petite**	**ferme**	**le**	**voile**
	Art	*Nom*	*Nom*	*Art*	*Nom féminin*
	Pronom	*Adj.*	*Verbe*	*Pronom*	*Verbe*
	Nom		*Adj.*		*Nom masculin*

Au niveau des étiquettes morphosyntaxiques, cette phrase peut avoir 108 interprétations (3*2*3*2*3). Mais deux seulement sont grammaticales : « Art Nom Verbe Art Nom masculin » et « Art Adj. Nom Pronom Verbe ».

3 Les constituants syntaxiques

3.1 Organisation de la phrase en syntagmes

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

Pourtant, l'existence d'une telle organisation hiérarchique ne fait pas de doute. Si une phrase n'était qu'une juxtaposition de mots, il ne devrait pas y avoir une différence d'interprétation entre les phrases suivantes :

(6) Mon frère déteste cet homme mesquin.

(7) Mon frère trouve cet homme mesquin.

Ces deux phrases sont formées de mots appartenant aux catégories syntaxiques identiques :
Déterminant Nom Verbe Déterminant Nom Adjectif
Le fait qu'elles ne s'interprètent pas de la même manière résulte de leur structure respective : les mots ne sont pas regroupés de la même manière dans (6) et (7) et n'entretiennent pas des relations identiques entre eux.

Dans (6), les mots de la séquence cet homme mesquin sont regroupés et forment un constituant :
C'est cet homme mesquin que mon frère déteste.
Cet homme mesquin, Mon frère le déteste.
Cet homme mesquin est détesté par mon frère.
- Qui ton frère déteste-t-il ?
- Cet homme mesquin.

Dans (7), les mots de la séquence cet homme mesquin peuvent être séparés en deux groupes et former ainsi deux constituants distincts :
C'est cet homme que mon frère trouve mesquin
Cet homme, mon frère le trouve mesquin
- Qui ton frère trouve-t-il mesquin ?
- Cet homme.

3.1.1 Tête de syntagme

Soit l'énoncé :

(8) le chien de ma voisine

L'entité désignée par ce groupe est un type de chien, pas une sorte particulière de voisine. On peut donc dire d'une certaine manière que dans ce syntagme, « chien » domine « voisine ». Se faire dominer pas son chien n'est d'ailleurs pas rare, quoique ceux-ci soient en principe plus dociles que les chats. Le constituant dominant d'un syntagme est appelé **tête** du syntagme.

On voit cette notion de syntagme sans la nommer à l'école primaire, lorsqu'on fait des « analyses » de phrases du type :

Tête de syntagme

```
L'étudiant    dort         sur son bureau
Sujet-------> verbe <----- C. C. L.


Paul          donne        une fleur         à Marie
Sujet ------> verbe <----- C.O.D.            C.O.I.
                    <------------------------
```

3.2 Les tests classiques

L'existence de composants dans cette structure hiérarchique est attestée par un certain nombre de faits syntaxiques : différents types de tests permettent de délimiter les constituants de la phrase.

3.2.1 La substitution

Elle consiste à remplacer un élément par un autre, opération permettant la segmentation des unités qui constituent la phrase, donc de dissocier les groupes.

3.2.1.1 La pronominalisation consiste à remplacer par un pronom l'ensemble du syntagme.
Exemple :

(9) Paul envoie à sa sœur la carte postale qu'il a achetée.
   a) Paul lui envoie la carte postale qu'il a achetée.
   b) Paul l'envoie à sa sœur.
   c) Paul la lui envoie.
   d) Paul l'envoie.

   La pronominalisation (9d) n'a pas le même sens que les trois autres : cela montre que « la carte postale » et « à sa sœur » sont deux syntagmes disjoints.

3.2.1.2 La variation paradigmatique entre composants de tailles différentes :
Exemple :

(10) a) Le chien de ma voisine mange.
   b) Le chien mange.

ont manifestement la même structure, ce qui impose de considérer que « chien de ma voisine » dans (10a) joue le même rôle que (i.e. est un constituant syntaxiquement équivalent à) « chien » dans (10b).

Pour identifier un syntagme verbal (SV), on peut le remplacer par un verbe intransitif. Exemple : « Jean met un chapeau moche » ⟶ « Jean dort » permet d'identifier « met un chapeau moche » comme un syntagme verbal.

3.2.2 Le déplacement

3.2.2.1 Détachement Exemple :

(11) a) Le directeur de l'usine reçoit Marc le matin.
   b) Le matin, le directeur de l'usine reçoit Marc.

3.2.2.2 Le clivage

(12) a) Le petit garçon écrit une lettre à sa sœur.
   b) C'est une lettre que le petit garçon écrit à sa sœur.
   c) C'est le petit garçon qui écrit une lettre à sa sœur.
   d) C'est à sa sœur que le petit garçon écrit une lettre.

3.2.2.3 La dislocation (à gauche ou à droite) Il s'agit de détacher un constituant en tête (ou en fin) de phrase, repris (ou annoncé) par un pronom. Ex :

(13) a) Pierre a dessiné les fleurs.
b) Les fleurs, Pierre les a dessinées.
c) Pierre les a dessinées, les fleurs.

(14) a) Pierre a dit que Jeanne était malade.
b) Pierre l'a dit, que Jeanne était malade.

(15) a) Pierre est arrivé à la gare.
b) Pierre y est arrivé, à la gare.

3.2.3 La transformation

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

3.2.3.1 Le passif

(16) Jean a cassé la boîte de Paul.
a) La boîte de Paul a été cassée par Jean.
b) *La boîte a été cassée par Jean de Paul.

(16) ne peut pas se transformer en (16b), ce qui montre que « la boîte de Paul » doit se déplacer en entier : c'est un syntagme.

3.2.3.2 L'Interrogatif

3.2.3.2.1 Interrogation partielle

Seul un syntagme peut servir de réponse à une interrogation partielle.
Exemple :

(17) Le petit garçon écrit une lettre à sa sœur.
a) Qui écrit une lettre à sa sœur ?
b) A qui le petit garçon écrit-il une lettre ?
c) Qu'écrit le petit garçon à sa sœur ?

3.2.3.2.2 Interrogation totale

Les réponses possibles sont « oui » ou « non ». Transformer une affirmation en interrogation totale peut donner un indice sur les constituants.
Exemple :

(18) a) L'homme qui est grand est dans la chambre.
b) L'homme qui est grand est-il dans la chambre ?
c) *L'homme qui est-il grand est dans la chambre ?

La position de « -il » est forcément après le second « est », ce qui montre ici que le premier « est » appartient à un constituant : « qui est grand » et que ce constituant n'est pas un constituant principal de la proposition. En effet, lors de la transformation d'une affirmative en interrogation totale, le « -il » se place après le « verbe principal » : cela montre que les constituants sont organisés selon une structure hiérarchique.

3.2.4 La Conjonction (de coordination)

Dans une phrase quelconque, il est possible de rajouter des éléments par conjonction ; cependant, cette possibilité est fortement contrainte, et l'on ne peut pas effectuer cette opération pour tous les groupes de mots. Ainsi, à partir de la phrase 19, on peut ajouter des coordinations selon (19a), (19b), (19c) et (19d), mais pas selon (19e) et (19f).

(19) *La fille de Minos se repose dans son île*

a) La fille de Minos et de Pasiphaé se repose dans son île.

b) La fille de Minos se repose et tricote dans son île.

c) La fille de Minos se repose dans son île ou dans son bateau.

d) Le fils et la fille de Minos se reposent dans leur île.

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

f) *Le fils de Minos se repose et lamente dans son île.

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

3.3 Organisation hiérarchique des syntagmes

Une phrase n'est pas une simple juxtaposition de mots, mais elle n'est pas une simple juxtaposition de syntagmes non plus. Pour former une phrase, les différents syntagmes entrent dans des réseaux de relations. Prenons la phrase suivante :

(20) Marie pense au calme.

Cette phrase est ambiguë, entre les deux interprétations suivantes :

1. Marie pense à quelque chose ; ce quelque chose c'est le calme.
2. Marie pense et elle le fait au calme.

Cette ambiguïté résulte de la relation que le syntagme prépositionnel « au calme » entretient avec le verbe : dans la première interprétation, c'est un complément (dépendant) du verbe (i.e. complément d'objet indirect dans l'analyse grammaticale). Dans la seconde interprétation, au calme porte sur le prédicat : il en indique localisation spatiale.

Dans l'exemple que nous venons de voir, l'ambiguïté résulte de la réalisation facultative du complément du verbe penser. Mais le même type d'ambiguïté peut être le résultat d'une ambiguïté lexicale (i.e. polysémie d'une unité lexicale ou homonymie entre plusieurs unités lexicales) :

(21) J'aspire au calme

(Publicité pour une marque d'aspirateur)

Pour obtenir une analyse syntaxique adéquate d'une phrase, il faut non seulement identifier ses constituants, mais aussi les relations que ces derniers entretiennent les uns avec les autres. La grammaire traditionnelle parle, dans ce cas, de fonctions grammaticales.

Autres exemples à méditer :

(22) Un étudiant de linguistique aux cheveux longs

(23) Un magazine de Paris sur la mode

(24) Une boutique de mode de Paris

Ces trois syntagmes semblent composés de la même façon : Dét N SP SP (i.e. déterminant, nom, syntagme prépositionnel, syntagme prépositionnel). Remarquez comme l'inversion des deux SP donne des résultats différents : elle rend la première phrase bizarre, change le sens de la deuxième, et est impossible dans la troisième. Remarquez aussi que dans *Un magazine sur la mode de Paris* et dans *Une boutique de mode de Paris*, le mot *mode* semble plus lié tantôt à droite, tantôt à gauche. Si l'on se contente de dire que *de mode* et *de Paris* sont des compléments du nom *boutique*, on ne rend donc pas compte de la totalité de l'intuition linguistique. Un moyen d'en rendre mieux compte est de représenter les syntagmes de manière plus ou moins hiérarchisée.

4 Grammaires génératives et règles de récriture

La plupart des modèles syntaxiques utilisés en TAL sont dérivés plus ou moins directement du concept de grammaires génératives, que nous exposons ici. Dans l'optique générative, une grammaire a pour but de pouvoir **engendrer** *l'ensemble infini des phrases d'une langue*, et ce à l'aide d'un ensemble fini de règles.

Définition d'une phrase

Définition d'une phrase :

La phrase est l'unité privilégiée en syntaxe, l'unité maximale de la description. Elle est une suite d'unités *hiérarchisées* ; ces unités hiérarchisées sont liées entre elles par des règles de "récriture" ou règles "syntagmatiques".

4.1 Règles de réécriture

Pour la grammaire générative, la phrase est un axiome de base, représenté par une suite de symboles. Ces symboles sont engendrés à partir d'un symbole initial noté P (en français), ou S (en anglais).

Exemple de règles de récriture

Règles

1) P $\longrightarrow$ SN SV
2) SV $\longrightarrow$ V
3) SV $\longrightarrow$ V SN
4) SN $\longrightarrow$ Dét Adj. N

Une grammaire constituée de ces règles de récriture peut générer une phrase par une succession de récritures d'une suite de symboles, en partant du symbole initial P :

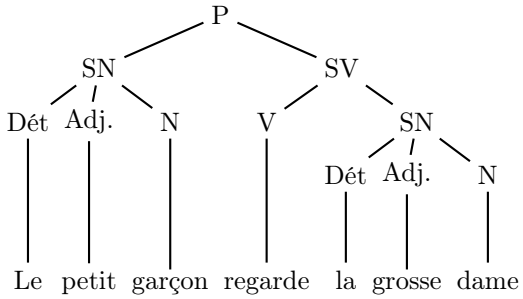
Dérivation

		P						
1	$\Longrightarrow$	SN			SV			
4	$\Longrightarrow$	Dét	Adj.	N	SV			
3	$\Longrightarrow$	Dét	Adj.	N	V	SN		
4	$\Longrightarrow$	Dét	Adj.	N	V	Dét.	Adj.	N
	$\Longrightarrow$	Le	petit	garçon	regarde	la	grosse	dame

La succession des étapes de récriture qui permettent de produire une phrase est appelée une **dérivation**. On peut représenter la dérivation précédente par un arbre.

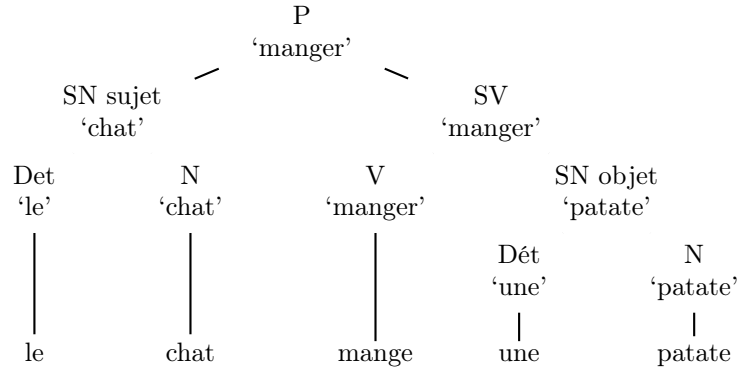
4.2 Les arbres syntaxiques

Représentation sous forme d'arbre



Traditionnellement, le résultat de l'analyse syntaxique est un arbre, forme qui permet de représenter les frontières de constituants, ainsi que les relations de dominance qu'ils entretiennent.

Autre représentation sous forme d'arbre



Au niveau le plus haut de l'arbre, on trouve un nœud étiqueté P, associé au concept <manger>. Ce nœud couvre toute la phrase, traduisant le fait que cet énoncé parle de l'action de manger. Au niveau juste inférieur, on trouve deux **constituants principaux**, l'un étiqueté SN (syntagme nominal), correspondant au constituant « le chat », et associé au concept <chat>, l'autre étiqueté SV, associé à <manger>. C'est donc un chat qui mange ! La lecture se poursuit selon le même schéma en descendant récursivement les branches de l'arbre syntaxique.

4.2.1 Vocabulaire :

SN suj. Est un nœud **non-terminal.**
V est un nœud **préterminal**. (domine le matériel lexical)
Mange est un nœud **terminal**.
P est le nœud **racine.**

Relations de dominance (immédiate ou non) :
P **domine** SN suj., SV, V et SN obj.
P **domine immédiatement** SN suj. et SV.
V est un nœud **non-branchant** (i.e. il domine directement le matériel lexical). Un nœud préterminal est toujours non-branchant (mais l'inverse n'est pas vrai).

4.3 Les parenthèses étiquetées

Lorsque, pour diverses raisons, la représentation en arbre syntaxique n'est pas nécessaire ou praticable, on a recours à un autre de type de représentation, qui comprend les mêmes renseignements : la représentation en parenthèses étiquetées.

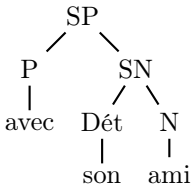
Il s'agit d'assigner une parenthèse (ou crochet) à chaque nœud d'une phrase et d'étiqueter cette parenthèse à l'aide du syntagme ou de la catégorie correspondant. Un SN, par exemple, qui se représente dans un arbre comme : SN

se représentera à l'aide de la parenthèse étiquetée suivante : $[_{SN}$

Tout ce qui se trouve à la droite d'un crochet ouvert est contenu à l'intérieur du nœud correspondant et un crochet fermé indique que le nœud est complet. Si un SP contient un P et un SN :

Il apparaîtra comme : $[_{SP}$ P SN]

Le SP plus développé suivant :



serait donc : $[_{SP}$ $[_{P}$ avec] $[_{SN}$ $[_{Dét}$ son] $[_{N}$ ami]]]

La représentation en parenthèses étiquetées de la phrase : Son chien mange un os dans la cuisine.
$[_{Ph}$ $[_{SN}$ $[_{Dét}$son] $[_{N}$chien]] $[_{SV}$ $[_{V}$mange] $[_{SN}$ $[_{Dét}$un] $[_{N}$os]]] $[_{SP}$ $[_{P}$dans] $[_{SN}$ $[_{Dét}$la] $[_{N}$cuisine]]]]

5 Quelques difficultés du traitement syntaxique

La conception d'analyseurs syntaxiques fiables et rapides est un problème redoutablement ardu. Le syntacticien est en effet confronté à une double contrainte :

Principales difficultés
- lutter contre la prolifération des ambiguïtés ;
- décrire des phénomènes complexes (et subtiles).

Or dans la pratique, ces deux contraintes sont largement contradictoires.

5.1 Ambiguïté lexicale

Ce point a déjà été évoqué : de très nombreuses formes graphiques correspondent à plusieurs « entrées lexicales » différentes, comme les suivantes :

- porte : nom féminin, verbe conjugué
- mousse : nom féminin, nom masculin, verbe conjugué
- la : déterminant, ou pronom personnel féminin singulier, ou nom masculin (note de musique)

Si l'on se limite aux simples catégories syntaxiques de base, environ 50% des mots d'un texte sont ambigus, c'est-à-dire qu'ils correspondent possiblement à plusieurs catégories morphosyntaxiques. Conséquence directe : une phrase de 20 mots a environ 2^{10} interprétations différentes au niveau des étiquettes des feuilles de l'arbre syntaxique.

5.2 Phénomènes de translation

Il s'agit d'un phénomène grammatical qui permet à une forme de changer de catégorie morphosyntaxique. Ainsi, tous les adjectifs, participes passés sont rendus ambigus si l'on veut que la grammaire soit capable de prendre en compte ce phénomène. Ex :

- vache : il est vraiment vache (emploi adjectival d'un nom)
- vert : mangez du vert
- blessé : il est blessé (vient du participe passé : il a blessé), le blessé

5.3 Schémas de sous-catégorisation

De surcroît, la description des phénomènes syntaxiques requiert bien souvent des descriptions lexicales bien plus précises que les simples étiquettes morphosyntaxiques. Prenons l'exemple du verbe « parler ». Ce verbe a en fait quatre variantes syntaxiques, correspondant à des schémas verbaux (schémas de sous-catégorisation) différents. Ces quatre variantes sont attestées par les phrases suivantes :

- Jean parle.
- Jean parle à Marie.
- Jean parle de Paul.
- Jean parle de Paul à Marie.

Ce comportement particulier de « parler » diffère très fortement de celui d'un verbe comme « courir », pour lequel seule la première des constructions précédentes est possible. Cette propriété du verbe parler doit figurer dans sa description lexicale. Sinon, il ne sera pas possible à la fois d'accepter la seconde construction pour « parler » et de la refuser pour « courir ». Du coup, pour chaque occurrence du verbe « parler », il y aura systématiquement quatre interprétations possibles, correspondant aux quatre schémas ci-dessus. Cette ambiguïté est d'ailleurs parfois parfaitement justifiée comme dans « je parle à la maîtresse de Marie », énoncé pour lequel il est impossible, au niveau syntaxique, de choisir entre l'usage du deuxième schéma ou celui du dernier.

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

5.4 Ambiguïtés de rattachement

L'ambiguïté lexicale est aggravée par les ambiguïtés purement syntaxiques, en particulier par les ambiguïtés de rattachement, dont la phrase « je parle à la maîtresse de Marie » est un cas prototypique. Le problème est le suivant : un groupe nominal (resp. verbal) introduit par une préposition (resp. conjonction) peut jouer le rôle de complément du nom comme celui de complément du verbe, ou encore de complément circonstanciel. A une même phrase peuvent donc correspondre plusieurs structures arborées différentes. Ce phénomène est illustré par les quelques exemples qui suivent :

(25) Elle mange une pomme avec les doigts / Elle mange une pomme avec la peau

(26) Elle mange une glace à la fraise / Elle mange une glace à la plage

(27) C'est la fille du cousin qui boit

(28) Il a parlé de déjeûner avec Paul.

En résumé, les phénomènes syntaxiques à décrire sont souvent complexes, et demandent des descriptions lexicales et syntaxiques très fines, qui aggravent plus qu'ils ne résolvent le problème de l'ambiguïté. Ceci explique peut-être pourquoi il n'existe à l'heure actuelle, en dépit de trente années de recherche intensive 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.