

# Syntaxe théorique et formelle

Volume 1 : modélisation, unités,  
structures

Sylvain Kahane

Kim Gerdes

Textbooks in Language Sciences



No series description provided

# Syntaxe théorique et formelle

Volume 1 : modélisation, unités,  
structures

Sylvain Kahane

Kim Gerdes

Kahane, Sylvain & Kim Gerdes. 2020. *Syntaxe théorique et formelle : Volume 1 : modélisation, unités, structures* (Textbooks in Language Sciences). Berlin : Language Science Press.

This title can be downloaded at :

<http://langsci-press.org/catalog/book/241>

© 2020, Sylvain Kahane & Kim Gerdes

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0) :

<http://creativecommons.org/licenses/by/4.0/> 

ISBN : **no digital ISBN**

**no print ISBNs!**

ISSN : 2364-6209

**no DOI**

Source code available from [www.github.com/langsci/241](https://www.github.com/langsci/241)

Collaborative reading : [paperhive.org/documents/remote?type=langsci&id=241](https://paperhive.org/documents/remote?type=langsci&id=241)

Cover and concept of design : Ulrike Harbort

Fonts : Libertinus, Arimo, DejaVu Sans Mono, Source Han Serif

Typesetting software : Xe<sub>La</sub>TeX

Language Science Press

Xhain

Grünberger Str. 16

10243 Berlin, Germany

[langsci-press.org](http://langsci-press.org)

Storage and cataloguing done by FU Berlin

Freie Universität  Berlin

# Table des matières

<b>Avant-propos</b>	<b>iii</b>
1 De quoi parle ce livre ?	iii
2 Arbre de dépendance	v
3 Les questions qui nous préoccupent	vi
4 À qui s'adresse cet ouvrage ?	vii
5 La syntaxe et les autres domaines de la linguistique	ix
6 Grammaire et lexique	xi
7 Notations	xii
8 Le lexique : un cabinet de curiosités	xii
9 Le plan du livre	xiv
10 Les termes grammaire, syntaxe et topologie	xvii
11 Commentaires sur le plan	xviii
12 Notions, termes, concepts et définitions	xix
13 Présentation de l'ouvrage	xxii
14 Remerciements	xxiii
 <b>I Modéliser la langue</b>	 <b>3</b>
<b>1 La langue : L'objet d'étude de la linguistique</b>	<b>7</b>
1.1 Parler une langue	7
1.2 La langue comme correspondance sens-texte	8
1.3 Sons et textes	10
1.4 Langue et variations	11
1.5 Sens et intention communicative	13
1.6 Mots et pensée	15
1.7 Sens lexicaux et traduction	15
1.8 Langue, linguistique et modélisation	16
1.9 Langue et parole, compétence et performance	17
1.10 La planification	19
1.11 Corpus et introspection	22
1.12 Acceptabilité	23

1.13	Parler et comprendre . . . . .	25
<b>2</b>	<b>Produire un énoncé : La syntaxe mise en évidence par un exemple</b>	<b>35</b>
2.1	Analyse et synthèse . . . . .	35
2.2	Les observables : textes et sens . . . . .	35
2.3	Partir d'un sens . . . . .	36
2.4	Graphe et arbre . . . . .	38
2.5	Les composantes du sens . . . . .	40
2.6	Choisir des unités lexicales . . . . .	41
2.7	Les quatre moyens d'expression du langage . . . . .	42
2.8	Contraintes syntaxiques sur les choix lexicaux . . . . .	43
2.9	Règles et exceptions . . . . .	44
2.10	Structure hiérarchique . . . . .	44
2.11	Du sens au texte : de 3D à 1D . . . . .	46
2.12	Ce que la langue nous force à dire . . . . .	47
2.13	Les sens grammaticaux . . . . .	48
2.14	Choisir le verbe principal . . . . .	49
2.15	Les contraintes de la grammaire . . . . .	50
2.16	D'où viennent les règles de la grammaire ? . . . . .	51
<b>3</b>	<b>La modélisation : Préciser l'objectif de notre étude</b>	<b>59</b>
3.1	Définition . . . . .	59
3.2	Modèle d'une langue ou modèle de la langue . . . . .	60
3.3	Modélisation et théorie . . . . .	62
3.3.1	Un exemple de modélisation en physique . . . . .	62
3.3.2	La modélisation de la langue . . . . .	63
3.3.3	Du modèle à la théorie . . . . .	63
3.4	Un exemple — l'accord — de la description à l'explication . . . . .	64
3.5	Modèle déclaratif . . . . .	66
3.6	Modèle génératif, équatif et transductif . . . . .	66
3.7	Modèle symbolique . . . . .	68
3.8	Calcul symbolique et grammaires catégorielles . . . . .	69
3.9	Modularité et stratification . . . . .	71
3.10	La Théorie Sens-Texte . . . . .	72
3.11	Modélisation des langues et ordinateur . . . . .	73

# Avant-propos

DE LA SYNTAXE *Ou la manière de joindre ensemble les parties d'oraison* [= les parties du discours ou catégories lexicales] *selon leurs divers régimes.*

Ces diverses parties font pour ainsi dire par rapport à une langue, ce que font les matériaux, par rapport à un édifice : quelque bien préparés qu'ils soient ils ne feront jamais un palais ou une maison, si on ne les place conformément aux règles de l'architecture. C'est donc la syntaxe qui donne la forme au langage, et c'est la partie la plus essentielle de la grammaire.

Buffier1709, nous modernisons l'orthographe

## 1 De quoi parle ce livre ?

En commençant cet ouvrage, nous souhaitons écrire un ouvrage d'introduction à la SYNTAXE DE DÉPENDANCE. La plupart des ouvrages récents en syntaxe s'appuient sur l'ANALYSE EN CONSTITUANTS qui a dominé la seconde moitié du 20<sup>e</sup> siècle. Et si les grammaires de dépendance ont connu un renouveau et un développement extraordinaire depuis le début du 21<sup>e</sup> siècle, jusqu'à supplanter quasi totalement les grammaires de constituants dans le domaine du traitement automatique des langues (TAL), elles ne sont encore que sporadiquement enseignées à l'université et l'unique ouvrage de référence reste souvent l'ouvrage fondateur de Lucien Tesnière publié en **Tesnière1959**. Les *Éléments de syntaxe structurale* de Tesnière sont un incontestable monument de la littérature scientifique en linguistique, dont on ne peut que recommander la lecture, mais cet ouvrage ne peut évidemment pas prendre en compte les développements importants qu'a connus le domaine depuis 60 ou 80 ans. (Tesnière est mort en 1954, il a été très malade après la guerre et ses idées ont peu évolué depuis l'édition de son polycopié *Esquisses de syntaxe structurale* distribué aux élèves de l'École normale d'institutrices de Montpellier en 1943, voire de son article de 1934 *Comment construire une syntaxe.*)

Au final, le livre que vous avez entre les mains n'est pas un manuel sur la syntaxe de dépendance, dans le sens où il **ne** souhaite **pas** livrer de recettes qui permettront au lecteur d'**apprendre** à associer un arbre de dépendance à n'importe quel énoncé et de discuter les différentes analyses possibles d'une construction donnée **dans un cadre préconçu**. L'objectif de ce livre est au contraire de **s'interroger sur le cadre lui-même**, de mettre en question la validité d'une approche de la syntaxe en termes de dépendance et au-delà de cela de **définir les principes** mêmes qui doivent présider à une construction théorique en syntaxe.

Il en découle que cet ouvrage n'est pas vraiment un ouvrage d'introduction : même s'il tente d'élaborer son objet à partir de rien et qu'il est donc en théorie accessible sans pré-requis, cet ouvrage fait certainement appel à une **maîtrise du raisonnement scientifique** qui ne s'acquière qu'avec l'expérience. On pourra comparer, toutes choses égales par ailleurs, une tentative de ce genre à celle du groupe de mathématiciens français rassemblés sous le pseudonyme de Nicolas Bourbaki qui rédigea un ouvrage de construction des mathématiques à partir de rien (le premier tome démarre par la construction des entiers à partir du seul ensemble vide). L'ouvrage de Nicolas Bourbaki, s'il est élémentaire au sens premier du terme (comme le souligne son titre, *Éléments de mathématiques*), s'avère d'une lecture bien difficile pour des non-mathématiciens.

Ce livre n'est pas non plus vraiment un ouvrage d'introduction à la syntaxe de dépendance, puisqu'une grande partie de l'ouvrage est consacrée à **définir l'objet même d'un ouvrage de syntaxe** et que la dépendance n'y est introduite qu'après une discussion détaillée sur les **unités de base de la syntaxe**. De plus, une large place est faite aux autres **représentations possibles de l'organisation syntaxique** et à la comparaison entre les différents modes de représentation. Nous pensons notamment que ceux qui travaillent en syntaxe de constituants en apprendront beaucoup sur les représentations qu'ils ont l'habitude d'utiliser et sur les **choix qui président à de telles représentations**.

Nous allons préciser l'objectif de cet ouvrage. Avant cela, le lecteur qui n'est pas familier avec la notion d'arbre de dépendance pourra consulter l'encadré qui suit. Dans la suite de l'ouvrage nous mettrons souvent des portions de texte en exergue de cette façon.



## 2 Arbre de dépendance

Tout au long de cet ouvrage, nous proposerons de petits encadrés. Certains anticipent un peu sur la suite de l'ouvrage ; celui-ci permet à un lecteur totalement néophyte d'avoir une première idée de ce qu'est la SYNTAXE DE DÉPENDANCE. De manière générale, les encadrés contiennent des informations complémentaires, généralement plus techniques ou à visée historique, qui ne sont pas essentielles à la compréhension du texte principal.

L'arbre de dépendance est une représentation de la structure syntaxique devenue traditionnelle après la publication en 1959 de l'ouvrage de Lucien Tesnière, *Éléments de syntaxe structurale*, et les différents travaux qui ont suivi, notamment ceux des pragois autour de Petr Sgall, ceux des Russes autour d'Igor Mel'čuk, ainsi que des travaux en Allemagne, en Angleterre ou aux États-Unis (mais étonnamment aucun travail significatif en France jusqu'aux années 1990).

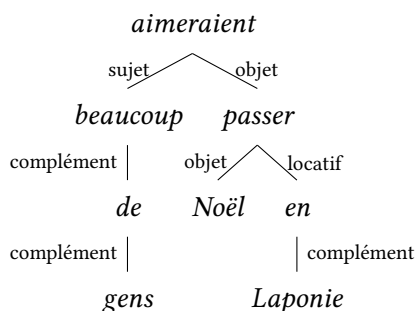
Dès le début de son ouvrage, Tesnière dit :

« Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des **connexions**, dont l'ensemble forme la charpente de la phrase. »

(Voir l'encadré ?? pour un *Historique des notions de dépendance et de tête*, où l'on verra que cette idée est déjà dans un article de l'*Encyclopédie* par Dumas en 1754 et que Tesnière a eu de nombreux prédécesseurs.) Tesnière ajoute ensuite que ces connexions sont orientées, liant un gouverneur à un dépendant, et forment ainsi une structure hiérarchique. C'est ce que Tesnière appelait un STEMMA et qu'on appelle aujourd'hui un ARBRE DE DÉPENDANCE (voir l'encadré ?? sur *Graphe et arbre*). Ainsi l'arbre de dépendance généralement proposé pour une phrase comme :

- (1) Beaucoup de gens aimeraient passer Noël en Laponie.

est :



Dans cette représentation, les mots dépendent les uns des autres. Le mot le plus important, le verbe principal, est *aimeraient*, qui occupe la racine de l'arbre et est placé tout en haut (l'arbre « pousse » à l'envers). Les dépendances sont étiquetées par des relations syntaxiques. Ainsi l'arbre nous dit que le sujet de *aimeraient* est le groupe *beaucoup de gens* et que le mot le plus important de ce groupe est *beaucoup*, qui se retrouve ainsi lié à *aimeraient*.

Nous ne justifierons pas ici cette représentation, dont on peut d'ailleurs contester certains choix. Cela sera très largement discuté dans cet ouvrage et en particulier dans le Chapitre ?? sur *Tête et dépendance*. Il s'agit juste de donner un premier exemple en vue de la discussion qui va suivre.

### 3 Les questions qui nous préoccupent

En écrivant cet ouvrage, nous nous sommes posé un grand nombre de questions auxquelles il nous a semblé qu'il fallait répondre avant de pouvoir présenter de façon objective la syntaxe de dépendance.

- Est-il justifié de représenter la structure syntaxique par un arbre de dépendance ? Par une structure arborescente ? Par des dépendances ? Jusqu'à quel point une représentation basée sur la dépendance est-elle ou non équivalente à d'autres modes de représentation et notamment aux arbres de constituants ?
- Quel est le statut de la structure syntaxique ? Est-ce un objet de la langue ou bien un artefact de la modélisation des langues ? Quel rôle souhaite-t-on donner à de telles structures à l'intérieur du modèle d'une langue ?
- Les dépendances syntaxiques sont-elles entre les mots ? Quelles sont les unités minimales de la syntaxe ? Comment définir le mot et quel rôle joue-t-il dans la syntaxe s'il n'est pas l'unité minimale de la syntaxe ?
- Les dépendances syntaxiques s'arrêtent-elles à la frontière de la phrase ? La notion de phrase est-elle légitime ? Y a-t-il une unité maximale de la syntaxe ?
- De quelles propriétés des énoncés cherche-t-on à rendre compte par un arbre de dépendance ? De quelles propriétés ne rend-on pas compte ? Comment encoder les propriétés qui ne sont pas prise en compte par l'arbre de dépendance ? Quelles sont les autres structures que l'on peut associer à un énoncé ? Quels rapports y a-t-il entre les différentes représentations de la structure d'un énoncé ?
- Finalement, qu'appelle-t-on la syntaxe ? Et en quoi est-il possible ou non, nécessaire ou non, d'introduire une structure syntaxique pour rendre compte

des propriétés syntaxiques d'un énoncé ? La modélisation du lien entre signifiant et signifié a-t-elle besoin d'une structure syntaxique intermédiaire ?

Toutes ces questions nous amèneront à commencer par rappeler les objectifs de notre discipline, la linguistique, et de sa sous-discipline, la syntaxe. Ces objectifs sont pour nous de **construire des modèles** des différentes langues du monde au sein d'une théorie de la langue. Notre première partie sera donc consacrée à définir les objectifs de la modélisation des langues et à montrer l'existence d'un ensemble de propriétés qui relèvent de ce que nous appelons la syntaxe.

On définit traditionnellement la syntaxe comme « l'étude de l'organisation des mots dans la phrase ». Une telle définition est problématique, puisqu'elle suppose que l'on peut définir les notions de mot et de phrase avant de définir ce qu'est la syntaxe. Dans cet ouvrage, la notion de mot ne sera définie qu'au début de la quatrième partie et l'unité maximale de la syntaxe ne sera discutée que dans la sixième et dernière partie de l'ouvrage.

Avant d'étudier l'organisation syntaxique, nous tenterons de **caractériser la syntaxe** et notamment les unités minimales de la syntaxe, que nous contrasterons avec les unités minimales de la morphologie et de la sémantique. Ce sera notre deuxième partie.

La troisième partie montrera comment définir une structure qui rend compte des principales propriétés syntaxiques des énoncés. Différentes structures seront présentées et la représentation par un arbre de dépendance sera particulièrement discutée.

Les trois parties suivantes s'apparentent davantage à un manuel traditionnel. Nous y présenterons différentes caractéristiques des langues et notamment du français et nous présenterons les structures qui en rendent le mieux compte. On notera néanmoins que les catégories syntaxiques et autres parties du discours, qui sont généralement introduites très tôt dans les ouvrages de syntaxe, ne seront réellement définies que dans les quatrième et cinquième parties, quand la question de l'organisation des unités aura été largement discutée.

Nous précisons le plan de cet ouvrage dans la section 9.

## 4 À qui s'adresse cet ouvrage ?

Cet ouvrage aborde la syntaxe comme une **composante d'un modèle linguistique**. L'idée même que la langue puisse être modélisée, comme peut l'être le mouvement des planètes ou le développement du fœtus, n'est pas nécessairement acceptée par tous ceux qui s'intéressent aux langues et prennent du plaisir à les apprendre ou les étudier. Cet ouvrage souhaite montrer qu'on peut déga-

ger de manière méthodique les propriétés des langues, mettre de l'ordre dans la forêt vierge que constitue chaque langue et élaborer un objet théorique qui reproduise certaines propriétés d'un locuteur qui parle et que nous appelons un modèle d'une langue.

Il existe en linguistique, comme ailleurs en sciences humaines, des courants théoriques variés, plus ou moins d'accord entre eux. Cet ouvrage ne se situe pas précisément dans un courant dominant en linguistique, mais il puise largement dans le courant structuraliste qui s'est développé depuis un siècle, des travaux pionniers de Ferdinand de Saussure aux travaux actuels en linguistique formelle, en passant par les travaux précurseurs d'Otto Jespersen, ceux des distributionnistes américains, Leonard Bloomfield en tête, et ceux de Lucien Tesnière et d'Igor Mel'čuk en syntaxe de dépendance. Les auteurs de ce livre ont une formation initiale en mathématiques et ont travaillé dans le domaine du traitement automatique des langues, des grammaires formelles et de la modélisation mathématique des langues et ils enseignent la linguistique à tous les niveaux universitaires. Bien que cet ouvrage ne traite pas directement de formalisation mathématique et d'implémentation informatique, il se place dans le cadre d'une **approche déductive de la langue** dont l'objectif est de construire des modèles qui peuvent être formalisés et implémentés pour simuler un locuteur humain. Ce n'est néanmoins pas l'objectif de ce livre de présenter des modèles de la langue ; ce livre se contente d'introduire, de la façon la plus rigoureuse possible, les notions nécessaires à l'étude de la syntaxe, en se concentrant sur les structures syntaxiques et non sur les règles de la grammaire.

Cet ouvrage a une visée à la fois **scientifique** et **pédagogique** : il a été élaboré avec l'objectif de fournir une base pour l'enseignement de la syntaxe à l'université et de présenter des notions fondamentales pour l'étude des langues, aussi diverses soient-elles. Nous espérons qu'il montre qu'il est utile d'enseigner la syntaxe pour comprendre le fonctionnement des langues et mieux les enseigner et les apprendre. Cet ouvrage souhaite également montrer que la syntaxe est un domaine de recherche vivant et que de nombreuses questions restent ouvertes, même pour des langues très étudiées comme le français.

Cet ouvrage écrit en français est évidemment destiné aux francophones et à ceux qui apprennent le français. Le français constituera donc la langue que nous étudierons par défaut. Bien que cet ouvrage ne soit pas une grammaire méthodique du français, il constitue une bonne introduction à la **syntaxe du français**. Mais il souhaite aussi fournir les outils nécessaires à l'étude d'autres langues, même très éloignées du français. À chaque fois que cela sera utile, nous montrerons le caractère exotique du français dans la **diversité des langues** et présen-

terons pour la notion étudiée un fonctionnement différent de celui du français dans une autre langue.

## 5 La syntaxe et les autres domaines de la linguistique

La syntaxe n'est qu'une partie d'un modèle linguistique, bien que dans beaucoup de théories linguistiques, elle occupe une place centrale ou même dominante par rapport aux autres domaines. Avant de préciser ce qu'est la syntaxe, nous allons situer celle-ci parmi les autres domaines de la linguistique. Nous allons le faire en répondant à différentes questions que l'on peut se poser concernant la langue.

Comment fonctionne l'esprit ? Comment fonctionne le raisonnement ? Comment les modéliser, les imiter ?

Ce genre de questions appartient à la **psychologie**, la **neurologie**, la **logique**, l'**intelligence artificielle**, mais on peut espérer que de nouvelles connaissances en linguistique éclaireront aussi ces sujets. En quelque sorte, la linguistique peut être considérée comme un sous-domaine de chacune des disciplines citées : elle couvre les questions liées à la langue qui se posent dans ces sciences. Parfois, on classe toute cette thématique sous le terme de **sciences cognitives**.

Que veut dire le dernier énoncé que j'ai entendu ? Comment ai-je pu en extraire le sens ? Comment combiner le sens des mots pour former le sens des énoncés ? Ou en quoi la phrase « Paul et Marie porte un chapeau » se distingue-t-elle de la phrase « Paul et Marie porte une machine à laver » ?

Ces questions caractérisent la **sémantique**. Même si elles ne relèvent pas directement de la syntaxe, nous les aborderons à plusieurs reprises, pour mieux délimiter la syntaxe, mais aussi parce que la syntaxe s'articule directement avec la sémantique.

Pourquoi dit-on telle ou telle phrase dans un contexte particulier ? Quelle est la signification d'un énoncé dans un contexte ? Dans des termes plus techniques : quel est le but de l'acte de langage et comment est-il poursuivi ? Concrètement : pourquoi doit-on donner à la question « Vous avez l'heure ? » la réponse « Il est trois heures » et pas la réponse « Oui » ?

Ces questions font partie de la **pragmatique** et se situent au-delà de la sémantique. Nous n'y toucherons pas.

Dans quel ordre doit-on placer les mots ? Quand doit-on utiliser un verbe ou un nom ? A-t-on le droit de coordonner des pronoms interrogatifs ? Quel type de structure forment les mots assemblés en une phrase ?

Celles-ci et beaucoup d'autres questions que nous étudions dans ce livre concernent la **syntaxe**.

Comment représenter et récupérer l'information sur le fonctionnement de chaque mot ? Pourquoi dit-on une peur bleue, mais vert de peur ?

Ces informations sont dans le lexique. La **lexicologie** est la science qui étudie le lexique. Nous aborderons ces questions dans la mesure où les particularités lexicales ont une influence sur la syntaxe et où la frontière entre les constructions syntaxiques et le lexique proprement dit est plus que mouvante. La question est discutée avec plus de détails dans la section suivante.

Comment sont formés les mots ? De quelles entités sont-ils formés ? Pourquoi peut-on dire désespéré et non désattendu ? Comment se forme la 2<sup>e</sup> personne singulier du passé simple pour le verbe aimer ?

La **morphologie** tente de répondre à ces questions. Nous verrons qu'une partie de ces questions relèvent pour nous de la syntaxe (par exemple la conjugaison des verbes).

Quels sont les sons d'une langue ? Quelles sont les combinaisons possibles de ces sons ? Par exemple, pourquoi les Espagnols mettent-ils des e devant chaque mot dont l'équivalent français commence par sp comme especial ? Pourquoi les Français prononcent-ils ze au lieu de the quand ils parlent anglais ? Dans la phrase « Le fromage je n'aime vraiment pas », pourquoi la mélodie monte-t-elle sur le mot fromage ? Comment est représentée la prononciation d'un mot dans le cerveau ?

La **phonologie** s'interroge sur ces questions. Nous les aborderons dans la mesure où la structure phonologique des énoncés nous offre de nombreux indices sur la structure syntaxique des énoncés et où la langue parlée constitue, du notre point de vue, un bien meilleur sujet d'étude que la langue écrite.

Comment sont réellement formés les sons de la langue ? Comment se distingue un son d d'un son z ? Pourquoi y a-t-il une différence entre un son k prononcé devant i et un son k prononcé devant u ? Comment se distingue le son a de l'allemand du son a du français.

S'il s'agit de répondre à ces questions en termes de fonctionnement de la langue et de l'humain produisant les sons, on se trouve dans la **phonétique** (articulatoire), si c'est en termes purement techniques et physiques, c'est plutôt l'**acoustique** et le **traitement du signal**, un domaine relevant de la physique, qui sont concernés.

Il existe d'autres domaines en linguistique, notamment tout ce qui concerne la langue dans ses variations dans le temps et dans l'espace : sociolinguistique, géolinguistique, dialectologie, diachronie, origine du langage, génétique des langues,

grammaticalisation, créolisation, productivité lexicale, acquisition de la langue par l'enfant, apprentissage d'une langue seconde. Il existe aussi de nombreux domaines d'application : le traitement automatique des langues ou TAL (traduction automatique, recherche d'information sur le web ou dans des bases de données, aide aux handicapés, synthèse de la parole), l'enseignement de la langue (Français Langue Étrangère ou FLE pour les apprenants langue seconde, la didactique des langues pour les écoliers francophones), etc. Dans les champs interdisciplinaires, nous trouvons la psycholinguistique, la sociolinguistique, la neurolinguistique, la linguistique textuelle, la linguistique mathématique, etc.

## 6 Grammaire et lexique

Lorsqu'un locuteur veut énoncer une idée dans sa langue, il doit trouver dans son LEXIQUE les unités lexicales qui correspondent le mieux aux entités dont il veut parler. Mais, pour former la phrase, le locuteur a besoin d'autres éléments linguistiques qui sont contraints de diverses façons par la langue. L'étude de ces éléments et des contraintes que la langue impose à un locuteur s'appelle la GRAMMAIRE de cette langue.

En fait, la frontière entre lexique et grammaire n'a rien d'évident. La description d'une langue est la description de chacune des unités de la langue et de la façon dont elles se combinent. Parmi ces unités, on trouve des unités lexicales prototypiques comme CHEVAL ou MANGER, tandis que d'autres sont des unités grammaticales incontestables comme le temps imparfait (*Le cheval mangeait*) ou les différentes réalisations syntaxiques du pluriel (*Les chevaux mangeaient*). Mais d'autres unités, tout en partageant des propriétés avec les unités lexicales prototypiques ont aussi un fonctionnement grammatical, comme CHOSE ou FAIRE (*La **chose** que je préfère, c'est manger ; Ce que je préfère **faire**, c'est manger*). Inversement, des unités qui ont un fonctionnement a priori grammatical, comme la préposition À dans *Elle parle à Pierre*, auront un fonctionnement beaucoup plus lexical lorsqu'elles commutent avec d'autres unités : *Elle est à la maison, dans la maison, devant la maison, derrière la maison*, etc.

La distinction entre lexique et grammaire est orthogonale à la partition du modèle linguistique entre morphologie, syntaxe et sémantique. Toutes les unités, qu'elles soient lexicales ou grammaticales, possèdent une forme, un sens et une combinatoire qu'il faut décrire (voir la section ?? sur *Signifié, signifiant, syntactique*). La SYNTAXE est l'étude de la combinatoire des unités lexicales et grammaticales et tout particulièrement des combinaisons libres obéissant à des règles générales. La syntaxe se trouve à mi-chemin de la SÉMANTIQUE qui s'intéresse

au sens des unités lexicales et grammaticales et des énoncés qu'elles forment en se combinant et de la MORPHOLOGIE qui s'intéresse à la **forme** et la structure des unités que la syntaxe combine.

## 7 Notations

Nous notons nos *exemples linguistiques* en italiques. Pour les UNITÉS LEXICALES, nous utilisons des petites capitales. Lorsqu'il s'agit d'unités lexicales multi-mots comme  $\boxtimes$ POMME DE TERRE $\boxtimes$ , nous utilisons des balises angulaires. Les unités grammaticales sont quant à elles généralement désignées par des termes métalinguistiques : présent, singulier, féminin, etc. Le sens d'une unité lexicale ou d'une portion de texte est indiqué en guillemets simples : 'cheval', 'pomme de terre', 'le cheval mange'. La signification d'une unité grammaticale est également noté entre guillemets, mais en mettant le terme en petites capitales : 'SINGULIER'. On ne confondra pas le sens grammatical 'PRÉSENT' qui signifie 'ayant lieu maintenant' avec le sens lexical 'présent'.

## 8 Le lexique : un cabinet de curiosités

Bien que la syntaxe et la grammaire soient au centre de cet ouvrage, on ne peut pas ne pas évoquer la complexité lexicale dans un tel ouvrage. Nous allons en donner trois exemples.

Chaque verbe impose à ses compléments une construction particulière : *manger quelque chose*, *parler à quelqu'un de quelque chose*, *donner quelque chose à quelqu'un*, *compter sur quelqu'un*, *aller quelque part*, *poser quelque chose quelque part*, etc. Ces constructions se comptent en dizaines. Même lorsque ces constructions semblent similaires, comme *parler à quelqu'un* et *penser à quelqu'un*, elles peuvent différer par leur comportement : ainsi, *à Marie, je lui parle, j'y pense* ou *je pense à elle*, mais on ne pourra pas dire *\*je lui pense* ou *??je parle à elle* (pour l'utilisation des symboles \* et ??, voir la section 1.12 sur l'Acceptabilité). Dans certains cas, le verbe contraint tellement son complément que seules quelques formes sont acceptables. C'est le cas par exemple de la tournure verbale *y comprendre quelque chose*, qui n'est possible qu'avec les compléments suivants : *Je n'y comprends rien*, *Je n'y comprends pas grand-chose*, *Que puis-je y comprendre?*, *Y comprends-tu quelque chose?* et *J'y comprends que dalle*. Il est impossible d'avoir un groupe nominal référentiel comme complément : *\*J'y comprends une chose intéressante*. La liste des compléments possibles de cette acception de COMPRENDRE constitue ainsi un véritable cabinet de curiosités avec un pronom in-



terrogatif (QUOI et sa forme atone *que*), un pronom négatif (RIEN), deux pronoms indéfinis —  $\square$ QUELQUE CHOSE $\square$  qui n'est possible ici qu'avec l'interrogation et GRAND-CHOSE qui est toujours accompagné de la négation — et enfin  $\square$ QUE DALLE $\square$ . Notons que d'autres tournures verbales possèdent quasiment la même complémentation : *Ça ne rime à rien*, *Ça ne rime pas à grand-chose*, *À quoi ça rime?*, mais pas *\*Ça rime à une chose intéressante* ou *\*Ça rime à faire ça*.

Les exceptions lexicales sont encore plus nombreuses quand on se rapproche de la grammaire. Le français possède par exemple plusieurs éléments négatifs qui se construisent avec *ne* : *Je ne dors pas*, *Je ne dors plus*, *Je ne dors jamais*, *Je ne dors nulle part*, *Je ne mange rien*, *Je ne parle à personne*, *Je n'ai aucun problème*, *Je n'ai qu'une idée*. Chacun de ces éléments possède des propriétés syntaxiques différentes : par exemple JAMAIS peut être déplacé, mais pas PAS ou PLUS : *Jamais je ne dors* vs *\*Plus je ne dors*. JAMAIS et PLUS peuvent être combinés, mais pas JAMAIS et PAS : *Jamais plus je ne dormirai*, *Je ne dormirai plus jamais* vs *\*Je ne dormirai pas jamais*, *\*Je ne dormirai jamais pas*. Notons encore que RIEN et PERSONNE se placent différemment par rapport au verbe : *Je n'ai vu personne*, *Je n'ai rien vu*. Sans aller plus loin, on aura compris que chacun de ces éléments négatifs nécessitera une étude séparée simplement pour déterminer ses propriétés combinatoires, c'est-à-dire sa « syntaxe ». Il en va de même de chacun des pronoms interrogatifs ou de chacun des pronoms relatifs et ainsi de la plupart des unités lexicales ayant un rôle grammatical.

Terminons par l'exemple des CONSTRUCTIONS, ainsi que l'on nomme les configurations qui possèdent un rôle grammatical. Il existe en français une construction très employée, le présentatif  $\square$ IL Y A ... QU- $\square$ , pratiquement obligatoire à l'oral lorsque le sujet est indéfini :

(2) Il y a quelqu'un qui nous regarde depuis la fenêtre.

(3) Il y a des choses que j'ai achetées là-bas sur la table.

Le présentatif peut être combiné avec la restriction en  $\square$ NE ... QUE $\square$  :

(4) Il n'y a que des choses que j'ai achetées là-bas sur la table.

La combinaison du présentatif avec la restriction s'applique à des compléments indirects :

(5) Il n'y a qu'à un endroit qu'on les trouve.

(6) Il n'y a qu'à elle que je pense.

alors que le présentatif seul ne le peut pas :

(7) \*Il y a à un endroit **qu'**on les trouve.

(8) \*Il y a à quelqu'un **que** je pense.

Lorsque le présentatif s'applique à un complément avec possessif, on peut exprimer cette possession dans la forme même du présentatif :

(9) Il y a mon frère **qui** doit venir.

(10) J'ai mon frère **qui** doit venir.

Enfin, la même configuration peut être utilisée pour introduire un complément de temps avec le sens 'depuis' :

(11) Il y a une semaine **qu'**on ne s'est pas vu.

Dans ce cas, elle possède une variante, mais celle-ci n'est possible que pour les compléments de temps :

(12) Ça fait une semaine **qu'**on ne s'est pas vu.

(13) \*Ça fait des choses **que** j'ai achetées là-bas.

Comme on le voit, on retrouve pour ces constructions de nombreuses **idiosyncrasies** qui justifient de leur donner une description détaillée, au même titre que les autres unités lexicales.

## 9 Le plan du livre

Ce livre est divisé en six parties que nous avons esquissées à la section 3 sur *Les questions qui nous préoccupent*. Nous allons préciser le plan du livre.

La première partie explique en quoi consiste une **LANGUE** et la **MODÉLISATION** de cette langue (Chapitre 1) et quelles sont les caractéristiques du modèle linguistique que nous construisons (Chapitre 3). Cette première partie permet donc de comprendre quel est le cadre théorique de cet ouvrage et avec quel objectif nous souhaitons mener notre étude de la langue et de sa syntaxe. La modélisation est illustrée par l'exemple de la production d'un énoncé (Chapitre 2).

La deuxième partie pose la question des unités minimales de la langue (Chapitre ??). Étudier la combinatoire des unités qui constituent les énoncés ne peut

se faire qu'après avoir identifié les unités qui se combinent et notamment les unités minimales. Nous montrons qu'il est nécessaire de considérer trois types d'unités minimales : les MORPHÈMES ou unités minimales de forme (Chapitre ??), les SÉMANTÈMES ou unités minimales de sens (Chapitre ??) et les SYNTAXÈMES ou unités minimales de la syntaxe, c'est-à-dire de la combinatoire libre. La distinction de trois types d'unités résulte de la non-correspondance entre les unités de forme et de sens. Prenons un exemple :

(14) *Les étudiants m'ont donné un coup de main.*

Il y a dans cet énoncé plusieurs sémantèmes qui sont exprimés par une combinaison de morphèmes : *coup de main* bien sûr, qui ne signifie pas ici un coup de la main, mais aussi *étudiant*, qui combine le radical du verbe ÉTUDIER avec le morphème *-ant* ou encore le passé composé exprimé par le verbe AVOIR combiné avec le morphème de participe passé *-é*.

La troisième partie introduit les UNITÉS SYNTAXIQUES et la façon dont celles-ci se combinent pour former la STRUCTURE SYNTAXIQUE. On y définit la syntaxe comme l'étude des COMBINAISONS LIBRES d'unités et on caractérise plus précisément le syntaxème (Chapitre ??). Nous montrons que les différentes fragmentations d'un énoncé en unités syntaxiques définissent un graphe que nous appelons la STRUCTURE DE CONNEXION et qui décrit les combinaisons entre syntaxèmes (Chapitre ??). On peut en plus hiérarchiser cette structure en considérant la notion de TÊTE d'une unité syntaxique et obtenir ainsi une STRUCTURE DE DÉPENDANCE (Chapitre ??). On montre comment représenter cette structure de manière plus ou moins équivalente par un ARBRE DE CONSTITUANTS (Chapitre ??). On s'intéresse ensuite au lien d'une part entre la structure et le texte et d'autre part entre la structure syntaxique et le sens. Le premier cas concerne l'« ordre des mots », c'est-à-dire à la façon dont les syntaxèmes s'ordonnent les uns par rapport aux autres et se regroupent pour former des CONSTITUANTS TOPOLOGIQUES au sein de la STRUCTURE TOPOLOGIQUE (Chapitre ??). Le deuxième cas concerne l'INTERFACE SÉMANTIQUE-SYNTAXE, c'est-à-dire à la façon dont les sémantèmes se combinent, ce que décrit la STRUCTURE SYNTAXIQUE PROFONDE, qui rend compte de la distinction entre ACTANT et MODIFIEUR et des restructurations parfois complexes entre la représentation sémantique et la structure syntaxique (Chapitre ??).

Les trois parties suivantes présentent les trois grands domaines de la syntaxe, que nous appelons nanosyntaxe, microsyntaxe et macrosyntaxe, et les principales notions de la syntaxe seront présentées : le mot, les catégories flexionnelles, les catégories lexicales, les fonctions syntaxiques, la phrase. Les princi-

pales constructions seront également étudiées, et notamment les listes, l'extraction et l'organisation des énoncés autour d'un noyau.

La quatrième partie de ce livre est donc consacrée à la NANOSYNTAXE ou MORPHOSYNTAXE. Elle présente les combinaisons de SYNTAXÈMES possédant une très grande cohésion, dont les composantes sont indissociables et se situent à l'intérieur ou à la frontière des mots. Elle inclut la SYNTAXE FLEXIONNELLE et la syntaxe des PARTICULES, c'est-à-dire la syntaxe de tous les éléments qui sont des marqueurs grammaticaux et qui possèdent très peu d'indépendance syntaxique. Nous montrons en particulier que le MOT n'est qu'un degré particulier dans l'échelle de cohésion des combinaisons de syntaxèmes en unités syntaxiques, même s'il constitue une unité naturelle et l'unité qui a été privilégiée pour la transcription écrite de nombreuses langues (Chapitre ??). Nous terminons cette partie par une première classification des unités minimales de la syntaxe et introduisons les CATÉGORIES FLEXIONNELLES (Chapitre ??) et les CATÉGORIES NANOSYNTAXIQUES de LEXÈMES (Chapitre ??).

La cinquième partie est consacrée à la MICROSyntaxe, c'est-à-dire la syntaxe de rection : la RECTION se caractérise par une relation hiérarchique avec des contraintes de réalisation imposées par un gouverneur à ses dépendants. Elle constitue la syntaxe *par excellence*. Nous introduisons la SYNTAXE DE DÉPENDANCE DE SURFACE et la distinction entre les propriétés fonctionnelles et catégorielles des éléments d'un énoncé. Nous étendons le classement des syntaxèmes à l'ensemble des unités syntaxiques et étudions les CATÉGORIES MICROSyntaxiques et le rôle de la TRANSLATION (Chapitre ??). Le classement des syntagmes nous amène à introduire la notion de RELATION SYNTAXIQUE. Les différentes FONCTIONS SYNTAXIQUES que peut remplir une unité sont caractérisées et notamment la fonction *sujet*, qui pose problème, dès qu'on prend en compte les langues ergatives (Chapitre ??). Une attention particulière est portée aux LISTES ou ENTASSEMENTS PARADIGMATIQUES : nous regroupons sous ce terme différents phénomènes, allant de la coordination à la reformulation, où plusieurs éléments viennent occuper une même position régie (Chapitre ??). L'étude détaillée de l'EXTRACTION et du rôle complexe joué par des éléments tels que les pronoms relatifs constitue le dernier chapitre de cette partie (Chapitre ??).

La sixième partie est consacrée à la MACROSyntaxe, c'est-à-dire l'étude de tout ce qui se situe au-delà de la microsyntaxe, notamment les éléments associés au noyau central de l'énoncé sans pour autant être régis. Nous y définissons l'UNITÉ ILLOCUTOIRE qui constitue l'unité minimale du discours et étudions son organisation interne en NOYAU et ADNOYAUX. Nous montrons pourquoi la notion traditionnelle de PHRASE est problématique, notamment parce qu'il y a de la rection

au-delà des limites de l'unité illocutoire et qu'à l'inverse des unités non régies peuvent venir s'insérer à l'intérieur d'une unité illocutoire (Chapitre ??).

## 10 Les termes grammaire, syntaxe et topologie

Le terme grec *grammatikē technē* (γραμματική τέχνη) désignait « l'art des lettres », *lettre* dans le sens de « l'écrit » ; c'était donc l'étude de l'écrit (et l'étude de sa lecture). Elle s'est ensuite développée en science de l'interprétation des textes, ce que, aujourd'hui, on classe plutôt sous le terme de PHILOGIE et que l'on distingue de la GRAMMAIRE. Il y a 3000 ans, bien avant les Grecs, il existait déjà des études grammaticales (au sens moderne du terme) de langues telles que le sanskrit ou le chinois et au 5<sup>e</sup> siècle avant notre ère, le grand linguiste indien Pāṇini a développé une analyse systématique de la nanosyntaxe du sanskrit. Les *Institutiones grammaticae*, écrites au 6<sup>e</sup> siècle par le grammairien latin Priscien, auront une influence déterminante sur le développement de la grammaire en Europe. Beaucoup de termes grammaticaux encore en utilisation aujourd'hui proviennent de l'étude du latin, *lingua franca* jusqu'à la fin du moyen-âge.

Le terme *syntaxis* est également grec : il est composé de *syn* (συν) 'ensemble' et de *taxis* (τάξις) 'ordre, arrangement'. La SYNTAXE a désigné l'étude de l'ordre des mots, puis plus largement l'étude de l'organisation des mots dans la phrase. Le terme allemand pour syntaxe est *Satzlehre*, tout simplement la 'science de la phrase'. Dans la conception traditionnelle, l'analyse syntaxique est limitée d'un côté par le mot, et de l'autre par la phrase : les structures en deçà du mot ne font pas partie de la syntaxe, elles appartiennent à la morphologie ; le discours, l'enchaînement de plusieurs phrases, n'est pas le sujet de la syntaxe, mais de la linguistique des textes ou ANALYSE DU DISCOURS. Notre définition de la syntaxe ne présuppose ni la notion de mot, ni celle de phrase, et considère la MORPHOLOGIE comme l'étude de la **combinaison des signifiants des signes** (voir la section ?? sur *Signifié, signifiant, syntactique*).

Les termes *microsyntaxe* et *macrosyntaxe* ont été forgés en 1990 par les linguistes qui se sont intéressés aux productions orales spontanées (notamment à Aix-en-Provence autour de Claire Blanche-Benveniste et à Fribourg en Suisse autour d'Alain Berrendonner) et ont vu la difficulté que pouvait poser une segmentation en phrases comme à l'écrit. Sur le même modèle, nous proposons le terme *nanosyntaxe*, à la place du terme *morphosyntaxe*, pour compléter la partition de l'étude de la syntaxe. Nous introduisons le terme *syntaxème* pour nommer les **unités minimales** de la syntaxe, sur le même modèle que les termes *morphème* et *sémantème*, désignant respectivement les unités minimales de forme et de sens.

Dans l'usage du terme *syntaxe*, on est passé de l'étude de l'ordre des mots à, aujourd'hui, l'étude de la combinaison des syntaxèmes et aux structures hiérarchiques qui en résultent. Une des conséquences est qu'il fallait réintroduire un nouveau terme pour l'étude de l'**ordre des mots** et des syntaxèmes. Nous avons adopté le terme *topologie*, également d'origine grecque : la TOPOLOGIE est l'étude des *topos* (τόπος), c'est-à-dire l'étude des lieux. Le terme a été introduit au 19<sup>e</sup> siècle par les linguistes décrivant l'ordre des mots des langues germaniques à l'aide de gabarits de places : le **modèle topologique** de l'allemand décrit la façon dont la phrase allemande peut être décomposée en cinq **champs**, les deuxième et quatrième champs modélisant les positions réservées aux verbes (voir le Chapitre ?? sur *La topologie*).

## 11 Commentaires sur le plan

Le plan de cet ouvrage amène plusieurs commentaires et permet déjà de se faire une idée des principaux partis pris.

On constatera tout d'abord que nous considérons trois structures « syntaxiques » — la structure topologique, la structure de dépendance de surface et la structure syntaxique profonde — là où la plupart des approches n'en considèrent qu'une : la structure syntagmatique ou analyse en constituants immédiats. Nous aurons plusieurs fois l'occasion de justifier le fait de séparer les informations qui peuvent l'être et donc de dissocier les différents modes d'organisation des différents types d'unités qui apparaissent dans un énoncé.

Autre point : nous donnons une place importante aux interfaces, c'est-à-dire à la correspondance entre les différents niveaux de représentation de l'énoncé. Ceci est principalement dû à notre objectif de modélisation : nous ne souhaitons pas seulement montrer comment un énoncé est structuré, mais aussi comment un locuteur produit ces différentes structures et quels rôles elles jouent dans la production des énoncés. Modéliser la langue, c'est pour nous modéliser comment une personne parle, c'est-à-dire comment elle produit des énoncés dans sa langue en fonction du message qu'elle souhaite communiquer.

La plupart des ouvrages de syntaxe et des cours de linguistique à l'université commencent par l'étude des catégories syntaxiques ou parties du discours, c'est-à-dire la caractérisation de ce qu'est un verbe, un nom, un adjectif, etc. Nous pensons pour notre part qu'une bonne définition des catégories syntaxiques ne peut se faire qu'après avoir dégagé la structure des énoncés et que la caractérisation des catégories repose sur l'analyse distributionnelle des unités syntaxiques à l'intérieur des combinaisons complexes dans lesquelles elles entrent. Autrement

dit, on ne peut caractériser les catégories syntaxiques d'une langue par l'étude du simple enchaînement linéaire des unités dans la chaîne parlée : il faut prendre en compte les relations plus complexes qui lient les éléments d'un énoncé et dont l'ordre des mots en surface n'est qu'une projection. La description des liens syntaxiques est au cœur de la syntaxe de dépendance et elle occupe une place centrale dans ce livre. Les catégories syntaxiques seront définies en deux chapitres (4.3 et 5.1) distribués dans les parties consacrées à la nanosyntaxe et à la microsyntaxe.

Cette présentation assez systématique des notions utiles à la syntaxe nous a obligé à définir de nouveaux concepts ou à revoir certains concepts traditionnels. Selon les cas, nous avons décidé d'utiliser un terme déjà usuel dans un sens un peu différent ou bien nous avons forgé un nouveau terme. Parmi les néologismes que contient cet ouvrage, on notera le terme *syntaxème* utilisé pour nommer les unités minimales de la syntaxe (qui curieusement ne sont jamais nommées) et *nanosyntaxe* pour désigner la syntaxe des éléments qui possèdent peu d'autonomie syntaxique. Par ailleurs, d'autres termes peu fréquents dans les manuels de syntaxe occupent ici une position plus centrale, comme *sémanème*, *topologie*, *macrosyntaxe* ou *syntaxe profonde*.

## 12 Notions, termes, concepts et définitions

Nous allons illustrer la distinction entre notions, termes, concepts et définitions à partir d'un exemple. Considérons la DÉFINITION suivante :

- (15) Une *phrase* est un segment de texte qui se trouve entre deux ponctuations majeures successives.

Un TERME est introduit : *phrase*. Ce terme est associé à un CONCEPT. Le concept est un objet abstrait, conceptuel, qui n'est accessible qu'à travers la définition qui le caractérise, à savoir « un segment de texte qui se trouve entre deux ponctuations majeures successives ». En associant ce terme et ce concept, nous construisons une notion. Une NOTION est donc un concept nommé ou un terme défini et associé à un concept.

La distinction entre le terme et le concept qui lui correspond est essentielle, puisqu'un même terme peut être associé par différents auteurs à différents concepts. Évidemment, une fois que nous avons associé le terme *phrase* à un concept, nous pouvons parler de « la notion de phrase », mais il faut être conscient qu'il s'agit d'un raccourci pour désigner « la notion que nous avons nommée *phrase* et qui

ne doit pas être confondue avec une autre notion que d'autres ont pu également nommer *phrase* et qui peut à l'inverse avoir été nommée autrement par d'autres ».

Il y a plusieurs raisons pour lesquelles, en linguistique, la terminologie n'est pas bien stabilisée et un même terme tend à désigner des concepts divers. Une première raison est que beaucoup de ces termes (*mot, phrase, nom, adverbe, sujet*, etc.) s'appliquent à des notions qui sont enseignées dès l'école et reçoivent donc des définitions simplifiées et facilement accessibles, qui deviennent inopérante lorsqu'un véritable cadre théorique est développé. Une deuxième raison est qu'il n'est pas possible de définir proprement de telles notions sans se doter d'un appareil conceptuel complexe et qu'il n'y a pas aujourd'hui de théorie consensuelle sur la nature de la langue et sur les primitives conceptuelles, c'est-à-dire les notions de base à partir desquels des notions plus complexes pourront être définies.

Dans cet ouvrage, nous allons nous attacher à introduire un appareillage théorique rigoureux. Nous introduirons un grand nombre de concepts auxquelles nous associerons bien sûr des termes. Nous avons fait le choix d'utiliser autant que possible les termes courants en linguistique, même lorsque nous décidions de définir une notion un peu différente de la tradition. C'est par exemple le choix que nous avons fait pour le terme *syntaxe*, auquel nous donnons une acception différente de la tradition. Lorsque nous avons introduit des concepts nouveaux, nous nous sommes permis d'utiliser un terme vacant s'il n'était pas trop éloigné : c'est ce que nous avons fait avec le terme *substantif*, qui désignait avant ce qu'on appelle aujourd'hui *nom* et que nous avons attribué à une notion du même ordre, mais différente de celle du nom (voir Chapitre ?? sur *Les catégories micro-syntaxiques*). Lorsqu'aucun terme ne se présentait, nous nous sommes résolus à forger un nouveau terme. Nous avons alors cherché à régulariser la terminologie. C'est ce qui nous a amené à introduire le terme *syntaxème* à côté des termes *morphème* et *sémantème*, ou à introduire le terme *nanosyntaxe* à côté des termes *microsyntaxe* et *macrosyntaxe*.

On peut être en désaccord avec la définition d'une notion. Il est important de voir que ce désaccord peut se situer à trois niveaux bien différents. Revenons sur la notion qui illustre cette section. On peut être en désaccord :

- au **niveau proprement définitionnel** : on peut considérer que notre définition n'en est pas vraiment une, car elle comprend des termes qui n'ont pas été eux-mêmes définis. Qu'appelle-t-on une ponctuation majeure ? Le point-virgule est-il une ponctuation majeure ? Qu'entend-on exactement par un segment de texte ? S'agit-il juste d'une chaîne de caractères ou bien



la phrase est-elle un signe linguistique avec un sens associé au texte proprement dit ?

- au **niveau théorique ou conceptuel** : la notion définie est-elle intéressante d'un point de vue théorique ? Certainement pas si on n'élimine pas le cas des points qui suivent une abréviation, comme dans *Georges W. Bush*. Et au-delà, les unités définies par la ponctuation sont-elles bien des unités linguistiques pertinentes ?
- au **niveau purement terminologique** : est-ce bien à ce concept que l'on veut associer le terme *phrase* ? Ou inversement, est-ce bien par le terme *phrase* que l'on veut désigner ce concept ?

Les problèmes terminologiques sont secondaires : un mauvais choix de terme ne remet pas en cause une théorie. Mais ils peuvent être catastrophiques du point de vue pédagogique et rendre incompréhensible une construction théorique valable par ailleurs. Lorsqu'on introduit un concept, on a principalement deux options terminologiques : utiliser un terme existant, avec le risque que d'autres auteurs l'utilisent avec une acception différente (c'est le cas avec le terme *phrase*), ou bien forger un nouveau terme, avec le risque d'avoir des termes « barbares » et difficiles à mémoriser. On peut par exemple proposer d'appeler le concept défini plus haut *phrase graphique*. C'est ce que nous ferons dans la suite de l'ouvrage.

Une *phrase graphique* est une unité de l'écrit, un segment de texte qui se trouve entre deux ponctuations majeures successives. Les ponctuations majeures sont le point (à l'exception des points utilisés dans les abréviations), le point d'interrogation, le point d'exclamation et les trois petits points. Le point virgule segmente une phrase en sous-phrases.

Les questions théoriques et conceptuelles sont les plus importantes : il est bien sûr crucial d'introduire les bons concepts. C'est l'ensemble des concepts introduits qui définit le cadre théorique à partir duquel un modèle d'une langue particulière pourra être élaboré. La notion de phrase que nous avons introduite est-elle vraiment la plus intéressante du point de vue linguistique ? Nous verrons que non, ne serait-ce que parce que l'écrit est une transcription de la langue et qu'une notion de l'ordre de la phrase existe indépendamment de la possibilité d'écrire ou pas (voir la section 1.3 sur les *Sons et textes*).


Les problèmes définitionnels sont immenses. En général, la définition d'une notion fait appel à d'autres notions. Il faut donc comprendre quels sont les concepts qui doivent être définis avant les autres. Dans notre exemple, nous avons défini la phrase à partir de la ponctuation. Mais comment un locuteur sait-il où mettre des ponctuations majeures quand il écrit ? Produit-on des unités de l'ordre de la phrase lorsqu'on parle ? Si oui, il existe une unité de l'ordre de la phrase plus


fondamentale que celle que nous venons de définir et qui ne se définit pas en fonction de l'écrit et encore moins en fonction de la ponctuation. Si beaucoup d'ouvrages de linguistique commence par la définition de la phrase, nous considérons pour notre part qu'il s'agit d'une notion complexe qui ne peut être définie qu'après avoir introduit la notion de cohésion syntaxique. C'est la raison pour laquelle les notions proches de la phrase seront seulement discutées à la fin de cet ouvrage, dans la sixième et dernière partie. Nous verrons de ce point de vue que derrière la phrase graphique se cache en fait deux types d'unités différentes, ce qui explique qu'il y a différentes façons de ponctuer une même production linguistique.


## 13 Présentation de l'ouvrage


Cet ouvrage est volontairement découpé en sections n'excédant généralement pas une page. Nous avons essayé de donner à chacune de ces sections autant d'autonomie que possible, de manière à rendre une lecture non linéaire de l'ouvrage aussi facile que possible. Il est néanmoins évident que ce livre a été organisé selon un ordre mûrement réfléchi et que de nombreuses sections ne peuvent être lues sans avoir lu avant les sections qui introduisent certaines notions préalables indispensables.


Certaines sections sont encadrées et présentées dans un style particulier. Ces **encadrés** sont des prolongements du texte principal de différentes natures. Nous avons cinq principaux types d'encadrés :

 des encadrés d'**éclairage**, prolongeant des points abordés dans le texte principal ;

 des encadrés **historiques**, sur l'origine de certains termes ou de certains concepts ;




 des encadrés **techniques**, présentant une élaboration de certaines notions, notamment du côté de la modélisation mathématique ;

 des encadrés de **typologie linguistique**, montrant diverses réalisations d'un phénomène ou d'une propriété au travers de la diversité des langues ;

 des encadrés sur le **français**, qui reste la langue privilégiée pour illustrer notre propos.

A ces encadrés qui figurent dans le corps du texte s'ajoutent encore quatre autres types d'encadrés placés en fin de chapitre :

 un encadré d'**exercices** ;

-  un encadré contenant des **éléments de correction** de nos exercices ;
-  un encadré de **lectures additionnelles**, comprenant en particulier les références citées au cours du chapitre ;
-  un éventuel encadré de **citations originales**, lorsque nous avons cité des auteurs qui n'avaient pas écrit en français.

## 14 Remerciements

Nous remercions les collègues et doctorants qui ont lu des parties du manuscrit et nous ont fait des commentaires qui nous ont parfois amené à réécrire des parties importantes : Nicolas Mazziotta, Marie-Sophie Pausé, Paola Pietrandrea, Rafaël Poiret.

Nous remercions nos étudiants de la licence de sciences du langage et du master TAL sur qui nous avons testé une grande partie du contenu de cet ouvrage et qui par leurs réactions parfois critiques nous ont permis d'améliorer grandement le texte et d'ajuster le plan de l'ouvrage.

Le contenu de cet ouvrage a fait l'objet de plusieurs articles et communications dans des colloques internationaux, notamment aux conférences bi-annuelles MTT (Meaning-Text Theory), créée en 2003 par Sylvain Kahane et Alexis Nasr, puis Depling (Dependency Linguistics) créée en 2011 par Kim Gerdes, Eva Hajičová et Leo Wanner. Nous remercions les collègues qui nous ont permis de développer nos idées en relisant et critiquant nos articles lors des soumissions ou en nous posant des questions lors des présentations.

Nous remercions également les auteurs qui nous ont précédé et dont la lecture nous a inspiré. Le travail scientifique est cumulatif et la part d'innovation est toujours plus faible qu'on ne le pense. Nous remercions en particulier Claire Blanche-Benveniste, José Deulofeu, Nicolas Mazziotta, Igor Mel'čuk et Alain Polguère avec qui nous avons eu la chance de collaborer et de discuter différents points qui sont développés dans l'ouvrage.



### Exercices

**Exercice 1** Quelles sont, à côté de la syntaxe, les autres composantes d'un modèle linguistique ?

**Exercice 2** Les notions de *syntaxe* et *grammaire* sont souvent confondues. Pouvez-vous donner un élément du modèle d'une langue qui relève de la grammaire mais pas de la syntaxe ? Et un élément qui relève de la syntaxe mais pas de la grammaire ?

**Exercice 3** Qu'appelons-nous la topologie ? L'interface sémantique-syntaxe ? la nanosyntaxe ?

**Exercice 4** Pourquoi pensons-nous qu'il est difficile de commencer un ouvrage de syntaxe en définissant les parties du discours ?

**Exercice 5** De quel type est le premier encadré de cet avant-propos ? Que représente le symbole choisi ?

**Exercice 6** Les mots *syntaxe* et *grammaire* ne sont pas seulement utilisés pour désigner les concepts considérés dans cet ouvrage.

a) Distinguer les emplois du mot *grammaire* dans les phrases suivantes :

- Il ne parle pas vraiment le swahili, parce qu'il ne connaît que quelques mots, et il ne connaît pas du tout la grammaire.
- Ça ne se dit pas comme ça, j'ai regardé dans une grammaire.
- Les grammaires de dépendance proposent un formalisme pour représenter les relations qu'entretiennent les mots entre eux.
- C'est incroyable comment un bébé apprend vite la grammaire.
- C'est parce que la grammaire est innée.
- La grammaire française ne permet l'inversion du sujet que dans des cas très restreints.
- On vous enseigne de ne plus faire de fautes de grammaire.

- Les composants fondamentaux des systèmes d'information géographique se conjuguent dans une sorte de grammaire des formes géographiques.
  - Le cinéma n'est pas une langue dont il suffirait d'apprendre la grammaire et le vocabulaire.
  - La seule grande grammaire italienne disponible en français, celle de Jacqueline Brunet, est à juste titre descriptive (et non normative).
  - Après tout, la rigueur de la pensée, on l'apprend avec la grammaire, la philosophie ...
- b) Distinguer les emplois du mot *syntaxe* dans les phrases suivantes :
- Chaque moteur de recherche a sa propre syntaxe.
  - Dans chacun des domaines de la linguistique (syntaxe, phonologie, morphologie et sémantique lexicale), nos connaissances ne sont ni apprises ni données.
  - Françoise Morvan s'efforce d'inventer une syntaxe française avec parfois des tournures bretonnes.
  - J'ai déjà eu l'occasion d'attirer l'attention sur la multiplicité des coquilles, fautes de syntaxe, fautes d'orthographe que l'on trouve dans ce journal.
  - Les rédacteurs du Nouvelliste ne maîtrisent qu'approximativement le français, ils se perdent dans la logique de la syntaxe et se noient dans l'abus de vocabulaire.
  - J'ai essayé de trouver la bonne syntaxe pour le film.
  - Du point de vue de la syntaxe, certaines évolutions sont visibles et prévisibles. Ainsi, le remplacement de "nous" par "on" avec un pluriel. "Mon fils et moi, on est allés au cinéma".
  - Envoyez un SMS selon la syntaxe suivante : le mot-clé 'METEO' 'espace' puis le numéro du département dont vous souhaitez connaître la météo.

## *Avant-propos*



## Lectures additionnelles

Nous recommandons bien sûr la lecture de l'incontournable ouvrage de Lucien Tesnière, *Éléments de syntaxe structurale* et tout particulièrement la première partie. On pourra avant cette lecture lire l'introduction écrite par Sylvain Kahane et Timothy Suborne pour la traduction anglaise de 2015.

En plus de cet ouvrage et de nombreux articles de recherche, quelques livres existent sur la syntaxe de dépendance : Richard Hudson a publié son introduction à la *Word Grammar* en 1984 et plus récemment, en 2006, *Language Networks : The New Word Grammar*. En 1986, est paru l'ouvrage *The meaning of the sentence in its semantic and pragmatic aspects* de Petr Sgall, Eva Hajičová et Jarmila Panevová sur le modèle pragois, qui a conduit au développement de la première banque d'arbre en dépendance, le *Prague Dependency Treebank*. Il a été suivi par *Dependency syntax : Theory and practice* d'Igor Mel'čuk, une œuvre fondatrice, mais qui est davantage dédiée à la présentation de la Théorie Sens-Texte, l'une des principales approches théoriques basées sur la syntaxe de dépendance, qu'à la définition de la dépendance. Le récent ouvrage d'*Introduction à la linguistique* (2014) d'Igor Mel'čuk et Jasmina Miličević, et notamment le second tome consacré à la syntaxe, est à notre connaissance le premier manuel général de linguistique basé sur la dépendance et nous en recommandons la lecture. Notre ouvrage partage en grande partie le point de vue du livre de Mel'čuk et Miličević, bien qu'il ne souhaite pas se placer a priori dans un cadre théorique donné, mais le construire de manière raisonnée. En chinois, Haitao Liu a présenté les grammaires de dépendance en 2009 dans son livre intitulé comme le livre de Mel'čuk 1988, *Théorie et pratique de la grammaire de dépendance*. Dans un cadre formel proche de la grammaire de dépendance, on trouve l'ouvrage de Joan Bresnan sur la *Lexical Functional Syntax* publié en 2001. On peut encore citer, l'ouvrage de Denis Costauoc et Françoise Guérin de 2007, *Syntaxe fonctionnelle : Théorie et exercices*, basé sur les travaux d'André Martinet, qui sans se placer réellement dans le cadre de la syntaxe de dépendance, a une approche constructiviste qui se rapproche

de la notre.


Joan **Bresnan**2001 *Lexical Functional Syntax*, Blackwell, Oxford.

Denis Costaouec & Françoise **Guérin**2007 *Syntaxe fonctionnelle : Théorie et exercices*, Presses Universitaires de Rennes.

Richard **Hudson**1984 *Word Grammar*, Blackwell, Oxford.

Richard **Hudson**2006 *Language Networks : The New Word Grammar*, Oxford University Press.

Sylvain Kahane & Timothy **Osborne**2015 Translators' introduction, in Lucien Tesnière, *Elements of structural syntax*, Benjamins, pages xxix–lxxiv (45 pages).

Haitao **Liu**2009  (Théorie et pratique de la grammaire de dépendance).

Igor **Mel'čuk**1988 *Dependency syntax : Theory and practice*, SUNY.

Igor Mel'čuk & Jasmina Milićević (**Mel'čuk****Milićević**2014 *Introduction à la linguistique*, 3 volumes, Hermann, Paris.

Petr Sgall, Eva Hajičová & Jarmila **Panevová**1986 *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Lucien **Tesnière**1959 *Éléments de syntaxe structurale*, Klincksieck, Paris.



## Corrections des exercices

**Corrigé 1** Les principales composantes d'un modèle linguistique sont la sémantique, la syntaxe, la morphologie et la phonologie. On peut ajouter à cela la pragmatique pour l'étude des liens entre le sens linguistique et intentions du locuteur et la phonétique pour l'étude des liens entre les



sons de la langue et le signal sonore.

**Corrigé 2** La grammaire inclut toutes les composantes du modèle et donc aussi la sémantique et la morphologie. Chacune de ces composantes coupe à travers la grammaire et le lexique. Ce qui relève de la syntaxe sans être de la grammaire est l'étude de la combinatoire des unités lexicales ou de la structure syntaxique interne des locutions. Dans cet ouvrage, nous nous intéressons à la partie grammaticale de la syntaxe (ou, autrement dit, à la partie syntaxique de la grammaire).

**Corrigé 3** Toutes ces notions seront étudiées en détail dans cet ouvrage. Mais il peut être bon de fixer les termes dès maintenant. La topologie est l'étude de l'interface entre syntaxe et ordre linéaire des mots. L'interface sémantique-syntaxe est, comme son nom l'indique l'étude de la correspondance entre sémantique et syntaxe. La nanosyntaxe est la partie de la syntaxe qui s'intéresse aux combinaisons d'unités linguistiques les plus cohésives, celles de l'ordre du mot. Voir les Sections 0.0.8-10.

**Corrigé 4** Parce que la définition des parties du discours repose sur une définition préalable de la structure syntaxique (voir section 11).

**Corrigé 5** L'encadré ?? est un encadré d'éclairage. Le symbole utilisé est une loupe.

**Corrigé 6** Comme tous les termes linguistiques, *grammaire* et *syntaxe* peuvent désigner un domaine de la linguistique (4, 5, 7, 11, 13, 15, 16, 18) ou la grammaire ou la syntaxe d'une langue en particulier (1, 6, 14). Par extension, on parle aussi de la grammaire ou syntaxe de systèmes sémiotiques autres que la langue (8,9,12,17,19). Dans *grammaire de dépendance* (3), le terme désigne un modèle linguistique complet, plutôt que la seule grammaire. On utilise aussi le mot *grammaire* pour désigner un livre de grammaire (2, 10).



# Todo list

unnumbered section . . . . . 23



Première partie

**Modéliser la langue**



## Présentation

Cette première partie est consacrée à la modélisation des langues en général. Elle est divisée en trois chapitres. Le premier chapitre essaye de définir la langue, notre objet d'étude, et précise les caractéristiques que nous retenons dans notre modélisation. Le deuxième chapitre montre à travers l'étude de la production d'un énoncé où se situent la syntaxe et la grammaire dans un modèle complet de la langue. Le troisième chapitre caractérise le type de modèles que nous adoptons dans cet ouvrage.





# 1 La langue : L'objet d'étude de la linguistique

## 1.1 Parler une langue

Pour comprendre ce qu'est une langue, il faut comprendre à quoi elle sert. Savoir **utiliser une langue**, c'est être capable de parler dans cette langue et de comprendre ceux qui nous parlent. **Parler une langue**, c'est être capable de verbaliser n'importe quelle idée, c'est-à-dire **transformer un sens en un son** dont notre interlocuteur pourra lui-même extraire un sens (voir dans l'encadré qui suit le schéma proposé par Saussure). Si le sens de départ – celui pensé par le locuteur – et le sens d'arrivée – celui construit par le destinataire – sont suffisamment proches, alors, la **communication** peut être considérée comme réussie.

La **LANGUE** est donc avant tout un objet qui se trouve dans le **cerveau** des locuteurs et que l'on peut modéliser par une **correspondance entre des sens et des sons**. Plus exactement, nous modélisons la langue par une correspondance entre des représentations sémantiques et des textes. Dans la suite, nous utiliserons le terme **TEXTE** pour désigner les productions langagières qu'elles soient orales, écrites ou gestuelles. Ce terme a l'avantage de ne pas présupposer quelle est la nature du médium utilisé pour communiquer. Il a également l'avantage sur le terme *son* de renvoyer à une représentation du son et non au son lui-même. Un texte sonore est ainsi la représentation que nous avons du son dans notre cerveau lorsque nous produisons des sons afin de communiquer. L'étude même de la représentation du son – la **PHONOLOGIE** – ne sera abordée que dans la mesure où elle interfère avec la syntaxe. Nous nous intéressons à la modélisation du mécanisme cognitif (situé dans le cerveau) et nous laissons de côté le mécanisme physique qui permet la production du son à partir du texte (cela concerne la phonétique articulatoire), ainsi que le mécanisme auditif qui permet le décodage du son.

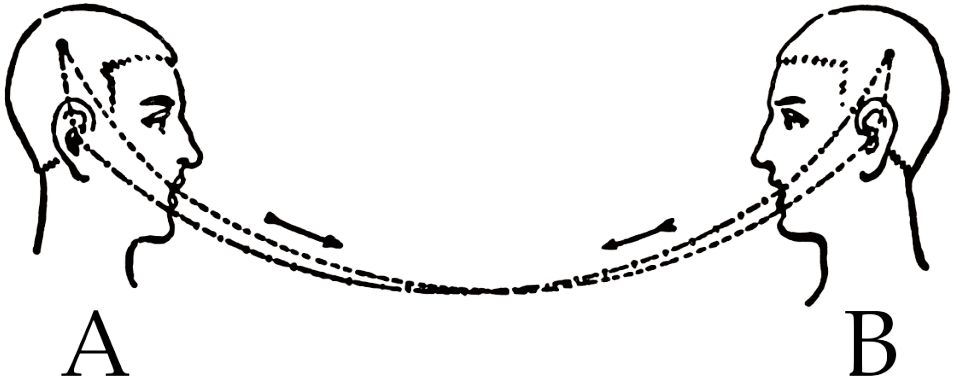
Considérer que la langue est une correspondance, c'est faire abstraction des mécanismes propres à la production ou la compréhension d'un texte. Le locuteur utilise la correspondance du sens vers le texte, tandis que le destinataire effectue le chemin inverse. Nous supposons implicitement que les deux directions, sens

vers texte et texte vers sens, utilisent le même ensemble de connaissances et un mécanisme commun, qui constituent la langue proprement dite.

## 1.2 La langue comme correspondance sens-texte

On trouve déjà dans le *Cours de linguistique générale* de Ferdinand de Saussure, publié en 1916, la conception de la langue comme un objet qui fait se correspondre des sens et des « sons ». Voici ce qu'il écrit :

« Pour trouver dans l'ensemble du langage la sphère qui correspond à la langue, il faut se placer devant l'acte individuel qui permet de reconstituer le circuit de la parole. Cet acte suppose au moins deux individus ; c'est le minimum exigible pour que le circuit soit complet. Soient donc deux personnes A et B, qui s'entretiennent :



Le point de départ du circuit est dans le cerveau de l'une, par exemple A, où les faits de conscience, que nous appellerons concepts, se trouvent associés aux représentations des signes linguistiques ou images acoustiques servant à leur expression. Supposons qu'un concept donné déclenche dans le cerveau une image acoustique correspondante : c'est un phénomène entièrement *psychique*, suivi à son tour d'un procès *physiologique* : le cerveau transmet aux organes de la phonation une impulsion corrélative à l'image ; puis les ondes sonores se propagent de la bouche de A à l'oreille de B : procès purement *physique*. Ensuite, le circuit se prolonge en B dans un ordre inverse : de l'oreille au cerveau, transmission physiologique de l'image acoustique ; **dans le cerveau association psychique de cette image psychique avec le concept correspondant**. Si B parle à son tour ce nouvel acte suivra — de son cerveau à celui de A — exactement la même marche que le premier

et passera par les mêmes phases successives. » (Saussure1916 : 27-28) [c'est nous qui soulignons]

Saussure insiste, quelques lignes plus loin, sur le fait que l'association entre sens et son qui a lieu dans le cerveau ne se fait pas avec le son lui-même, mais avec une représentation de ce son dans le cerveau, que Saussure nomme l'**image acoustique** :

« [Il faut] distinguer les parties physiques (ondes sonores) des physiologiques (phonation et audition) et psychiques (images verbales et concepts). Il est en effet capital de remarquer que l'image verbale ne se confond pas avec le son lui-même et qu'elle est psychique au même titre que le concept qui lui est associé. » (Saussure1916 : 28-29)

Leonard Bloomfield1933, le père de la linguistique américaine, va dans le même sens :

« On produit de nombreuses sortes de bruits vocaux dont on utilise la variété : sous certains types de stimuli, on produit certains sons vocaux et nos compagnons, entendant les mêmes sons, font la réponse appropriée. Pour le dire brièvement, dans la parole humaine, des sons différents ont des sens différents. **Étudier la coordination de certains sons avec certains sens, c'est étudier la langue.** » [nous soulignons]

Le fait de considérer que l'objet d'étude de la linguistique est de modéliser la correspondance sens-son décrite par Saussure (et appelée par lui association concept-image acoustique) peut être imputé à Žolkovskij et Mel'čuk qui posent en 1967 à Moscou les bases de la Théorie Sens-Texte, dont le nom même est tout à fait explicite sur ce point. Dans son cours au Collège de France en 1997 intitulé *Vers une linguistique Sens-Texte*, Igor Mel'čuk pose qu'un modèle d'une langue est une correspondance multivoque entre un ensemble de sens et un ensemble de textes, où les textes désignent les images acoustiques des phrases de la langue. Le caractère multivoque est dû au fait qu'un même sens peut être exprimé par différents textes (qui sont alors des « paraphrases » les uns des autres) et qu'un texte peut avoir plusieurs sens (c'est-à-dire être sémantiquement ambigu).

Cette conception de la langue comme une correspondance sens-son est maintenant partagée par la plupart des théories, y compris par la grammaire générative de Noam Chomsky, qui pose, depuis le *programme minimaliste* élaboré dans les années 1990, qu'un modèle linguistique doit relier sens et sons.

### 1.3 Sons et textes

Il n'est jamais inutile de rappeler que les langues sont avant tout orales (ou gestuelles comme les langues des signes) et que l'écrit n'est qu'une TRANSCRIPTION, nécessairement imparfaite et partielle, des productions orales, même s'il tend à avoir son autonomie et à acquérir sa propre codification. D'après Leonard Bloomfield<sup>1933</sup>, « L'écrit n'est pas la langue, mais simplement une façon d'enregistrer la langue au moyen de marques visibles. Une langue est la même quel que soit le système d'écriture utilisé pour l'enregistrer, exactement comme une personne est la même quelle que soit la façon dont on prend son image. » (Nous traduisons en français toutes les citations. Voir les citations originales en fin de chapitre.) La même idée est reprise par H. A. Gleason<sup>1961</sup> : « Une langue écrite est typiquement un reflet, indépendant sous quelques aspects seulement, des langues parlées. En tant qu'image de la parole réelle, elle est inévitablement imparfaite et incomplète. [...] La linguistique doit commencer par une étude approfondie de la langue parlée avant d'étudier la langue écrite. C'est vrai pour des langues avec une longue tradition écrite, comme l'anglais, autant que pour les langues de tribus isolées qui n'ont jamais envisagé la possibilité d'une écriture. » Cent ans avant (dans son ouvrage sur la langue kavi publié en 1836), le philosophe et linguiste Wilhelm von Humboldt écrivait que « la langue, comprise dans son essence réelle, est quelque chose de constant et à la fois, à tout moment, quelque chose de passager. Même sa conservation par l'écriture n'est jamais autre chose qu'un stockage ressemblant à une momie, qui nécessite qu'on cherche à s'imaginer à nouveau le discours vivant. » Cette primauté de l'oralité est constitutive des sciences du langage et la distingue fondamentalement des lettres et de la philologie (l'étude des textes écrits). Les scientifiques n'en ont néanmoins réellement pris conscience qu'au début du vingtième siècle avec les possibilités nouvelles que donnait l'enregistrement des sons et l'intérêt croissant pour les langues sans tradition écrite et notamment les langues amérindiennes ; la linguistique s'était jusque-là, à l'exception des travaux des missionnaires qui avaient pour mission d'enseigner la Bible dans des langues inconnues, essentiellement développée par l'étude de langues mortes dont on ne conservait que des traces écrites que l'on souhaitait déchiffrer.

Certaines caractéristiques essentielles des langues sont dues au fait que ce sont des modes de communication oraux. La principale de ces caractéristiques est que la communication orale impose une production linéaire : la CHAÎNE PARLÉE est **unidimensionnelle**. Les sons doivent être produits les uns à la suite des autres. L'écrit n'imposerait pas cela. La communication écrite fait d'ailleurs un grand

usage de schémas bidimensionnels souvent complexes et la présentation globale d'un texte écrit (titres, paragraphes, encadrés, etc.) joue un rôle non négligeable. Néanmoins, l'écrit traditionnel, et parce qu'il est au départ une transcription de l'oral, a une organisation linéaire, les mots devant se lire à la suite les uns des autres. On notera qu'aujourd'hui, avec l'internet, les textes contiennent une multitude de liens vers d'autres textes et qu'une production textuelle faite de plusieurs pages web n'est plus totalement linéaire : il a une structure en réseau et on parle alors d'HYPERTEXTE. Cette organisation, qui permet différents parcours du texte, existe déjà en partie dans les écrits traditionnels par la présence de notes ou d'encadrés.

Les langues des signes n'ont pas de contraintes de linéarité, puisque les gestes peuvent se développer dans toutes les dimensions de l'espace et différentes parties du corps être utilisées simultanément (mains, position de la tête, regard, orientation du buste, etc.). Il y a néanmoins une contrainte temporelle dans la communication qui veut qu'il y ait une certaine séquentialité des signes.

Le caractère linéaire de la chaîne parlée est lui-même en partie contourné par l'utilisation de la PROSODIE : les locuteurs ajoutent à la suite des sons distinctifs élémentaires qu'ils produisent — les PHONÈMES — de l'information en modulant leur mélodie et en jouant sur l'intensité du signal et la durée des phonèmes. Ceci permet non seulement de communiquer certaines émotions, mais aussi de structurer la chaîne parlée en faisant apparaître des regroupements ou des changements de plan. On retrouve à l'écrit une transcription de la prosodie par la PONCTUATION (point, virgule, :, ?, ! On peut imaginer qu'un langage purement écrit aurait développé un tout autre système et on voit que la ponctuation n'est rien d'autre qu'un marquage partiel et imparfait de la prosodie. On notera que, avec le récent développement du dialogue par écrit, le système de ponctuation s'est enrichi des smileys, émojis et autres émoticônes (☺, ☹ et plein d'autres).

Nous aurons plusieurs fois l'occasion dans cet ouvrage de montrer que les faits de langue se comprennent bien mieux lorsqu'on se place du côté de l'oral et que notre objet d'étude est d'abord la langue parlée, même si, par commodité, dans un ouvrage écrit, il est souvent plus facile de donner des exemples écrits.

## 1.4 Langue et variations

On observe de nombreuses variations entre locuteurs d'une même langue : accent, choix lexicaux, constructions grammaticales, etc. Ces variations sont dues à de multiples facteurs : le degré d'apprentissage de la langue, l'époque à laquelle vit le locuteur, le lieu où il vit, le contexte social dans lequel il s'exprime (la fa-

mille, le travail, la radio ...), etc. Ainsi l'allemand est pris entre deux langues très proches, le néerlandais et le suisse allemand et on observe un continuum de variétés d'allemand lorsqu'on se rapproche de ces deux zones linguistiques. Pour le français, on observe également des différences de parler suivant les régions, entre le nord et le sud de la France (le système phonologique, notamment, est différent), mais aussi entre la France, la Belgique, la Suisse, le Québec ou les pays francophones d'Afrique. Et à chaque fois, on observe un continuum de parler entre divers extrêmes. Il en va de même pour l'évolution des langues : la limite entre le latin et les langues romanes d'aujourd'hui (italien, espagnol, catalan, français, roumain, corse, etc.) n'est pas une frontière tranchée : le latin a évolué de génération en génération jusqu'à perdre ses désinences casuelles (qu'on ne retrouve dans aucune des langues romanes), puis divers groupes de locuteurs du haut-latin se sont retrouvés isolés au Moyen-Age et ont développés les dialectes que sont nos langues d'aujourd'hui. Le français, ancienne langue d'oïl, a subi l'influence de locuteurs d'origine germanique et la syntaxe de l'ancien français est beaucoup plus proche de celle des langues germaniques d'aujourd'hui (allemand, néerlandais, scandinave) que du latin classique. La situation est identique partout. Si l'on appelle chinois aussi bien le chinois classique que le mandarin actuel, ces langues n'en sont pas moins éloignées que le latin et l'italien. Il existe d'ailleurs aujourd'hui en Chine au moins quatre langues différentes, qui bien que partageant la même écriture, sont plus différentes entre elles que ne le sont les langues romanes entre elles.

D'une certaine façon, ces variations ne nous intéressent pas : nous décrivons un système unique, celui d'un locuteur donné à un moment donné ou au moins la langue d'un groupe de locuteurs qui communiquent usuellement entre eux. Ce qui nous intéresse avant tout, c'est la cohérence interne de ce système. Néanmoins, les variations possibles de ce système peuvent parfois nous intéresser : elles nous permettent en particulier de comprendre certaines « bizarreries » du système qu'on ne saurait expliquer sans prendre en compte le fait qu'il s'agit d'un système en évolution. Par exemple, la non-correspondance entre les unités de forme et les unités de sens (question que nous développerons en long et en large dans la partie 2 consacrée aux *Unités de la langue*) ne peut s'expliquer sans prendre en compte comment de nouveaux termes sont créés par les locuteurs et comment d'autres disparaissent. Il faut d'ailleurs distinguer deux types de changements dans les langues : les changements lexicaux et les changements grammaticaux. Tout locuteur modifie constamment son vocabulaire, acquérant de nouvelles unités lexicales et cessant d'en utiliser certaines. Par contre, il est peu probable qu'une fois l'enfance passée et l'acquisition complète de la gram-

maire effectuée des changements se produisent dans le système grammatical. Il est plus raisonnable de penser que les changements grammaticaux ont lieu lors du passage de relais d'une génération à l'autre : pour une raison ou une autre, l'apprenant va faire une analyse différente des productions langagières de ses « instructeurs » et construire une grammaire différente de la leur. Le cas le plus radical est celui d'un groupe d'apprenants d'une autre langue maternelle qui va projeter sur la langue qu'il apprend des constructions de sa langue maternelle. Les langues romanes et tout particulièrement l'ancien français sont ainsi des formes du latin dont une partie de l'évolution est due à l'assimilation d'apprenants de langue germanique.

## 1.5 Sens et intention communicative

Notre définition de la langue n'est compréhensible que si on s'entend quelque peu sur ce qu'est le sens. Pour cela nous allons partir de la définition que donne Leonard Bloomfield dans *Language*, son ouvrage fondateur de 1933.

Pour Bloomfield, il y a ACTE DE LANGAGE lorsque Jill a faim, qu'elle voit une pomme dans un arbre et qu'au lieu de grimper dans l'arbre la cueillir, elle produit un son avec son larynx, sa langue et ses lèvres et que c'est Jack qui cueille la pomme et la lui apporte. Autrement dit, face à un stimulus S (Jill a faim et voit une pomme), il y a deux voies pour arriver à la réaction R :

- la voie directe  $S \rightarrow R$ , où Jill grimpe dans l'arbre,
- et la voie indirecte  $S \rightarrow r \text{ --- } s \rightarrow R$ , où une réaction linguistique  $r$  se substitue à la réaction mécanique de Jill et où le stimulus  $s$  qu'elle provoque donne la réaction  $R$  chez Jack.

Reprenons la citation de Bloomfield, donnée dans l'encadré ?? sur *La langue comme correspondance sens-texte*, sous ce nouvel éclairage :

« On produit de nombreuses sortes de bruit vocaux dont on utilise la variété : sous certains types de stimuli, on produit certains sons vocaux et nos compagnons, entendant les mêmes sons, font la réponse appropriée. Pour le dire brièvement, dans la parole humaine, des sons différents ont des sens différents. Étudier la coordination de certains sons avec certains sens, c'est étudier la langue. » (Bloomfield 1933 : 27)

Jusque-là, nous sommes parfaitement d'accord. Reste à définir le sens :

« En produisant une forme linguistique, un locuteur incite son interlocuteur à répondre à une situation ; cette situation et la réponse qu'elle déclenche

sont le *sens linguistique* de la forme. Nous supposons que chaque forme linguistique a un sens constant et défini, différent du sens de n'importe quelle autre forme linguistique de la même langue. » (Bloomfield 1933 : 165)

Là, nous ne sommes plus en accord. Nous pensons qu'il faut absolument séparer la situation du sens linguistique. Dans la même situation, Jill peut produire des énoncés de sens différents comme « *Pourrais-tu me cueillir cette pomme?* » ou « *Apporte-moi cette pomme!* » ou des énoncés moins coercitifs comme « *J'ai faim.* » ou « *Regarde cette belle pomme!* », qui peuvent néanmoins amener la même réaction de Jack. Inversement, dans une tout autre situation, par exemple face à une nature morte dans un musée, Marie peut dire à Pierre « *Regarde cette belle pomme!* » et cet énoncé a, pour notre définition du sens, le même sens que l'énoncé « *Regarde cette belle pomme!* » de Jill, qui procédait pourtant d'intentions complètement différentes.

Autrement dit, nous distinguons clairement trois objets :

- le CONTEXTE D'ÉNONCIATION, c'est-à-dire les caractéristiques extérieures de la situation où est produit l'énoncé : qui parle à qui ? où et quand ? dans quelles circonstances ? etc.
- les INTENTIONS COMMUNICATIVES du locuteur, c'est-à-dire les buts que se fixe le locuteur, les informations qu'il souhaite communiquer sur tel ou tel objet dans le contexte, etc.
- le SENS LINGUISTIQUE de l'énoncé, c'est-à-dire, indépendamment du contexte et des intentions du locuteur, le contenu de son message, les éléments de sens qu'il a choisis pour communiquer l'information, désigner tel objet, etc. Le sens linguistique, tel que nous l'envisageons, est très proche du texte, puisqu'il contient déjà les sens des différentes unités du texte.

La phase d'élaboration du contenu d'un message, c'est-à-dire le passage d'intentions communicatives dans un contexte donné à un sens linguistique s'appelle la PLANIFICATION du message. On considère généralement que l'étude de la planification ne relève pas de la linguistique et que cette étape reste relativement indépendante du langage, dans la mesure où toutes les langues permettent d'exprimer à peu près tous les sens (comme le montre l'absence d'obstacles majeurs à la traduction d'une langue à l'autre, sauf lorsqu'il s'agit de concepts absents d'une culture à l'autre). Laurence Danlos, l'une des pionnières de la génération automatique de textes, propose d'appeler la planification le *Quoi dire*, qu'elle oppose au *Comment le dire*, qui constitue la langue proprement dite.

Nous savons bien sûr que la planification est en partie guidée par le stock lexical que chaque langue propose et nous allons voir un peu plus loin (section 1.10) que la planification joue quand même un rôle dans l'organisation des énoncés et



la syntaxe, même si dans cet ouvrage nous étudierons essentiellement le passage du sens au texte. L'étude même des sens linguistiques — la SÉMANTIQUE — ne sera abordée que dans la mesure où elle interfère avec la syntaxe.

## 1.6 Mots et pensée

La question se pose de savoir si on peut penser sans mots, c'est-à-dire si on peut manipuler des concepts sans les verbaliser, même dans sa tête. Nous pensons que oui. Le bricoleur qui répare son moteur pense à toutes les pièces du moteur qu'il manipule sans avoir nécessairement une idée de comment les appeler et il effectue un grand nombre d'actions bien réfléchies qu'il aurait bien du mal à expliquer en mots. Plus on va vers une pensée abstraite, plus on pourrait penser que les sens doivent se confondre avec les mots. Pourtant, le mathématicien qui fait une démonstration n'a pas toujours les mots pour exprimer les concepts qu'il manipule et une part de son activité est justement d'isoler ces concepts et de leur donner des noms : ensemble, fonction, continuité, etc. Ici il est clair que le concept précède dans la pensée le terme qui lui correspond. Nous faisons l'hypothèse qu'il existe une forme abstraite de pensée sans mots et que la langue est l'ensemble des connaissances qui nous permet d'exprimer cette pensée. Cette hypothèse reste aujourd'hui plus ou moins invérifiable et est controversée. Pour une discussion très lisible de ces idées, voir l'ouvrage de Steven Pinker, *Comment fonctionne l'esprit*, Éditions Odile Jacob, 2000.

## 1.7 Sens lexicaux et traduction

Les sens exprimables par des mots simples varient d'une langue à l'autre, ce qui élimine tout espoir de traduction exacte. Par exemple, l'anglais possède une unité lexicale MELON qui dénote aussi bien le melon que la pastèque et le français ne possède pas d'équivalent. Le verbe ESPERAR de l'espagnol couvre à la fois les sens 'espérer' et 'attendre' du français et il existe un continuum entre les deux sens : 'esperar' ne sera donc pas ambigu mais « sous-spécifié » en ce qui concerne le degré de joie et de certitude qu'a la personne qui « *espera* ». À l'inverse, le verbe AIMER couvre ce que l'anglais exprime avec LOVE et LIKE, car encore une fois, les locuteurs du français considère un continuum entre ces sens. Il existe aussi des sens lexicaux propres à une langue : par exemple, le nom allemand SCHADENFREUNDE désigne la joie que procure le dépit mérité des autres et n'a pas d'équivalent en français, bien que le concept puisse être universellement compréhensible. De la même façon, un verbe français comme s'ACCOUDER n'aura

pas de meilleur équivalent dans beaucoup d'autres langues (par exemple en anglais ou en allemand) qu'une traduction littérale de 's'appuyer sur les coudes'.

## 1.8 Langue, linguistique et modélisation

Nous allons définir un certain nombre de termes dont nous aurons besoin et que nous avons déjà commencé à utiliser.

Un **LOCUTEUR** ou **SUJET PARLANT** est quelqu'un qui parle, c'est-à-dire quelqu'un qui cherche à communiquer avec d'autres gens en produisant des paroles ou un texte écrit. Il s'adresse à des **DESTINATAIRES** ou **INTERLOCUTEURS**.

Une **LANGUE** est un système de signes conventionnels partagés par un certain nombre de personnes et qui leur permet de communiquer entre elles. Ces signes s'assemblent pour former des mots, des phrases et des discours. Une langue est à la fois un **objet individuel** — c'est l'ensemble des connaissances stockées dans notre cerveau qui nous permettent de parler (dans cette langue) — et un **objet collectif et social**, puisque ces connaissances sont partagées par un certain nombre de personnes, qui sont les locuteurs de cette langue.

La **FACULTÉ LANGAGIÈRE** est l'aptitude que nous avons à apprendre et utiliser les langues. Le cerveau est l'organe de la faculté langagière (avec le système phonatoire que le cerveau commande pour produire des sons). Lorsqu'on parle de **LA LANGUE**, et non plus d'une langue particulière, on

fait généralement référence à la faculté langagière et ce qu'elle va imposer comme traits communs à l'ensemble des langues possibles.

La LINGUISTIQUE est la science qui étudie les langues du monde. Comme beaucoup de scientifiques, nous considérons que la linguistique inclut l'étude de la faculté langagière, c'est-à-dire l'étude de la production langagière et de l'apprentissage d'une langue par l'enfant. La linguistique est alors quasiment une branche de la psychologie.

La linguistique produit des MODÈLES DES LANGUES et de la faculté langagière. Ces modèles doivent être capables de simuler un ACTE DE LANGAGE, c'est-à-dire la façon dont un locuteur produit un texte à partir d'un sens qu'il veut exprimer et la façon dont son interlocuteur reconstruit un sens à partir de ce texte.

Tout modèle se situe dans un CADRE THÉORIQUE. Décider que la langue est une correspondance est un choix théorique. Décider ce qui est mis en correspondance par la langue, c'est-à-dire ce que sont la représentation sémantique et le texte relève aussi de choix théoriques. C'est le cadre théorique qui caractérise notamment l'objet d'étude. Pour reprendre une formule de **Saussure**1916 devenue fameuse, « Bien loin que l'objet précède le point de vue, on dirait que c'est le point de vue qui crée l'objet. ».

## 1.9 Langue et parole, compétence et performance

Saussure oppose deux notions fondamentales qu'il nomme **LANGUE** et **PAROLE** :

« Entre tous les individus ainsi reliés par le langage, il s'établira une sorte de moyenne : tous reproduiront, — non exactement, mais approximativement — les mêmes signes unis aux mêmes concepts. [...] »

La **langue** n'est pas une fonction du sujet parlant, elle est le produit que l'individu enregistre passivement. [...] Elle est la partie sociale du langage, extérieure à l'individu, qui à lui seul ne peut ni la créer ni la modifier ; elle

## 1 La langue : L'objet d'étude de la linguistique

n'existe qu'en vertu d'une sorte de contrat passé entre les membres de la communauté. [...]

La **parole** est au contraire un acte individuel de volonté et d'intelligence, dans lequel il convient de distinguer :

1° les combinaisons par lesquelles le sujet parlant utilise le code de la langue en vue d'exprimer sa pensée personnelle ;

2° le mécanisme psycho-physique qui lui permet d'extérioriser ces combinaisons. »

(Saussure1916 : 29)

La parole, au sens de Saussure, couvre deux notions qu'il convient de séparer. Nous préférons parler de PRODUCTIONS LANGAGIÈRES pour la première notion, c'est-à-dire les énoncés réellement produits par des sujets parlants, tandis qu'on préférera appeler la deuxième notion la FACULTÉ LANGAGIÈRE. Alors que la parole est un objet individuel, la LANGUE est un objet social par excellence, mais c'est aussi la trace qu'a imprimée cet objet dans le cerveau de chacun de nous, objet collectif, donc, qui n'existe que par la somme de ses traces individuelles.

L'opposition entre compétence et performance proposée par Chomsky1965 nous renvoie à l'opposition entre langue et parole, mais il convient de les distinguer : la compétence ne se confond pas avec la langue, ni la performance avec la parole. La COMPÉTENCE désigne notre compétence passive à savoir utiliser la langue, mais aussi à l'acquérir. Elle se divise en une compétence innée, qui peut se confondre avec la faculté langagière, et une compétence acquise, qui peut se confondre avec la langue en tant que trace individuelle d'une langue dans notre cerveau.

La PERFORMANCE est l'usage proprement dit de la langue. La PAROLE est le produit de la compétence et de la performance : nous avons la compétence de produire des énoncés supposément parfaits, mais divers facteurs (notre état émotionnel, des éléments qui vont nous distraire, la recherche d'un message approprié, etc.) vont faire que notre énoncé ne sera pas aussi parfait qu'il aurait pu l'être. Ceux qui veulent décrire la langue, comme nous, vont essayer de séparer ce qui relève d'un MANQUE DE COMPÉTENCE de ce qui relève d'une ERREUR DE PERFORMANCE. La chose est loin d'être évidente, notamment lorsqu'on touche aux LIMITATIONS MÉMOIRELLES. Par exemple, Chomsky a beaucoup insisté sur le caractère RÉCURSIF de la langue : une proposition peut contenir une proposition subordonnée (*La personne [que le chien a mordu] est à l'hôpital*) et un tel enchâssement peut être itéré (*La personne [que le chien [auquel le garçon a donné un os] a mordu] est à l'hôpital*). Mais cette dernière phrase est difficilement compréhensible et une insertion de plus dépasse nos capacités d'analyse en situation de

communication ordinaire (*La personne que le chien auquel le garçon qui habite au coin de la rue a donné un os a mordu est à l'hôpital*). Défaut de compétence ou de performance ? (Voir Exercice 3.)

## 1.10 La planification

Nous voudrions montrer ici que la planification (voir définition dans la section 1.5 sur *Sens et intention communicative*) peut avoir des incidences non négligeables sur la nature du texte produit et que dans l'absolu il faut l'inclure dans notre objet d'étude. En particulier, on ne peut pas en général considérer que le locuteur construit le contenu de son message avant de le transformer en un énoncé, c'est-à-dire qu'il planifie complètement avant de produire un énoncé. Les choses ne se passent généralement pas comme ça et le locuteur élabore le contenu de son message au fur et à mesure de l'énonciation. Les productions orales présentent en particulier de nombreux indices de la planification en cours. (C'est moins net à l'écrit, puisque le scripteur a la possibilité de revenir en arrière pour corriger sa production.) Comme le dit Claire **Blanche-Benveniste** 1990, « Lorsque nous produisons des discours non préparés, nous les composons au fur et à mesure de leur production, en laissant des traces de cette production. [...] L'étude de ces traces est en elle-même un sujet d'observations ; on y voit la production de langage en train de se faire. [...] Une observation attentive permet de voir comment nous procédons, quelles unités nous utilisons pour faire avancer nos discours, quelles tenues en mémoire nous avons, à la fois pour les morceaux déjà énoncés et pour ceux que nous projetons d'énoncer. On peut ainsi observer comment se fait la mise au point des syntagmes, la recherche des « bonnes dénominations », et le travail constant d'évaluation que nous faisons sur nos propres discours. »

Nous allons montrer trois phénomènes qui illustrent l'influence de la planification sur la structure de l'énoncé. Les exemples qui suivent sont des retranscriptions fidèles de productions orales attestées ; dans ces transcriptions figurent absolument tous les mots prononcés par le locuteur, y compris les bribes et répétitions dues aux hésitations du locuteur.

Les premiers indices de la planification en cours sont les nombreuses amorces que le locuteur fait et auxquelles il renonce momentanément ou définitivement. Dans l'exemple suivant, la locutrice semble ne pas trouver tout de suite la bonne formulation ; elle hésite et répète *les* pour se donner du temps, amorce la production de *les capitales*, mais s'y prend quand même à deux fois pour finalement proposer une reformulation par *les grandes villes* :

À chaque fois, on obtient une BRIBE, c'est-à-dire un segment inachevé qui est

## 1 La langue : L'objet d'étude de la linguistique

et je voulais pas aller à Addis Abeba puisque **les les les les c- les capitales les grandes villes** ne me disaient rien du tout

ensuite corrigé par le locuteur. La disposition suivante du texte, dite ANALYSE EN GRILLE, permet de mettre en évidence l'ENTASSEMENT (voir le Chapitre ??) dans la même position syntaxique de la bribe et du segment qui vient la remplacer :

- (1) et je voulais pas aller à AA puisque les  
les  
les  
les c-  
les capitales  
**les grandes villes** ne me disaient rien du tout

Plus étonnant est le phénomène de la GREFFE bien étudié par José Deulofeu (voir la partie 6). L'exemple suivant en fourni deux :

- (2) on avait critiqué le le journal de je crois que c'était le Provençal on l'avait critiqué par rapport à ou le Méridional par rapport à la mort de comment il s'appelle ... pas Coluche l'autre

À deux reprises, ici, le locuteur ne trouve pas le nom qu'il cherche et il vient greffer un énoncé qui pourrait fonctionner comme un énoncé autonome et qui est ici inséré dans l'énoncé principal dont il assure la complétion. Nous reprenons le texte précédent en le disposant selon les principes de l'analyse en grille et en indiquant les greffes en gras.

- (3) on avait critiqué le  
le journal de  
je crois que c'était le Provençal  
on l'avait critiqué par rapport à  
ou le Méridional  
par rapport à la mort de **comment il s'appelle**  
pas Coluche  
l'autre

On notera que la première greffe s'entrelace avec l'énoncé principal, ce qui semble montrer que le locuteur poursuit en parallèle la planification de son énoncé principal et de la greffe. On voit à nouveau ici un fonctionnement par entassement de segments similaires ou identiques (*on avait critiqué le journal - on l'avait critiqué, par rapport à - par rapport à*).

Un dernier phénomène illustre bien le fait que la planification a lieu en même temps que l'énonciation : les parenthèses. On appelle PARENTHÈSE tout énoncé qui vient s'insérer dans l'énoncé principal et qui est marqué par un changement de registre (une modification de l'intonation) qui le détache nettement de l'énoncé principal. Dans le texte suivant, nous avons indiqué les parenthèses entre parenthèses et nous avons directement disposé le texte en grille :

- (4) donc pour essayer un petit peu de sortir cette personne de la misère  
 (car c'est vraiment un petit peu semblable aux Misérables de  
 Victor-Hugo)  
 nous essayons tant bien que mal de lui faire comprendre que **sa cabane**  
 dans quelques années (entre parenthèses, elle a 79 ans)  
 quand elle aura des difficultés (ce qu'on espère pas)  
 des difficultés à se déplacer ou à évoluer (c'est-à-dire qu'il y a  
 énormément d'escaliers à monter pour arriver à sa cabane)  
 donc le jour où elle ne pourra plus se déplacer  
 ou qu'elle sera malade un petit peu plus sévèrement,  
 on essaye de lui faire comprendre qu'elle ne pourra plus vivre dans cette  
 cabane

On peut résumer le contenu de ce texte ainsi : comme cette personne a 79 ans et qu'il y a énormément d'escaliers pour arriver à sa misérable cabane, il faut lui faire comprendre qu'elle ne pourra plus y vivre dans quelques années. On voit que deux informations essentielles ('elle a 79 ans' et 'il y a énormément d'escaliers') ont été ajoutées à la volée, ce qui a finalement obligé le locuteur à abandonner sa première proposition (**sa cabane**, en gras dans le texte, est le sujet d'un verbe originalement planifié qui ne vient jamais) et à reprendre par une proposition équivalente (*nous essayons tant bien que mal de lui faire comprendre que sa cabane* → *on essaye de lui faire comprendre qu'elle ne pourra plus vivre dans cette cabane*). Il est probable qu'à la première lecture (qui aurait normalement dû être une écoute) de ce texte, vous ne vous êtes pas rendu compte que la première proposition avait été laissée inachevée : ceci montre le caractère très naturel de telles constructions et le fait que le destinataire est habitué à « corriger » les « erreurs » dues à la planification.

Les linguistes considèrent généralement que la planification est hors de leur objet d'étude et que la langue constitue uniquement le passage du contenu du message, c'est-à-dire le sens, à un énoncé. Les exemples précédents montrent que la planification, ou plutôt les problèmes de planification, laisse de nombreuses

traces en surface et qu'il est donc difficile d'en faire abstraction, surtout si on étudie les productions orales. À l'écrit, par contre, les défauts de planification sont gommés par les passages successifs du rédacteur et la possibilité d'interrompre la rédaction pendant la planification. Ceci est encore une raison de préférer l'étude de l'oral à celle de l'écrit, car on trouve à l'oral davantage d'indices de la façon dont les locuteurs « travaillent », alors les corrections successives sont invisibles à l'écrit.

## 1.11 Corpus et introspection

Il existe deux moyens d'obtenir des faits de langue pour le linguiste. Le premier est de collecter des textes déjà produits. Un ensemble de textes est appelé un **CORPUS**. Le plus grand corpus disponible est le web et les moteurs de recherche constituent un assez bon moyen de récolter les données que l'on cherche, même s'il faut savoir trier ces données selon le type de page (site d'information, blog, forum, chat, etc.) et le type d'auteur (locuteur natif, génération automatique, traduction automatique, etc.). Il existe des corpus plus spécialisés, comme les bibliothèques numériques d'ouvrages classiques, les archives des grands journaux, les encyclopédies en ligne ou les revues scientifiques. Les linguistes constituent des corpus pour leur besoin, notamment des corpus de productions orales dont les textes sont minutieusement retranscrits à l'écrit (nous en avons donné des exemples dans la section précédente). Certains de ces corpus concernent des populations particulières : enfants en phase d'acquisition, apprenants d'une seconde langue, aphasiques, etc. Tous ces paramètres constituent le **GENRE** de la production textuelle. Un bon corpus doit comporter ce type d'informations, qu'on appelle les **MÉTADONNÉES**, c'est-à-dire les données qui concernent le corpus et se trouvent à côté des données proprement dites. En plus des métadonnées, certains corpus sont agrémentés d'annotations diverses permettant une meilleure étude des structures des énoncés. On trouve également des corpus alignés de différentes langues (qu'on appelle corpus multilingues ou bitextes) très utiles pour développer des modèles pour la traduction.

Le deuxième moyen d'étude est **L'INTROSPECTION**. Il s'agit de construire artificiellement des énoncés et d'en faire juger l'**ACCEPTABILITÉ** par des locuteurs natifs. Ce moyen permet de tester toutes les variantes imaginables d'un phénomène et surtout de vérifier les limites d'un phénomène en produisant des énoncés jugés inacceptables. Une autre raison qui peut justifier le recours à l'introspection est que, sur corpus, on rencontre beaucoup d'erreurs de performance, qui font que certains énoncés seraient jugés inacceptables même par ceux qui les ont produits.



Il est donc nécessaire de garder un esprit critique et de savoir filtrer les résultats.

Un énoncé produit dans des conditions normales de production est dit **ATTESTÉ**, par opposition à un énoncé **CONSTRUIT** par le linguiste. Même si nous ne rejetons pas l'introspection et l'appel au jugement des locuteurs, nous considérons que l'étude des corpus restent le meilleur moyen d'accéder aux données et d'éviter de passer à côté de phénomènes importants.

## 1.12 Acceptabilité

Parmi toutes les phrases bizarres qu'un linguiste rencontre ou construit, il est souvent difficile de classer les phrases en bonnes et mauvaises. On constate plutôt une gradation de « qualité » qu'un jugement binaire en bon et mauvais.

Considérons les énoncés construits suivants :

- (5) a. C'est un film que je sais que tu n'hésiteras pas une seconde à regarder.
- b. C'est un film que je me demande quand tu regarderas.
- c. C'est un film que je ne sais pas si tu accepteras que je regarde.
- d. C'est un film que je me demande jusqu'où tu es prêt à regarder.
- e. C'est un film que je dormais quand tu regardais.

On ressent facilement que la phrase (A) est meilleure que la phrase (B), qui elle-même est meilleure que (C). On constate que la phrase (E) est clairement inacceptable, tandis qu'on peut se demander si (D) l'est ou pas. Il est d'usage de noter l'**ACCEPTABILITÉ** des énoncés par des symboles allant de l'absence de symbole signifiant l'acceptabilité au symbole \* signifiant l'inacceptabilité, en passant par les symboles ? (léger doute), ?? (doute sérieux) et ?\* (inacceptabilité probable). Pour nos exemples, nous aurions donc :

- (6) a. C'est un film que je sais que tu n'hésiteras pas une seconde à regarder.
- b. ?C'est un film que je me demande quand tu regarderas.
- c. ??C'est un film que je ne sais pas si tu accepteras que je regarde.
- d. ?\*C'est un film que je me demande jusqu'où tu es prêt à regarder.
- e. \*C'est un film que je dormais quand tu regardais.

Les marques d'acceptabilité ne sont pas absolues et doivent plutôt être interprétées comme relatives (c'est-à-dire qu'un énoncé marqué ?? est plus acceptable qu'un énoncé marqué ?\*).

**Acceptabilité et grammaticalité**

## 1 La langue : L'objet d'étude de la linguistique

Lorsqu'on étudie la syntaxe, il est important de faire abstraction des problèmes qui viennent de la sémantique. Comparons les deux énoncés suivants :

- (7) a. D'élégants chevaux blancs courent librement.
- b. D'incolores idées vertes dorment furieusement.

Évidemment, l'énoncé (2), traduit d'un célèbre exemple de Noam Chomsky (Colorless green ideas sleep furiously), est plus que bizarre et il est difficile de trouver un contexte, autre que poétique, où cet énoncé aurait un sens approprié. Pourtant d'un strict point de vue syntaxique, l'énoncé (2) est identique à (1) et peut être jugé comme tout à fait GRAMMATICAL.

Voici un autre exemple (attesté) d'une phrase grammaticale, mais incompréhensible, car trop complexe et appartenant à un langage de spécialité dont le vocabulaire nous est inhabituel :

- (8) À mon avis, à l'exception de l'effet des éventuels redressements que j'aurais pu juger nécessaires si j'avais pu m'assurer de l'intégralité des produits dont il est question au paragraphe précédent, ces états financiers présentent fidèlement, à tous égards importants, la situation financière de la société au 31 décembre 1996 ainsi que les résultats de son exploitation et l'évolution de sa situation financière pour la période de douze mois terminée à cette date selon les principes comptables généralement reconnus.

Lorsqu'un énoncé est grammatical, mais jugé INAPPROPRIÉ, nous utilisons le symbole #. Ainsi en réponse à la question « *Qui a mangé les framboises?* », la réponse suivante sera jugée inappropriée :

- (9) #C'est les framboises que Pierre a mangées.

Autre exemple : quand on parle des expressions figées, on peut relever que **BRISER LA GLACE** (au sens de 'dissiper la gêne') se passive, mais pas **PERDRE LES PÉDALES** (au sens de 'perdre le contrôle de soi-même') :

- (10) Marie a brisé la glace.
- (11) La glace a été brisée par Marie
- (12) Marie a perdu les pédales.

(13) # Les pédales ont été perdues par Marie.

Cette dernière phrase est grammaticale, mais elle a perdu le sens figé (la seule lecture possible est littérale : on parle vraiment de pédales et d'une perte) et elle n'a donc plus le sens approprié.

Enfin, lorsqu'une combinaison morphologique est jugée inappropriée, car absente du stock lexical actuel, nous utilisons le symbole ° : °*tournement*, °*bravitude*, °*aspire-poussière*, °*loup-chien*, etc.

## 1.13 Parler et comprendre

Si les textes (oraux ou écrits) sont le principal moyen d'accès à la langue, il ne faut pas oublier que la description des textes n'est pas notre finalité. Un texte est une production langagière et c'est bien la production du texte qui, derrière le texte, nous intéresse.

Une langue est une correspondance entre sens et textes et donc modéliser la langue ce n'est pas seulement modéliser les textes de cette langue, mais modéliser la correspondance entre sens et textes. Or il y a deux façons d'effectuer cette correspondance : soit on passe du sens au texte, c'est-à-dire qu'on parle, soit on passe du texte au sens et l'on est dans une situation d'analyse et de décodage du texte.

La plupart des études partent des textes et donc modélisent la langue dans le sens de l'ANALYSE. Nous pensons, à la suite d'Igor Mel'čuk, qu'il est préférable d'étudier la production et de travailler dans le sens de la SYNTHÈSE (on parle encore de GÉNÉRATION DE TEXTES quand la production est automatisée). Ceci peut paraître parfois délicat, car pour étudier la production, il faut partir d'une intention communicative et donc commencer par construire un sens. Mais c'est le seul moyen de comprendre quelles sont les contraintes qui s'exercent sur le locuteur et modèlent la langue. Nous allons illustrer notre propos en étudiant un exemple de production dans le chapitre suivant.



## **Exercices**

**Exercice 1** S'intéresse-t-on aux variations individuelles entre locuteurs d'une même langue ?

**Exercice 2** Le symbole ?\* veut-il dire qu'on ne croit pas que la phrase est agrammaticale ?

**Exercice 3** Nous avons terminé l'encadré ?? en nous demandant si le fait que la phrase *La personne que le chien auquel le garçon qui habite au coin de la rue a donné un os a mordu est à l'hôpital* était incompréhensible mettait en évidence un problème de compétence ou de performance. Quel élément de réponse peut-on apporter ?

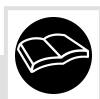
**Exercice 4** Quel phénomène intéressant observe-t-on dans les énoncés suivants du point de vue de la planification ?

mais voilà ça c'est une sorte de mélange très très très étrange et qui moi me je sais pas pourquoi me renverse littéralement

vous allez passer devant la poste qui sera à votre droite et un peu plus loin euh je sais pas à quel niveau c'est exactement Habitat et euh la Chambre de de Commerce à votre gauche

**Exercice 5** La société humaine repose sur la croyance. Cette idée est très bien développée dans *Sapiens - Une brève histoire de l'humanité* de Yuval Noah Harari. La question dépasse très largement la croyance religieuse. Notre société fonctionne car nous, humains, croyons en un même système de lois, nous croyons dans les valeurs de notre société, comme l'éducation ou la solidarité, nous croyons en la valeur de l'argent, alors même qu'il peut s'agir d'un simple bout de papier imprimé ou d'un nombre dans la mémoire d'un ordinateur de notre banque. En quoi la notion de croyance (ainsi définie) concerne la linguistique et la langue ?

*1 La langue : L'objet d'étude de la linguistique*



## Lectures additionnelles

Les ouvrages de Ferdinand de Saussure 1916 et de Leonard Bloomfield 1933 sont des monuments de la linguistique, dont la lecture reste toujours incontournable. Le livre de Wilhelm von Humboldt de 1836 est un ouvrage précurseur, mais difficile à lire aujourd'hui. La lecture des ouvrages de Claire Blanche-Benveniste est une excellente plongée dans les données véritablement attestées et ce qu'elles révèlent sur la syntaxe du français parlé et la langue en général. Le livre de Henry A. Gleason 1955 est une introduction très pédagogique aux bases de la linguistique et de la syntaxe, qui a étonnamment peu vieilli. Le cours donné par Igor Mel'čuk au collège de France complétera les ouvrages de ce linguiste déjà mentionné dans l'*Avant Propos*. Les lecteurs intéressés par l'histoire des sciences pourront consulter les premiers travaux sur la Théorie Sens-Texte de Žolkovskij et Mel'čuk (Žolkovskij Mel'čuk 1967), traduits en français en 1970. Les bases de la génération automatique de textes sont développées dans l'ouvrage fondateur de Laurence Danlos 1987. Nous avons également mentionné les ouvrages de Steven Pinker et de Yuval Noah Harari, qui ne concernent pas directement notre sujet, mais s'inscrivent dans une approche scientifique des sciences humaines que nous suivons. Nous avons également mentionné le programme minimaliste de Noam Chomsky, présenté dans son ouvrage de 1995.

Claire Blanche-Benveniste 1990 *Le français parlé : études grammaticales*, Éditions du CNRS, Paris.

Claire Blanche-Benveniste 2000 *Approche de la langue parlée en français*, collection L'essentiel, Ophrys.

Leonard Bloomfield 1933 *Language*, New York (traduction française : *Le langage*, Payot, 1970).

Noam Chomsky 1995 *The minimalist program*, MIT Press.

Laurence Danlos 1987 *Génération automatique de textes en langues naturelles*, Cambridge University Press.

Henry A. Gleason 1955 *An introduction to descriptive linguistics*, Holt, Rinehart and Wilston.

Wilhelm von **Humboldt**1836 Über die Verschiedenheit des menschlichen Sprachbaus und seinen Einfluss auf die geistige Entwicklung des Menschengeschlechts (traduction en français par Pierre Caussat : Introduction à l'œuvre sur le kavi et autres essais, Seuil, 1974).

Yuval Noah **Harari**2014 Sapiens : A brief history of humankind (traduction française : Sapiens : Une brève histoire de l'humanité, Albin Michel, 2015).

Mel'čuk **Igor**1997 *Vers une linguistique Sens-Texte*, Leçon inaugurale, Collège de France.

Steven **Pinker**1996 *How the mind works* (traduction française : *Comment fonctionne l'esprit*, Odile Jacob, 2000)

Ferdinand de **Saussure**1916 *Cours de linguistique générale*, Paris.

Aleksander Žolkovskij & Igor Mel'čuk (1967) O semanticeskom sinteze [Sur la synthèse sémantique (de textes)], *Problemy Kybernetiki* [Problèmes de Cybernétique], 19, 177-238 (traduction française : 1970, *T.A. Information*, 2, 1-85).





## Citations originales

Citations de la section ??.

Bloomfield (1933 : 27) :

Man utters many kinds of vocal noise and make use of the variety : under certain types of stimuli he produces certain vocal sounds and his fellows, hearing the same sounds, make the appropriate response. To put it briefly, in human speech, different sounds have different meanings. To study this co-ordination of certain sounds with certain meanings is to study language.

**Citations de la section 1.3. Bloomfield 1933 :**

Writing is not language, but merely a way of recording language by means of visible marks. [...] A language is the same no matter what system of writing may be used to record it, just as a person is the same no matter how you take his picture.

**Gleason 1961 :**

A written language is typically a reflection, independent in only limited ways, of spoken languages. As a picture of actual speech, it is inevitably imperfect and incomplete. [...] Linguistics must start with thorough investigation of spoken language before it proceeds to study written language. This is true for language with long histories of written literature, such as English, no less than those of isolated tribes which have never known of the possibility of writing.

**von Humboldt 1836 :**

Die Sprache in ihrem wirklichen Wesen aufgefasst ist etwas beständig und in jedem Augenblicke Vorübergehendes. Selbst ihre Erhaltung durch die Schrift ist immer nur eine unvollständige mumienartige Aufbewahrung die es doch erst wieder bedarf dass man dabei den lebendigen Vortrag zu versinnlichen sucht.

**Citations de la section 1.5.**

**Bloomfield (1933 : 165) :**

By uttering a linguistic form, a speaker prompts his hearers to respond to a situation; this situation and the response to it are the linguistic meaning of the form. We assume that each linguistic form has a constant and defining meaning, different from the meaning of any other linguistic form in the same language.



## Corrections des exercices

**Corrigé 1** Certains linguistes s'intéressent beaucoup aux variations individuelles, notamment les chercheurs en acquisition des langues ou en socio-linguistique. Mais dans cet ouvrage, nous ne nous y intéresserons pas vraiment : nous voulons modéliser un système linguistique particulier et pas les variations entre deux systèmes. Qui plus est, nous privilégierons dans notre étude les règles qui sont communes au plus grand nombre de locuteurs. Et en même temps, nous savons qu'il existe des variations et nous voulons que le système que nous proposons soit capable de les saisir.

**Corrigé 2** Non, il signifie qu'on n'est pas complètement sûr que la phrase soit grammaticale. Voir section 1.12.

**Corrigé 3** Précisons pour commencer que le problème ne se situe pas du côté de la compréhension. Le problème n'est pas que cette phrase ne sera pas comprise, le problème est qu'elle ne pourra pas être produite par un locuteur dans une situation normale de communication. Pour la produire, nous avons fait un travail de linguiste et pas de locuteur ordinaire, en appliquant des insertions successives de relatives. À partir de là, nous considérons que les locuteurs n'ont pas la compétence de produire de tels énoncés et que cela n'a rien à voir avec la performance. Mais on peut aussi estimer (comme Noam Chomsky) que cette limitation n'est pas à proprement parler une propriété de la langue et qu'elle relève d'une modélisation des capacités cognitives en général. Si le modèle linguistique est combiné avec un tel modèle, on peut ne pas tenir compte des limitations mémorielles dans la modélisation des langues. Quoi qu'il en soit, il faut, à notre avis, les prendre en compte quelque part dans un modèle de la compétence.

**Corrigé 4** Ces deux énoncés présentent une parenthèse (*je sais pas pourquoi* dans le premier et *je sais pas à quel niveau c'est exactement* dans le

second). Il s'agit d'un énoncé syntaxiquement indépendant à l'intérieur d'un autre énoncé. Cela illustre la capacité du locuteur à commenter ce qu'il dit (comme si quelqu'un d'autre venait interrompre son propos) sans pour autant perdre le fil de ce qu'il dit.

**Corrigé 5** Toute langue repose sur un système de conventions que les locuteurs doivent accepter. De la même façon que nous acceptons qu'un billet de 20€ a pour tous nos condisciples deux fois plus de valeur qu'un billet de 10€, nous acceptons que le texte *un chien* désignera bien un chien pour nos condisciples. L'apprentissage d'une langue repose donc sur la croyance que les conventions linguistiques seront acceptées par nos condisciples. Pour qu'une langue existe, il faut qu'un groupe d'humains accepte de croire au système de conventions qu'elle suppose. On peut citer **Saussure**<sup>1916</sup> : « [La langue] est la partie sociale du langage, extérieure à l'individu, qui à lui seul ne peut ni la créer ni la modifier ; elle n'existe qu'en vertu d'une sorte de contrat passé entre les membre de la communauté. » On considère généralement qu'il existe deux types d'entités : les entités objectives, qui appartiennent au monde réel, et les entités subjectives, que je crée dans mon cerveau. Il existe en fait un troisième type d'entités, qu'Harari appelle les ENTITÉS INTER-SUBJECTIVES, et qui n'existent qu'à travers un accord entre les individus d'une communauté : Zeus, l'euro (la monnaie), la France, Google (la société) ou encore la langue française. Toutes ces choses n'existent que parce que les humains pensent qu'elles existent. Si les humains disparaissent, ces choses disparaîtront avec eux.

## 2 Produire un énoncé : La syntaxe mise en évidence par un exemple

### 2.1 Analyse et synthèse

Ce chapitre est entièrement consacré à l'étude de la production d'un énoncé ou plus exactement à la production d'une famille d'énoncés concurrents exprimant plus ou moins le même sens. Il est habituel de commencer l'étude d'une langue en analysant les textes (éventuellement oraux) produits dans cette langue. L'étude procède alors dans le sens de l'ANALYSE, c'est-à-dire du texte vers le sens. Nous pensons, à la suite de Lucien Tesnière et d'Igor Mel'čuk, qu'il est préférable d'étudier la langue dans le sens de la SYNTHÈSE, c'est-à-dire du sens vers le texte.

Pour comprendre le fonctionnement de la langue, il est donc nécessaire d'étudier la façon dont un énoncé est produit par un locuteur et les opérations qui conduisent à la production de cet énoncé. Nous mettrons en évidence un ensemble de contraintes auxquelles doit obéir le locuteur lors de l'énonciation et qui constitue la grammaire de la langue. Parmi ces contraintes, nous verrons que la langue nous impose une structuration hiérarchique qui constitue le cœur de la syntaxe.

### 2.2 Les observables : textes et sens

Un modèle se construit à partir d'OBSERVABLES, puisqu'il modélise le fonctionnement de quelque chose de réel, que l'on peut observer. Néanmoins, un modèle est amené à faire des hypothèses sur la façon dont ces observables sont produits et à construire ainsi un certain nombre d'objets qui ne sont pas observables.

Pour ce qui est de la langue, deux choses sont réellement observables : les textes et les sens. Pour les textes, c'est assez évident, même si la question de savoir quelle est la représentation d'un texte dans notre cerveau reste un problème. Pour les sens, c'est plus complexe : on ne peut pas observer le sens en tant que tel, mais on peut s'assurer qu'un texte est compréhensible et qu'il est compris. L'un des outils d'observation est la paraphrase : on peut demander à un locuteur de

reformuler un texte ou lui demander si deux textes ont le même sens. On peut ainsi considérer, à la suite d'Igor Mel'čuk, que « avoir le même sens » est un observable et **définir le sens** comme un invariant de paraphrases, c'est-à-dire comme ce qui est commun à tous les textes qui ont le même sens.

Nous faisons l'hypothèse qu'il y a entre le sens et le texte un niveau d'organisation syntaxique. Cette organisation n'est pas directement observable. Dans la suite, nous serons amenés à construire un grand nombre d'objets linguistiques et notamment des représentations syntaxiques. Il s'agit de **constructions théoriques** et de rien d'autre. On ne prouve leur « existence » qu'à l'intérieur d'une théorie. Elles existent dans le modèle, mais cela ne prouve pas qu'elles existent dans l'objet réel que nous modélisons. On ne peut pas réfuter directement leur existence. On ne peut que réfuter l'adéquation du modèle avec l'objet modélisé. Néanmoins, si le modèle est adéquat et suffisamment économique, alors les objets construits par le modèle acquièrent une part de réalité, deviennent plus tangibles.

### 2.3 Partir d'un sens

Considérons un locuteur du français qui veut exprimer un certain sens que nous allons essayer de décrire sans le verbaliser directement par une phrase. Ce sens concerne une personne  $x$  appelée Ali et un événement  $e$  concernant  $x$ . Cet événement est une maladie et la durée de cette maladie est de deux semaines. Nous pouvons représenter ce sens par la « formule » suivante :

$$\begin{aligned}x &: \text{'Ali'} \\e &: \text{'malade'}(x) \\&\text{'durer'}(e, \text{'2 semaines'})\end{aligned}$$

Figure 2.1 – Description formelle d'un sens

Dans cette représentation, il y a quatre éléments de sens considérés : 'Ali', 'malade', 'durer' et '2 semaines'. Ce dernier élément est en fait la combinaison de deux sens : l'unité de temps 'semaine' et le prédicat 'deux' qui la quantifie. Ces éléments de sens sont des sens de mots du français. Ils peuvent être exprimés de façons variées : par exemple, le sens 'durer' peut aussi bien être exprimé par le nom DURÉE que le verbe DURER ou encore la préposition PENDANT comme nous le verrons plus loin. Remarquons que nous distinguons les UNITÉS LEXICALES, notées en majuscule (DURÉE), de leur sens, noté entre guillemets simples ('durer'). Un sens exprimable par une unité lexicale est appelé un SENS LEXICAL. On utilise parfois le terme *sémantème* pour désigner un sens lexical, mais nous réserverons

ce terme pour des signes linguistiques élémentaires (voir chapitres 2.3 et 2.4).

Certains sens lexicaux fonctionnent comme des PRÉDICATS et possèdent des ARGUMENTS : le sens 'malade' possède toujours un argument (*quelqu'un est malade*), le sens 'durer' possède deux arguments (*quelque chose dure quelque temps*). Le sens 'Ali', quant à lui, renvoie à une entité du monde (ou plus exactement à la représentation mentale qu'en a le locuteur) et n'a pas d'argument. Dans la formule ci-dessus, les variables  $x$  et  $e$  nous ont servi à indiquer que certains sens étaient arguments d'autres :  $x$  désigne le sens 'Ali' et 'malade'( $x$ ) signifie que l'élément désigné par  $x$  est l'argument de 'malade'.

Les RELATIONS PRÉDICAT-ARGUMENT entre les sens lexicaux constituent la STRUCTURE PRÉDICATIVE. La structure prédictive exprime le CONTENU INFORMATIONNEL.

Nous nous contenterons de cette définition sommaire de la structure prédictive (voir compléments dans l'encadré ?? sur *Les composantes du sens*). Une définition plus précise sera donnée au Chapitre ?? sur *La structure syntaxique profonde*.

Nous avons présenté la structure prédictive par une FORMULE, mais on peut aussi la représenter par un GRAPHE (voir l'encadré ?? qui suit) que nous appelons le GRAPHE SÉMANTIQUE. Les nœuds du graphe sémantique sont les sens lexicaux 'Ali', 'malade', 'durer', 'semaine' et 'deux' et les arêtes représentent les RELATIONS PRÉDICAT-ARGUMENT. Les arêtes sont matérialisées par des flèches. Les arêtes sont étiquetées par des chiffres permettant de distinguer les différents arguments d'un prédicat : 1 pour le premier argument, 2 pour le deuxième, etc. L'ordre dans lequel les arguments sont numérotés est l'ORDRE DE SAILLANCE, dont nous reparlerons au Chapitre ?? sur *Les relations syntaxiques*.

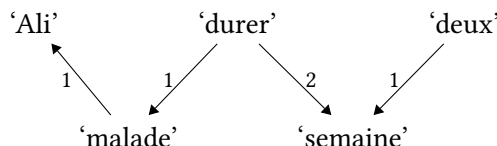


Figure 2.2 – Graphe sémantique

## 2.4 Graphe et arbre

Graphe et arbre sont des structures mathématiques très utilisées en sciences. Elles sont utilisées en linguistique pour la représentation du sens et de la structure syntaxique.

Un GRAPHE est une structure liant ensemble des éléments. Les éléments sont appelés les NŒUDS du graphe et les liens les ARÊTES. Un graphe est dit CONNEXE lorsque pour chaque couple de nœuds du graphe, il existe un ensemble d'arêtes formant un chemin connectant un nœud à l'autre.

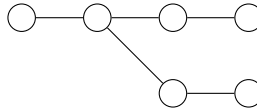


Figure 2.3 – Graphe connexe

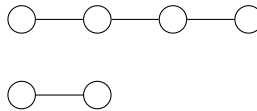


Figure 2.4 – Graphe non connexe

Un graphe est dit ACYCLIQUE s'il n'existe aucun chemin partant d'un nœud et revenant à ce nœud sans emprunter deux fois la même arête.

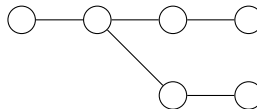


Figure 2.5 – Graphe acyclique

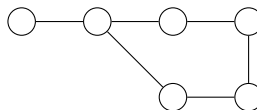


Figure 2.6 – Graphe avec cycle

Un ARBRE est un graphe connexe et acyclique dont un nœud, qu'on appelle la RACINE, est pointé. Cela revient à orienter les arêtes à partir de ce point. On peut donc aussi définir un graphe comme un cas particulier de graphe orienté.



Un **GRAPHE ORIENTÉ** est un graphe dont les arêtes sont orientées, c'est-à-dire distinguent un nœud **SOURCE** et un nœud **CIBLE** de l'arête. On représente généralement l'orientation par une flèche allant de la source à la cible.

Un **ARBRE** est donc aussi un graphe orienté connexe pour lequel chaque nœud est la cible d'une seule arête à l'exception d'un nœud, la racine de l'arbre. Tout nœud autre que la racine de l'arbre possède ainsi un **GOUVERNEUR** qui est l'unique nœud qui le prend pour cible. On représente traditionnellement les arbres « à l'envers » avec la racine en haut. Les nœuds qui ne sont le gouverneur d'aucun nœud, c'est-à-dire qui n'ont pas de **DÉPENDANTS**, sont appelés les **FEUILLES** de l'arbre. Un chemin orienté allant de la racine ou d'un nœud intérieur à une feuille est appelé une **BRANCHE**. Du fait que, par convention, chaque arête est toujours orientée vers le bas (la cible est en dessous de la source), il n'est pas nécessaire d'utiliser une deuxième convention et d'indiquer l'orientation par une flèche.

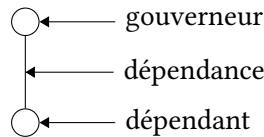


Figure 2.7 – Dépendance

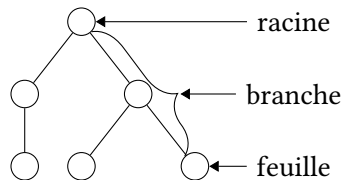


Figure 2.8 – Branche

La structure syntaxique d'une phrase est généralement représentée par un arbre dont les nœuds sont les unités lexicales.

Notons encore qu'il existe une notion d'acyclicité plus restrictive pour les graphes orientés : un graphe orienté est dit **ACYCLIQUE** s'il n'existe aucun chemin orienté permettant de partir d'un nœud et de source en cible de revenir au même nœud.

La **structure prédictive** d'un énoncé peut être formalisée par un **graphe orienté connexe et acyclique** (encore appelé **DAG**, de l'anglais *directed acyclic graph*) dont les nœuds sont les sens lexicaux et grammaticaux et les arêtes sont les dépendances sémantiques ou relations prédicat-argument entre ces sens.

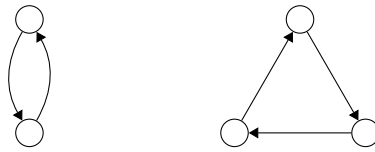


Figure 2.9 – Graphes avec cycle orienté

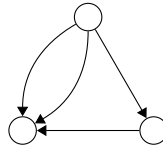


Figure 2.10 – Graphe sans cycle orienté

## 2.5 Les composantes du sens

Nous avons déjà expliqué dans le chapitre précédent la distinction que nous faisons entre le SENS LINGUISTIQUE et les INTENTIONS COMMUNICATIVES du locuteur. Reste à savoir, bien que cette question dépasse le cadre de cet ouvrage consacré à la syntaxe, comment modéliser le sens linguistique. Nous pensons que la structure prédicative (voir la section 2.3 *Partir d'un sens*) exprime une partie et une partie seulement du sens linguistique : il s'agit de l'**information pure** contenue dans le message — quels sont les **sens de base** que nous souhaitons utiliser et comment ils se **combinent**.

Au contenu informationnel s'ajoute au moins quatre autres types de contenus :

- La STRUCTURE COMMUNICATIVE, dont nous discuterons plus loin dans ce chapitre, indique ce qui est réellement informatif pour le destinataire, c'est-à-dire ce qu'on suppose qu'il sait déjà, ce qu'on souhaite souligner ou au contraire mettre en arrière-plan, etc. On appelle aussi cette structure l'EMBALLAGE DE L'INFORMATION, de l'anglais *information packaging*. Le terme le plus employé actuellement est *structure informationnelle*, de l'anglais *information structure*, mais nous éviterons absolument ce terme qui est une source de confusion évidente avec le *contenu informationnel*.
- La STRUCTURE RHÉTORIQUE indique le style (familier, poétique, humoristique, etc.) avec laquelle l'information sera communiquée (voir section suivante).
- La STRUCTURE ÉMOTIONNELLE indique quelles sont les émotions liées à cette information. Elle a surtout un impact sur la prosodie, mais peut influencer certains choix lexicaux (des termes injurieux par exemple).

- Par ailleurs, le contenu informationnel ne nous informe que si nous pouvons l’ancrer dans la réalité (ou plus exactement la représentation du monde que les interlocuteurs construisent dans leur cerveau à partir de la réalité), c’est-à-dire si on peut décider, par exemple, si ‘Ali’ renvoie à un objet du monde que nous connaissons déjà ou pas et si oui lequel. Les liens entre le contenu informationnel et le monde (ou plus précisément le CONTEXTE D’ÉNONCIATION, c’est-à-dire la partie du monde concernée par l’énonciation) constitue la STRUCTURE RÉFÉRENTIELLE. La référence joue surtout un rôle dans la planification et le choix de l’information permettant de s’assurer que l’interlocuteur identifiera le bon référent, mais une fois le message élaboré, la structure référentielle a peu d’incidence sur la réalisation du message. Ce qui compte surtout, c’est si le locuteur présente l’information comme nouvelle ou non et ceci appartient à la structure communicative.

Ce découpage du sens (à l’exception de la structure émotionnelle) et la représentation du contenu informationnel par un graphe sémantique ont été proposés par Žolkovskij et Mel’čuk en 1965 dans leur introduction à la Théorie Sens-Texte.

Notons encore qu’il existe des relations d’équivalence entre les structures prédicatives : une configuration de sens lexicaux peut être remplacée par un seul sens ou une autre configuration sans modifier le sens global. Un sens lexical peut être ainsi décomposé à la manière de ce que l’on fait quand on donne une définition d’un des sens d’un mot dans un dictionnaire. On peut considérer, à la suite d’Anna Wierzbicka (*Lingua Mentalis*, Academic Press, 1980), qu’il existe un ensemble d’unités minimales de sens à partir desquelles peuvent être définis tous les autres sens lexicaux. (De tels ensembles contenant une cinquantaine de sens minimaux ont été proposés.) Un contenu informationnel correspond ainsi à un ensemble de structures prédicatives équivalentes, dont les sens lexicaux sont plus ou moins décomposés.

## 2.6 Choisir des unités lexicales

Comment peut-on exprimer le sens que nous avons considéré en français ? Nous allons nous intéresser à deux formulations possibles de ce sens, suffisamment différentes pour illustrer notre propos.

- (1) La maladie d’Ali a duré deux semaines.
- (2) Ali a été malade pendant deux semaines.

L’analyse de la production de ces deux énoncés va nous permettre de mettre en évidence de nombreuses règles de grammaire, appartenant à la langue en général

ou spécifiques au français.

La première chose qui va déterminer la nature du texte que nous produisons est la façon dont nous réalisons chacun des éléments de sens du message que nous souhaitons communiquer. Pour simplifier, nous considérons que chaque sens va être réalisé par une unité lexicale : par exemple, 'malade' peut être réalisé par l'adjectif MALADE ou le nom MALADIE ; 'durer' peut être réalisé par le nom DURÉE, le verbe DURER ou les prépositions DURANT ou PENDANT.

Comment se font les CHOIX LEXICAUX ? Les choix lexicaux dépendent bien sûr du sens que l'on veut exprimer : par exemple, pour parler de l'ingestion d'un aliment, on devra choisir entre BOIRE et MANGER selon la nature de cet aliment, mais aussi entre MANGER, DÉGUSTER ou DÉVORER selon la façon dont on l'a ingéré ou encore entre MANGER et BOUFFER selon la familiarité avec laquelle nous avons l'habitude de nous adresser à notre interlocuteur. Cependant, la subtilité du sens que nous voulons exprimer n'est pas le seul facteur qui contraint les choix lexicaux. Ils existent d'autres contraintes qui relèvent de la grammaire et dont nous allons parler maintenant.

## 2.7 Les quatre moyens d'expression du langage

Lorsque nous parlons, nous avons quatre moyens à notre disposition pour exprimer du sens.

- Le premier moyen, ce sont les **mots** ou plus exactement les **unités lexicales et grammaticales** (voir la partie 2 consacrée aux *Unités de la langue*).
- Le deuxième moyen, c'est la façon de combiner les mots et en particulier l'**ordre** dans lequel nous les mettons : « *Ali regarde Zoé* » ne veut pas dire la même chose que « *Zoé regarde Ali* ». Plus subtilement, « *À Paris, Ali travaille le lundi* » n'est pas exactement synonyme de « *Le lundi, Ali travaille à Paris* » (le premier énoncé n'implique pas qu'Ali travaille tous les lundis et qu'il est à Paris tous les lundis où il travaille).
- Le troisième moyen d'expression du langage est la **prosodie**. La séquence *tu viens* peut être prononcée de bien des façons et avoir autant de sens différents. Prononcée avec une voix forte et autoritaire et un accent montant sur la première syllabe, ce sera un ordre : « *TU VIENS!* ». Prononcée d'une voix suave avec une courbe mélodique montante, ce sera une question ou une invitation : « *Tu viens?* ». Prononcée d'une voix neutre avec une courbe mélodique descendante ce sera une simple constatation : « *Tu viens.* ».
- Le quatrième moyen à notre disposition dépasse le simple usage de la voix. Une énonciation en face à face s'accompagne toujours de **mimiques faciales**

et de gestes divers. Ceux-ci sont beaucoup plus codifiés et beaucoup plus riches qu'on ne le pense généralement. Ils vont accompagner la parole et parfois se combiner avec elle. Ainsi « *Ton pull* » suivi d'un geste avec le pouce levé est un énoncé équivalent à « *Ton pull est vraiment super* ». Et une moue désapprobatrice en prononçant « *Jean?* » en dira beaucoup plus qu'un long discours sur la confiance qu'on met en Jean.

## 2.8 Contraintes syntaxiques sur les choix lexicaux

À peu près n'importe quel énoncé en français possède un verbe principal et ce verbe est conjugué. Cela signifie que l'un des sens de notre message devra être lexicalisé par un verbe, par exemple 'durer' par DURER. L'autre possibilité est de lexicaliser 'malade' par un verbe, mais comme il n'existe pas de verbe \*MALADER en français, nous lexicalisons 'malade' par un adjectif et nous en faisons une tournure verbale ÊTRE MALADE grâce au verbe ÊTRE. Le verbe ÊTRE n'a donc aucune contribution sémantique ici ; il a juste un rôle grammatical, qui est d'assurer que l'élément principal de la phrase est bien un verbe et donc de faire d'un adjectif l'équivalent d'un verbe. Ce rôle très particulier du verbe ÊTRE lui vaut le nom de COPULE.

Une fois choisi l'élément principal de la phrase, les éléments lexicaux choisis ensuite se voient imposer un certain nombre de choses, à commencer par leur PARTIE DU DISCOURS (les principales parties du discours du français sont nom, verbe, adjectif et adverbe). L'élément principal de la phrase est un verbe conjugué comme on vient de le dire (ceci sera justifié dans le chapitre consacré à la tête dans la Partie 3). Les éléments qui vont dépendre de ce verbe devront ensuite être soit des noms, soit des adverbes selon la relation qu'ils entretiennent avec ce verbe.

Commençons par le cas de la phrase (1) (*La maladie d'Ali a duré deux semaines*) dont le verbe principal est DURER : le sens 'malade', qui est le premier argument du sens 'durer' devra être réalisé comme le sujet de DURER et devra être un nom. La lexicalisation de 'malade' par un nom donne ainsi MALADIE. Le sens 'semaine' sera également réalisé par un nom, qui sera un complément direct du verbe. Nous discuterons dans le chapitre sur les fonctions syntaxiques de la Partie 6 de ces compléments de mesure qui ressemblent à des compléments d'objet directs mais n'en possède pas toutes les bonnes propriétés.

Le cas de la phrase (2) (*Ali a été malade pendant deux semaines*) est plus complexe. Son élément principal est la tournure verbale ÊTRE MALADE. L'unique argument de 'malade' est 'Ali', qui sera donc réalisé comme sujet de la tournure

verbale et sera un nom, ALI en l'occurrence. Le sens 'durer' est lui aussi directement lié à 'malade' et devra donc être réalisé comme un dépendant direct de la tournure verbale. Mais contrairement aux cas précédents, 'durer' n'est pas un argument de l'élément principal : c'est même l'inverse, c'est lui qui prend l'élément principal comme argument sémantique. Dans un tel cas, l'élément de sens doit être réalisé par un groupe adverbial. C'est ainsi que le sens 'durer' peut être réalisé par la préposition PENDANT. Le deuxième argument de 'durer' est réalisé par un nom, qui est le complément de la préposition. La préposition et son complément, *pendant deux semaines*, forme un groupe de distribution équivalente à un adverbe (comme *longtemps* par exemple).

## 2.9 Règles et exceptions

La plupart des règles ont des exceptions. C'est le cas de la règle qui veut que l'élément principal d'une phrase soit un verbe conjugué. Il existe en effet quelques éléments lexicaux particuliers qui ne sont pas des verbes mais ont la propriété de pouvoir être l'élément principal d'une phrase, comme l'adverbe HEUREUSEMENT ou le bizarre BONJOUR :

*Heureusement qu'il y en a !*

*Sinon, bonjour le chômage des linguistes.*

Par ailleurs, il existe des contextes où la règle ne s'applique pas, notamment en réponse à une question (*Tu viens à la fac demain? Oui à 10h pour le cours de syntaxe*) ou encore pour les titres (*Nouvel incident diplomatique entre la France et l'Allemagne*). Cela ne signifie pas que la règle est fausse, mais qu'il faut bien préciser quand elle s'applique. Quant aux exceptions, il faut les **lister** et inclure ces éléments dans une classe particulière d'élément que nous appelons les **prédicatifs** et les **locutifs** (voir le Chapitre ?? sur *Les catégories microsyntactiques*).

## 2.10 Structure hiérarchique

Lors du passage du sens au texte, tout se passe comme si on suspendait le graphe sémantique par l'un de ses nœuds dont on décide de faire le verbe principal de l'énoncé et qu'on parcourait le graphe à partir de ce nœud pour les autres lexicalisations. Nous pouvons illustrer cela par les schémas suivants : à gauche nous représentons le graphe sémantique suspendu par un de ses nœuds (visé ici par une grosse flèche blanche) et à droite nous avons la structure hiérarchique correspondante où les sens lexicaux ont été lexicalisés.

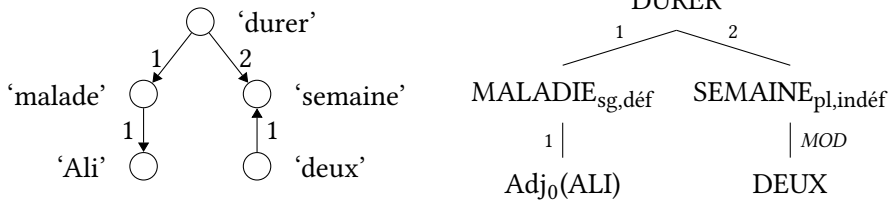


Figure 2.11 – Hiérarchisation à partir de ‘durer’

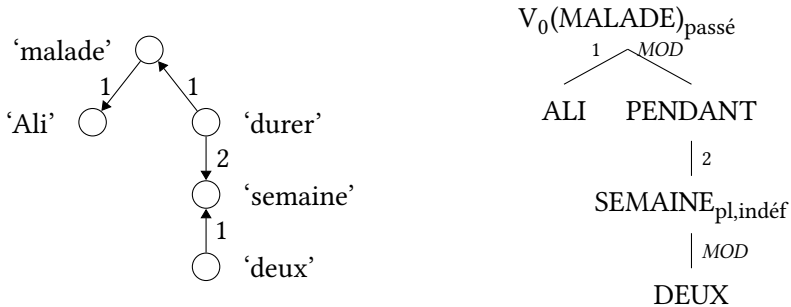


Figure 2.12 – Hiérarchisation à partir de ‘malade’

La structure que nous obtenons s’appelle un **ARBRE DE DÉPENDANCE SYNTAXIQUE PROFOND** (voir le Chapitre ?? entièrement consacré à *La syntaxe profonde*). Chaque nœud de l’arbre est occupé par une unité lexicale correspondant à un sens. D’autres sens (non considéré ici) donne des éléments grammaticaux qui vont se combiner aux unités lexicales : temps pour les verbes (présent, passé, etc.), nombre (singulier, pluriel) et définitude (défini, indéfini) pour les noms. Le nœud au sommet est appelé la **RACINE** de l’arbre (voir l’encadré ?? sur *Graphe et arbre*). Tous les nœuds à l’exception de la racine dépendent d’un autre nœud appelé leur **GOUVERNEUR** (syntaxique profond). À l’inverse, les nœuds qui dépendent d’un autre nœud en sont appelés les **DÉPENDANTS** (syntaxiques profonds). Le lien entre deux nœuds est appelé une **DÉPENDANCE** (syntaxique profonde).

Chaque dépendance est étiquetée en fonction de son parcours : les relations sémantiques qui ont été parcourues dans le sens de la flèche (du prédicat vers l’argument) donne une **DÉPENDANCE SYNTAXIQUE ACTANCIELLE** (que nous numérotions comme la dépendance sémantique correspondante), tandis que les relations sémantiques qui ont été parcourues à contre-courant donne une **DÉPENDANCE SYNTAXIQUE MODIFICATIVE** (étiquetée *MOD*).

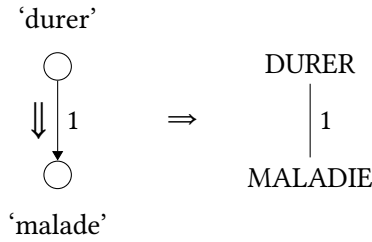


Figure 2.13 – Dépendance actancielle

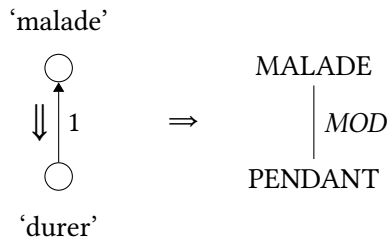


Figure 2.14 – Dépendance modificative

Comme nous l’avons expliqué dans la section précédente, chaque unité lexicale se voit imposer sa partie du discours par son gouverneur, la racine étant un verbe conjugué. Ainsi dans la phrase (1), ALI, qui dépend du nom MALADIE, devrait être réalisé par un adjectif, ce que nous indiquons par la notation  $\text{Adj}_0(\text{ALI})$ . C’est la préposition DE qui assurera cette TRANSLATION (*la maladie d’Ali*) et permettra à ALI d’être complément du nom (voir la section du Chapitre ?? sur *La translation*). De la même façon, la notation  $\text{V}_0(\text{MALADE})$  indique que MALADE occupe une position où un verbe est attendu et c’est la copule ÊTRE qui assurera la translation d’adjectif en verbe.

## 2.11 Du sens au texte : de 3D à 1D

Nous avons mentionné dans la section 1.3 *Sons et textes* le caractère unidimensionnel de la chaîne parlée. Le sens lui est localisé dans notre cerveau : un certain nombre de zones de notre cerveau s’activent simultanément (ou les unes après les autres) et se mettent en réseau : le sens est donc fondamentalement un objet au moins tridimensionnel (quadridimensionnel si l’on prend en compte la dimension temporelle et le caractère dynamique de la construction du sens). Le passage du sens au texte s’accompagne donc d’une réduction de dimensionnalité, un pas-



sage de la dimension 3 à la dimension 1, une suite ordonnée d'unités élémentaires. Il semble que la langue effectue ce changement de dimension en deux étapes :

- le passage de la dimension 3 à la dimension 2 est une hiérarchisation du sens, le passage d'un graphe à un arbre ;
- le passage de la dimension 2 à la dimension 1 est la linéarisation de ce graphe.

La phase de hiérarchisation s'accompagne d'une phase de « lexicalisation », c'est-à-dire de choix de signes linguistiques élémentaires, des unités lexicales et des unités grammaticales, qui se contraignent les unes les autres. La phase de linéarisation s'accompagne d'une « morphologisation » des signes, c'est-à-dire de combinaison des signifiants selon des règles morpho-phonologiques propres à chaque langue. Une telle architecture est à la base de la Théorie Sens-Texte, que nous évoquerons dans l'encadré ?? éponyme.

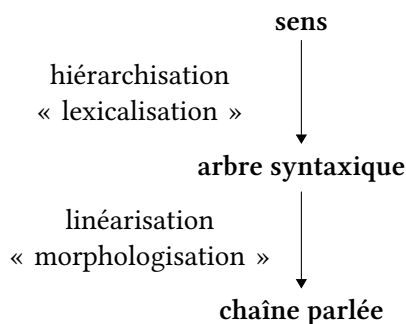


Figure 2.15 – Please provide a caption

## 2.12 Ce que la langue nous force à dire

En produisant les phrases (1) et (2) (section 2.6), nous avons exprimé plus que le sens de départ. Celui-ci ne contenait pas d'information temporelle sur le moment de la maladie d'Ali, mais nous avions besoin d'une telle information pour conjuguer le verbe principal. C'est la **grammaire** du français qui nous **contraint**, en nous obligeant à ajouter une flexion au verbe principal de la phrase, à situer l'événement dont nous voulons parler par rapport au moment où nous parlons. Nous devons décider si la maladie a eu lieu avant maintenant (*Ali a été malade pendant deux semaines*) ou si l'événement aura lieu après maintenant (*Ali sera malade pendant deux semaines*) ou encore s'il est en cours (*Ali est malade depuis deux semaines*).

De même, en français, on ne peut pas lexicaliser un sens par un nom sans préciser le nombre (*Ali a mangé **une** pomme* vs *Ali a mangé **des** pommes*) et sans préciser si la chose est déjà connue ou non (*Ali a mangé **la** pomme* vs *Ali a mangé **une** pomme*) (voir Chapitre ??). C'est l'une des raisons pour lesquelles nos deux phrases de départ (1) et (2) ne sont pas parfaitement synonymes, puisque, dans (1), 'malade' est lexicalisé par un nom et est donc accompagné d'un article défini, lequel présuppose que la maladie d'Ali est connue des interlocuteurs au moment où la phrase est prononcée, ce qui n'est pas le cas en (2).

En conclusion, il n'est pas possible en français de communiquer uniquement notre sens de départ ! Nous devons y ajouter des informations et en particulier situer le fait dont nous parlons (la maladie d'Ali) dans le temps.

## 2.13 Les sens grammaticaux

Dans un article de 1959 sur la traduction, le grand linguiste d'origine russe Roman Jakobson remarque que « Les langues diffèrent essentiellement par ce qu'elles *doivent* communiquer et pas par ce qu'elles *peuvent* communiquer. » Il explique par exemple que la traduction de la phrase anglaise *I hired a worker* en russe nécessiterait deux informations supplémentaires : le verbe en russe devra indiquer si l'embauche a été complétée ou pas, tandis que le nom en russe devra indiquer le genre et donc s'il s'agit d'un travailleur ou d'une travailleuse. À l'inverse, la phrase russe n'aura pas à choisir entre article défini ou indéfini (*une* ou *la travailleuse*), ni entre les temps verbaux *hired* vs. *have hired*. Sur cette nécessité de faire des choix dès qu'il s'agit de catégories grammaticales obligatoires, Jakobson renvoie à Franz Boas 1938 : « Nous devons choisir parmi ces aspects, et l'un ou l'autre doit être choisi. »

Il existe quantités de sens plus ou moins curieux qui sont ainsi imposés par la grammaire d'une langue. Par exemple, il existe une langue amérindienne, le *nootka*, parlée sur l'île de Vancouver au Canada, où le verbe s'« accorde » avec son sujet en fonction de particularités physiques du référent sujet ; on doit nécessairement choisir entre l'une des sept possibilités suivantes : normal, trop gros, trop petit, borgne, bossu, boiteux, gaucher (Sapir 1915 ; Mel'čuk 1993). On ne peut donc pas dire ce que fait une personne donnée sans dire si cette personne est normale ou si elle est affublée d'une des six particularités physiques retenues par la grammaire de cette langue.

Si la nature des sens qui sont grammaticalement exprimés au travers des langues du monde est assez vaste, elle est quand même assez homogène. D'une part, il s'agit généralement de sens abstraits ayant une **importance cognitive fon-**

damentale et qui ont de surcroît une importance culturelle énorme, puisqu'ils apparaissent dans pratiquement chaque phrase et donc modèlent la pensée des locuteurs à chaque instant. D'autre part, ces sens s'associent naturellement à certaines parties du discours : les langues ont ainsi généralement une classe d'éléments lexicaux qui varient en temps et/ou aspect et désignent des procès (et correspondent grosso modo à nos verbes) et une classe d'éléments lexicaux qui varient en nombre et désignent des entités (et correspondent à nos noms).

On trouvera dans le volume 2 du *Cours de morphologie générale* d'Igor Mel'čuk publié en 1993 une typologie de tous les sens qui doivent être exprimés obligatoirement dans une langue au moins. Nous donnons ici l'exemple du **respect** en japonais.

Un Japonais ne peut pas dire « *Pierre est malade* » sans préciser deux choses qui, au premier abord, peuvent sembler étonnantes à un locuteur du français :

- est-ce que Pierre est quelqu'un de respectable ou non ?
- est-ce que la personne à qui je parle est quelqu'un de respectable ou non ?

Ces deux informations doivent obligatoirement figurer dans le choix lexical de l'adjectif et la conjugaison de la copule :

1. Pieru-san-wa go-byoki **desu**.
2. Pieru-wa byoki **desu**.
3. Pieru-san-wa go-byoki **da**.
4. Pieru-wa byoki **da**.

La respectabilité envers Pierre est exprimée par le suffixe SAN et par le choix d'une forme polie GO-BYOKI de l'adjectif 'malade'. La respectabilité envers l'interlocuteur est exprimée par la forme polie de la copule, *desu*, opposée à la forme neutre *da*. Ces deux marques de respectabilité sont indépendantes, ce qui nous donne quatre formes possibles.

La respectabilité envers l'interlocuteur peut être comparée au **vouvoiement** en français. Mais alors que le choix entre *tu* et *vous* se limite au cas, évitable, où l'on interpelle directement son interlocuteur, le choix entre la forme verbale respectueuse et la forme familière se pose pour chaque phrase en japonais.

## 2.14 Choisir le verbe principal

Nous savons que la structure syntaxique est une structure hiérarchique et que la construction de cette structure commence par le choix d'un élément sémantique pour être la racine de l'arbre syntaxique et donc le verbe principal de

l'énoncé produit. Le choix de l'élément principal de la phrase peut être guidé par les choix lexicaux (quel sens lexical peut donner un verbe) ou grammaticaux (sur quel sens lexical veut-on ou peut-on ajouter tel ou tel sens grammatical). Mais ce choix est avant tout guidé par ce dont on est en train de parler et ce qu'on veut dire. Ainsi le même message, selon que nous sommes en train de parler d'Ali ou de sa maladie sera exprimé différemment : si l'on est en train de parler d'Ali et que l'information nouvelle que l'on veut communiquer est sa maladie, on choisira plutôt (2) (*Ali a été malade pendant deux semaines*), alors que si l'on est en train de parler de sa maladie et que l'information nouvelle que l'on veut communiquer est seulement sa durée, on peut préférer (1) (*La maladie d'Ali a duré deux semaines*). Le verbe principal de la phrase est ainsi généralement l'élément central du RHÈME, c'est-à-dire l'élément de sens que l'on souhaite prioritairement communiquer, tandis que le sujet du verbe principal est généralement le THÈME, c'est-à-dire ce dont on parle, ce sur quoi porte le rhème. L'indication des rhème et thème ne fait pas partie du contenu informationnel, représenté par la structure prédicative, mais dépend de la façon dont on communique le contenu informationnel, ce que nous avons appelé la **structure communicative** (voir l'encadré ?? sur *Les composantes du sens*). Formellement, la délimitation des thème et rhème se surajoute à la structure prédicative en indiquant quelle zone constitue le rhème et quelle autre le thème.

## 2.15 Les contraintes de la grammaire

Reprenons rapidement la liste des contraintes qui nous ont été imposées lors de la production de nos deux phrases. La principale contrainte est la structure hiérarchique de l'énoncé. De cette contrainte majeure découlent des contraintes sur les parties du discours des unités lexicales choisies : la racine de l'arbre syntaxique devra être un verbe, les actants d'un verbe des noms, les modifieurs d'un verbe des adverbes et les dépendants d'un nom des adjectifs (pour les notions d'actant et de modifieur, voir le Chapitre ?? sur la *Syntaxe profonde*). Ainsi les sens 'malade' et 'durer' peuvent recevoir différentes lexicalisations (MALADE vs MALADIE, DURER vs PENDANT), mais ces différents choix ne sont pas totalement indépendants : il serait par exemple difficile de faire une phrase naturelle avec les unités lexicales *malade* et *durer* exprimant notre sens de départ. (On peut en faisant deux phrases : « *Pierre a été malade. Ça a duré deux semaines.* ».)

Quand les unités lexicales n'appartiennent pas à la partie du discours attendues, d'autres unités lexicales sont introduites, comme la copule ÊTRE ou la préposition DE, pour « masquer » le mauvais choix catégoriel. En fonction de leur

## 2.16 D'où viennent les règles de la grammaire ?

partie du discours, les unités lexicales se voient assignées des unités grammaticales, telles que le temps pour les verbes et le nombre et la définitude pour les noms. Il existe encore d'autres contraintes. Par exemple, en français, un verbe conjugué, à l'exception de l'impératif, doit toujours avoir un sujet et ce sujet devra se placer devant lui (il existe quelques cas où le sujet peut être « inversé » mais pas ici). Enfin, le verbe devra s'accorder avec son sujet. De même, le nom devra avoir un déterminant (article ou autre) exprimant la définitude, accordé en nombre avec le nom placé avant lui. Ces différentes contraintes – la nature de l'élément principal d'une phrase, la partie du discours des dépendants, les unités grammaticales obligatoires comme la conjugaison pour les verbes ou l'article pour les noms, l'accord, le placement des mots les uns par rapport aux autres –, sont des **RÈGLES GRAMMATICALES** du français. La **GRAMMAIRE** d'une langue, ce sont toutes les contraintes que nous impose cette langue quand nous parlons.

L'expression même des contraintes grammaticales repose sur les « connexions » entre différentes parties de l'énoncé. Ces connexions constituent, comme nous le verrons, le squelette des **RELATIONS SYNTAXIQUES** entre les mots, les groupes de mots, mais aussi des unités plus petites que les mots. En mettant en évidence les contraintes que ces unités, que nous appellerons les **UNITÉS SYNTAXIQUES**, s'imposent les unes aux autres, nous avons mis en évidence l'existence d'une **STRUCTURE SYNTAXIQUE**. L'étude de cette structure, qui modélise la façon dont les signes linguistiques se combinent les uns avec les autres, constitue l'objet central de la **SYNTAXE** et donc de cet ouvrage.

## 2.16 D'où viennent les règles de la grammaire ?

D'où viennent ces règles ? Ces règles n'ont pas été inventées par des grammairiens ; elles existent depuis beaucoup plus longtemps que les grammairiens et elles sont pour la plupart indépendantes d'eux. Aussi le terme « régularité » correspondrait-il peut-être mieux que « règle » au caractère « naturel » de ces contraintes imposées par la grammaire de la langue. Le linguiste cherche juste à décrire la grammaire qu'utilisent naturellement les locuteurs natifs d'une langue.

Il nous faut quand même dire quelques mots de la **GRAMMAIRE NORMATIVE**, élaborée par les « grammairiens » et dont l'objectif est de prescrire le « bon » usage du français. L'institutionnalisation de la normativité n'est pas universelle, mais plutôt une particularité française. Dans des traditions différentes, en Allemagne ou aux États-Unis par exemple, on considère davantage qu'en France qu'un locuteur natif sait parler sa langue : on n'enseigne que très peu de grammaire à l'école et il n'y a pas d'Académie nationale chargée de protéger la langue.

### Quel est le bon français ?

La distinction entre grammaire descriptive et grammaire normative est importante quand on tente de répondre à la question suivante : la phrase suivante est-elle grammaticale ?

- (3) Après qu'il se soit assis, elle s'est mise à parler.

La réponse est non selon les grammaires normatives qui estiment que APRÈS QUE doit être suivi de l'indicatif, et oui, si on se base sur une grammaire descriptive, étant donné qu'il s'agit d'un énoncé contenant des structures attestées régulièrement. Une grammaire descriptive nous informera qu'en français contemporain, la conjonction APRÈS QUE est suivie soit de l'indicatif, soit du subjonctif. La description tentera peut-être d'expliquer quand on utilise l'un ou l'autre. De plus, elle notera probablement que l'utilisation de l'indicatif a un caractère plus écrit à cause de la prescription de la grammaire normative.

Pour la grammaire normative seule l'utilisation de l'indicatif est correcte. Pour ce jugement, les grammaires normatives se basent sur :

- un certain conservatisme : « ce qui est vieux est meilleur » ;
- et sur un certain élitisme : « la langue des riches, puissants et instruits est meilleure ».

La linguistique ne s'intéresse pas à ces jugements esthétiques. Les règles linguistiques sont toujours descriptives et jamais prescriptives.

### Quand les grammairiens ont gagné

La grammaire normative peut avoir une influence sur le comportement linguistique des locuteurs d'une langue. En effet, une prescription peut être acceptée par la langue et donc entrer dans les régularités de la langue. Dans ce cas-là, la linguistique s'intéresse bien entendu à cette régularité.

Un exemple de ce phénomène concerne l'accord au masculin. Aujourd'hui, les francophones accordent naturellement au masculin des éléments coordonnés composés de noms de genre différents. On dit donc :

- (4) Arrivèrent alors un homme et cinquante femmes très **élégants**.

et non :

- (5) \*Arrivèrent alors un homme et cinquante femmes très **élégantes**.

si on veut qualifier les 51 personnes d'élégantes. Mais cette règle est une invention machiste en correspondance avec la pensée de l'époque où la règle a été prescrite : « [La hiérarchie entre le masculin et le féminin] remonte au XVII<sup>e</sup> siècle lorsqu'en 1647 le célèbre grammairien Vaugelas déclare que « *la forme masculine a prépondérance sur le féminin, parce que plus noble* ». Dorénavant, il faudra

écrire : « *Les légumes et les fleurs sont frais* » et faire en sorte que l'adjectif s'accorde au masculin, contrairement à l'usage de l'époque qui l'aurait accordé au féminin. En effet, au Moyen Âge, on pouvait écrire correctement, comme Racine au XVII<sup>e</sup> siècle : « *Ces trois jours et ces trois nuits entières* » - l'adjectif « *entières* » renvoyant alors à « *nuits* » autant qu'à « *jours* ». » (Agnès Callamard, « Droits de l'homme » ou « droits humains » ?, *Le sexisme à fleur de mots*, Le monde diplomatique, mars 1998, page 28)



## Exercices

**Exercice 1** Chercher des couples de paraphrases dont les verbes principaux correspondent à des sens différents (comme dans l'exemple de la section 2.6 où les verbes principaux de notre couple de paraphrase correspondent au sens 'durer' vs. 'malade'). On pourra ensuite proposer la structure prédicative commune à ce couple de paraphrase, puis les structures syntaxiques profondes des deux phrases proposées.

**Exercice 2** Quels sont les quatre moyens d'expression du langage ? (Le premier étant les mots.)

**Exercice 3** Expliquez en quoi la phrase « Excellent, ce café ! » dévie des règles générales du français et proposez une description.

**Exercice 4** Quel problème pose la traduction de *son sac* en anglais ? Qu'est-ce que cela illustre comme différence grammaticale entre les deux langues ?

**Exercice 5** Pour vous rendre compte de la complexité du choix entre l'article défini et l'article indéfini, essayez d'expliquer (de manière claire et compréhensible pour un apprenant du français, japonais par exemple)

quand il faut dire « *Où puis-je trouver les toilettes ?* » et quand faut-il choisir l'article indéfini « *Où puis-je trouver des toilettes ?* ».

**Exercice 6** Il ne faut pas croire que le système de politesse du français est beaucoup plus simple que celui du japonais. Essayez d'expliquer (de manière claire et compréhensible pour un apprenant du français par exemple)

- quand et avec qui on peut utiliser les verbes *bouffer*, *manger* ou *dîner* pour parler d'une personne en train de prendre son repas du soir ?
- quelles sont les règles (tutoiement, utilisation du nom/prénom) dans la situation suivante : un professeur tutoie normalement sa collègue, Mme Marie Dupont. Maintenant, il la présente et s'adresse à elle devant un groupe d'étudiants.



### Lectures additionnelles

Si l'on est intéressé par l'inventaire des significations morphologiques que les langues nous obligent à communiquer, on pourra consulter le deuxième volume du monumental cours de morphologie d'Igor Mel'čuk. La citation de Jakobson de la section ?? est extraite d'un article très lisible. On peut aussi lire le très bon article publié dans le New York Times Magazine par Guy Deutscher. L'article d'Edward Sapir sur le nootka se trouve dans sa sélection d'écrits. Les autres sujets abordés dans ce chapitre seront largement développés dans la suite de l'ouvrage.

Franz Boas<sup>1938</sup> Language, *General Anthropology*, Boston.

Roman Jakobson<sup>1959</sup> On linguistic aspects of translation, *On translation*, 3, 30-39.

Guy Deutscher, Does your language shape how you think ?, article du



*New York Times Magazine* du 26 août 2010. (<http://www.nytimes.com/2010/08/29/magazine/29language-t.html>)

Igor Mel'čuk (??) *Cours de morphologie générale*, 5 volumes, Presses de l'Université de Montréal/CNRS.

Edward Sapir 1973 *Selected Writings of Edward Sapir*, édites par D. Mandelbaum, University of California Press, Berkeley.



## Citations originales

Citations de la section ??.

Jakobson (1959) :

*Languages* differ essentially in *what* they must convey and not in *what* they may convey.

Boas (1938 :132) :

We have to choose between these aspects, and one or the other must be chosen.



## Corrections des exercices

**Corrigé 1** Il existe de nombreux exemples. Nous en proposons deux :

- (6) Ali pleure parce que Zoé est partie.
- (7) Le départ de Zoé a fait pleurer Ali.
- (8) Les Chinois consomment de plus en plus de viande.
- (9) La consommation de viande des Chinois a augmenté.

Pour la représentation sémantique et les représentations syntaxiques profondes, voir le Chapitre ???. Vous pouvez néanmoins vous essayer dès maintenant à proposer des représentations pour les exemples ci-dessus.

**Corrigé 2** Voir la section ??.

**Corrigé 3** Nous avons dit que le français demandait que la racine de l'arbre soit un verbe, mais il est également possible que ce soit un adjectif. Dans ce cas, l'argument de l'adjectif ne peut pas être réalisé comme sujet : pour qu'il le soit, il faut ajouter la copule (*ce café est excellent*), qui devient la tête syntaxique. Il est néanmoins possible d'ajouter l'argument de l'adjectif comme un élément détaché (voir la partie 6 sur la macrosyntaxe).

**Corrigé 4** Le français *son sac* peut être traduit en anglais par *his bag* ou *her bag* selon que le propriétaire est un homme ou une femme. Cela illustre le fait que le nom en français a la catégorie du genre et que le déterminant possessif doit s'accorder en genre avec le nom qu'il détermine (*son sac* vs. *sa valise*), alors que l'anglais qui n'a pas de genre nominal, distingue pour les référents d'un pronom possessif s'il s'agit d'un non humain (*its*), d'un humain féminin (*her*) ou d'un humain masculin (*his*). Le français

n'exprime pas cette information. Une langue peut exprimer les deux informations en même temps, comme l'allemand (*sein Koffer, ihr Koffer, seine Tasche, ihre Tasche*, où *Koffer* est un nom masculin qui signifie 'valise', *Tasche* un nom féminin qui signifie 'sac', *sein* signifie 'his' et *ihr* signifie 'her').

**Corrigé 5** L'article défini présuppose qu'il existe des toilettes : on l'emploie quand on sait qu'on est dans un endroit où il y a des toilettes. Au contraire, on préfère utiliser l'indéfini quand on n'est pas sûr qu'il y ait des toilettes à proximité ou qu'on pense qu'il peut y en avoir plusieurs.

**Corrigé 6** Cette situation est complexe puisqu'on s'adresse simultanément à Marie Dupont et aux étudiants, ce qui peut nous amener à vouvoyer Marie Dupont, alors qu'on la tutoie dans d'autres conditions.



## 3 La modélisation : Préciser l'objectif de notre étude

### 3.1 Définition

Ce chapitre tente de caractériser ce qu'est un modèle linguistique et précise le type de modélisation dans lequel s'inscrit notre présentation de la syntaxe. Il n'est pas essentiel à la lecture de la suite de l'ouvrage, mais permet de lever certains présupposés méthodologiques.

Nous appelons **MODÈLE** un objet construit par le scientifique afin de **simuler les propriétés de l'objet d'étude**. L'objet d'étude est déterminé par le **CADRE THÉORIQUE**. Notre position théorique est de considérer que nos objets d'étude sont les langues vues comme des correspondances entre sens et textes. Ce que nous entendons par sens et textes est également déterminé par le cadre théorique (voir les sections 1.3 sur *Sons et textes* et 1.5 sur *Sens et intention communicative*). Le sens est représenté dans le modèle par un objet du modèle que nous appelons la **REPRÉSENTATION SÉMANTIQUE**. Un modèle, dans ce cadre théorique, devra donc être capable de construire pour chaque représentation sémantique tous les textes correspondants et pour chaque texte toutes les représentations sémantiques correspondantes.

Un modèle permet de faire des **PRÉDICTIONS** : par exemple, pour un sens, ou plus exactement pour une représentation sémantique, qu'on n'a jamais envisagé avant, on doit être capable de construire les textes qui lui correspondent. Il en découle qu'un modèle est **FALSIFIABLE** : on peut évaluer si les prédictions du modèle correspondent à notre cadre théorique et dire par exemple si tel texte ne peut pas correspondre à tel sens et si un texte donné ne peut correspondre à aucun sens et n'est donc pas un texte de la langue. On peut ainsi construire un contre-exemple, c'est-à-dire une association sens-texte qui n'est pas prédit par le modèle, mais qui appartient à notre objet d'étude, ou, inversement, une association prédite par le modèle, mais qui n'appartient pas à la langue (telle que nous l'envisageons dans notre cadre théorique). L'**ADÉQUATION** du modèle aux données est sa capacité à faire de bonnes prédictions.

Il faut distinguer la **falsifiabilité d'un modèle** et la **RÉFUTABILITÉ de la théo-**

rie dans laquelle se situe ce modèle. Il relève des choix théoriques de prendre en compte ou pas tel ou tel phénomène, comme par exemple les limitations mémorielles ou les erreurs de performance (voir la section ?? sur *Langue et parole, compétence et performance*). On peut réfuter le choix théorique de prendre ou non en compte de tels phénomènes. Mais une fois le choix théorique fait, le modèle doit s'y conformer. Évidemment, la frontière entre théorie et modèle est mouvante et il est tentant d'adapter la théorie aux résultats du modèle. (Voir compléments dans le corrigé de l'exercice 2.)

La **PORTÉE** du modèle mesure l'ambition théorique du modèle. Par exemple un modèle qui ne prend pas en compte la sémantique aura une portée moins grande qu'un modèle proposant une représentation du sens. La portée peut également concerner les données : un modèle qui prend en compte les erreurs de performance aura une portée plus grande qu'un modèle qui se limite aux productions idéales du locuteur.

La **COUVERTURE** du modèle est l'ensemble des données de l'objet d'étude qui ont été décrites par le modèle. Plus sa couverture est grande, meilleur est le modèle.

L'**ÉCONOMIE** du modèle mesure la quantité de paramètres nécessaires à la modélisation d'un phénomène. Plus un modèle est économique, meilleur il est. Ceci repose sur l'idée qu'il y a une certaine économie dans les organismes naturels et que par le biais de la sélection naturelle les systèmes les moins économiques tendent à être éliminés.

La **FLEXIBILITÉ** du modèle est sa capacité à être adapté à un grand nombre d'objets par la modification d'un minimum de paramètres. Dans le cas des langues naturelles, il s'agit de pouvoir rendre compte des variations dialectales entre locuteurs, ainsi que des variations entre les différents états de langue d'un même locuteur au cours de l'apprentissage.

## 3.2 Modèle d'une langue ou modèle de la langue

Chaque langue du monde est un objet d'étude et notre objectif premier est de construire des modèles pour chaque langue. Il y a néanmoins quelque chose de commun dans le fonctionnement des différentes langues et l'on peut, lorsqu'on a étudié suffisamment de langues particulières, être tenté d'extraire ce noyau commun que l'on peut appeler LA LANGUE ou la FACULTÉ DE LANGAGE. Les propriétés communes à toutes langues sont dites **UNIVERSELLES**.

Certains linguistes, comme Noam Chomsky, pensent qu'il existe une grammaire commune à toutes les langues qu'on appelle la **GRAMMAIRE UNIVERSELLE**.

et que cette grammaire est innée. Il est en effet assez légitime de penser que, de même que nos mains et nos pieds se sont spécialisés pour réaliser des tâches particulières, une partie de notre cerveau, qui est l'organe du langage, doit s'être spécialisé pour cette tâche. On va alors chercher dans l'élaboration du modèle à bien distinguer PRINCIPES et PARAMÈTRES : les principes sont les universaux innés de la langue que chaque être humain va paramétrer lors de l'acquisition de sa langue. Ces paramètres constituent les IDIOSYNCRASIES de chaque langue. Le terme vient du grec *idios* 'propre' et *synkrisis* 'constitution'. En médecine ou en psychologie, le terme réfère souvent à un comportement propre à une personne, une réaction qui diffère de personne à personne. En linguistique, on s'intéresse peu aux spécificités langagières d'une personne, mais on s'intéresse aux spécificités d'une langue par rapport à la (faculté humaine de) langue. L'opposition entre principes et paramètres évoque une machine générique qui n'est pas construite différemment pour chaque utilisateur, mais qui permet à l'utilisateur quelques ajustements personnels. Comme pour une machine, ces ajustements personnels — les paramètres — ne sont pas libres et sont interdépendants : on ne peut choisir que certaines valeurs et le choix d'une valeur peut limiter les choix dont on dispose pour une autre valeur.

Une modélisation d'une langue est meilleure quand elle permet facilement le paramétrage d'un sous-langage, d'un dialecte ou d'un registre de langue : on aspire à une analyse du français où un petit changement dans les paramètres — un « paramétrage » — nous donne les grammaires du français écrit journalistique, oral de conversation, de Marseille, de banlieue parisienne et même les différences d'acceptation individuelle. Un tel modèle est meilleur qu'un modèle où ces variantes du français nécessitent des descriptions tout à fait distinctes.

#### Exemples de propriétés universelles de la langue et de spécificités du français

- Chaque langue parlée a un système phonologique avec un nombre fini de phonèmes. Mais le français a des voyelles nasales (*en, on, in*) que la majorité des langues ne possèdent pas.
- Toutes les langues ont des syllabes, c'est-à-dire que les locuteurs coarticulent certains sons. Le français permet l'enchaînement de trois consonnes, comme à l'initiale du mot *structure*, mais d'autres langues ne le permettent pas, comme par exemple le japonais ou le yoruba (une langue très diffusée en Afrique de l'Ouest).
- Toutes les langues ont une structure syntaxique hiérarchique. Le français possède des mots, les pronoms relatifs, qui peuvent à la fois marquer la subordination et jouer un rôle dans la subordonnée (par exemple dans *le livre que Pierre lit*, QUE est à la fois un subordonnant et le complément d'objet

direct du verbe LIRE), mais la majorité des langues n'ont aucun mot comme ça.

- Si une langue distingue singulier et pluriel pour les objets non animés, alors, il faut aussi faire cette distinction pour les objets animés, alors que l'inverse n'est pas vrai. Il existe des langues qui possèdent un pluriel grammatical seulement pour les objets animés, par exemple le japonais.
- Le fait de posséder un système flexionnel, comme la conjugaison des verbes en français ou la déclinaisons des noms en latin, n'est pas universel. En effet, le chinois n'a de flexion pour aucune classe de mots.

Quels paramétrages sont possibles ? Beaucoup de langues, mais pas toutes, obligent le locuteur à indiquer s'il parle d'un seul ou de plusieurs objets, mais aucune langue n'oblige le locuteur à indiquer la couleur de l'objet décrit. Dans aucune langue, le sens d'un mot dépend de la position de ce mot dans la phrase (début, 2<sup>e</sup> place, 3<sup>e</sup> place, etc.) et il est probable que les humains sont incapables d'apprendre une telle langue pourtant imaginable (en français, on a quelque chose qui s'apparente à ça avec la différence d'interprétation qu'il y a pour certains adjectifs entre la position avant et après le nom : *un grand savant* vs *un savant grand*, *un jeune marié* vs *un marié jeune*). Un enfant qui apprend à parler n'est donc pas obligé de peser toutes les possibilités théoriques qu'offre la communication dans l'absolu, mais il doit seulement « paramétrer » les choix qui lui sont offerts. La question de savoir quelles sont les contraintes innées qui président à ces choix reste un important sujet de recherche en linguistique.

### 3.3 Modélisation et théorie

Pour mieux comprendre ce qu'est la modélisation, nous allons faire un parallèle avec la physique.

#### 3.3.1 Un exemple de modélisation en physique

Tout le monde peut observer les marées au bord de l'océan et les décrire. On verra ainsi que les marées respectent un cycle régulier d'un peu plus de 12 heures entre deux marées hautes, ainsi qu'un cycle d'environ 14 jours entre deux grandes marées et encore un cycle annuel pour l'amplitude des grandes marées, le sommet étant atteint pour les marées d'équinoxe deux fois par an. On pourra ainsi relever très précisément chaque jour l'heure des marées et leur hauteur (leur coefficient) et noter que ces hauteurs varient selon une courbe plus ou moins sinusoïdale. Jusque-là, on a décrit le phénomène des marées, mais on ne l'a pas



modélisé. Si la description est suffisamment poussée et qu'on dispose de beaucoup de données, on pourra, par des méthodes statistiques, **prédire** le moment et la hauteur des marées suivantes.

On peut pousser la description jusqu'à noter une certaine corrélation entre les mouvements de la mer et les mouvements respectifs de la terre, de la lune et du soleil (la terre fait un tour sur elle-même en 24 h, la lune fait le tour de la terre en 28 jours et la terre fait le tour du soleil en un an). Dès que l'on applique la théorie de la gravitation de Newton et que l'on considère que la lune comme le soleil attirent suffisamment les mers pour les faire bouger par rapport à la croûte terrestre, on obtient un nouveau modèle des marées. Ce modèle permet non seulement de prédire les dates et hauteurs des marées des années à l'avance, mais il permet aussi d'**expliquer** le phénomène des marées et de comprendre la superposition des trois rythmes sinusoïdaux.

#### 3.3.2 La modélisation de la langue

Considérons maintenant la langue. Tout le monde peut observer des productions langagières et les décrire. On peut compter les occurrences de chaque mot, regarder dans quels contextes elles apparaissent, avec quels autres mots avant et après, etc. On peut même noter des corrélations complexes comme la forme du verbe et la présence de tel ou tel pronom à tel endroit. Si on a pris soin de noter les circonstances dans lequel le texte a été énoncé, on peut pousser la description jusqu'à faire des corrélations entre le contexte d'énonciation et le texte produit. On obtient ainsi une DESCRIPTION qui, combinée avec un modèle statistique, peut avoir une bonne VALEUR PRÉDICTIVE. Certains parlent de MODÈLES DESCRIPTIFS et de MODÈLES PRÉDICTIFS, mais, de notre point de vue, la modélisation commence réellement lorsqu'on se place dans un certain cadre théorique et que l'on fait des hypothèses sur le fonctionnement de la langue. On parle alors vraiment de MODÈLE (THÉORIQUE). Celui-ci sera d'autant meilleur qu'il aura une VALEUR EXPLICATIVE, c'est-à-dire qu'il nous permettra de comprendre non seulement comment nous construisons nos énoncés, mais aussi pourquoi nous devons respecter ce type de contraintes, comment nous apprenons à parler, pourquoi nous faisons tel lapsus, etc. On obtient ainsi un MODÈLE EXPLICATIF.

#### 3.3.3 Du modèle à la théorie

On peut pousser encore plus loin la comparaison entre physique et linguistique. Le phénomène des marées n'est pas seulement une application de la théorie de la gravitation. L'histoire de la pomme de Newton est plaisante, mais il est

clair qu'elle n'a été qu'un déclencheur. Newton ne cherchait pas à résoudre le problème de la chute des pommes, mais souhaitait surtout comprendre l'origine du mouvement des planètes et le phénomène des marées et c'est l'observation de ces phénomènes qui lui a permis d'élaborer et de valider la théorie de la gravitation. Dès qu'on fait l'hypothèse que les masses s'attirent, tous ces problèmes — la pomme, les planètes et les marées — trouvent une solution simple. C'est ce qu'on appelle un RAISONNEMENT PAR ABDUCTION, où on pose une hypothèse A, car elle est l'explication la plus simple à une observation C. C'est par l'observation des marées et de la chute des pommes qu'on peut émettre l'hypothèse que les masses s'attirent.

Il en va exactement de même pour la théorie linguistique : ce sont tous les phénomènes observés dans les langues qui nous permettent d'élaborer, par abduction, une théorie linguistique. C'est de cette façon que, dans la section 2 *Produire un énoncé*, nous avons été amenés à faire l'hypothèse qu'il existe une structure syntaxique hiérarchique sous-jacente aux énoncés, sur laquelle s'appuient les contraintes linguistiques. Nous allons illustrer à nouveau ce point dans l'encadré qui suit sur les phénomènes d'accord.

### 3.4 Un exemple — l'accord — de la description à l'explication

Nous allons illustrer notre propos précédent par un exemple, celui des règles d'accord.

**Étape 1. Description de la concordance des formes du nom et du verbe.** On peut remarquer, en français, que la forme du verbe varie en fonction du sujet : *Marie dort* vs *Marie et Pierre dorment*, *Marie finit de manger* vs *Marie et Pierre finissent de manger*, etc. On poussera la description jusqu'à noter que les segments qui peuvent aller dans l'environnement — *dort* et — *finit de manger* sont les mêmes (*Marie, elle, mon amie, la dame*, etc.) et donner un nom à cette classe : les groupes nominaux singuliers.

**Étape 2. Énoncer la règle d'accord.** Voici la règle : en français, il y a deux nombres — singulier et pluriel — et le verbe conjugué s'accorde en nombre avec son sujet. Cette règle paraît élémentaire (il y en a des plus complexes comme la règle d'accord du participe passé en français), mais elle ne l'est pas tant, car elle suppose que l'on sait définir le sujet d'un verbe et le reconnaître. L'histoire de la linguistique montre que ce ne fut pas chose facile et que définir correctement la notion de sujet, c'est déjà élaborer un début de théorie syntaxique. Modéliser cette règle par exemple pour implémenter un correcteur automatique capable

### 3.4 Un exemple – l'accord – de la description à l'explication

d'assurer l'accord du verbe avec le sujet est encore une autre affaire. Aujourd'hui les correcteurs grammaticaux ne sont généralement pas capables de retrouver les sujets des deux verbes de la phrase « *La pièce dans laquelle veulent jouer les enfants **est** trop petite* » (et ils le seraient encore moins s'il y avait une faute d'orthographe).

**Étape 3. Pourquoi il y a des règles d'accord dans les langues.** Pour bien modéliser la règle d'accord, il faut comprendre quel est son rôle dans le système. Les règles d'accord servent à marquer les relations entre les mots de manière à aider le destinataire à reconstruire la structure de l'énoncé et son sens. Il y a plusieurs manières de marquer ces relations :

1. Une **marque** qui dépend de la nature de la relation : c'est la **rection**. Il peut s'agir d'un CAS, c'est-à-dire d'une marque sur le dépendant, comme en latin (*Petrus* 'Pierre' → *Petri cani* 'le chien de Pierre'), d'une préposition, c'est-à-dire d'un mot, comme en français (*le chien **de** Pierre*) ou d'une marque sur le gouverneur comme en wolof (*xaj bi* 'le chien', lit. chien le → *xaju Peer bi* 'le chien de Pierre').
2. Une **marque d'accord** : cette marque peut se trouver sur le dépendant et reprendre une caractéristique du gouverneur (accord de l'adjectif avec le nom) ou l'inverse (accord du verbe avec son sujet).
3. Un **ordre fixe** : la position du dépendant par rapport au gouverneur est très contrainte. Par exemple, en anglais, un adjectif précède le nom dont il dépend et l'objet direct suit le verbe dont il dépend.
4. La **prosodie** : dans la phrase ambiguë *Pierre regarde la fille avec un télescope*, le groupe prépositionnel *avec un télescope* peut dépendre de *regarde* ou de *la fille*. Un contour prosodique regroupant *la fille avec un télescope* désambiguïsera la phrase.

Les différentes techniques peuvent se combiner, comme en français où le sujet possède une position contrainte (devant le verbe en général), déclenche un accord du verbe et varie en cas lorsqu'il s'agit d'un pronom (***il** dort*, *Marie **le** regarde*, *Marie **lui** parle*).

Présenter l'accord comme nous venons de le faire, c'est prendre une position théorique : nous considérons que les mots se connectent entre eux et forment une structure hiérarchique et que les phénomènes d'accord en découlent normalement. De la même façon, la théorie de la gravitation fait l'hypothèse que les masses s'attirent et montre que les marées découlent naturellement de cette hypothèse.

### 3.5 Modèle déclaratif

Comme nous l'avons dit, un modèle linguistique doit être capable d'associer une représentation sémantique à des textes et vice versa. On distingue dans le modèle l'ensemble des **connaissances** nécessaires pour effectuer cette association de la **PROCÉDURE** qui permet d'activer ces connaissances et de réaliser l'association. Un modèle qui sépare connaissances et procédure est dit **DÉCLARATIF**. Sinon le modèle est dit **PROCÉDURAL**. On appelle généralement **GRAMMAIRE FORMELLE** un modèle déclaratif d'une langue. Le terme *grammaire* employé comme ceci inclus la description du lexique de la langue.

Supposer que les langues possèdent des modèles déclaratifs est une hypothèse forte et difficile à vérifier. Elle repose sur l'idée que les connaissances qui permettent de parler une langue et de la comprendre sont peu ou prou les mêmes. Une grammaire commune à la production et à l'analyse est dite **RÉVERSIBLE**. On aura ensuite des procédures distinctes pour la production et l'analyse. La procédure d'analyse est plus tolérante, puisque les locuteurs comprennent des énoncés qu'ils ne seraient pas capables de produire ; les contraintes de la grammaire devront donc être relâchées en analyse.

Un modèle déclaratif est normalement un **MODÈLE DE LA COMPÉTENCE**. En effet, les erreurs de performance doivent être imputées à des problèmes rencontrés lors de l'activation des connaissances. Néanmoins opter pour un modèle déclaratif et une séparation entre connaissances et procédures ne signifie pas que nous rejetons la procédure hors du modèle. Notre modèle de la compétence doit être inclus dans un modèle complet de la langue prenant en compte la mise en œuvre du modèle déclaratif. Ce modèle complet est un **MODÈLE DE LA PERFORMANCE**.

### 3.6 Modèle génératif, équatif et transductif

Noam Chomsky a révolutionné la linguistique en 1957 dans son ouvrage *Syntactic structures* en définissant son objet étude comme un problème mathématique. Il pose qu'une langue est l'ensemble potentiellement infini des phrases grammaticales de cette langue et qu'une grammaire est un système mathématique comportant un nombre fini de règles capable de générer l'ensemble des phrases grammaticales d'une langue. Ce courant sera appelé la **GRAMMAIRE GÉNÉRATIVE**. Plus tard, Chomsky renoncera à la présentation générative de la langue, mais on continuera d'appeler son école de pensée la grammaire générative, ce qui prête souvent à confusion. La naissance de la grammaire générative a coïncidé avec la naissance de la cybernétique et de l'informatique théorique et les deux

mouvements se sont fortement influencés, langues naturelles et langages de programmation ayant été vus comme des objets de natures similaires.

Dans sa version initiale, la grammaire générative ne traite pas le sens, considéré comme difficilement accessible. Les phrases, notamment dans les versions formalisées du modèle, sont traitées comme de simples suites de caractères, c'est-à-dire, pour reprendre notre terminologie, uniquement des textes. La grammaire ne se fixe alors comme principal objectif que de générer les textes d'une langue. En fait les grammaires proposées par Chomsky offrent une analyse syntaxique des phrases et définissent donc, indirectement, une correspondance entre textes et représentations syntaxiques. D'autres chercheurs ont proposé, comme Aravind Joshi en 1975, des GRAMMAIRES D'ARBRE permettant de générer simultanément une phrase et sa représentation syntaxique, construisant ainsi les premières grammaires de correspondance génératives.

Une GRAMMAIRE DE CORRESPONDANCE est une grammaire définissant la **correspondance entre deux ensembles de structures**, par exemple des représentations sémantiques et des textes (voir l'encadré ?? sur *La Théorie Sens-Texte*). On peut utiliser trois types de procédures pour définir une correspondance avec une grammaire de correspondance : procédure générative, équative ou transductive. Une PROCÉDURE GÉNÉRATIVE est une procédure qui va générer la correspondance, c'est-à-dire l'ensemble des couples en correspondance ; au lieu de générer uniquement un texte, la grammaire génère simultanément le texte et son sens. Une PROCÉDURE ÉQUATIVE est une procédure qui va vérifier pour chaque couple de structures qu'on lui proposera si ces structures se correspondent ; cela suppose qu'on fournisse un texte et un sens et la procédure permettra de vérifier que ce texte et ce sens peuvent être associés par la grammaire. Une PROCÉDURE TRANSDUCTIVE est une procédure qui à chaque fois qu'on lui propose une structure est capable de construire toutes les structures qui lui correspondent ; dans ce cas, on fournit soit un texte, soit un sens et la procédure construit les sens correspondant au texte ou les textes correspondant au sens. Dans le Chapitre 2, nous avons adopté une procédure transductive pour associer un graphe sémantique à des arbres syntaxiques.

Ces trois types de procédure peuvent être utilisés pour présenter une même grammaire. Ce qui distingue ces trois procédures est le nombre de structures au départ : 0, 1 ou 2. Dans tous les cas, on a un couple de structures à l'arrivée. Dans la procédure générative, on part de rien et on génère simultanément les deux structures en correspondance. Dans les procédures transductives, on a une structure au départ et on produit l'autre. Dans la procédure équative, on a les deux structures dès le départ et on vérifie qu'elles se correspondent. Nous sché-

matisons ci-dessous les trois procédures. Supposons qu'on veuille associer des graphes sémantiques, représentés par des  $\boxtimes$ , à des arbres syntaxiques, représentés par des  $\triangle$ . Ce qui distingue les trois procédures, ce sont les structures données au départ : nous les schématisons en noir, tandis que les structures à construire par la procédure sont en blanc :

- (1) procédure générative :  $\boxtimes \Leftrightarrow \triangle$
- (2) procédures transductives :  $\star \Rightarrow \triangle$  ou  $\boxtimes \Leftarrow \boxtimes$
- (3) procédure équative :  $\star \Leftrightarrow \boxtimes$

Pour des raisons historiques que nous venons de rappeler, la procédure générative est souvent privilégiée. La procédure équative est généralement la procédure la plus élégante pour présenter un modèle déclaratif et elle tend à se généraliser sous le nom de GRAMMAIRES DE CONTRAINTES. Mais des trois procédures, c'est la procédure transductive qui est descriptivement la plus pertinente, car c'est ce type de procédure que les locuteurs utilisent quand ils parlent : ils doivent à partir d'un sens lui faire correspondre un texte et, à l'inverse, quand ils écoutent quelqu'un qui parle, ils partent d'un texte et doivent lui donner un sens.

### 3.7 Modèle symbolique

Une des propriétés remarquables des langues est l'utilisation d'un petit nombre de sons — les phonèmes — pour construire les signifiants de tous les éléments lexicaux de la langue. Cela met en évidence notre capacité à catégoriser, c'est-à-dire à identifier, dans la multitude de signaux de parole auxquels nous sommes confrontés lors de l'apprentissage de notre langue, un nombre fini de SYMBOLES, c'est-à-dire d'éléments qui ont une portée symbolique et qui possède une valeur d'interprétation. Plus généralement, on peut penser que les fonctions supérieures du cerveau, celles qui sont liées à la cognition et à la pensée consciente, se caractérisent par la capacité à catégoriser et à manipuler des symboles.

On appelle MODÈLE SYMBOLIQUE OU MODÈLE DISCRET OU ENCORE MODÈLE ALGÈBRIQUE un système basé uniquement sur la **manipulation algébrique** d'un **nombre fini de symboles**. Par manipulation algébrique, on entend des opérations mathématiques permettant de combiner des configurations de symboles pour créer de nouvelles configurations (voir encadré ci-dessous sur le calcul symbolique). Un modèle qui ne « discrétise » pas est dit CONTINU. La question se pose de savoir si un modèle linguistique doit ou non être symbolique. Les arguments contre les modèles symboliques sont assez nombreux :

- l’acceptabilité des énoncés n’est pas binaire : il semble y avoir un continuum entre les énoncés acceptables et les énoncés inacceptables ;
- les catégories syntaxiques sont assez floues ; on trouve de nombreux éléments à la frontière de plusieurs catégories ;
- les unités lexicales sont généralement polysémiques et il est difficile de déterminer combien de sens peut avoir exactement une unité lexicale ;
- la prosodie, qui joue un rôle non négligeable dans l’expression du langage, semble difficilement catégorisable (même si la ponctuation est une forme de catégorisation de certains contours prosodiques).

Malgré cela, on arrive à fournir des modèles symboliques des langues assez satisfaisants et la plupart des modèles théoriques sont basés sur des règles manipulant des symboles.

On peut rendre compte des différents niveaux d’acceptabilité en modifiant un peu un modèle symbolique. Deux directions au moins ont été envisagées. La première, exploitée par la Théorie de l’Optimalité de Alan Prince et Paul Smolensky1993, consiste à traiter les règles comme des contraintes éventuellement violables. De plus, les contraintes peuvent être rangées par ordre d’importance. Plus un énoncé viole de contraintes et plus ces contraintes sont importantes, moins il est acceptable.

L’autre direction consiste à pondérer les règles. On peut alors associer un score à chaque énoncé en fonction des règles qui ont permis de le produire. Cette technique est surtout utilisée en analyse pour désambiguïser : lorsque plusieurs analyses sont possibles pour un énoncé, on privilégie celle qui a le meilleur score. Des grammaires de ce type sont généralement construites en pondérant les règles selon leur probabilité d’apparition dans un corpus syntaxiquement annoté calculée à partir d’une analyse statistique de leurs occurrences. On obtient ainsi un MODÈLE STOCHASTIQUE, dont la base reste un modèle symbolique.

Nous montrerons dans cet ouvrage les différents problèmes que pose l’identification des unités d’une langue et leur catégorisation.

### 3.8 Calcul symbolique et grammaires catégorielles

Le premier linguiste mathématicien montrant qu’on pouvait vérifier la bonne formation d’une phrase par un CALCUL SYMBOLIQUE est probablement le polonais Kazimierz Ajduckiewicz en 1935. Voici son idée. Pour montrer que « *Pierre dort* » est une phrase nous allons associer à chaque mot une catégorie complexe :

- *Pierre* forme à lui seul un groupe nominal, nous lui associons la valeur GN ;

### 3 La modélisation : Préciser l'objectif de notre étude

- *dort* peut former une phrase P à condition qu'on le combine avec un GN : nous lui associons la valeur.

Nous pouvons maintenant calculer la valeur associée à *Pierre dort* :

- (4) *Pierre dort*  
 $GN \cdot = P$

Cette valeur est calculée en combinant les catégories associées à chaque mot et en simplifiant comme on le fait avec des fractions ordinaires. L'analyse d'une phrase devient un CALCUL ALGÈBRE simulé au calcul numérique ( $a \cdot \frac{b}{a} = b$ ).

Nous avons construit un début de grammaire capable de vérifier que chaque verbe conjugué a un sujet. Une grammaire associant ainsi des catégories complexes à chaque mot est appelée une GRAMMAIRE CATÉGORIELLE. Il s'agit du premier exemple de GRAMMAIRE FORMELLE CONNU.

Ce calcul a été ensuite repris par Yehoshua Bar-Hillel en 1953, qui a montré que si l'on veut modéliser les contraintes d'ordre, il fallait distinguer ce qu'on combine à droite de ce qu'on combine à gauche. On va donc associer à *dort* la catégorie  $GN \backslash P$  (qui se lit « GN sous P ») indiquant que le GN avec lequel le verbe doit se combiner pour former une P doit se trouver à gauche. Analysons *Pierre mange une banane* avec cette grammaire. Les mots de cette phrase ont les catégories *Pierre* := GN, *banane* :=  $D \backslash GN$ , *un* := D et *mange* :=  $(GN \backslash P) / GN$ . Le calcul est le suivant :

- (5) *Pierre mange une banane*
- (6)  $GN \cdot (GN \backslash P) / GN \cdot D \cdot D \backslash GN$   
 $GN \cdot (GN \backslash P) / GN \cdot GN$   
 $GN \cdot GN \backslash P$   
P

Cette séquence de mots est bien reconnue comme une phrase par notre grammaire. De plus, la structure du calcul est un arbre que l'on peut interpréter comme la structure syntaxique de la phrase.

Joachim Lambek a montré en 1958 que la règle de combinaison des catégories pouvait être interprétée comme une inférence logique. On peut en effet voir  $GN \backslash P$  comme une implication  $GN \rightarrow P$  à interpréter comme « si on me donne un GN à gauche, je formerai un P ». La règle de combinaison devient alors : « de GN et de  $GN \rightarrow P$ , je déduis P », ce qui n'est autre que le *modus ponens*, la règle de déduction de base de la logique (« de  $p$  et de  $p \rightarrow q$ , je déduis  $q$  »). L'analyse d'une



phrase devient maintenant un CALCUL LOGIQUE. Le calcul devient une *preuve* que la suite de mots considérée au départ est bien une phrase. Il est intéressant de remarquer que la logique ainsi construite diffère de la logique classique, puisqu'elle est **sensible aux ressources** (donner deux GN n'est pas équivalent à en donner un seul) et à l'**ordre** ( $GN \rightarrow P$  et  $P \leftarrow GN$  ne se comportent pas pareil, l'un attend un GN avant et l'autre après). Une telle logique, appelée LOGIQUE LINÉAIRE, a des applications dans des domaines très éloignés de la linguistique et notamment en robotique où les actions doivent être effectuées dans un ordre bien précis.

Dans l'encadré ?? sur la *Grammaire de réécriture*, nous présenterons un autre exemple de grammaire formelle qui a marqué la deuxième moitié du 20<sup>e</sup> siècle : les grammaires de réécriture hors-contexte introduites par Noam Chomsky en 1957.

### 3.9 Modularité et stratification

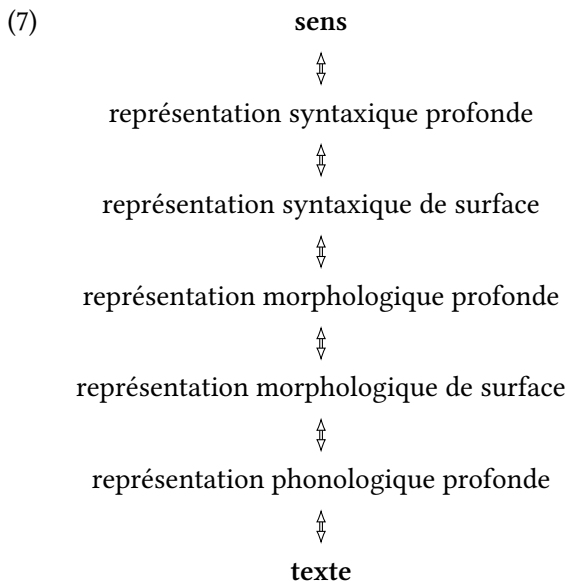
Un MODULE est un sous-système du modèle suffisamment autonome pour effectuer seul une partie des calculs nécessaires à la production d'un énoncé. Les travaux en neurologie semblent accréditer l'idée que le cerveau possède un fonctionnement modulaire et que, lors de la production d'un énoncé, différentes aires cérébrales sont sollicitées avec des tâches différentes. Néanmoins les connaissances sur l'architecture du cerveau sont encore insuffisantes pour que se dégage une vision claire des différents modules que devrait avoir un modèle linguistique et sur la façon dont ces modules coopèrent.

Dans la suite de cet ouvrage, nous montrerons qu'il existe plusieurs types d'unités linguistiques et plusieurs modes d'organisation de ces unités. Une architecture possible est alors de considérer que chaque niveau d'organisation fournit une STRATE et que le passage du sens au texte se fait en plusieurs modules qui permettent de passer d'une strate à l'autre, l'un à la suite de l'autre. Un modèle de ce type est dit STRATIFICATIONNEL. La façon la plus simple d'utiliser un modèle stratificationnel est d'avoir une séquence de modules qui fonctionnent **à la suite l'un de l'autre** : un premier module prend en entrée le sens et fournit au deuxième module une structure complète de la strate suivante et ainsi de suite. Une telle architecture est dite LINÉAIRE ou en PIPELINE. Cette architecture s'oppose à une ARCHITECTURE DISTRIBUÉE où toutes les strates peuvent communiquer. Même dans une architecture linéaire, on peut faire que tous les modules fonctionnent simultanément et que chaque module traite les données que lui fournissent les autres modules **au fur et à mesure** qu'elles arrivent. Une telle procédure est dite INCRÉMENTALE. Le modèle que nous défendons est stratifié, modulaire, incrémen-

tal et en grande partie linéaire. Nous n'avons aujourd'hui aucun moyen de valider ou d'invalider une telle architecture.

### 3.10 La Théorie Sens-Texte

La Théorie Sens-Texte est la théorie qui a le plus fortement inspiré les auteurs de cet ouvrage. Il s'agit d'un modèle développé autour d'Igor Mel'čuk à partir de 1965, d'abord en Union Soviétique, puis au Canada et en Europe. Ce modèle est fortement stratifié : il suppose l'existence de 5 niveaux de représentation intermédiaires entre le sens et le texte, soit 7 niveaux en tout. Le système est divisé en 6 modules permettant de passer d'un niveau à l'autre. Il s'agit d'une architecture linéaire que nous présentons ci-dessous :



Chaque  $\Leftrightarrow$  représente un module effectuant la correspondance entre deux niveaux de représentation adjacents. On trouvera une présentation et une justification de cette architecture dans le cours donné par Igor Mel'čuk en 1997 au collège de France, *Vers une linguistique Sens-Texte*, disponible en ligne.

Dans cet ouvrage, nous présentons les quatre niveaux supérieurs de représentation. Notre présentation peut être assez différente de celles que l'on trouve dans les travaux d'Igor Mel'čuk. Notre objectif est de justifier toutes les structures que nous introduirons, quitte parfois à remettre en question le statut ou la nature des

représentations utilisées en Théorie Sens-Texte ou dans d'autres théories comparables, notamment la *Lexical Functional Grammar* (LFG) de Joan Bresnan et Ronald Kaplan, qui possède également une architecture stratificationnelle.

## 3.11 Modélisation des langues et ordinateur

Le développement « à la main » d'un modèle linguistique est un travail considérable. Un dictionnaire de français courant possède 60 000 mots et l'on évalue à près d'un million le nombre d'unités lexicales si l'on y inclut les expressions figées et que l'on compte les différentes acceptions de chaque lexème. Le lexique des constructions grammaticales est encore mal connu et certainement sous-évalué. La combinatoire de n'importe quelle description sérieuse d'un phénomène grammatical est tellement importante qu'il est difficile de voir la description dans son ensemble. L'ordinateur est alors le seul moyen pour combler cette difficulté. Les ordinateurs possèdent aujourd'hui des capacités de calcul suffisantes pour tester la plupart des modèles imaginables (et théoriquement raisonnables).

Aujourd'hui les modèles capables d'effectuer des correspondances entre sens et textes sont très fragmentaires. On trouve par exemple des modèles de ce type assez performants, mais limités à la production de bulletins météo et manipulant donc un vocabulaire restreint. Les modèles informatiques couvrants se limitent à la correspondance entre textes et représentations syntaxiques de surface et ils font aujourd'hui pour cette tâche d'analyse superficielle près d'une erreur par phrase en moyenne.

La validation des théories linguistiques n'est pas le seul intérêt de l'implémentation des modèles : le développement de modèles informatiques possède un réel intérêt économique, puisque la langue est au centre de toutes les activités sociales humaines. Le développement de modèles informatiques et de ressources formelles tels que lexiques, grammaires ou corpus annotés s'appelle la LINGUISTIQUE COMPUTATIONNELLE. La linguistique computationnelle est incluse dans le domaine plus vaste du TRAITEMENT AUTOMATIQUE DES LANGUES ou TAL, qui s'intéresse également à toutes les applications que l'on peut développer à partir de tels modèles. Les applications industrielles du TAL constituent l'INGÉNIERIE LINGUISTIQUE.

Le TAL ne nécessite pas toujours de grandes connaissances en linguistique et certaines méthodes de traitement du langage n'ont pas grand-chose à voir avec la modélisation des langues. Par exemple, pour décider en quelle langue est une page sur la Toile le moyen le plus simple et le plus performant est de faire une

statistique des séquences de trois lettres et de comparer avec les statistiques pour les langues que l'on souhaite reconnaître. Nul besoin de lexique et encore moins de connaissances en grammaire. De la même manière, il est imaginable de faire une étude statistique sur les marées des cent dernières années et en déduire de bonnes prédictions sur les marées à venir. Par contre, il semble clair qu'une telle « modélisation » ne nous avance pas dans la compréhension du phénomène des marées et du rôle de la gravitation (voir la section 3.3 sur *Modélisation et théorie*).

Aujourd'hui, beaucoup de modèles en TAL se basent avant tout sur des DONNÉES STATISTIQUES apprises sur de grand corpus, car les résultats sont souvent meilleurs et moins coûteux qu'avec des données entièrement développées à la main qui s'avèrent généralement incomplètes. La très grande capacité de mémoire de l'ordinateur fait qu'il est souvent plus facile de travailler avec une énorme liste de données (par exemple toutes les constructions possibles pour tous les verbes d'une langue) que de dégager et d'implémenter les régularités derrières (avec leurs exceptions qui existent toujours). Il est possible que les modèles les plus efficaces pour les applications informatiques ne soient pas les modèles les plus satisfaisants du point de vue de leur « explicativité ». À l'heure actuelle, les traducteurs automatiques les plus performants sont ceux qui interrogent de grands corpus bilingues pour rechercher tous les fragments du texte pour lesquels on a déjà une traduction et qui choisissent la plus fréquente (**modèle de traduction**). Ainsi les traductions sont combinées et lissées à l'aide de corpus unilingue encore plus grands (**modèle de langue**). De tels systèmes de traduction statistiques obtiennent de meilleurs résultats que les systèmes qui cherchent à simuler un traducteur humain, c'est-à-dire qui cherchent à calculer une représentation sémantique du texte à traduire pour générer à partir de celle-ci un texte dans une autre langue.



## Exercices

**Exercice 1** Quelle différence de statut faisons-nous entre le sens et la représentation sémantique ?

**Exercice 2** Quelle est la différence entre falsifiabilité et réfutabilité ?

**Exercice 3** Quels seraient des modèles respectivement descriptif, prédictif et explicatif du réchauffement climatique ?

**Exercice 4** Nous avons discuté de la règle d'accord du verbe avec son sujet et du fait que cette règle reposait sur une définition préalable du sujet. Comme définiriez-vous le sujet syntaxique pour le français ?

**Exercice 5** L'énoncé *La plupart sont verts* remet-il en cause la règle d'accord en nombre du verbe avec son sujet ? Comment résoudre le problème ?

**Exercice 6** Le français possède deux genres (on dit **le** soleil et **la** lune ou **une** armoire et **un** tabouret), l'allemand en possède trois (féminin, masculin et neutre), l'anglais aucun (on a juste un marquage du sexe dans les pronoms), les langues bantoues peuvent avoir jusqu'à une vingtaine de classes nominales différentes. Ce phénomène est facile à décrire, mais qu'est-ce qu'un modèle explicatif pourrait en dire ? Pourquoi les langues peuvent avoir ou ne pas avoir des classes d'accord différentes pour les noms ?

**Exercice 7** Pourquoi le fait qu'on puisse distinguer une question d'une assertion par la seule prosodie (une intonation montante pour « *Tu viens ?* » et descendante pour « *Tu viens.* ») met-il en défaut une architecture linéaire comme celle de la Théorie Sens-Texte ?



### Lectures additionnelles

Sur la distinction entre falsifiabilité et réfutabilité, on pourra lire l'article de **Lakatos1968** qui fait lui-même référence aux débats entre Popper et Kuhn suite à la réfutation de la théorie de la gravitation de Newton et à sa résolution par la théorie de la relativité d'Einstein.

La distinction entre modèle et théorie est discutée dans **Chomsky1957**, dont la lecture est incontournable si l'on veut comprendre pourquoi cet ouvrage a marqué un basculement de la linguistique dans le domaine des sciences. Pour Chomsky, une théorie donne un formalisme grammatical et les modèles sont les grammaires particulières que l'on peut définir avec ce formalisme.

Le raisonnement par abduction a été dégagé par le philosophe américain, Charles S. Peirce (1839-1914). En 1903, il en donne la formulation suivante : « The surprising fact, C, is observe ; But if A were true, C would be a matter of course ; Hence, there is reason to suspect that A is true. ».

La Théorie Sens-Texte est présentée dans la plupart des ouvrages d'Igor Mel'čuk. Voir les ouvrages dont nous avons déjà parlé dans les trois précédents chapitres.

Noam **Chomsky1957** *Syntactic structures*, MIT Press (traduction française de M. Bradeau, 1969, *Structures syntaxiques*, Seuil).

Imre **Lakatos1968** *Criticism and the Methodology of Scientific research Programmes*, *Proceedings of the Aristotelian Society*, New Series, vol. 69, Blackwell, pp. 149-186.

Charles S. **Peirce1903** *Harvard lectures on pragmatism*, *Collected Papers* v. 5, paragraphs 188-189.



## Corrections des exercices

**Corrigé 1** Tout énoncé linguistique a un sens. Le sens appartient à la langue. La représentation sémantique appartient au modèle de la langue et modélise le sens.

**Corrigé 2** La falsifiabilité est une propriété des modèles. Lorsqu'un modèle est faux (c'est-à-dire qu'il fait une mauvaise prédiction), on peut essayer de le réparer en changeant des paramètres. La réfutabilité est une propriété des théories. Pour réfuter une théorie, il faut montrer que tous les modèles qu'elle propose sont faux, ce qui est très difficile, voire impossible. La réfutabilité d'une théorie se fait donc généralement en proposant une nouvelle théorie dont l'un des modèles prend mieux en compte les données qui posent problème à la théorie précédente.

**Corrigé 3** Un modèle descriptif du réchauffement climatique serait une description des relevés de températures en divers point du globe dans les années ou les siècles qui précèdent qui montrerait une augmentation de température. Une analyse statistique des variations de température permettrait de faire des prédictions sur l'évolution de la température dans les années ou siècles à venir (avec l'hypothèse théorique que la température d'une année est corrélée aux températures des années qui précèdent). Un modèle explicatif tenterait de rechercher les causes des variations de température et de corréliser ces variations avec un certain nombre de paramètres, comme la consommation d'énergie par les humains. Le modèle permet alors d'affiner la prédiction en fonction de l'évolution de ces paramètres et donc de prédire que la température augmentera encore plus vite si la consommation d'énergie augmente.

**Corrigé 4** La notion de sujet syntaxique sera définie dans le Chapitre ??.

L'accord du verbe est une des propriétés définitoires du sujet en français,

avec la position privilégiée avant le verbe ou l'emploi de pronom comme *il* ou *on*.

**Corrigé 5** *La plupart* est utilisé en français comme pronom pluriel masculin ou féminin. Il s'agit d'une forme figée, d'un sémantème (voir le Chapitre ??), où *la* n'est plus un marqueur du singulier féminin. Il suffit donc de déclarer *la plupart* comme un pronom pluriel dont le genre dépend de son antécédent pour assurer l'accord selon les règles habituelles.

**Corrigé 6** L'origine des genres en français (et dans les autres langues indo-européennes) est un marquage des sexes pour les noms d'êtres sexués qui s'est propagé à tous les noms par régularisation du système. Pour la plupart des noms, il n'a aucune signification (même si des études montrent que l'existence des genres a une influence sur la représentation mentale et que les locuteurs d'une langue où *mort* est féminin comme le français personnifient plus naturellement la mort par une femme que par un homme à l'inverse des locuteurs d'une langue comme l'allemand où *Tod* est masculin). Devoir apprendre des genres pour des noms où cela n'a pas de sens a un coût cognitif. Mais les genres vont permettre de renforcer le marquage des relations syntaxiques au travers des accords en genre. Nous avons vu que l'accord est un des moyens de marquer l'existence d'une relation syntaxique, voire la nature de cette relation (l'accord du verbe avec son sujet permet de caractériser cet élément en tant que sujet).

**Corrigé 7** L'interrogation réalisée par la seule prosodie est un exemple d'un sens qui est réalisé directement au niveau phonologique sans aucune incidence sur la syntaxe. La description de cette construction peut être faite par une correspondance directe entre sémantique et phonologie, alors qu'un modèle en pipeline obligerait à passer par la syntaxe et donc à introduire un élément fictif au niveau syntaxique.



