

---

## Introduction : le projet

---

Le projet décrit dans cet ouvrage s'inscrit au cœur de la linguistique : il a pour but de décrire avec exhaustivité et une précision absolue l'ensemble des phrases d'une langue susceptibles d'apparaître dans les textes écrits<sup>1</sup>. Ce projet répond à deux besoins : il s'agit d'une part de fournir aux linguistes des outils pour les aider à décrire les langues de façon exhaustive (la *linguistique descriptive*<sup>2</sup>), et d'autre part de construire des logiciels capables de traiter automatiquement des textes rédigés en langue naturelle (le *Traitement automatique des langues*, TAL).

– Un projet linguistique a besoin d'un cadre théorique et méthodologique (comment décrire tel ou tel phénomène linguistique, comment organiser les différents niveaux de description), des outils formels (comment rédiger chaque description), des outils de développement pour tester et gérer chaque description, et des outils d'ingénierie pour partager, accumuler et maintenir des ressources linguistiques en grande quantité<sup>3</sup>.

– Les applications potentielles de la linguistique descriptive pour le TAL sont nombreuses : correcteurs orthographiques, moteurs de recherche intelligents, extracteurs et annotateurs d'information, producteurs automatiques de résumés, traducteurs automatiques, etc. Ces applications ont une utilité économique potentielle considérable : il est donc important que les linguistes s'approprient ces technologies et puissent y contribuer.

---

1. On peut transcrire des discours oraux en représentant orthographiquement des prononciations particulières, comme « chui » ou « jsui » pour *je suis*. (Chevalier 2004) a étudié la variation morpho-phonologique dans les corpus oraux français de diverses régions du Canada en utilisant les mêmes outils d'analyse que pour l'écrit.

2. J'emploierai les termes « linguistique » et « linguistique descriptive » de façon interchangeable.

3. (Saussure 1916) est l'une des premières tentatives de rationaliser la description des langues.

Pour le moment, on doit réduire le projet linguistique global de la description de *tous* les phénomènes liés à l'utilisation de la langue à un projet bien plus modeste : ici, on se contentera de chercher à décrire l'ensemble des phrases que l'on peut écrire ou lire dans des textes en langue naturelle. Le but est donc simplement de concevoir un système capable de distinguer les deux séquences suivantes :

(a) *Luc mange une pomme*

(b) *Luc pomme mange une*

La séquence (a) est une phrase correcte ; la séquence (b) n'est pas une phrase.

Ce projet constitue la base obligatoire de tout projet de description linguistique plus ambitieux. Il serait vain en effet de vouloir formaliser le style des textes (stylistique), l'évolution d'une langue au cours des siècles (étymologie), la variation d'une langue selon les classes sociales (sociolinguistique), les phénomènes cognitifs mettant en œuvre l'apprentissage ou la compréhension d'une langue (psycholinguistique), etc. si l'on ne dispose pas d'un modèle même rudimentaire capable de caractériser les phrases de la langue.

Si l'ensemble des phrases avait une taille finie, c'est-à-dire s'il existait un nombre maximal de phrases dans une langue, alors on pourrait toutes les recenser et les ranger dans une base de données : pour vérifier si une séquence arbitraire de mots est une phrase, il suffirait de consulter cette base de données : la séquence est une phrase si on la trouve dans la base ; elle n'est pas une phrase sinon. Malheureusement, l'ensemble des phrases d'une langue naturelle est infini. Pour s'en convaincre, raisonnons par l'absurde : imaginons pour un instant qu'il y ait  $n$  phrases en français.

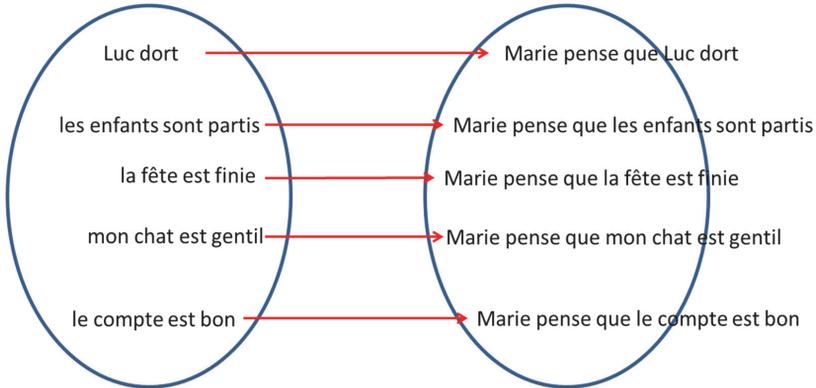
– A partir d'un ensemble fini de phrases initiales, on peut construire un deuxième ensemble de phrases en ajoutant devant chacune des phrases initiales la séquence *Marie pense que*, par exemple :

*Luc dort* → *Marie pense que Luc dort*

*La fête est finie* → *Marie pense que la fête est finie*

Par ce simple mécanisme, on vient de doubler le nombre de phrases, comme on le voit en figure 1.1. On peut généraliser ce mécanisme en utilisant d'autres verbes que le verbe *penser*, par exemple :

*Marie (affirme | crie | dit | pense | prétend | rêve | ...) que* Phrase.



**Figure 1.1.** On peut doubler la taille de n'importe quel ensemble de phrases

Il existe plusieurs centaines de verbes qui pourraient être utilisés ici<sup>4</sup>. De même, on peut remplacer *Marie* par plusieurs milliers de noms humains<sup>5</sup> :

*(Le directeur | l'instituteur | l'employée | la voisine | ...) pense que Phrase.*

Quelle que soit la taille  $n$  d'un ensemble initial de phrases, on peut donc multiplier la taille de cet ensemble par plusieurs centaines de milliers simplement en insérant devant chacune des phrases initiales des séquences comme *Marie pense que*, *L'instituteur affirme que*, *La voisine déclare que*, etc.

– La langue dispose d'autres mécanismes permettant de faire exploser le nombre de phrases. Par exemple, à partir de  $n$  phrases initiales, on peut construire  $n \times n$  phrases en combinant toutes les phrases de départ deux à deux, et en insérant dans chaque paire de phrases le mot *et*, par exemple :

*Il pleut + Luc dort → Il pleut et Luc dort*

On peut ici aussi généraliser ce mécanisme en utilisant plusieurs centaines de connecteurs<sup>6</sup>, par exemple :

*Il pleut (alors que | donc | mais | pour autant | quand | tant et si bien que | ...) Luc dort.*

4. La table 6 du *Lexique-Grammaire des Verbes* recense 480 verbes de structure *N0 V que Phrase* (exemple *Luc voit qu'il pleut*). La table 9 recense 400 verbes de structure *N0 V que Phrase à N2* (exemple *Luc dit qu'il pleut à Léa*), voir (Gross 1975).

5. Le dictionnaire DEM recense 16 635 noms humains, voir (Dubois 2010).

6. Voir (Piot 1993).

Ces deux mécanismes (imbrication de phrases, et utilisation de connecteurs) peuvent être appliqués plusieurs fois de suite, comme on le voit ci-dessous :

*Marie pense que Luc rêve que Jean espère que Paul dort  
Il pleut alors que Paul dort et que Marie attend mais il fera beau dès que le  
jour se lèvera*

On dit alors que ces mécanismes sont *récurifs* ; l'ensemble des phrases que l'on peut construire avec des mécanismes récurifs est infini. Il n'est donc pas possible de définir l'ensemble des phrases *in extenso*<sup>7</sup>. Pour caractériser cet ensemble, il faut donc trouver un autre moyen.

### 1.1. Caractériser un ensemble de taille infinie

Les mathématiciens savent depuis longtemps comment définir des ensembles de taille infinie. Par exemple, pour définir l'ensemble des entiers naturels  $\mathbb{N}$ , on peut utiliser les deux règles énoncées ci-après :

- (a) Chacun des dix éléments de l'ensemble  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  est un entier naturel ;
- (b) tout mot qui peut s'écrire sous la forme  $xy$  est un entier naturel si et seulement si ses deux constituants  $x$  et  $y$  sont des entiers naturels.

Ces deux règles constituent une définition formelle de l'ensemble des entiers naturels. Elles permettent de distinguer les entiers naturels de tout autre objet (nombre décimal ou autre). Par exemple :

– le mot « 123 » est-il un entier naturel ? Grâce à la règle (a), on sait que « 1 » et « 2 » sont des entiers naturels. La règle (b) nous permet d'en déduire que « 12 » est un entier naturel. Grâce à la règle (a), on sait que « 3 » est un entier naturel ; puisque « 12 » et « 3 » sont des entiers naturels, alors la règle (b) nous permet de déduire que « 123 » est un entier naturel ;

– en revanche, le mot « 2,5 » n'est pas un entier naturel. En effet, la règle (a) nous permet de déduire que « 2 » est bien un entier naturel, mais elle ne s'applique pas à la virgule « , ». La règle (b) ne peut s'appliquer qu'à deux entiers naturels, donc elle ne s'applique pas non plus à la virgule puisqu'elle n'est pas un entier naturel. Dans ce cas, « 2, » n'est pas un entier naturel, donc « 2,5 » n'est pas non plus un entier naturel.

---

7. (Gross 1975) montre qu'il est facile de construire un ensemble de  $10^{50}$  phrases sans même faire intervenir des opérations récurives. Ce nombre, comparable au nombre d'atomes contenus dans le soleil, est considéré comme infini dans la plupart des domaines scientifiques.

Il y a une similarité intéressante entre cette définition de l'ensemble des entiers naturels  $\mathbb{N}$  et le problème de la caractérisation des phrases d'une langue :

– la règle (a) décrit *in extenso* l'ensemble fini des chiffres que l'on doit utiliser pour former des entiers naturels valides. Cette règle ressemble à un dictionnaire dans lequel on recenserait tous les mots qui constituent le vocabulaire d'une langue ;

– la règle (b) explique qu'on peut combiner les chiffres pour construire une infinité d'entiers naturels. Cette règle ressemble aux règles de grammaire qui précisent comment combiner les mots pour construire une infinité de phrases.

Pour décrire une langue naturelle, on procédera donc de la façon suivante : d'une part, on définira *in extenso* le vocabulaire qui contient l'ensemble fini des mots d'une langue ; d'autre part, on recensera les règles qui permettent de combiner les éléments du vocabulaire pour construire des phrases.

## 1.2. L'ordinateur et la linguistique

L'ordinateur est un outil crucial pour le projet linguistique, pour au moins quatre raisons :

– d'un point de vue théorique : l'ordinateur est une machine mathématique que l'on peut utiliser pour vérifier l'appartenance d'un élément à un ensemble défini mathématiquement. Pour le projet linguistique, on vérifiera par exemple si une séquence de mots est une phrase valide dans une langue, à partir d'un vocabulaire et d'une grammaire<sup>8</sup> ;

– d'un point de vue méthodologique : l'ordinateur impose un cadre pour décrire aussi bien les objets linguistiques (par exemple les mots) que les règles d'utilisation de ces objets (par exemple les règles syntaxiques). La façon dont on décrit des phénomènes linguistiques n'est pas neutre et doit être cohérente avec le système : toute incohérence dans la description produit inévitablement une erreur (un « bug ») ;

– lorsque l'on a entré des descriptions linguistiques dans un ordinateur, celui-ci peut les utiliser pour explorer automatiquement des corpus de textes de taille très importante, afin d'extraire de ces corpus des exemples ou des contre-exemples qui valident (ou non) ces descriptions. On peut ainsi utiliser l'ordinateur comme un instrument scientifique (c'est l'approche de la *linguistique de corpus*), comme le télescope pour l'astronomie ;

– décrire une langue nécessitera un grand effort de description : on utilisera un logiciel pour aider au développement des bases de données contenant les nombreux

---

8. Nous verrons au chapitre 7 qu'il est possible de vérifier automatiquement si une phrase est correcte dans une langue en utilisant des grammaires génératives.

objets linguistiques ainsi que les nombreuses règles de grammaire, tout comme les logiciels de Conception assistée par ordinateur (CAO) utilisés par les ingénieurs pour concevoir des ponts, des voitures, des circuits électroniques, etc.

Finalement, la description de certains phénomènes linguistiques permet de construire des logiciels du TAL. Par exemple : si l'on dispose de la liste des mots d'une langue, on peut construire un vérificateur orthographique ; si l'on dispose de la liste des règles de conjugaison, on peut construire un conjugueur automatique ; la liste des règles morphologiques et phonologiques permet de proposer des corrections orthographiques lorsque l'ordinateur a détecté des erreurs ; la liste des termes simples et composés permet de construire un indexeur automatique ; si l'on dispose d'un dictionnaire et d'une grammaire bilingue, on peut construire un traducteur automatique, etc. L'ordinateur est donc devenu un outil central pour la linguistique, au point qu'opposer « linguistes informaticiens » aux « linguistes purs » n'a plus vraiment de sens.

### 1.3. Les niveaux de la formalisation

Lorsque l'on caractérise un phénomène à l'aide de règles mathématiques on le *formalise*. La formalisation d'un phénomène linguistique consiste donc à le décrire à l'aide d'objets et de règles linguistiques dans un ordinateur. Les langues sont compliquées à décrire, en partie parce que les interactions entre leur système phonologique et leur système d'écriture ont multiplié le nombre d'objets à traiter, et donc le nombre de niveaux de règles de combinaison. On peut distinguer cinq niveaux fondamentaux de phénomènes linguistiques : à chacun de ces niveaux correspondra donc un niveau de formalisation.

Pour étudier un texte écrit, on a accès aux lettres de l'alphabet plutôt qu'aux mots : il faut donc décrire le lien entre l'alphabet et les formes orthographiques qu'on veut traiter (*orthographe*). Ensuite, il faut établir un lien entre les formes orthographiques et les éléments du vocabulaire correspondants (*morphologie*). Les éléments du vocabulaire sont recensés et décrits en général dans un lexique qui doit aussi donner accès à toutes les ambiguïtés potentielles (*lexique*). Les éléments du vocabulaire se combinent pour construire des unités plus grandes (les syntagmes) qui elles-mêmes se combinent pour constituer des phrases : il faut donc établir des règles de combinaison (*syntaxe*). Enfin, les liens entre les éléments de sens (les sèmes) qui forment les prédicats transcrits en phrases élémentaires, ainsi que les liens entre prédicats dans une phrase complexe doivent être établis (*sémantique*).

## 1.4. Hors sujet

Nous n'utilisons pas toujours la langue pour représenter et communiquer directement et simplement les informations : parfois, nous jouons avec la langue pour créer des effets sonores (par exemple en poésie), parfois nous jouons avec les mots, parfois, nous laissons implicites certaines informations « évidentes » car relevant de la culture partagée par les interlocuteurs (anaphores), parfois nous exprimons une idée pour en fait en suggérer une autre (métaphore), parfois, nous utilisons la langue pour communiquer des affirmations sur le monde réel ou dans des domaines scientifiques (y compris en linguistique), et même parfois nous disons l'inverse de ce que nous voulons vraiment dire (ironie), etc.

Il est important de bien distinguer les problèmes que l'on peut résoudre à l'intérieur d'un cadre d'analyse strictement linguistique, de ceux qui, pour être résolus, nécessiteraient l'accès à des informations qui dépendent d'autres domaines.

### 1.4.1. Poésie, jeux de mots

Les littéraires, les poètes et les auteurs de jeux de mots s'autorisent souvent à construire des textes qui violent les contraintes syntaxiques ou sémantiques de la langue. Par exemple, considérons le texte suivant<sup>9</sup> :

*Jamais, avez-vous dit, tandis qu'autour de nous  
Résonnait de Schubert la plaintive musique ;  
Jamais, avez-vous dit, tandis que, malgré vous,  
Brillait de vos grands yeux l'azur mélancolique.*

Dans cette poésie, les compléments de nom *de Schubert* et *de vos grands yeux* ainsi que l'adjectif *plaintive* sont déplacés : en français standard, on écrirait *la musique plaintive de Schubert* plutôt que *de Schubert la plaintive musique*. On écrirait aussi *l'azur mélancolique de vos grands yeux* plutôt que *de vos grands yeux l'azur mélancolique*.

L'association nom adjectif *azur mélancolique* ne correspond pas à une association standard, car l'adjectif *mélancolique* qualifie une personne (ou une histoire, un événement, etc.), mais pas une couleur. Pour appréhender cette association, il faudrait calculer des interprétations comme les suivantes :

- soit c'est la personne aux grands yeux couleur azur qui est mélancolique ;
- soit ce sont les yeux couleur azur qui provoquent la mélancolie chez l'auteur ;
- soit c'est le « jamais » exprimé par la personne aux grands yeux couleur azur qui provoque la mélancolie chez l'auteur.

---

9. Extrait du poème *Jamais* d'Alfred de Musset.

Même si l'on avait l'ambition de formaliser ces interprétations dans un « analyseur poétique », il faudrait commencer par décrire la langue standard, ne serait-ce que pour détecter que *azur mélancolique* n'est pas une association standard, et donc qu'il faut effectuer une analyse poétique. En effet, si un système ne perçoit même pas que *azur mélancolique* n'est pas une association standard, il ne cherchera pas la vraie interprétation de l'adjectif *mélancolique*.

#### 1.4.2. *Stylistique et rhétorique*

La stylistique étudie les manières de formuler les phrases dans un discours. On étudie par exemple dans un texte l'utilisation de litotes, métaphores et métonymie (les « figures de style »), l'ordre des éléments dans une phrase, et celui des phrases entre elles pour articuler un discours, ainsi que l'utilisation d'anaphores. (Karabétian 2002) présente les enjeux de ce domaine à la frontière entre analyse du discours, linguistique, rhétorique et études littéraires. Voici quelques exemples de phénomènes que l'on ne peut pas traiter dans un cadre strictement linguistique :

Litote :	<i>Luc n'est vraiment pas bête</i>
Métaphore :	<i>Luc est un véritable éléphant</i>
Métonymie :	<i>La table a éclaté de rire</i>

En fait, la phrase *Luc n'est vraiment pas bête* signifie ici que Luc est très intelligent... En quelque sorte, cette phrase ne dit pas ce qu'elle exprime !

Pour comprendre le sens de la phrase *Luc est un véritable éléphant*, il faut savoir quelle « propriété caractéristique » des éléphants est utilisée dans la métaphore. Les éléphants ont plusieurs réputationes : ils sont gros, forts, maladroits, ils ont une grande mémoire, ils ont peur des souris, ils sont une espèce en danger, ils ont de grandes oreilles, ils aiment faire des bains de boue, ils habitent en Afrique ou en Inde, etc. Luc est-il maladroit ? A-t-il peur des souris ? Aime-t-il les bains de boue ? A-t-il une bonne mémoire ? Pour le savoir il faudrait connaître Luc...

Pour comprendre le sens de la phrase *La table a éclaté de rire*, il faut inférer qu'il y a des personnes positionnées autour d'une table (lors d'un repas, ou d'une réunion de travail) et que ce sont ces personnes qui ont éclaté de rire. Le nom *table* n'est ni un nom humain collectif (comme *groupe*, *colonie*), ni un nom de lieu contenant typiquement des personnes (comme *salle de réunion* ou *amphithéâtre*), ni un nom d'organisation (comme *association* ou *banque*) : il n'est donc pas possible d'interpréter la métonymie à l'aide des seules propriétés lexicales associées au nom *table*.

### 1.4.3. Anaphores, résolution de coréférence, désambiguïstation sémantique

Coréférence : *Luc a invité Paul à dîner. Il a apporté une bouteille de vin.*

Anaphore : *Phelps est rentré. Le champion est reparti avec 6 médailles.*

Ambiguïté sémantique : *Il y a une table ronde en salle B17.*

Pour comprendre que dans la phrase *Il a apporté une bouteille de vin*, il réfère à *Paul* et non à *Luc*, il faut savoir que d'habitude, c'est l'invité qui apporte une bouteille de vin.

Pour comprendre que *Le champion* est une référence à *Phelps*, il faut savoir que *Phelps* est un champion. Noter qu'on aurait pu utiliser des dizaines d'autres noms dans cette anaphore : l'Américain, le médaillé, le recordman, le surdoué des piscines, le jeune homme, le nageur, l'ex-étudiant de l'Université d'Etat de Floride, le solitaire, le philanthrope, etc.

Pour lever l'ambiguïté de la séquence *table ronde* (entre « une table de forme ronde » et « une réunion »), il faudrait avoir accès à un contexte plus grand que la phrase.

ATTENTION.— Je ne dis pas qu'il est toujours impossible de traiter la poésie, les jeux de mots, les litotes, les métaphores, les métonymies, les coréférences, les anaphores et les ambiguïtés sémantiques ; je dis seulement que, dans le cas général, ces phénomènes dépassent le cadre du projet présenté dans cet ouvrage. Il existe ainsi des cas « chanceux » où des logiciels peuvent résoudre automatiquement certains de ces phénomènes. Par exemple, dans la séquence suivante :

*Luc a invité Léa à dîner. Elle a apporté une bouteille de vin*

une simple vérification du genre du pronom permettrait d'attacher *Elle* à *Léa*. Noter qu'inversement, on peut aisément construire un logiciel qui, à partir des deux phrases : *Luc a invité Anne à dîner* et *Anne a apporté une bouteille de vin* produira la phrase *Celle-ci a apporté une bouteille de vin*. De même, dans la phrase suivante :

*Les collègues ont assisté à la table ronde*

un logiciel pourrait calculer automatiquement que le nom *table ronde* réfère à une réunion s'il dispose d'un dictionnaire dans lequel le verbe *assister* est décrit comme nécessitant un complément d'objet indirect abstrait.

#### 1.4.4. *Calculs extralinguistiques*

Considérons les énoncés suivants :

- a) *Deux euros plus trois euros font quatre euros ;*
- b) *Jacques Chirac était déjà président en 1736 ;*
- c) *Dieu a sept lettres ;*
- d) *Cette phrase est fausse.*

Ces énoncés sont exprimés à l'aide de phrases valides linguistiquement puisqu'elles respectent les contraintes orthographiques, morphologiques, syntaxiques et sémantiques de la langue. En revanche, elles expriment des affirmations incorrectes du point de vue mathématique (a), historique (b), de l'orthographe<sup>10</sup> (c) ou de la logique (d). Il serait impossible d'intégrer à la description de la langue la description de tous les domaines du savoir. Le projet décrit dans cet ouvrage ne traite pas de tels phénomènes<sup>11</sup>.

Il faut cependant mentionner que ces problèmes « extralinguistiques » ont tout de même des applications potentielles en TAL : par exemple il existe des logiciels capables de résumer des textes automatiquement : ces outils ont besoin de résoudre des métaphores, métonymies, coréférences et anaphores et doivent produire des textes stylistiquement agréables. Ces applications ne prétendent pas décrire les langues (ni même comprendre le sens des textes produits), mais réussissent souvent à produire des résumés utiles<sup>12</sup>. Cet ouvrage n'abordera donc pas les problèmes liés à l'analyse de la poésie, à la stylistique, à la rhétorique, à la résolution d'anaphores ou à la levée d'ambiguïté sémantique.

#### 1.5. Structure de cet ouvrage

Le projet décrit dans ce chapitre a pour but de décrire très précisément, mathématiquement (c'est-à-dire de formaliser) les langues naturelles, plus précisément les phrases qui peuvent apparaître dans des textes écrits. Dans le chapitre 2, je situe ce

---

10. Dans son introduction à (Harris 1976), Maurice Gross analyse cette phrase à partir de la phrase de base *Le mot « Dieu » a quatre lettres*. Le statut « strictement linguistique » de cette phrase n'est pas différent de celui de la phrase *Dieu a sept lettres*, même si cette dernière exprime un fait orthographique faux.

11. Certains linguistes théoriciens avancent des exemples de type (c) ou (d) pour démontrer la nécessité d'utiliser des grammaires non restreintes pour décrire les langues.

12. (Mani 2001) présente le domaine du résumé automatique et ses problématiques : comment un humain s'y prend-il pour résumer un document, quelles sont les informations « importantes » qu'il faut inclure dans un résumé, comment construire un résumé cohérent, comment résumer des documents multimédias (vidéo, graphique, etc.) ? Comment évaluer la qualité des résumés.

projet par rapport à d'autres projets tels que le *Traitement automatique des langues* (TAL), la linguistique formelle et l'informatique linguistique.

Le nombre de phrases d'une langue est infini, mais il est tout de même possible de le définir en s'appuyant sur l'ensemble fini des unités linguistiques de base, ce que je ferai dans la première partie de cet ouvrage. Le chapitre 3 montre comment formaliser l'alphabet des langues ; le chapitre 4 présente comment formaliser le vocabulaire ; le chapitre 5 montre les différences essentielles entre les dictionnaires traditionnels et les dictionnaires nécessaires au projet de formalisation linguistique ; le chapitre 6 présente enfin les dictionnaires électroniques dont le but est de formaliser le vocabulaire de la langue.

Il faut ensuite se doter de mécanismes pour décrire comment les éléments linguistiques de base (lettres, morphèmes ou mots) se combinent pour construire des éléments de niveau plus haut (mots, formes, ou phrases) : la seconde partie de cet ouvrage introduit les notions de *langage formel*, de *grammaire générative* et de *machine*. Je présente ces notions ainsi que la hiérarchie de Chomsky-Schützenberger au chapitre 7 ; les chapitres 8 à 11 présentent les quatre types de langages/grammaires/machines.

La troisième partie (chapitres 12 à 16) est consacrée à l'analyse linguistique automatique des textes. Dans le chapitre 12, je présente la structure d'annotation des textes (TAS) qui permet de représenter de façon unifiée tous les résultats produits par les analyses linguistiques. Le chapitre 13 est consacré à l'analyse lexicale automatique. Les chapitres 14, 15 et 16 présentent trois niveaux d'analyse syntaxique : analyse locale, analyse structurelle et analyse transformationnelle. Le chapitre 16 montre en particulier qu'un analyseur transformationnel automatique peut produire une analyse sémantique strictement linguistique, c'est-à-dire une analyse de sens fondée uniquement sur la langue (sans inférence dans le monde réel).

## 1.6. Exercices

1) Sur le modèle de la définition de  $\mathbb{N}$  vue en section 1.1, caractériser l'ensemble  $\mathbb{D}$ , c'est-à-dire l'ensemble des nombres décimaux qui contiennent une partie entière, une virgule et une partie décimale de longueur finie.

2) Soit la phrase : *Le grand-père a été relevé de ses fonctions*. Décrire informellement les analyses linguistiques de cette phrase aux niveaux orthographique, lexical, morphologique, syntaxique et sémantique.

3) Soit le texte : *Luc a invité Paul à dîner. Le viticulteur a, comme d'habitude, sabré le champagne*. A qui réfère le nom *viticulteur* : à *Luc* ou à *Paul* ? Le calcul de la référence peut-il être effectué grâce à des informations linguistiques ? Si oui, lesquelles ?

### 1.7. Liens Internet

La page Wikipedia suivante présente le domaine de la linguistique :

[fr.wikipedia.org/wiki/Linguistique](http://fr.wikipedia.org/wiki/Linguistique)

La page suivante présente les applications du TAL :

[fr.wikipedia.org/wiki/Traitement\\_automatique\\_du\\_langage\\_naturel](http://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel)

L'*Association pour le traitement automatique des langues* (ATALA) a un site *web* :

[www.atala.org](http://www.atala.org)

Le logiciel NooJ, qui sera utilisé dans cet ouvrage pour formaliser les phénomènes linguistiques est gratuit, open source, et est disponible à partir du site :

[www.nooj4nlp.net](http://www.nooj4nlp.net)