

## Chapitre 1

# Mots et niveau lexical

### 1.1. Introduction

Le niveau lexical du traitement des langues naturelles correspond aux traitements centrés sur la notion de mot. Le matériau de base de l'ingénierie linguistique étant le texte, parlé ou écrit, et le texte étant fait de mots, on conçoit facilement que le niveau lexical soit fondamental. Il est en fait relié, de près ou de loin, à la totalité de ce vaste domaine. Nous passerons donc en revue des sujets très variés.<sup>1</sup>

Cet examen sera l'occasion d'exercer une évaluation critique à différents niveaux. Nous envisagerons les principaux traitements du point de vue de leurs objectifs applicatifs comme du point de vue des résultats obtenus. Nous nous intéresserons ensuite aux outils de base qui ont été élaborés en vue de ces traitements et qui font partie de l'acquis des recherches : il s'agit de données linguistiques, des dictionnaires par exemple, et d'outils méthodologiques, tels que des formalismes. Enfin, pour donner un aperçu des perspectives de recherche actuellement ouvertes, nous évoquerons brièvement les principaux cadres méthodologiques qui constituent l'arrière-plan de tous ces travaux.

---

<sup>1</sup> Une version précédente de ce chapitre a été publiée sous forme d'article dans la revue *Traitement automatique des langues* (vol. 38, n° 2, *État de l'art*, 1997). Le contenu de cet article a été entièrement refondu et actualisé en vue de l'édition du présent ouvrage.

## 2 Ingénierie des langues

Nous illustrerons notre propos par un échantillon de références bibliographiques, sélectionnées pour leur qualité, pour leur originalité ou pour leur représentativité, et nous tenterons de fonder sur des jugements scientifiques un bilan critique constructif. Nous avons choisi de ne pas structurer ce chapitre à partir des tendances et des modes scientifiques du domaine. Mentionnons toutefois que les considérations “sociologiques” de ce type sont souvent encore perçues comme des éléments très importants, peut-être en raison de l'histoire relativement courte et rapide du traitement des langues naturelles, de ses liens évidents avec la linguistique et l'informatique, et de ses prolongements applicatifs.

### **1.2. Définition du niveau lexical et enjeux applicatifs**

Une profusion de traitements sur les mots ont déjà été mis en œuvre au moins une fois. S'y ajoutent ceux dont rêvent les utilisateurs, les industriels et les chercheurs. Ces traitements se laissent toutefois classer sous quelques rubriques : étiqueter, corriger, retrouver, diviser et quelques autres. Sous chacune de ces rubriques, nous définissons brièvement les applications.

#### ***1.2.1. Détection et correction d'erreurs***

Les applications industrielles de la détection et de la correction d'erreurs orthographiques vont des logiciels distribués au grand public, comme les traitements de texte, jusqu'à des systèmes spécifiquement destinés aux professionnels et parfois développés à façon : correction de textes pour l'édition, recherche de noms ou mots-clés éventuellement erronés... On peut rattacher à ce type d'applications celles qui font intervenir la reconnaissance de variantes de mots : par exemple, la recherche de relations entre de nouveaux noms commerciaux et des noms commerciaux existants, ou le dédoublement de fichiers de noms ou d'adresses. Le dédoublement consiste à rechercher les noms de personnes ou les adresses qui figurent plusieurs fois dans une base de données, même sous des formes légèrement différentes, et à éliminer les doublons.

#### ***1.2.2 Documentation, indexation et moteurs de recherche***

Il s'agit de localiser des textes ou portions de textes : paragraphe, article, livre... L'utilisateur poursuit un but de documentation. La requête de l'utilisateur, quelle que soit la forme qu'elle prend, est confrontée au contenu d'une base de documents.

L'attente des industriels dans ce domaine est considérable, étant donné le marché que représente la gestion documentaire. Ce marché s'étend encore avec le développement du réseau informatique mondial : les documents électroniques

accessibles dans les sites web constituent un champ de recherches documentaires et de veille technologique vaste et en pleine expansion.

Le lien entre la requête de l'utilisateur et un document se fait par l'intermédiaire de descripteurs : mots, éléments de mots ou groupes de mots, censés indiquer de quoi parle le document. Dans les domaines techniques, les termes techniques constituent bien sûr des candidats privilégiés au statut de descripteurs, en raison de leur contenu informatif élevé et de leur utilisation relativement normalisée. Bien que les termes techniques apparaissent souvent comme des "groupes de mots", comme  *fibre optique*, remarquons dès maintenant qu'il est parfaitement légitime ici de les considérer comme des mots au même titre que, par exemple,  *autocommutateur*, à ceci près qu'il s'agira de mots composés <sup>2</sup> (GROSS, 1986b). D'un point de vue linguistique, il existe plusieurs notions de "mot", en relation complexe les unes avec les autres ; aucune ne correspond exactement à la notion typographique de mot simple délimité par des séparateurs, alors que seule cette dernière est simple à reconnaître dans les textes, en raison de sa définition purement formelle.

Par ailleurs, la requête, et éventuellement une description du document, peuvent se présenter sous l'aspect de formules qui comportent des descripteurs et qui se conforment à un langage d'indexation plus ou moins étroitement formalisé.

### 1.2.3. *Quelques autres traitements*

L'**accentuation automatique** consiste à rétablir les accents et autres signes diacritiques, comme la cédille en français, dans les textes non accentués ou accentués de façon incomplète. Ce type de traitement s'adresse par exemple aux textes en français typographiés sans accents dans les années 1970. De nos jours, la plupart des textes saisis en français le sont de façon soignée. Cependant, ce n'est pas toujours le cas des mess<sup>3</sup>ages envoyés par courrier électronique. De plus, la norme internationale veut que les lettres accentuées apparaissent effectivement avec leurs accents même lorsqu'elles sont en majuscules, mais cette règle n'est pas encore tout à fait passée dans les mœurs, même dans les textes typographiés de façon soignée. Au-delà de ce problème précis, l'accentuation automatique est comparable à d'autres traitements spécifiquement destinés aux textes corrompus, que la perte d'information vienne du mode de saisie : saisie manuelle, lecture optique... ou d'un canal de transmission.

---

<sup>2</sup> On parle aussi d'unités multi-mots ou d'expressions figées.

<sup>3</sup> En versification, une césure est une limite rythmique à l'intérieur d'un vers ; en musique, il s'agit d'un repos suspensif dans une phrase musicale.

#### 4 Ingénierie des langues

En typographie, le découpage des mots en fin de ligne est appelé **division** par les professionnels, et non “césure”. La division automatique des mots est une application classique dans l'édition ou la typographie. En français, elle correspond souvent, mais pas toujours, au découpage en syllabes.

La **phonétisation** est la transcription phonétique du texte écrit. Les applications de ce traitement se situent dans le domaine de la parole : synthèse de messages vocaux et reconnaissance de la parole. Il existe toutefois une application au traitement du texte écrit : la correction orthographique par phonétisation.

#### 1.2.4. Étiquetage lexical

Les traitements auxquels nous avons fait référence ci-dessus, pour donner de bons résultats, doivent presque tous faire appel à une étape au cours de laquelle on étiquette les mots par des informations linguistiques. Il s'agit donc d'une opération de base.

Nous regrouperons sous le terme d'étiquetage lexical l'ensemble des techniques qui concourent à passer d'un texte brut, exempt d'informations linguistiques, à une séquence de mots étiquetés par des informations linguistiques, au premier rang desquelles les informations morphologiques et grammaticales. Cette définition inclut donc la délimitation des mots, la morphologie et la levée des ambiguïtés lexicales.

À première vue, les applications industrielles d'une telle tâche ne sautent pas aux yeux, et en effet, elles sont toutes indirectes. Cependant elles sont remarquablement variées : pratiquement toutes les applications qui mettent en jeu du texte verraient leurs performances s'améliorer par l'intégration de meilleurs systèmes d'étiquetage lexical, y compris des applications relevant de la syntaxe ou de la parole. C'est d'ailleurs un des acquis sur lesquels s'accorde la communauté scientifique du domaine : après une période de décantation des techniques destinées à effectuer un traitement donné, on arrive à la conclusion qu'une étape préalable d'étiquetage lexical de qualité résoudrait bien des difficultés.

Compte tenu des enjeux applicatifs, l'objectif de l'étiquetage lexical est de reconnaître et d'étiqueter des formes pertinentes, qui ont un statut d'unité de base, c'est-à-dire bien souvent celles auxquelles est attaché un sens. Les difficultés de l'étiquetage lexical sont variables suivant les langues :

– les informations linguistiques sur les mots ne sont pas déductibles de leur forme (ainsi, en français, les mots terminés par *-s* sont souvent au pluriel, comme *raisons*, mais certains peuvent être au singulier, comme *stimulus* et *prends* ; pour d'autres, la notion de pluriel n'est pas pertinente, comme *après* ; et certains mots non terminés par *-s* peuvent aussi être au pluriel, comme *réseaux*) ;

- la taille du vocabulaire de mots simples au sens typographique, c'est-à-dire le nombre de mots distincts susceptibles d'apparaître dans les textes, se compte en centaines de milliers pour le français et l'anglais, et en centaines de millions pour des langues agglutinantes telles que le hongrois et le coréen (PROSZEKY, 1996) ;
- toutes les langues présentent des ambiguïtés lexicales, c'est-à-dire que certaines formes peuvent recevoir plus d'une étiquette, comme *règle* qui est nom ou verbe ;
- certains systèmes d'écriture n'ont pas de séparateurs, ce qui pose le problème de la délimitation des mots (SPROAT, 1996) ; ce problème se pose, en français et en anglais, pour les mots composés qui la plupart du temps ne sont pas graphiquement marqués par des séparateurs, comme *fibres optiques* ;
- certains systèmes d'écriture ne sont pas normalisés, c'est-à-dire qu'il existe plusieurs façons correctes d'orthographier un texte donné, en raison de lettres facultatives, de séparateurs facultatifs, ou parce qu'ils permettent de passer facultativement d'un alphabet à un autre comme en japonais.

#### **1.2.5. Traitement des mots inconnus**

Le terme de mot inconnu est généralement employé pour désigner les mots corrects dont on ne dispose pas dans un dictionnaire. Le traitement des mots inconnus consiste à en effectuer l'étiquetage lexical. Le problème n'a donc de sens que dans le cadre d'un système d'étiquetage lexical donné. Les applications du traitement des mots inconnus se confondent avec celles de l'étiquetage lexical, dont il est le prolongement. Dans les systèmes d'étiquetage lexical qui n'utilisent ni dictionnaire ni corpus étiqueté, le problème des mots inconnus ne se pose pas particulièrement : tous les mots sont considérés comme inconnus *a priori*.

#### **1.2.6. Un exemple développé : les concordanciers**

L'élaboration de concordances consiste à rechercher dans un texte toutes les occurrences d'un mot ou d'un autre motif linguistique, puis à les présenter, une par ligne, chacune dans son contexte. Les applications relèvent de la lexicographie, de l'apprentissage des langues et de l'exploitation de bases de données littéraires. L'élaboration du dictionnaire *COBUILD* (COBUILD, 1987), par exemple, a systématiquement fait appel à la recherche d'exemples dans des concordances (SINCLAIR, 1991). Un concordancier est un logiciel de construction de concordances.

Examinons un peu plus en détail cette application. Bien qu'elle soit peu connue des amateurs, elle offre des exemples concrets de problèmes typiques du niveau lexical, vus du point de vue de l'utilisateur du traitement des langues naturelles.

La difficulté de la construction de concordances réside essentiellement dans le mécanisme de sélection des mots à partir du motif demandé par l'utilisateur (GARRIGUES, 1997). Ce motif peut être :

(1) une simple forme de surface, comme *résolu*, ou une séquence de motifs de ce type ;

(2) une expression rationnelle sur les lettres, comme *[rR]éso\**, qui reconnaîtra tout mot commençant par  *réso* ou  *Rés*, ou une séquence de motifs de ce type ;

(3) un motif défini par des critères linguistiques : un lemme (<*résoudre*> pour reconnaître toutes les formes fléchies de ce verbe), une catégorie grammaticale ou des traits flexionnels (<*V:S2p*> pour reconnaître tous les verbes au subjonctif, 2e personne du pluriel)... ou une séquence de motifs de ce type.

Les motifs des types (1) et (2) peuvent être confrontés aux occurrences de mots du texte par des fonctions d'appariement de chaînes de symboles. Ils peuvent rendre des services, en particulier dans les langues à morphologie pauvre, comme l'anglais. Dans le cas des langues romanes, ils ne permettent pas de simuler les motifs du type (3) : par exemple, le motif *[rR]éso\** reconnaîtra aussi bien *résolution* et les formes du verbe *résonner* que celles de *résoudre*. Les concordanciers du commerce se rangent dans cette catégorie.

La construction de concordances à partir de motifs du type (3) suppose un étiquetage lexical préalable du texte (SILBERZTEIN, 1993). Inversement, toutes les informations qu'il est possible d'obtenir par un étiquetage lexical peuvent être exploitées dans la construction de concordances, par exemple la délimitation des noms composés. Les concordanciers du type (3) sont encore absents du commerce, bien que leur faisabilité soit incontestable. Certains concordanciers commercialisés utilisent même comme argument de vente le fait qu'ils ne comportent pas de dictionnaire.

Au terme de cette illustration du niveau lexical à travers les principaux traitements concernés, retenons l'importance stratégique particulière que revêtent deux types d'applications :

- d'une part, la gestion documentaire, en raison du fait que les documents sont, dans les entreprises, la principale matière première qui suscite des besoins de traitements linguistiques ;
- d'autre part, et surtout, l'analyse lexicale, de plus en plus indispensable à l'amélioration des performances qualitatives d'autres applications sur les textes.

### 1.3. Outils actuels : formalismes et données

De nombreux outils conceptuels fondamentaux ont été mis au point pour le traitement des langues naturelles. Les outils fondamentaux pertinents au niveau lexical, donc les briques de base du traitement des langues naturelles, sont bien sûr :

- les techniques, méthodes et algorithmes conçus pour effectuer des tâches spécifiques ;
- les formalismes et modèles formels qui servent de cadre aux calculs et autres traitements ;
- mais aussi les données spécifiques au langage : dictionnaires électroniques, grammaires, collections de textes.

En effet, au cours des quelques décennies qui nous séparent de la naissance du traitement automatique des langues naturelles, les acteurs du domaine ont plus ou moins progressivement pris conscience de l'enjeu que constituent les données spécifiques au langage : dictionnaires électroniques, grammaires, textes, par opposition aux processus de calcul eux-mêmes. Les spécialistes de l'apprentissage pensent que les données ne peuvent être qu'incomplètes et que seules des procédures de calcul peuvent en compenser les manques, mais ils considèrent comme un résultat de leurs travaux les données obtenues par apprentissage. Quelle qu'en soit l'origine, le contenu des données est indispensable au fonctionnement de nombreuses applications ; leur volume peut être tellement considérable qu'il est nécessaire d'en tenir compte à la conception des applications ; le format dans lequel elles sont exprimées n'est généralement pas indépendant des algorithmes et des programmes qui les traiteront. Tous ces caractères font des données linguistiques un élément central dans la conception d'un traitement, et même un élément structurant dont le volume et le format doivent être pris en compte en priorité, dès les prises de décisions fondamentales.

Au début de l'histoire du domaine, rien n'orientait les spécialistes vers des conclusions aussi contraires à l'intuition (LOCKE, 1955 ; CECCATO, 1962), mis à part certains linguistes (GROSS, 1972). D'une part, chacun pouvait constater que n'importe quel être humain mémorise et pratique sa langue maternelle sans effort apparent de mémoire. D'autre part, la technologie de traitement des langages artificiels avait atteint en quelques années un niveau largement suffisant pour les applications dont elle est la base. Or, elle avait atteint ce niveau en définissant des propriétés abstraites et des algorithmes, mais sans mettre particulièrement l'accent sur le format que doivent prendre les données, ni sur leur volume. Enfin, aux yeux de nombreux linguistes, la description du lexique est une activité peu valorisée, comme l'est la taxonomie aux yeux de nombreux biologistes.

C'est plutôt l'insuffisance qualitative et quantitative des données linguistiques disponibles qui a peu à peu attiré l'attention sur elles. Au fur et à mesure que l'intérêt pour les applications réelles s'est précisé, il s'est avéré que la construction de

systèmes à grande échelle nécessitait l'emploi de données linguistiques formelles (BOITET, 1982). Ces données y ont gagné un nom, celui de "ressources linguistiques". Étant donné le coût de la constitution de données fiables, utilisables et réutilisables, la communauté scientifique internationale a pris conscience de l'existence d'un "goulot d'étranglement" dans l'acquisition des connaissances linguistiques, et on a reconnu comme résultats intermédiaires à part entière les données linguistiques indispensables aux traitements.

Cette section est donc consacrée aux techniques et méthodes de base, aux formalismes et modèles formels, ainsi qu'aux différents ensembles de données linguistiques formelles existantes. Notre but étant de présenter un état de l'art, nous examinerons les perspectives de développement dont ces outils peuvent constituer le support, et les performances atteintes par les systèmes réalisés, et nous les confronterons aux attentes des industriels pour la réalisation d'applications.

### ***1.3.1. Étiquetage lexical***

L'objectif de l'étiquetage lexical d'un texte est de faciliter ou même de rendre possible des traitements applicatifs. C'est pourquoi nous commençons notre tour d'horizon par l'étiquetage lexical : les autres types de traitements y feront bien souvent référence.

Les différentes méthodes d'analyse lexicale en usage utilisent comme données linguistiques :

- un dictionnaire et des grammaires de levée d'ambiguïtés (KOSKENNIEMI, 1983 ; SILBERZTEIN, 1993 ; OFLAZER, 1996),
- un corpus de textes étiquetés (CHURCH, 1988 ; DERMATAS, 1995), éventuellement accompagné d'informations telles que des schémas de règles établis à la main (BRILL, 1995),
- un corpus non étiqueté mais accompagné d'informations linguistiques, par exemple un jeu d'étiquettes lexicales et un ensemble de relations étiquette-suffixe (LEVINGER, 1995),
- un corpus non étiqueté (MACMAHON, 1996) ; dans ce cas, le jeu d'étiquettes lui-même est construit automatiquement par des calculs statistiques et son contenu est à peu près imprévisible.

Les systèmes à dictionnaire accèdent directement aux données ; les systèmes à corpus font appel à un apprentissage statistique pour deviner l'étiquette de chaque mot. Certains systèmes utilisent à la fois un dictionnaire, pour recenser les étiquettes possibles, et un corpus, pour lever des ambiguïtés.

Les systèmes à dictionnaire permettent une délimitation et un étiquetage des mots composés, ce qui constitue un avantage du point de vue applicatif dans la mesure où ce sont des unités significatives qui sont ainsi reconnues. La faisabilité de cette stratégie est illustrée par le système INTEX, avec lequel les mots composés sont délimités, étiquetés en tant que tels et indexés avec succès (SILBERZTEIN, 1993).

Les systèmes à apprentissage à partir de corpus de textes sont par nature statistiques ; cependant, les résultats de l'apprentissage peuvent prendre une forme numérique ou symbolique, et leur réutilisabilité dépend de ce paramètre (cf. 1.3.4). Ces systèmes se sont développés en nombre croissant depuis le milieu des années 1980, car la construction des données nécessaires à leur réalisation est moins coûteuse qu'un dictionnaire de bonne qualité accompagné de grammaires de levée d'ambiguïtés.

Pour l'étiquetage lexical comme pour de nombreuses autres activités consistant à associer à des éléments connus (le texte) des éléments pris parmi un stock (les étiquettes), les résultats obtenus ne correspondent pas toujours exactement aux résultats désirés. Cet écart se mesure par le bruit et le silence. On peut définir le taux de bruit comme la proportion d'étiquettes non désirées parmi les étiquettes présentées, et le taux de silence comme la proportion d'étiquettes non présentées parmi les étiquettes désirées. De façon équivalente, on peut définir et utiliser des taux "positifs" qui font le complément à 100 % par rapport aux précédents : le taux de précision (proportion d'étiquettes désirées parmi les étiquettes présentées), complémentaire du taux de bruit, et le taux de rappel (proportion d'étiquettes présentées parmi les étiquettes désirées), complémentaire du taux de silence. Pour homogénéiser les évaluations quantitatives dans ce chapitre, nous emploierons systématiquement ces formules lorsqu'il sera question de bruit et de silence.

L'évaluation quantitative des performances d'un outil d'étiquetage lexical devrait logiquement refléter l'écart entre les résultats obtenus et les résultats corrects, et notamment mettre en jeu l'estimation des taux de bruit et de silence, dont les définitions sont simples et générales. Toutefois, il existe toujours des obstacles sérieux à l'élaboration de procédures d'évaluation adéquates qui permettraient des comparaisons entre tous les systèmes.

Le premier obstacle concerne la définition des étiquettes désirées, ou exactes. En effet, l'écart entre les résultats obtenus et les résultats désirés est mesuré en prenant comme références des étiquetages considérés comme corrects, et constitués d'échantillons de textes étiquetés. Les systèmes fondés sur l'apprentissage statistique ne permettent de délimiter les mots composés que de façon très partielle et leurs auteurs répugnent à inclure l'étiquetage des mots composés dans leurs objectifs, et donc dans leurs étiquetages de référence.

Un second obstacle tient à la variété des jeux d'étiquettes utilisés. Le calcul du bruit et du silence repose sur des comptages d'étiquettes. Or ces comptages donnent des valeurs différentes en fonction du jeu d'étiquettes, même pour une langue donnée. On constate de grandes disparités dans la taille des jeux d'étiquettes utilisés, et donc dans leur granularité : lorsque le nombre d'étiquettes croît, elles sont plus précises et plus informatives. Dans les systèmes à corpus, un jeu de 36 étiquettes pour l'anglais est souvent utilisé (MARCUS, 1993) ; des jeux de 10 à 500 étiquettes environ sont utilisés pour diverses langues. Dans les systèmes à dictionnaire, qui répertorient pour chaque mot toutes les étiquettes lexicales a priori envisageables, il n'y a pas de limite technique à la richesse du jeu d'étiquettes. Ainsi, les valeurs calculées du bruit et du silence dépendent du jeu d'étiquettes considéré (LAPORTE, 1996) : les résultats d'un même système de levée d'ambiguïtés peuvent même recevoir une évaluation quantitative plus flatteuse lorsqu'ils sont transférés dans un jeu d'étiquettes plus pauvre, c'est-à-dire de taille inférieure et de granularité plus grossière. De plus, la mise en correspondance effective de jeux d'étiquettes distincts n'est pas une opération simple (ADDA, 1997), et n'a jamais été tentée dans le cas où l'un des deux jeux comporte des étiquettes de mots composés.

Face à ces difficultés, peut-on mettre au point des méthodes d'évaluation fondées sur de tout autres principes ? Pour des applications effectives qui ont des utilisateurs directs, on peut partir de l'idée que la qualité d'un système est proportionnelle au confort qu'il apporte à son utilisateur. Mais l'étiquetage lexical est essentiellement envisagé comme une facilitation d'autres applications, comme l'accentuation ou la phonétisation. Quant à celles-ci, il existe des habitudes relativement normalisées d'évaluation quantitative de leurs performances, mais elles ne reflètent guère le confort de l'utilisateur : il faudrait par exemple compter le nombre de phrases traitées sans erreurs plutôt que le nombre de mots... ou de symboles phonétiques, car pour l'utilisateur une erreur gêne la compréhension de toute une phrase.

Dans la pratique, la plupart des équipes de recherche ou de développement engagées dans le domaine relativement compétitif de l'étiquetage lexical effectuent des évaluations quantitatives de leurs résultats à partir de comptages d'étiquettes, mais les valeurs obtenues ne permettent pas de comparaisons valables, sauf entre systèmes techniquement très proches.

Le résultat de l'étiquetage lexical peut se présenter sous deux formes : avec ou sans la contrainte de ne présenter qu'une étiquette par mot. Dans les systèmes qui obéissent à cette contrainte, on atteint des résultats tels qu'un taux de silence de 3,4 % et un taux de bruit de 3,4 % également, avec un jeu de 36 étiquettes pour l'anglais et sans tenir compte des mots composés (BRILL, 1995). Dans les systèmes qui n'obéissent pas à cette contrainte, on observe des résultats tels que les suivants :

- un taux de silence de 1,0 % et un bruit de 30,7 %, avec un jeu de 36 étiquettes pour l'anglais et sans tenir compte des mots composés (BRILL, 1995),

– un taux de silence de 2,0 % et un bruit de 50 %, avec un jeu de 1.000 étiquettes pour le français (LAPORTE, 1996).

Un autre critère de qualité important est la facilité d'amélioration du système, que ce soit par l'utilisateur ou par un autre acteur chargé de la maintenance. Dans le cas des noms propres, qui constituent un stock linguistique à renouvellement fréquent, cette souplesse est d'autant plus nécessaire, mais elle est bien difficile à obtenir. Dans tous les systèmes existants, les améliorations font intervenir des opérations assez lourdes.

### ***1.3.2. Corpus de textes étiquetés***

Un corpus, ou collection de textes, peut être vu comme un échantillon d'une langue et c'est à ce titre que les corpus sont utilisés pour le traitement automatique des langues naturelles. Plus le corpus est étendu et varié, plus l'échantillon est représentatif. On peut parler de corpus écrits ou oraux mais nous nous intéresserons ici aux corpus écrits. Les corpus sont en quelque sorte complémentaires ou duaux par rapport aux descriptions formelles des langues : dictionnaires et grammaires. D'un côté, les corpus ont l'avantage d'être constitués d'usages réels ; de l'autre, les dictionnaires et les grammaires peuvent couvrir un vocabulaire et des phénomènes variés aux prix d'un encombrement beaucoup plus réduit. Autre différence fondamentale, un corpus de textes n'a pas réellement d'auteur bien défini, alors que le contenu des dictionnaires et grammaires dépend au plus haut point de leur auteur humain. Ainsi, le recours exclusif à des corpus de textes comme source de données sur les langues naturelles est parfois considéré comme un gage d'objectivité et d'indépendance par rapport au coûteux travail humain. Cette stratégie a cependant des limites intrinsèques : il est impossible de réunir un corpus exhaustif des usages attestés et potentiels d'une langue. En revanche, l'utilisation de corpus de textes comme source d'informations apporte une aide irremplaçable à la construction de dictionnaires et de grammaires.

Dans un corpus de textes, chaque mot peut être étiqueté d'informations grammaticales et morphologiques, en vue d'études et d'analyses ultérieures. Plusieurs corpus étiquetés connus sont très utilisés pour l'anglais (GARSIDE, 1987 ; MARSHALL, 1983 ; MARCUS, 1993). En ce qui concerne l'évaluation du contenu des corpus étiquetés, les jeux d'étiquettes sont concernés au premier chef (cf. 1.3.1). L'étiquetage de ce type de corpus ignore, pour l'essentiel, la notion de mot composé : chacun des éléments est étiqueté en tant que mot simple. En ce qui concerne la forme de l'étiquetage, le projet européen MULTEXT représente un effort de normalisation compatible avec la Text encoding initiative (TEI) (IDE, 1995).

On peut rattacher aux corpus étiquetés les corpus parallèles bilingues tels que le corpus Hansard (GALE, 1991) qui enregistre parallèlement en français et en anglais les débats du Parlement canadien.

### 1.3.3. Dictionnaires

Il ne sera pas question ici des dictionnaires au sens courant du terme, c'est-à-dire des dictionnaires dits éditoriaux ou conventionnels, mais des dictionnaires électroniques conçus pour servir de données dans des programmes de traitement des langues naturelles. Cette distinction ne se situe d'ailleurs pas au niveau du support matériel, puisque les dictionnaires éditoriaux peuvent se présenter sur papier comme sur support électronique, mais du contenu. La rédaction d'un dictionnaire éditorial s'adresse à un lecteur humain : elle fait appel à son intelligence et à son intuition pour rétablir l'information implicite par analogie, par application de règles générales qu'il est inutile de formuler précisément, par suggestion à partir d'exemples, ou tout simplement à partir de sa connaissance préalable de la langue. Au contraire, le contenu d'un dictionnaire électronique est destiné à l'exploitation informatique directe, et n'est constitué que d'informations codées et explicites ; les exemples éventuels sont à usage interne et ont un statut comparable à celui des commentaires dans le code source des programmes.

#### *Mots simples*

Les différents dictionnaires électroniques de mots simples peuvent être classés en fonction des informations linguistiques associées aux entrées : morphologiques, syntaxiques, sémantiques. Dans chaque catégorie, les principaux critères d'évaluation sont, d'une part, la couverture, c'est-à-dire le nombre d'entrées, et d'autre part, l'exactitude et la précision des informations linguistiques. Ce dernier critère est lié à la quantité et à la qualité des distinctions d'emplois, c'est-à-dire des distinctions entre entrées homographes. En effet, dans un dictionnaire formel, un mot ambigu est séparé en plusieurs emplois lorsque des informations linguistiques distinctes peuvent être associées aux emplois. Par exemple, au niveau morphologique, trois noms *débiteur* peuvent être distingués par leur féminin :

Mot	Code	Commentaire
<i>débiteur</i>	N1	pas de féminin
<i>débiteur</i>	N35	féminin : <i>débiteuse</i>
<i>débiteur</i>	N36	féminin : <i>débitrice</i>

**Tableau 1.1.** Séparation morphologique des emplois

Les codes du tableau 1.1 sont empruntés au DELAS (COURTOIS, 1990). Au niveau syntaxique, deux verbes *perquisitionner* peuvent être distingués par la forme de leur complément :

Mot	Code	Commentaire
<i>perquisitionner</i>	32R2	<i>La police a perquisitionné la maison</i>
<i>perquisitionner</i>	35L	<i>La police a perquisitionné dans la maison</i>

**Tableau 1.2.** *Séparation syntaxique des emplois*

Ces codes empruntés au lexique-grammaire du LADL (GROSS, 1975). Au niveau sémantique, deux noms *bélier* peuvent être distingués dans le cadre d'une classification des concrets (GROSS, 1994) :

Mot	Classifieur	Commentaire
<i>bélier</i>	animal	<i>Ses cornes ressemblent à celles d'un bélier</i>
<i>bélier</i>	outil	<i>Ils ont forcé l'entrée avec un bélier</i>

**Tableau 1.3.** *Séparation sémantique des emplois*

Les ambiguïtés étant réparties à tous les niveaux dans les langues du monde, la précision et l'exactitude du codage lexical se reflètent inmanquablement dans des distinctions d'emplois. Les données quantitatives liées aux distinctions d'emplois sont donc un des moyens d'évaluer la précision des informations fournies par un dictionnaire électronique.

Ces quelques remarques tout à fait générales sur les dictionnaires électroniques dessinent un cadre de raisonnement commun, indépendant du niveau d'analyse linguistique dont relève le contenu des dictionnaires. Cette généralité est propice à la réflexion sur les formats d'échange de dictionnaires électroniques : elle a été à la base des modèles élaborés par le projet GENELEX.

En ce qui concerne les informations grammaticales, morphologiques et phonétiques, les dictionnaires du LADL (COURTOIS, 1990) décrivent de 50 000 à 90 000 mots suivant les langues, sans compter les distinctions d'emplois ni les formes fléchies, ce qui représente une couverture raisonnable du vocabulaire. Ils incluent les informations permettant la flexion, avec une précision supérieure à celle observée dans les dictionnaires éditoriaux.

Au niveau syntaxique, le lexique-grammaire des verbes simples du français (GROSS, 1975) possède une couverture correcte du lexique des verbes (12 000 emplois), et code les principales propriétés distributionnelles et transformationnelles. Malgré la tendance récente et actuelle à la lexicalisation en syntaxe, il ne semble pas exister d'autres dictionnaires syntaxiques comparables.

Certains dictionnaires fondés sur une classification sémantique associent à chaque mot décrit un hyperonyme (mot plus général) et éventuellement d'autres relations sémantiques, par exemple la synonymie ou des relations plus vagues d'association, et sont qualifiés de thésaurus. Les thésaurus sont couramment utilisés dans des applications commerciales. Des dictionnaires sémantiques d'une couverture respectable sont en cours de construction par des méthodes artisanales (MEL'CUK, 1984-88-92 ; GROSS, 1994). Il existe aussi des dictionnaires sémantiques et thésaurus, de l'ordre de 50 000 mots, construits automatiquement à partir des nomenclatures et des définitions des dictionnaires éditoriaux (EVENS, 1988). Des procédés statistiques appliqués à des thésaurus ont également permis de produire des classes d'entrées sémantiques (GREFENSTETTE, 1994).

Un type d'information original est le degré de plausibilité d'emploi. La plausibilité d'emploi des mots est classiquement évaluée automatiquement par des calculs de fréquences dans un corpus ; la difficulté intrinsèque à cette méthode est que les résultats obtenus dépendent du corpus utilisé, aussi grand soit celui-ci, et ne prédisent pas les fréquences d'emploi dans un autre texte. En remplaçant ces calculs objectifs par des méthodes artisanales, on a pu obtenir des valeurs qui prédisent mieux la fréquence d'emploi dans des corpus nouveaux (GARRIGUES, 1992).

#### *Mots composés*

Les mots composés et expressions figées, c'est-à-dire les unités lexicales composées de plusieurs mots, ont fait l'objet de recensements dont les plus anciens sont dus à des auteurs de dictionnaires destinés à l'apprentissage des langues, surtout l'anglais. Dans le traitement automatique des langues, on a pu réutiliser de tels recensements (MILLER, 1990), mais on considère traditionnellement que les mots composés ont un caractère "exceptionnel" dans les langues, et donc ne revêtent qu'une importance marginale. Le choix de termes comme *idiom* ou *formes idiomatiques* reflète cette conception, par exemple chez W. Plath qui en préconisait cependant l'étude dès 1974. Or, la construction de dictionnaires morphologiques et syntaxiques de mots composés offrant une couverture non négligeable de la langue générale (GROSS, 1986a) a permis de démontrer l'importance numérique, d'une part, des entrées composées dans le lexique des langues, et d'autre part, des occurrences de mots composés dans les textes. Dans ce contexte, ce résultat prend le caractère d'une découverte scientifique. Nous avons d'ailleurs déjà évoqué, dans ce qui précède, plusieurs problèmes liés aux mots composés, car de tels problèmes

interviennent quasiment à tous les niveaux de traitement des textes en langues naturelles.

Les termes techniques constituent une portion stratégiquement importante du lexique des composés, en raison de leur contenu informatif élevé et de leur utilisation relativement normalisée. Des informations formelles peuvent figurer dans des dictionnaires terminologiques : domaine, informations grammaticales et morphologiques, traduction. Étant donné l'extension rapide du vocabulaire terminologique, les dictionnaires terminologiques formalisés n'offrent malheureusement pas une couverture suffisante des domaines techniques pour en permettre l'exploitation dans des applications telles que la documentation automatique.

La notion de collocation, ou rapprochement de mots simples statistiquement significatif par sa fréquence dans des textes, est corrélée à la notion de mot composé, mais distincte. Certaines collocations sont des mots composés fréquents, par exemple *sur le fond* ; d'autres correspondent à des structures libres fréquentes, comme *soumettre au vote*, ou à des associations syntaxiques fréquentes entre un mot plein et un ou plusieurs mots grammaticaux, comme *taux de* ou *avoir la responsabilité* (GROSS, 1996). Les algorithmes d'alignement de corpus bilingues parallèles permettent de constituer automatiquement des dictionnaires de collocations bilingues (SMADJA, 1996) et fournissent une aide à la construction de dictionnaires terminologiques bilingues (GAUSSIER, 1995).

#### ***1.3.4. Autres données utilisées pour l'étiquetage lexical***

Nous rangeons sous ce titre les automates et transducteurs finis qui effectuent des opérations liées à l'étiquetage lexical (KOSKENNIEMI, 1990 ; OFLAZER, 1996). Le principal avantage de ce type de données est que la souplesse inhérente aux automates finis permet de passer aisément d'une forme descriptive, compacte et lisible, à une forme opérationnelle qui permet un traitement efficace : le temps de traitement peut souvent être rendu linéaire par rapport à la taille du texte et indépendant de la taille de l'automate. Parallèlement à ces données écrites à la main, certains systèmes de règles (BRILL, 1995) et automates finis (ROCHE, 1995) pour l'analyse lexicale sont acquis initialement par apprentissage statistique, mais en principe exploitables et modifiables en dehors de l'outil qui les a créés, alors que les données issues de l'apprentissage statistique ne sont pas toujours réutilisables sans le réseau neuronal ou le système d'apprentissage qui les produit.

#### ***1.3.5. Traitement des mots inconnus***

Dans un système à dictionnaire, le traitement des mots inconnus n'est autre que le problème de la maintenance. L'exploitation d'une application entraîne la détection de mots absents du dictionnaire, mais ils doivent être codés avant d'être intégrés. Il est parfois possible d'anticiper cette détection grâce à une aide automatique au codage de nouveaux mots. Dans le cas des mots dérivés, qui constituent une proportion significative des mots absents des dictionnaires électroniques, des connaissances linguistiques sur les dérivations vivantes dans la langue permettent de mettre sur pied de telles aides (CLEMENCEAU, 1996). En-dehors de ce cas, c'est-à-dire essentiellement pour les mots composés et pour les noms propres, leur forme est imprévisible et leurs propriétés n'ont guère de raisons de se déduire de leur forme. Ainsi, la maintenance d'un dictionnaire électronique constitue une tâche sensiblement distincte de celle d'un dictionnaire éditorial. D'une part, elle doit être intégrée à l'exploitation de l'application informatique, car la détection et l'intégration des mots absents ne doivent pas être trop dissociées dans le temps. D'autre part, elle nécessite des compétences qui ne sont généralement pas celles des utilisateurs professionnels de l'application. Cette activité est encore peu représentée dans le monde professionnel.

Dans un système à corpus étiqueté, le problème des mots inconnus a été abondamment étudié. On tire parti de la morphologie interne du mot et de son contexte pour deviner son étiquette par des calculs probabilistes. Quelques résultats typiques : un système utilisant un jeu de 36 étiquettes pour l'anglais (BRILL, 1995) étiquette correctement 82 % des mots inconnus ; un système utilisant un jeu de 10 étiquettes pour l'anglais (DERMATAS, 1995) étiquette correctement 65 % des mots inconnus.

Un problème voisin, dans le cadre d'un système à corpus étiqueté, est celui des mots ambigus qui apparaissent dans le corpus, mais seulement avec une des étiquettes a priori possibles, par exemple les hapax ambigus (BAAYEN, 1996). Ce ne sont pas à proprement parler des mots inconnus, mais l'unique étiquette qui leur est attribuée dans le corpus n'est d'aucun secours pour lever leur ambiguïté. Le traitement de ces formes en est à ses balbutiements.

### ***1.3.6. Phonétisation***

Les méthodes de phonétisation font intervenir soit un dictionnaire phonétique, soit des règles contextuelles, soit un réseau connexionniste (SEJNOWSKI, 1987) ou une autre forme d'apprentissage automatique. Le lien avec l'étiquetage lexical est évident dans le premier cas, puisqu'on peut voir la phonétisation comme un étiquetage par des étiquettes constituées de formes phonétiques. Dans les deux autres cas, un étiquetage lexical améliore les performances du phonétiseur s'il permet, par exemple, de distinguer les homographes non homophones tels que

*poster* verbe et *poster* nom, ou de placer les liaisons. Les progrès futurs sont dépendants de l'étiquetage lexical. La distinction entre formes phonétiques, proches du son observable, et formes phonologiques, qui permettent le regroupement des variantes phonétiques, est parfois respectée. Cette distinction rend le modèle du problème plus rigoureux en séparant les deux difficultés de la phonétisation :

- le changement de système de notation, puisqu'on passe de l'orthographe à la représentation phonétique ;
- et la prédiction des variations phonétiques.

En ce qui concerne la phonétisation par règles contextuelles, le jeu de règles est généralement équivalent à une transduction rationnelle ; plusieurs auteurs préconisent l'utilisation indirecte (KOSKENNIEMI, 1983) ou directe (KAPLAN, 1994 ; LAPORTE, 1997) de transducteurs finis dans la mise en œuvre de ces règles.

Les évaluations publiées par les auteurs sont classiquement fondées sur le décompte des erreurs par symbole phonétique ou par mot. Comme nous l'avons déjà noté, cela ne donne pas une idée fidèle des performances des applications ni du confort de leur utilisateur, car une erreur gêne la compréhension de toute une phrase pour l'utilisateur. Les taux d'erreur publiés sont de l'ordre de 5 à 8 % de symboles phonétiques erronés pour le français (cf. BARTKOVA, 1994), la correspondance entre l'orthographe et la transcription phonétique étant particulièrement irrégulière en français. L'utilisation d'un dictionnaire phonémique permet d'obtenir une phonétisation avec 0,6 % de mots erronés. Cet écart est moins important dans le cas d'autres langues que le français et l'anglais.

Un autre critère d'évaluation, important pour les applications, est souvent négligé : la granularité du système de représentation phonétique. Un alphabet de granularité fine permet de formuler des transcriptions plus précises, alors qu'un alphabet de granularité plus grossière convient pour des transcriptions plus approximatives : ainsi, le voisement ou le non-voisement des *r* et des *l* peut être pris en compte ou non.

### ***1.3.7. Détection et correction d'erreurs***

Dans ce domaine on note plusieurs acquis méthodologiques. Tout d'abord, on distingue détection et correction d'erreurs, car la plupart des techniques de détection sont tout à fait indépendantes des techniques de correction. Seules certaines techniques aboutissent simultanément à la détection et à la correction.

En ce qui concerne la détection, on distingue les erreurs lexicales, c'est-à-dire productrices d'un mot qui ne fait pas partie du vocabulaire, par exemple *sugner* pour *signer*, et les erreurs non lexicales, dont la détection met nécessairement en jeu le

contexte : par exemple, *singer* pour *signer*. Dans les deux cas, la détection est souvent étroitement liée à un processus d'étiquetage lexical, et c'est l'échec de l'étiquetage d'un mot qui permet de repérer une erreur. Par exemple, au cours d'un étiquetage par dictionnaire, les erreurs lexicales sont détectées, car les mots correspondants ne sont pas trouvés dans le dictionnaire. De même, certaines procédures de levée d'ambiguïtés lexicales fournissent automatiquement une détection d'erreurs. Il s'agit des procédures bâties sur le principe de supprimer des hypothèses correspondant à des séquences grammaticales incorrectes, par exemple *le* (déterminant) *véhicule* (verbe). Lorsqu'une telle procédure en arrive à éliminer toutes les étiquettes grammaticales d'un mot, c'est que le texte n'admet aucune analyse correcte, et donc qu'il est erroné.

En ce qui concerne la correction, la typologie des causes d'erreurs orthographiques est variée. Les techniques mises en œuvre correspondent aux types d'erreurs : par proximité de chaînes, par phonétisation, par analyse syntaxique... ou des techniques correspondant à des modèles d'erreur plus spécifiques. Un système d'étiquetage lexical tolérant aux erreurs ne peut pas détecter certaines erreurs, mais il peut les corriger, car il peut associer à une forme erronée une forme correcte ressemblante (OFLAZER, 1996).

L'évaluation des détecteurs et correcteurs orthographiques dépend bien sûr de leurs performances. Les détecteurs orthographiques commerciaux pour le français distribués au grand public se situent, par leurs performances, notamment en deçà des possibilités techniques de détection d'erreurs lexicales, en tout cas pour le français, alors qu'il est aisé de détecter ces erreurs automatiquement avec un simple dictionnaire de formes fléchies. On note même couramment des erreurs lexicales comme la suivante, épinglée dans un quotidien du soir par un hebdomadaire satirique :

*(...) on attendait, qu'on le rêvasse ou qu'on le craignasse, l'avènement de la reconnaissance de la parole et la synthèse vocale (...)*

Quant aux détecteurs et correcteurs d'erreurs non lexicales, ils n'ont pas acquis une réputation de fiabilité suffisante auprès de leurs utilisateurs, et on peut parier que les progrès futurs seront tributaires de l'avancement des techniques d'étiquetage lexical et d'analyse syntaxique.

Ici encore, la facilité d'amélioration du système, en d'autres termes la prise en compte de nouvelles erreurs par rapport à celles déjà traitées, est un critère d'évaluation important. Comme nous l'avons dit à propos de l'étiquetage lexical, les noms propres posent des problèmes particulièrement préoccupants de ce point de vue.

### ***1.3.8. Accentuation automatique***

Un dictionnaire inverse, qu'il est facile de construire à partir d'un dictionnaire de formes fléchies, peut donner pour chaque forme non accentuée la liste des formes correctes correspondantes. Le problème se ramène alors à un choix entre ces formes dans le cas où il en existe plusieurs. L'ambiguïté spécifique introduite par l'absence d'accentuation affecte de l'ordre de 25 % des occurrences de mots dans un texte en français. Ce problème est donc très lié à la levée d'ambiguïtés lexicales. La solution peut s'appuyer sur un étiquetage lexical des formes candidates et des mots voisins.

De bons résultats ont été obtenus à l'aide d'un modèle probabiliste obtenu par apprentissage automatique dans un corpus (EL-BEZE, 1994). Ce système introduit les accents de manière correcte dans 97,6 % des occurrences de mots ambigus, soit 99,4 % des occurrences de mots, ou encore 87,2 % des phrases si l'on compte environ 20 mots par phrase.

### ***1.3.9. Documentation, indexation et moteurs de recherche***

Les langages d'indexation les plus formalisés fixent un lexique de descripteurs spécifique et une syntaxe de combinaison des descripteurs, par exemple une syntaxe booléenne. Ils nécessitent des opérations manuelles ou semi-automatiques d'indexation des documents et sont adaptés à l'environnement commercial des banques de données documentaires, où des documentalistes professionnels tiennent à jour la base en indexant les documents et formulent les requêtes pour les utilisateurs (LEWIS, 1996).

Dans le cas de figure diamétralement opposé, le langage d'indexation est libre. L'utilisateur est libre dans sa manière de formuler sa requête, tous les mots de la requête ou du document peuvent servir de descripteurs, et la syntaxe de combinaison des descripteurs se limite à la notion de liste. Beaucoup de banques de données commerciales offrent maintenant les deux types de langage d'indexation, contrôlé et libre. Le deuxième est incontournable dans les situations où il n'est pas question de tenir à jour manuellement l'indexation du stock de documents. C'est notamment le cas des moteurs de recherche dans le réseau informatique mondial. Un des obstacles au repérage de descripteurs dans les documents est l'ambiguïté lexicale des mots : la solution à ce problème passe par la levée des ambiguïtés lexicales, ce qui fait apparaître la gestion documentaire comme une application supplémentaire de l'étiquetage lexical.

L'évaluation d'un outil de documentation met en jeu l'estimation du taux de bruit (proportion de documents non pertinents parmi les documents présentés), ou de son taux complémentaire la précision, et du taux de silence (proportion de documents

non présentés parmi les documents pertinents), ou de son taux complémentaire le rappel.

Dans le cas d'un langage d'indexation contrôlé avec indexation des documents par des documentalistes, la qualité du service apporté à l'utilisateur dépend au premier chef de la qualité de l'indexation et de la formulation des requêtes.

Il existe des méthodes statistiques de documentation automatique avec langage d'indexation libre. Les descripteurs d'un document sont tirés soit du texte intégral, soit du titre et d'un résumé. Chaque descripteur est pondéré par une estimation de sa pertinence calculée à partir de statistiques sur les documents. La similarité entre la requête et le document est également estimée par un calcul statistique. Ces méthodes ont été testées à une échelle qui va jusqu'à quelques dizaines de milliers de documents. Les taux de bruit et de silence sont estimés de 40 à 70 % (SALTON, 1986).

#### ***1.3.10. Division***

Les méthodes de division automatique se fondent sur la reconnaissance de motifs dans les mots, lorsque cette division obéit à des règles suffisamment générales. Les mots qui font exception à ces règles peuvent être recensés dans un dictionnaire d'exceptions. Les positions dans le mot où la division est possible constituent alors une information lexicale sur le mot. Lorsque ces mots exceptionnels sont nombreux, par exemple en anglais, ou dans les applications qui mettent en jeu une importante proportion de noms propres, le problème de la division se rapproche de celui de l'étiquetage lexical, et les performances dépendent de la couverture lexicale du dictionnaire de mots exceptionnels.

Les outils de division automatique utilisés dans l'édition offrent des performances satisfaisantes, hormis en ce qui concerne les mots qui font exception aux règles et qui apparaissent de façon imprévisible, c'est-à-dire, en pratique, une certaine proportion de noms propres d'origine étrangère. On ne peut guère prévoir d'améliorations que par l'organisation d'une maintenance efficace du dictionnaire d'exceptions.

#### ***1.3.11. Conclusions***

Sans prétendre à l'exhaustivité, nous avons passé en revue une certaine variété de traitements automatiques qui peuvent être vus comme des traitements sur les mots. Les résultats appellent une conclusion mitigée. Nous n'en avons trouvé aucun pour lequel les performances des systèmes existants correspondent indiscutablement aux

attentes des industriels, sauf peut-être la division des mots en fin de ligne. Les données linguistiques formelles qui ont été élaborées jusqu'à présent sont loin de suffire pour la réalisation de toutes les applications. Cependant, elles forment un échantillon substantiel des données nécessaires à ces applications. De plus, la réutilisabilité de ces données est désormais perçue comme une notion cruciale. Enfin, dans plusieurs applications, les performances des systèmes commercialisés sont inférieures à celles qui ont été démontrées comme faisables dans le monde de la recherche, ce qui laisse présager des progrès techniques dans l'avenir.

#### **1.4. Directions de recherche actuelles**

Nous avons évoqué brièvement, dans ce qui précède, une variété d'outils : techniques, méthodes, données, qui interviennent au niveau lexical dans le traitement des langues naturelles. On est frappé par la diversité de ces outils et par le fait qu'ils se rattachent à plusieurs cadres méthodologiques. En fait, pour un objectif applicatif commun, ce sont des problèmes fondamentalement différents qui sont posés et résolus, ou partiellement résolus. Ainsi, dans le cas de l'étiquetage lexical, il s'agit pour les uns de décrire formellement le vocabulaire d'une langue, puis d'accéder aux résultats de cette description, tandis que d'autres se donnent pour objectif de spécifier le résultat voulu sur un échantillon, puis de simuler cette performance sur de nouveaux textes. Les problèmes posés n'ayant rien de commun, il est naturel que les techniques employées soient tout à fait différentes. Cette situation s'explique d'ailleurs en partie par la pluridisciplinarité inhérente au traitement des langues naturelles, qui a des liens évidents avec l'informatique et la linguistique, et aussi des prolongements applicatifs industriels.

Mais si l'on se tourne vers les perspectives et les points de recherche actuels, la question centrale qui se pose est celle des potentialités à plus long terme des différents problèmes posés et des cadres méthodologiques correspondants. Répondre à cette question est nécessaire pour la maturation du domaine, car elle figure au premier rang parmi les motivations de la stratégie des équipes pour orienter la recherche dans une voie rentable à long terme en termes scientifiques et applicatifs. C'est en partie possible aujourd'hui grâce à un recul relatif, mais la question n'est pas facile pour autant. L'évaluation des systèmes actuels, telle qu'elle est pratiquée, relève souvent de techniques d'ingénieur et n'est pas conçue pour refléter les potentialités des méthodes qui ont servi à leur construction : un intéressant article paru dans une revue de traitement de la parole (BOURLARD, 1996) relève à quel point cet écart entre évaluation et potentialités est parfois flagrant. L'évaluation des potentialités à long terme des méthodes est polémique, difficile à justifier, non mesurable quantitativement, et, dans notre domaine, remarquablement absente des discussions scientifiques.

Il n'est donc pas inutile de s'interroger sur les principaux cadres méthodologiques dans lesquels se situent les chercheurs, en vue d'en comprendre les motivations et d'en évaluer les perspectives de développement. Dans cette section, nous présentons une synthèse personnelle sur ce sujet, en essayant de caricaturer le moins possible la pensée de nos confrères. Nous distinguerons trois cadres méthodologiques : ceux qui privilégient la théorie, la description empirique, ou l'optimisation.

#### ***1.4.1. Méthodes à dominante théorique***

On peut caractériser les travaux à dominante théorique comme ceux qui ont pour objectif la découverte de faits aussi généraux et profonds que possible. Comme exemple de résultat de tels travaux, citons DATR (EVANS, 1996), un langage de représentation des données lexicales, implémenté en Prolog et en Lisp, indépendant de la langue, conçu pour tous les niveaux de l'analyse linguistique, et entièrement structuré sur les notions de règle et d'exception, elles aussi tout à fait générales. HPSG serait un autre exemple.

Les thèmes développés dans ce cadre méthodologique sont abstraits. Ils incluent un principe explicatif universel applicable aux langues du monde et impliquent l'élimination de la redondance de la représentation formelle des faits linguistiques. Dans le monde de la linguistique, ces thèmes rejoignent ceux que privilégient les écoles générativistes.

L'idée centrale en faveur des travaux à dominante théorique est que la découverte d'un fait général réduit à néant l'intérêt des faits isolés préalablement décrits, de même que la démonstration d'un théorème mathématique dépasse par sa portée l'observation de cas particuliers de ce théorème. Le but recherché est de fonder sur de tels faits la construction de formalismes universels de dépôt de données linguistiques.

Les principales réserves qu'on peut faire à propos de ce cadre méthodologique sont liées à son abstraction. La manipulation et la comparaison de modèles abstraits et généraux ressemblent parfois à l'étude de variantes de notation, dont on peut concevoir une variété potentiellement infinie. Les concepteurs de ces modèles abstraits laissent d'ailleurs à la charge des utilisateurs le travail de formaliser effectivement des données linguistiques dans leurs modèles. Ces études débouchent souvent sur des systèmes à petite échelle, alors que la plupart des applications réelles nécessitent l'élaboration de systèmes à grande échelle.

#### ***1.4.2. Travaux descriptifs empiriques***

Il s'agit des travaux qui privilégient la description formelle et explicite des phénomènes linguistiques à travers l'observation empirique. Par exemple, le générateur automatique de textes de L. Danlos (DANLOS, 1985) se fonde sur des faits linguistiques précis, obtenus essentiellement par l'observation de jugements d'acceptabilité. Ce cadre méthodologique, illustré tôt dans l'histoire du domaine par l'école de M. Gross et les travaux de N. Sager (SAGER, 1973), est centré sur la surface directement observable et sur le thème de la reproductibilité des observations. Dans le monde de la linguistique, on peut rattacher ce type d'objectifs et de méthodes à ceux de l'école harrissienne.

L'avantage de ce type de travaux est que le descripteur est libre d'adapter le système formel de description aux informations linguistiques : la précision et l'exactitude des informations recueillies n'ont donc d'autres limites que les capacités d'introspection de son auteur. Les dictionnaires électroniques les plus étendus en nombre d'entrées et les plus précis en informations ont d'ailleurs été construits de cette façon.

Le principal inconvénient de ce type de travaux est leur coût : ils ont en effet un caractère intrinsèquement artisanal et même rebelle à une industrialisation complète, au même degré, par exemple, que la description des espèces animales et végétales par l'observation. Étant donné ce coût, la prise de conscience de l'existence d'un "goulot d'étranglement" dans l'acquisition des connaissances linguistiques a eu pour conséquence l'apparition de méthodes plus industrielles, dans lesquelles l'exigence de la reproductibilité des observations disparaît au profit du thème de l'objectivité.

#### ***1.4.3. Méthodes privilégiant l'optimisation***

Ces méthodes font reposer l'optimisation des systèmes sur des techniques relevant du génie mathématique : statistiques, traitement du signal... ou de l'intelligence artificielle : logique floue..., ou des deux ensemble : réseaux connexionnistes. Le système d'accentuation automatique déjà cité (EL-BEZE, 1994) en est un exemple. Ce type de méthodologie présente un caractère empirique certain, mais d'une autre façon que le précédent : la pratique expérimentale repose sur un modèle ou des hypothèses (MERIALDO, 1995) qu'il ne s'agit pas de vérifier, car ils sont volontairement simplificateurs ; c'est l'exploitabilité de ces hypothèses pour une application spécifique qui est confirmée ou infirmée a posteriori par les performances obtenues. Par ailleurs, l'apprentissage automatique se situe relativement à l'écart de la linguistique et séduit plutôt les informaticiens purs. En effet, les systèmes d'intelligence artificielle reposent par définition sur des modèles symboliques ou numériques de réalités floues, généralement extra-linguistiques. Quant aux traitements statistiques, les modèles sur lesquels ils reposent comportent des hypothèses simplificatrices destinées à réduire la dimension de l'espace de calcul

en limitant le nombre de paramètres, par exemple en diminuant la taille de l'alphabet phonétique dans un phonétiseur, la taille du jeu d'étiquettes grammaticales ou la teneur du contexte grammatical dans un système de levée d'ambiguïtés lexicales, ou en faisant abstraction des mots composés pour ne considérer que les mots simples. Certaines expériences ont même été menées avec un modèle qui ne comporte aucun élément d'analyse linguistique et utilise comme seules données un corpus de textes non étiqueté (MACMAHON, 1996).

Ce choix a l'avantage d'aboutir à des traitements entièrement automatiques, donc moins coûteux, et de faire table rase d'une bonne partie des imperfections de l'analyse linguistique transmises par la tradition, "*pour produire des [données] plus neutres, plus objectives*" (GUTHRIE, 1996). Après tout, depuis le XIX<sup>e</sup> siècle, la priorité accordée à la notion d'objectivité et le rejet des superstitions ont permis la fondation des sciences expérimentales.

Mais ce choix comporte le risque de jeter le bébé avec l'eau du bain : le langage étant intrinsèquement interne à l'être humain, le rejet des méthodes subjectives comme moyen d'étude est nécessairement appauvrissant. Quoi qu'il en soit, il semble clair que dans des traitements de ce type, les chances de succès tiennent en grande partie à la qualité du modèle sous-jacent. Les critiques adressées aux systèmes probabilistes, par exemple, mettent souvent en cause la simplicité du modèle du problème.

#### **1.4.4. Méthodes mixtes**

Elles se développent depuis quelques années, surtout par combinaison entre des méthodes relevant de la description empirique et des méthodes privilégiant l'optimisation (BRILL, 1995 ; EL-BEZE, 1995). Une des articulations possibles entre ces deux cadres méthodologiques est la suivante : les résultats de travaux de description empirique sont utilisés comme modèle du problème pour un traitement probabiliste.

### **1.5. Conclusions**

Pour la plupart des grands types d'applications que nous avons passés en revue, l'évolution du domaine est favorable. Un acquis s'est constitué. Il regroupe des données et des méthodes sur lesquelles la communauté scientifique peut s'appuyer pour construire des applications. Certes, les performances des systèmes existants ne correspondent pas encore suffisamment aux attentes des industriels. Cependant, il serait simpliste de croire que ces performances nous permettent de juger directement des potentialités des méthodes qui ont servi à les construire.

Sur les cinquante années qui nous séparent de la naissance du domaine, son évolution peut être caractérisée par trois éléments, qui donnent l'impression d'une maturation progressive et d'un professionnalisme croissant.

D'une part, le volume des données traitées, et donc l'ambition des traitements, ont changé d'échelle.

D'autre part, on note une structuration progressive en sous-domaines : certains de ces sous-domaines ont désormais le statut de fournisseurs par rapport à des clients ou de briques de base par rapport à des produits intégrés ; des problèmes à l'origine considérés comme des tâches qui ne nécessitent guère d'investissement intellectuel, sont maintenant reconnus comme des problèmes à part entière méritant une solution spécifique. À cet égard, l'histoire de la phonétisation de textes est éloquent. Les systèmes réalisés il y a vingt ans tentaient d'apporter des solutions, nécessairement approchées, à des problèmes aussi difficiles et hétérogènes que l'étiquetage lexical, l'analyse syntaxique, l'application de règles de réécriture à des séquences de symboles... Chacun de ces problèmes a désormais acquis son autonomie, et les chercheurs n'en abordent qu'un à la fois, à moins qu'ils ne tentent un travail d'intégration de techniques existantes.

Enfin, le caractère pluridisciplinaire du domaine s'est affirmé. Il amène à collaborer entre eux des spécialistes de disciplines considérées comme bien éloignées les unes des autres, et traditionnellement rattachées à des modes de pensée qui ont tendance à s'exclure : sciences expérimentales, techniques d'ingénieur, humanités. On reconnaît maintenant volontiers que chacun a son rôle, même si les acteurs ne sont pas en mesure de s'accorder sur le rôle de chacun. C'est probablement dans cet aspect que le domaine est le plus en devenir.

## Références

- [ADDA 1997] ADDA Gilles, LECOMTE Josette, MARIANI Joseph, PAROUBEK Patrick, RAJMAN Martin. "Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français", *Actes des 1es Journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF*, Avignon.
- [BAAYEN 1996] BAAYEN Harald, SPROAT Richard. "Estimating lexical priors for low-frequency morphologically ambiguous forms", *Computational Linguistics* 22(2), pp. 155–166.
- [BARTKOVA 1994] BARTKOVA K., LARREUR D., METAYER I. "Choix et adaptation d'un phonétiseur pour la reconnaissance automatique de la parole", *20es Journées d'étude sur la parole, Trégastel (France)*, pp. 181-186.
- [BOITET 1982] BOITET Christian, NEDOBEJKINE N.. "Base lexicale : organisation générale et indexage", *Projet ESOPE, ADI, Rapport final*, partie D.
- [BOURLARD 1996] BOURLARD Hervé, HERMANSKY Hynek, MORGAN Nelson. "Towards increasing speech recognition error rates", *Speech communication* 18, Elsevier, pp. 205-231.
- [BRILL 1995] BRILL Eric. "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging", *Computational Linguistics* 21(4), pp. 543–565.
- [CECCATO 1962] CECCATO S. "Automatic translation of languages", *Automatic Translation of Languages*, 1966, London: Pergamon.
- [CHURCH 1988] CHURCH Kenneth. "A stochastic parts program and noun phrase parser for unrestricted text", *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, ACL.
- [CLEMENCEAU 1996] CLEMENCEAU David. "Finite-state morphology: inflections and derivations in a single framework using dictionaries and rules", *Finite-state language processing*, Cambridge, MA: MIT Press, pp. 67-98.
- [COBUILD 1987] COBUILD *dictionary of the English language* (première édition), Collin's.
- [COURTOIS 1990] COURTOIS Blandine. "Un système de dictionnaires électroniques pour les mots simples du français", *Langue française* 87, Larousse, pp. 11-22.
- [DANLOS 1985] DANLOS Laurence. *Génération automatique de textes en langues naturelles*, Paris : Masson.
- [DERMATAS 1995] DERMATAS Evangelos, KOKKINAKIS George. "Automatic stochastic tagging of natural language texts", *Computational Linguistics* 21(2), pp. 137-163.
- [EL-BEZE 1994] EL-BEZE Marc, MERIALDO Bernard, ROZERON Bénédicte, DEROUAULT Anne-Marie. "Accentuation automatique de textes par des méthodes probabilistes", *Technique et science informatiques* 13(6), pp. 797-815.
- [EL-BEZE 1995] EL-BEZE Marc, SPRIET Thierry. "Intégration de contraintes syntaxiques dans un système d'étiquetage probabiliste", *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 47-66, Paris : ATALA.
- [EVANS 1996] EVANS Roger, GAZDAR Gerald. "DATR: a language for lexical knowledge representation", *Computational Linguistics* 22(2), pp. 167-216.
- [EVENS 1988] EVENS M. (ed.). *Relational models of the lexicon*, Cambridge University Press.

- [GALE 1991] GALE William, CHURCH Kenneth. "A program for aligning sentences in bilingual corpora", *Proceedings of the 29th Annual meeting of the Association for Computational Linguistics*, Berkeley, CA.
- [GARRIGUES 1992] GARRIGUES Mylène. "Dictionnaires hiérarchiques du français. Principes et méthodes d'extraction", *Langue française* 96, Larousse.
- [GARRIGUES 1997] GARRIGUES Mylène. "Lemmatized concordances of complex utterances: application to language learning", *New technologies in language teaching and learning*, A.K. Korsvold, B. Röschoff, eds., Strasbourg : Council of Europe Publishing, pp. 87–98.
- [GARSIDE 1987] GARSIDE Roger, LEECH Geoffrey, SAMPSON Geoffrey. *The computational analysis of English*, Longman.
- [GAUSSIER 1995] GAUSSIER Éric, LANGE Jean-Marc. "Modèles statistiques pour l'extraction de lexiques bilingues", *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 133-155, Paris : ATALA.
- [GREFENSTETTE 1994] GREFENSTETTE Gregory. *Explorations in Automatic Thesaurus Discovery*, Kluwer international series in engineering and computer science, Dordrecht: Kluwer.
- [GROSS 1972] GROSS Maurice. "Notes sur l'histoire de la traduction automatique". *Langages* 28, Paris : Larousse, pp. 40-48.
- [GROSS 1975] GROSS Maurice. *Méthodes en syntaxe*, Paris : Hermann, 414 p.
- [GROSS 1986a] GROSS Maurice. *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*, Paris : ASSTRIL, 670 p.
- [GROSS 1986b] GROSS Maurice. "Lexicon-Grammar. The representation of compound words", *Proceedings of COLING-86*, Bonn, pp. 1-6.
- [GROSS 1994] GROSS Gaston. "Classes d'objets et description des verbes", *Langages* 115, Paris : Larousse, pp. 15-30.
- [GROSS 1996] GROSS Gaston. *Les expressions figées en français. Noms composés et autres locutions*, Paris : Ophrys, 161 p.
- [GUTHRIE 1996] GUTHRIE Louise, PUSTEJOVSKY James, WILKS Yorick, SLATOR Brian M.. The role of lexicons in natural language processing", *Commun. ACM* 39(1), pp. 63-72.
- [IDE 1995] IDE Nancy, VERONIS Jean. *The Text Encoding Initiative: background and context*, Dordrecht: Kluwer.
- [KAPLAN 1994] KAPLAN Ronald M., KAY Martin. "Regular models of phonological rule systems", *Computational Linguistics* 20(3), pp. 331-378.
- [KOSKENNIEMI 1983] KOSKENNIEMI Kimmo. *Two-level morphology: a general computational model for word-form recognition and production*, Publication no. 11, Department of General Linguistics, University of Helsinki.
- [KOSKENNIEMI 1990] KOSKENNIEMI Kimmo. "Finite-state parsing and disambiguation", in *Proceedings of COLING 90*.
- [LAPORTE 1996] LAPORTE Éric, SILBERZTEIN Max. "Ambiguity rates. Automatic analysis of French text corpora and computation of ambiguity rates for different tagsets", *GRAMLEX deliverables, October 1995 - June 1996*, Eric Laporte (ed.), Paris : LADL.

- [LAPORTE 1997] LAPORTE Éric. "Rational transductions for phonetic conversion and phonology", *Finite-state language processing*, Cambridge, MA: MIT Press, pp. 407-429.
- [LEVINGER 1995] LEVINGER Moshe, ORNAN Uzzi, ITAI Alon. "Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew", in *Computational Linguistics* 21(3), pp. 383-404.
- [LEWIS 1996] LEWIS David D., SPARCK JONES Karen. "Natural language processing for information retrieval", *Commun. ACM* 39(1), pp. 92-101.
- [LOCKE 1955] LOCKE W., BOOTH D., eds. *Machine Translation of Languages*, Cambridge: MIT Press.
- [MACMAHON 1996] MACMAHON John G., SMITH Francis J.. "Improving statistical language model performance with automatically generated word hierarchies", *Computational Linguistics* 22(2), pp. 217-247.
- [MARCUS 1993] MARCUS Mitchell P., SANTORINI Beatrice, MARCINKIEWICZ Maryann. 1993. "Building a large annotated corpus of English: the Penn Treebank", *Computational Linguistics* 19(2), pp. 313-330.
- [MARSHALL 1983] MARSHALL Ian. "Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus", *Computers and the Humanities*, pp. 139-150.
- [MEL'CUK 1984, 1988, 1992] MEL'CUK Igor. *Dictionnaire explicatif et combinatoire du français. Recherche lexico-sémantique* I, II, III. Montréal : Presses de l'Université de Montréal.
- [MERIALDO 1995] MERIALDO Bernard. "Méthodes probabilistes et étiquetage automatique", *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 7-22, Paris : ATALA.
- [MILLER 1990] MILLER George A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K.J.. "Introduction to WordNet : an on-line lexical database", *Journal of Lexicography* 3, pp. 235-244.
- [OFLAZER 1996] OFLAZER Kemal. "Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction", *Computational Linguistics* 22(1), pp. 73-89.
- [PROSZEKY 1996] PROSZEKY Gabor. "Morphological analyzer as syntactic parser", *Proceedings of COLING-96*, Copenhagen, pp. 1123-1126.
- [ROCHE 1995] ROCHE Emmanuel, SCHABES Yves. "Deterministic part-of-speech tagging with finite state transducers", *Computational Linguistics* 21(2), pp. 227-253.
- [SAGER 1973] SAGER Naomi. "The string parser for scientific literature", *Natural language processing*, New York: Academic press, pp. 61-87.
- [SALTON 1986] SALTON G. "Another look at automatic text-retrieval systems", *Commun. ACM* 29(7), pp. 648-656.
- [SEJNOWSKI 1987] SEJNOWSKI T.J., ROSENBERG C.R.. "Parallel networks that learn to pronounce English text", *Complex Systems* 1, pp. 145-468.
- [SILBERZTEIN 1993] SILBERZTEIN Max. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Coll. Informatique linguistique, Paris : Masson, 233 p.

[SINCLAIR 1991] SINCLAIR John McH. *Corpus, Concordance, Collocation*, Oxford Univ. Press.

[SMADJA 1996] SMADJA Frank, MCKEOWN Kathleen R., HATZIVASSILOGLOU Vasileios. "Translating collocations for bilingual lexicons: a statistical approach", *Computational Linguistics* 22(1), pp. 1-38.

[SPROAT 1996] SPROAT Richard, SHIH Chilin, GALE William, CHANG Nancy. "A stochastic finite-state word-segmentation algorithm for Chinese", *Computational Linguistics* 22(3), pp. 377-404.

## Index