LES MOTS
UN DEMI-SIECLE DE TRAITEMENTS SUR LES MOTS

ERIC LAPORTE*

Résumé - Abstract

Nous passons en revue les objectifs visés, les résultats atteints, les données construites et les méthodes utilisées dans les domaines du traitement automatique des langues où on peut considérer le mot comme l'unité fondamentale. L'ambition de cette évaluation critique est de contribuer à orienter les efforts de recherche et de développement dans une direction aussi prometteuse que possible en termes de résultats applicatifs.

We survey those domains of natural language processing where the notion of word can be considered as the fundamental unit. We examine the results aimed at, the results achieved, the data acquired and the methods used in these domains. Our ambition is that this critical evaluation could contribute to orientating research and development effort into directions leading to applicative results.

Mots-clés - Keywords

Analyse lexicale, correction orthographique, documentation automatique, données linguistiques, apprentissage automatique.

Word tagging, spelling correction, information retrieval, linguistic data, learning programs.

Nous tenons à rermercier les relecteurs pour leur contribution à la discussion et pour leurs intéressantes suggestions de forme et de fond.

* eric.laporte@univ-reims.fr ;

Université de Reims-Champagne-Ardenne - Moulin de la Housse - B.P. 1039 - 51687 Reims CEDEX 2 ;

CERIL - Institut Gaspard-Monge - Université de Marne-la-Vallée - 2, rue de la Butte-Verte - 93166 Noisy-le-Grand CEDEX.

Dans plusieurs domaines du traitement automatique des langues, on peut considérer le mot comme l'unité fondamentale, quoi que cela veuille dire. Il sera question ici des objectifs visés, des résultats atteints, des données construites et des méthodes utilisées dans ces domaines.

L'un des principaux intérêts d'une telle évaluation critique est bien sûr de contribuer à orienter les efforts de recherche et de développement dans une direction aussi prometteuse que possible en termes de résultats applicatifs. L'examen de sujets aussi variés que ceux qui seront abordés ici est l'occasion d'exercer cette évaluation critique à différents niveaux.

Nous passerons d'abord en revue les principaux traitements qu'on peut regrouper dans une contribution intitulée *Les Mots*, et nous les envisagerons, thème par thème, du point de vue de leurs objectifs comme du point de vue de leurs résultats. Nous nous intéresserons ensuite aux données linguistiques qui ont été élaborées en vue de ces traitements et qui font partie de l'acquis des recherches. Cette deuxième partie sera organisée en fonction du contenu et du mode de formalisation des données. Enfin, nous évoquerons brièvement les principaux cadres méthodologiques qui constituent l'arrière-plan de tous ces travaux.

Nous illustrerons notre propos par des références bibliographiques qui ne sauraient constituer autre chose qu'un échantillon limité de contributions, sélectionnées pour leur qualité, pour leur originalité ou pour leur représentativité - ce qui ne signifie pas que les autres ne soient ni bonnes, ni originales, ni représentatives. L'auteur d'un article de synthèse tel que celui-ci prend nécessairement trois risques en ce qui concerne la sélection des références et le niveau de détail auquel il traite les sujets :

- les orienter en fonction de l'équilibre des tendances représentées, donc des rapports de force politiques du moment,
- ou en fonction des modes scientifiques,
- ou de ses propres opinions scientifiques et de son esprit critique.

Dans notre domaine, modes et politique occupent le devant de la scène, plus peut-être que dans d'autres domaines scientifiques plus mûrs ; c'est pourquoi il nous nous sommes plus particulièrement attaché à contourner, autant que possible, les deux premiers écueils.

1. OBJECTIFS ET RESULTATS DES APPLICATIONS SUR LES MOTS

Une profusion de traitements correspondent à la définition ci-dessus, qu'il s'agisse de ceux qui ont déjà été mis en œuvre au moins une fois ou de 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, nous donnons quelques éléments de méthodologie, nous examinons les performances atteintes par les systèmes réalisés, et nous les confrontons aux attentes des industriels pour la réalisation d'applications.

1.1. Analyse lexicale

Nous regrouperons sous le terme d'analyse lexicale 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, l'étiquetage lexical et la levée des ambiguïtés lexicales. La délimitation des mots n'est pas le plus facile de ces problèmes : dans les langues comme le français et l'anglais, en raison des mots composés ; dans les langues agglutinantes, en raison de la taille du vocabulaire de mots typographiques (Proszeky G. 1996) ; et aussi dans les systèmes d'écriture sans séparateurs (Sproat R. et al. 1996).

A 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 et dont il ne sera pas question ici. C'est d'ailleurs un des acquis sur lesquels s'accorde la communauté scientifique du domaine : après une période de décantation ou de maturation 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. 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 K. 1983 ; Silberztein M. 1993 ; Oflazer K. 1996),
- un corpus de textes étiqueté (Church K. 1988 ; Dermatas E. & Kokkinakis G. 1995), éventuellement accompagné d'informations linguistiques telles que des schémas de règles établis à la main (Brill E. 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 M. et al., 1995),
- un corpus non étiqueté (MacMahon J. & Smith F. 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 aprentissage 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. Tout système d'apprentissage à partir de corpus de textes est par nature statistique ; cependant, les résultats de l'apprentissage peuvent prendre une forme numérique ou symbolique, et leur réutilisabilité dépend de ce paramètre (cf. section 2.3).

Comme toute activité consistant à associer à des éléments connus (le texte) des éléments a priori inconnus (les étiquettes), l'étiquetage lexical pose des problèmes de bruit et de silence. On peut définir le taux de bruit comme la proportion d'étiquettes inexactes parmi les étiquettes présentées, et le taux de silence comme la proportion d'étiquettes non présentées parmi les étiquettes exactes. 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 exactes parmi les étiquettes présentées), complémentaire du taux de bruit, et le taux de rappel (proportion d'étiquettes présentées parmi les étiquettes exactes), complémentaire du taux de silence. Pour homogénéiser les évaluations quantitatives dans cet article, 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 la quantité d'information exacte apportée, et notamment mettre en jeu l'estimation des taux de bruit et de silence. Toutefois, la simplicité et la généralité des définitions de taux ci-dessus sont trompeuses, car les valeurs des taux dépendent du jeu d'étiquettes considéré (Laporte E. & Silberztein M. 1996) : les résultats d'un même système de levée d'ambiguïtés recevront une évaluation quantitative d'autant plus flatteuse qu'ils seront exprimés avec un jeu d'étiquettes plus pauvre, c'est-à-dire de taille inférieure et de granularité plus grossière. On constate de grandes disparités dans la taille des jeux d'étiquettes utilisés. 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 ; de plus, il existe une tentative sérieuse d'étiqueter en tant que tels les mots composés (Silberztein M. 1993). Dans les systèmes à corpus, le jeu de 36 étiquettes de la Penn Treebank pour l'anglais est souvent utilisé (Marcus M. et al. 1993) ; des jeux de 10 à 500 étiquettes environ sont utilisés pour diverses langues. Pour la mise en correspondance de jeux d'étiquettes distincts, des méthodes ont été expérimentées (Adda G. et al. 1997), mais elles ne permettent pas d'évaluer ou de comparer de façon satisfaisante la segmentation, dans le cas où l'un des deux jeux comporte des étiquettes de mots composés.

La conception de procédures d'évaluation adaptées et permettant des comparaisons est ici une gageure. Pour des applications effectives qui ont des utilisateurs directs, on peut concevoir des méthodes d'évaluation à partir de l'idée que la qualité d'un système est proportionnelle au confort qu'il apporte à son utilisateur. Mais l'analyse lexicale est essentiellement envisagée 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 phonèmes, car pour l'utilisateur une erreur gêne la compréhension de toute une phrase. Telle qu'elle est pratiquée, l'évaluation des résultats sert plutôt, en fait, de repère abstrait aux équipes de recherche ou de développement engagées dans une compétition.

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 E. 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 E. 1995),
- un taux de silence de 2,0 % et un bruit de 50 %, avec un jeu de 1.000

étiquettes pour le français (Laporte E. & Silberztein M. 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.

1.2. Détection et correction d'erreurs orthographiques

Les applications industrielles de la détection et de la correction d'erreurs orthographiques vont de logiciels distribués au grand public, tels que 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, le dédoublonnage de fichiers de noms ou d'adresses, ou la recherche de relations entre de nouveaux noms commerciaux et des noms commerciaux existants.

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 ; mais 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 qui ne peuvent pas constituer l'analyse correcte d'un texte correct, 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. Ici encore, un système d'étiquetage lexical tolérant aux erreurs constitue ipso facto un correcteur, car il peut associer à une forme erronée une forme correcte ressemblante (Oflazer K. 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,

notoirement en-deçà des possibilités techniques de détection d'erreurs lexicales, en tout cas pour le français, en raison des limitations des données lexicales qu'ils utilisent. On note même couramment, dans les grands quotidiens nationaux français, des erreurs lexicales comme la suivante, épinglée par le *Canard enchaîné* :

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

qu'il est aisé de détecter automatiquement avec un simple dictionnaire de formes fléchies. 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. Traitement des mots inconnus

Le terme de mot inconnu est généralement employé pour désigner les mots corrects dont on ne dispose ni dans un dictionnaire ni dans un corpus étiqueté. 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.

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 H. & Sproat R. 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.

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. La maintenance du dictionnaire constitue ainsi une tâche nouvelle et tout à fait particulière. D'une part, elle nécessite des compétences qui ne sont généralement pas celles des utilisateurs professionnels de l'application. D'autre part, elle doit être intégrée à l'exploitation de l'application, car la détection et l'intégration des mots absents ne doivent pas être dissociées dans le temps. Ce type d'activité, comparable à celle des correcteurs dans l'édition, n'a pas encore fait son apparition dans le monde professionnel. D'un point de vue technique, des aides automatiques au codage de nouveaux mots ne sont pas toujours envisageables. 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 D. 1996). En-dehors de ce cas, c'est-à-dire essentiellement pour les noms propres, leurs propriétés n'ont guère de raisons de se déduire de leur forme.

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 probablilistes. Quelques résultats typiques : E. Brill (1995) étiquette correctement 82 % des mots inconnus, avec un jeu de 36 étiquettes pour l'anglais. E. Dermatas & G. Kokkinakis (1995) étiquettent correctement 65 % des mots inconnus, avec un jeu de 10 étiquettes pour l'anglais.

1.4. Accentuation automatique

L'accentuation automatique consiste à rétablir les accents et autres signes diacritiques dans les textes non accentués ou accentués de façon incomplète.

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.

M. El-Bèze et al. (1994) obtiennent de bons résultats en construisant un modèle probabiliste par apprentissage automatique dans un corpus. Leur 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.

1.5. 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 (1987), par exemple, a systématiquement fait appel à la recherche d'exemples dans des concordances (Sinclair J. 1991). Un concordancier est un logiciel de construction de concordances.

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 M. 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 *Réso,* ou une séquence de motifs de ce type, ou

(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 M. 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.

Ici encore, 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.

1.6. Documentation automatique

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'est encore étendu récemment 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 nouveau.

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. 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.

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 D. & Sparck Jones K. 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 l'indexation du stock de documents. C'est notamment le cas lorsqu'on désire balayer systématiquement le réseau informatique mondial à la recherche de sites concernés par un thème donné. 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 G. 1986).

1.7. Division en fin de ligne

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.

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 converture 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.

1.8. Phonétisation

La phonétisation est la transcription phonétique du texte écrit. Les applications 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.

Les méthodes de phonétisation font intervenir soit un dictionnaire phonétique, soit des règles contextuelles, soit un réseau connexionniste (Sejnowski T. & Rosenberg C. 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 ; elle aide alors à séparer le problème du passage d'un système de notation à un autre (l'orthographe et la représentation phonétique) et le problème 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 K. 1983) ou directe (Kaplan R. & Kay M. 1994 ; Laporte E. 1996) 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 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 K. et al. 1994), ce qui est évidemment insuffisant. L'utilisation d'un dictionnaire phonémique permet d'obtenir une phonétisation avec 0,6 % de mots erronés.

De plus, un autre critère d'évaluation important 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.

1.9. 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. Nous n'en avons trouvé aucun pour lequel les performances des systèmes existants correspondent indiscutablement aux attentes des industriels, sauf peut-être la division en fin de ligne. Cependant, 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.

Par ailleurs, nous avons noté 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.

Les pages qui précèdent nous ont amené à envisager les mots sur l'axe syntagmatique, c'est-à-dire les occurrences dans un texte, le matériau de base des applications. On peut voir les mots sur l'axe paradigmatique : les entrées dans un dictionnaire. Ces mots, et les informations lexicales qui leur sont associées, constituent des données linguistiques qui font partie des acquis de la recherche et ont été élaborées pour être utilisées dans des applications.

2. DONNEES

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 W. & Booth D. 1955 ; Ceccato S. 1962). 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.

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 Ch. & Nedobejkine N. 1982). Ces données y ont gagné un nom, celui de "ressources linguistiques". Etant 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 (Gross M. 1986b).

Cette section est consacrée aux différents ensembles de données linguistiques formelles existantes, et aux perspectives de développement dont elles peuvent constituer le support.

2.1. Dictionnaires de 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
débiteur	N1	pas de féminin
débiteur	N35	*débiteuse*
débiteur	N36	*débitrice*

(codes empruntés au DELAS, Courtois, 1990). Au niveau syntaxique, deux verbes *perquisitionner* peuvent être distingués par la forme de leur complément :

mot	code	commentaire
perquisitionner	32R2	*La police a perquisitionné la maison*
perquisitionner	35L	*La police a perquisitionné dans la maison*

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

mot	code	commentaire
bélier	animal	*Ses cornes ressemblent à celles d'un bélier*
bélier	outil	*Ils ont forcé l'entrée avec un bélier*

(cf. Gross G. 1994).

Les ambiguïtés étant réparties à tous les niveaux dans les langues du monde, la précision et l'exactitude du codage lexical se reflètent immanquablement dans des distinctions d'emplois.

Ces quelques remarques tout à fait générales sur les dictionnaires électroniques dessinent un cadre de raisonnement commun, indépendant du niveau d'analyse syntaxique 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 B. 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 raisonnnable 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 M. 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). Certains incluent d'autres relations sémantiques, la synonymie par exemple, 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 I. 1984-88-92 ; Gross G. 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 M. 1988). Des procédés statistiques appliqués à des thésaurus ont également permis de produire des classes d'entrées sémantiques (Grefenstette G. 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. M. Garrigues (1992), en remplaçant ces calculs par des méthodes artisanales, obtient des valeurs qui prédisent mieux la fréquence d'emploi dans des corpus nouveaux.

2.2. Dictionnaires de 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 et al. 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 M. 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. 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 distinctes. 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*. Les algorithmes d'alignement de corpus bilingues parallèles permettent de constituer automatiquement des dictionnaires de collocations bilingues (Smadja F. et al., 1996) et fournissent une aide à la construction de dictionnaires terminologiques bilingues (Gaussier É. & Langé J.M. 1995).

2.3. Machines abstraites

Nous rangeons sous ce titre les automates et transducteurs qui effectuent des opérations telles que l'étiquetage lexical (Koskenniemi K. 1990 ; Oflazer K. 1996) ou la phonétisation (Laporte É. 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 E. 1995) et automates finis (Roche E. & Schabès Y. 1995) pour l'analyse lexicale sont acquis par apprentissage statistique, mais 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.

2.4. Corpus étiquetés

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 sont utilisés pour l'anglais. Les plus connus sont probablement le corpus Brown (Garside R. et al. 1987), le corpus Lancaster-

Oslo-Bergen (LOB) (Marshall I. 1983), tous les deux d'un million de mots, et la Penn Treebank (Marcus M. et al. 1993).

Dans l'évaluation de ces corpus étiquetés, les jeux d'étiquettes sont concernés au premier chef (cf. plus haut, section 1.1). Pour simplifier, les jeux d'étiquettes sont essentiellement conçus pour les mots simples, et l'étiquetage laisse de côté la plupart des mots composés : chacun de leurs 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 N. & Véronis J.1995).

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

2.5. Conclusions

Ces résultats appellent une conclusion mitigée. Les données linguistiques formelles qui ont été élaborées jusqu'à présent sont loin de suffire pour la réalisation de toutes les applications que nous avons évoquées dans la première partie. Elles forment cependant un échantillon substantiel des données nécessaires à ces applications. De plus, la réutilisabilité de ces données est désormais perçue comme une notion cruciale.

3. LES TYPES DE METHODES ET LEURS POTENTIALITES

Nous avons évoqué brièvement, dans ce qui précède, une variété de techniques et de méthodes à propos des objectifs qu'elles visent ou des données qu'elles emploient. On est frappé par la diversité de ces techniques et par le fait qu'elles se rattachent à plusieurs cadres méthodologiques.

Ceci s'explique par la pluridisciplinarité inhérente au problème, mais il n'est 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. L'évaluation des potentialités à long terme des méthodes est polémique, difficile à justifier, non mesurable quantitativement, et, probablement pour cette raison, remarquablement absente des discussions scientifiques, alors qu'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 d'applications. En tous cas, l'évaluation des systèmes actuels, telle qu'elle est pratiquée, n'est pas conçue pour refléter les potentialités des méthodes qui ont servi à leur construction (cf. Bourlard H. et al. 1996).

Dans cette section, nous tentons une synthèse personnelle sur ce sujet, au risque de caricaturer la pensée de nos confrères. Nous distinguerons trois cadres méthodologiques : ceux qui privilégient la théorie, la description empirique, ou l'optimisation.

3.1. 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 R. & Gazdar G. 1996), un langage de représentation des données lexicales, dont il existe plusieurs implémentations en Prolog et en Lisp, indépendant de la langue, valable 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 l'universalité par rapport aux langues du monde et 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 génératives.

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 ressemble 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.

3.2. 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 (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 (cf. Sager N. 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. Etant 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 le thème de la reproductibilité des observations fait place à celui de l'objectivité.

3.3. 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 d'IBM France (El-Bèze M. et al. 1994) en est un exemple. Ce type de méthodologie présente un caractère empirique certain, mais d'une autre façon que le précédent : chaque expérience repose sur un modèle du problème et sur des hypothèses (Merialdo B. 1995) qui sont confirmés ou infirmés 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 modélisations informatiques de réalités floues, généralement extra-linguistiques. Quant aux traitements statistiques, les modèles sur lesquels ils reposent comportent toujours 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 J. & Smith F. 1996).

Ce choix a l'avantage d'aboutir à des traitements entièrement automatiques, donc peu 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 L. et al. 1996). Après tout, depuis le XIXe 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.

3.4. 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 E. 1995 ; El-Bèze M. & Spriet Th. 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.

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 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 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. A 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 contraint 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.

REFERENCES

Adda, Gilles, Josette Lecomte, Joseph Mariani, Patrick Paroubek, Martin Rajman. 1997. "Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français", in *Actes des 1es Journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF*, Avignon.

Baayen, Harald, Richard Sproat. 1996. "Estimating lexical priors for low-frequency morphologically ambiguous forms", in *Computational Linguistics* 22(2), pp. 155–166.

Bartkova, K., D. Larreur, I. Métayer. 1994. "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, Christian, N. Nedobejkine. 1982. "Base lexicale : organisation générale et indexage", *Projet ESOPE, ADI, Rapport final*, partie D.

Bourlard, Hervé, Hynek Hermansky, Nelson Morgan. 1996. "Towards increasing speech recognition error rates", in *Speech communication* 18, Elsevier, pp. 205-231.

Brill, Eric. 1995. "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging", in *Computational Linguistics* 21(4), pp. 543–565.

Ceccato, S. 1962. "Automatic translation of languages", in *Automatic Translation of Languages*, 1966, London: Pergamon.

Church, Kenneth. 1988. "A stochastic parts program and noun phrase parser for unrestricted text", in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, ACL.

Clemenceau, David. 1996. "Finite-state morphology: inflections and derivations in a single framework using dictionaries and rules", in *Finite-state processing of natural language*, Cambridge, MA: MIT Press.

COBUILD dictionary of the English language. 1987 (première édition), Collin's.

Courtois, Blandine. 1990. "Un système de dictionnaires électroniques pour les mots simples du français", in *Langue française* 87, Larousse, pp. 11-22.

Danlos, Laurence. 1985. *Génération automatique de textes en langues naturelles*, Paris : Masson.

Dermatas, Evangelos, George Kokkinakis. 1995. "Automatic stochastic tagging of natural language texts", in *Computational Linguistics* 21(2), pp. 137-163.

El-Bèze, Marc, Bernard Mérialdo, Bénédicte Rozeron, Anne-Marie Derouault. 1994. "Accentuation automatique de textes par des méthodes probabilistes", in *Technique et science informatiques* 13(6), pp. 797-815.

El-Bèze, Marc, Thierry Spriet. 1995. "Intégration de contraintes syntaxiques dans un système d'étiquetage probabiliste", in *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 47-66, Paris : ATALA.

Evans, Roger, Gerald Gazdar. 1996. "DATR: a language for lexical knowledge representation", in *Computational Linguistics* 22(2), pp. 167-216.

Evens, M. (ed.). 1988. *Relational models of the lexicon*, Cambridge University Press.

Gale, William, Kenneth Church. 1991. "A program for aligning sentences in bilingual corpora", in *Proceedings of the 29th Annual meeting of the Association for Computational Linguistics*, Berkeley, CA.

Garrigues, Mylène. 1992. "Dictionnaires hiérarchiques du français. Principes et méthodes d'extraction", in *Langue française* 96, Larousse.

Garrigues, Mylène. 1997. "Lemmatized concordances of complex utterances: application to language learning", in *New technologies in language teaching and learning*, A.K. Korsvold, B. Rüschoff, eds., Strasbourg : Council of Europe Publishing, pp. 87–98.

Garside, Roger, Geoffrey Leech, Geoffrey Sampson. 1987. *The computational analysis of English*, Longman.

Gaussier, Éric, Jean-Marc Langé. 1995. "Modèles statistiques pour l'extraction de lexiques bilingues", in *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 133-155, Paris : ATALA.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*, Kluwer international series in engineering and computer science, Dordrecht: Kluwer Academic Publisher.

Gross, Gaston. 1994. "Classes d'objets et description des verbes", in *Langages* 115, Paris : Larousse, pp. 15-30.

Gross, Maurice. 1975. *Méthodes en syntaxe*, Paris : Hermann, 414 p.

Gross, Maurice. 1986a. *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*, Paris : ASSTRIL, 670 p.

Gross, Maurice. 1986b. "Lexicon-Grammar. The representation of compound words", in *Proceedings of COLING-86*, Bonn, pp. 1-6.

Guthrie, Louise, James Pustejovsky, Yorick Wilks, Brian M. Slator. 1996. The role of lexicons in natural language processing", in *Commun. ACM* 39(1), pp. 63-72.

Ide, Nancy, Jean Véronis. 1995. *The Text Encoding Initiative: background and context*, Dordrecht: Kluwer.

Kaplan, Ronald M., Martin Kay. 1994. "Regular models of phonological rule systems", in *Computational Linguistics* 20(3), pp. 331-378.

Koskenniemi, Kimmo. 1983. *Two-level morphology: a general computational model for word-form recognition and production*, Publication no. 11, Department of General Linguistics, University of Helsinki.

Koskenniemi, Kimmo. 1990. "Finite-state parsing and disambiguation", in *Proceedings of COLING 90*.

Laporte, Eric. 1996. "Rational transductions for phonetic conversion and phonology", in *Finite-state processing of natural language*, Cambridge, MA: MIT Press.

Laporte, Eric, Max Silberztein. 1996. "Ambiguity rates. Automatic analysis of French text corpora and computation of ambiguity rates for different tagsets", in *GRAMLEX deliverables, October 1995 - June 1996*, Eric Laporte (ed.), Paris : LADL.

Levinger, Moshe, Uzzi Ornan, Alon Itai. 1995. "Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew", in *Computational Linguistics* 21(3), pp. 383–404.

Lewis, David D., Karen Sparck Jones. 1996. "Natural language processing for information retrieval", in *Commun. ACM* 39(1), pp. 92-101.

Locke W., D. Booth., eds. 1955. *Machine Translation of Languages*, Cambridge:

MIT Press.
MacMahon, John G., Francis J. Smith. 1996. "Improving statistical language model performance with automatically generated word hierarchies", in *Computational Linguistics* 22(2), pp. 217-247.
Marcus, Mitchell P., Beatrice Santorini, Maryann Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn Treebank", in *Computational Linguistics* 19(2), pp. 313-330.
Marshall, Ian. 1983. "Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus", *Computers and the Humanities*, pp. 139-150.
Mel'cuk, Igor. 1984, 1988, 1992. *Dictionnaire explicatif et combinatoire du français. Recherche lexico-sémantique I, II, III*. Montréal : Presses de l'Université de Montréal.
Merialdo, Bernard. 1995. "Méthodes probabilistes et étiquetage automatique", in *Traitement automatique des langues* 36(1-2), *Traitements probabilistes et corpus*, pp. 7-22, Paris : ATALA.
Miller, George A., R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. "Introduction to WordNet : an on-line lexical database", in *Journal of Lexicography* 3, pp. 235-244.
Oflazer, Kemal. 1996. "Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction", in *Computational Linguistics* 22(1), pp.73-89.
Proszeky, Gabor. 1996. "Morphological analyzer as syntactic parser", in *Proceedings of COLING-96*, Copenhagen, pp. 1123-1126.
Roche, Emmanuel, Yves Schabès. 1995. "Deterministic part-of-speech tagging with finite state transducers", in *Computational Linguistics* 21(2), pp.227–253.
Sager, Naomi. 1973. "The string parser for scientific literature", in *Natural language processing*, New York: Academic press, pp. 61-87.
Salton, G. 1986. "Another look at automatic text-retrieval systems", in *Commun. ACM* 29(7), pp. 648-656.
Sejnowski, T.J. & C.R. Rosenberg. 1987. "Parallel networks that learn to pronounce English text", in *Complex Systems* 1, pp. 145-468.
Silberztein, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Coll. Informatique linguistique, Paris : Masson, 233 p.
Sinclair, John McH. 1991. *Corpus, Concordance, Collocation*, Oxford Univ. Press.
Smadja, Frank, Kathleen R. McKeown, Vasileios Hatzivassiloglou. 1996. "Translating collocations for bilingual lexicons: a statistical approach", in *Computational Linguistics* 22(1), pp. 1-38.
Sproat, Richard, Chilin Shih, William Gale, Nancy Chang. 1996. "A stochastic finite-state word-segmentation algorithm for Chinese", in *Computational Linguistics* 22(3), pp. 377-404.