DOCUMENT RESUME

ED 317 072 FL 018 426

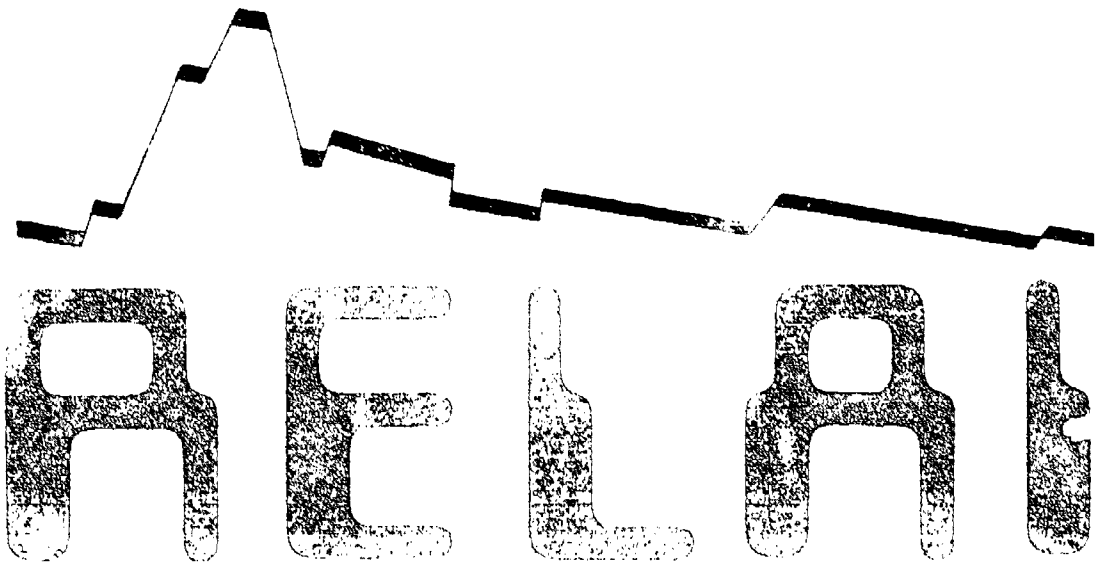
AUTHOR	Ouellon, Conrad, Comp.
TITLE	La Description des langues naturelles en vue d'applications linguistiques: Actes du colloque (The Description of Natural Languages with a View to Linguistic Applications: Conference Papers). Publication K-10.
INSTITUTION	Laval Univ., Quebec (Quebec). International Center for Research on Bilingualism.
REPORT NO	ISBN-2-89219-204-8
PUB DATE	89
NOTE	330p.; Papers presented at a colloqium (University of Laval, Quebec, Canada, December 7-9, 1988).
AVAILABLE FROM	International Center for Research on Bilingualism, Par Casault-Universite Laval, Quebec GIK 7P4, Canada.
PUB TYPE	Collected Works - Conference Proceedings (021)
LANGUAGE	French
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	Applied Linguistics; Computational Linguistics; *Computer Science; Computer Software; *Descriptive Linguistics; Foreign Countries; French; *Language Processing; *Language Research; Oral Language; Phonology; *Research Utilization; Spanish; Spelling
IDENTIFIERS	*Natural Languages; Parsing; Transcription

ABSTRACT

Presentations from a colloquium on applications of research on natural languages to computer science address the following topics: (1) analysis of complex adverbs; (2) parser use in computerized text analysis; (3) French language utilities; (4) lexicographic mapping of official language notices; (5) phonographic codification of Spanish; (6) electronic dictionaries; (7) specialized linguistic programs; (8) text difficulty; (9) linguistic variation and formalization in Quebec French; (10) French text generation software; (11) consequences for parsing of heterogeneity and insertion in sentences; (12) language utilities; (13) interaction of orthographic and phonological representations in reading; (14) automatic phoneticization of French texts; (15) computerized content analysis; (16) description of natural languages with a view to computer applications; (17) software to aid in the conception of deontic knowledge bases; (18) software for computer-assisted text generation; (19) organization of segment lengths in syllabic rhyme; (20) a language utility for public administration; (21) French prosody; (22) interactive treatment of documents; (23) transcription of oral corpuses from a comparative perspective; (24) a linguistic work station; (25) universal applicative grammar; (26) and the speed of synthesized speech. Other papers are abstracted only. (MSE)

PUBLICATION
K-10

(CIRB)
Centre international de recherche
sur le bilinguisme



Recherche en linguistique appliquée à l'informatique

Actes du Colloque

LA DESCRIPTION
DES LANGUES NATURELLES
EN VUE D'APPLICATIONS
LINGUISTIQUES

Montréal, [illegible]
[illegible] décembre [illegible]

sous la direction
de [illegible]

1984

Conrad OUELLON, présentation

Collaborateurs:

Esther Blais
Bernard Guay

Actes du Colloque

LA DESCRIPTION DES LANGUES NATURELLES EN VUE D'APPLICATIONS LINGUISTIQUES

Publication K-10

1989
Centre international de recherche sur le bilinguisme
International Center for Research on Bilingualism
Québec

Le Centre international de recherche sur le bilinguisme est un organisme de recherche universitaire qui reçoit une contribution du Secrétariat d'État du Canada pour son programme de publication.

Ont contribué de façon spéciale à la publication de ce bulletin, l'Université Laval et la Compagnie IBM du Canada.

Le Conseil de recherches en sciences humaines du Canada, grâce à une importante subvention, a permis que se tienne le Colloque sur **La description des langues naturelles en vue d'applications informatiques** *et qu'en soient publiés les Actes.*

The International Center for Research on Bilingualism is a university research institution which receives a supporting grant from the Secretary of State of Canada for its publication programme.

Laval University and the IBM Company of Canada have also contributed, in a special manner, to the publication of this bulletin.

The presentation of this Colloquium **La description des langues naturelles en vue d'applications informatiques** *and the publication of its proceedings, were made possible through a generous grant by the Social Sciences and Humanities Research Council of Canada.*

Dépôt légal (Québec) 4ième trimestre 1989
ISBN 2-89219-204-8

AVANT-PROPOS

Conrad Ouellon
CIRB

Il n'est guère utile d'insister sur les défis que pose le développement technologique à la survie du français comme langue véhiculaire de la science, de l'innovation. Mais doit-on les relever ces défis?

Les positions les plus diverses ont cours dans le monde francophone. Les uns prônent le maintien du statut du français, langue de culture, langue littéraire et refusent qu'on retouche le monument... D'autres par contre sont convaincus de la nécessité d'adapter le français à la technique; ils croient en la capacité de resourcement de la langue, en sa capacité de pouvoir encore exprimer les concepts scientifiques comme elle l'a toujours fait.

C'est ainsi que de nombreux chercheurs travaillent à l'établissement de ponts entre l'univers informatique et les utilisateurs francophones, à l'adaptation du matériel informatique aux spécificités de la langue française, à la description de la langue aussi bien sous sa forme orale qu'écrite pour qu'elle puisse bénéficier du traitement informatique. L'objectif visé par ces travaux, c'est d'abord l'accroissement des connaissances linguistiques; mais aussi et peut-être surtout, c'est, dans le contexte des technologies de l'information, de permettre que la langue française jouisse des mêmes avantages que la langue anglaise, que l'utilisateur francophone puisse disposer d'outils propres à ses spécificités linguistiques.

C'est dans le but de réunir les spécialistes préoccupés par cette question des industries de la langue qu'est née cette idée d'une rencontre sur *le traitement des langues naturelles en vue d'applications informatiques*. Le groupe RELAI (recherche en linguistique appliquée à l'informatique) rattaché au Centre international de recherche en aménagement linguistique (CIRB) de l'Université Laval a été heureux de s'associer au Laboratoire d'automatique documentaire et linguistique (LADL) du professeur Maurice Gross de l'Université de Paris VII pour organiser l'évènement.

Plus de 150 congressistes, provenant de la plupart des laboratoires et centres de recherche intéressés par les industries de la langue, se sont inscrits au Colloque **Le traitement des langues naturelles en vue d'appplications informatiques** qui s'est tenu à l'Université Laval les 7, 8 et 9 décembre 1988. On a noté une forte participation des étudiants gradués aux divers ateliers, ce qui démontre bien l'actualité et l'intérêt du thème retenu. Nous avons également apprécié la présence de nombreux chercheurs du LADL et du centre d'ATO de l'Université du Québec à Montréal.

Le Colloque **Le traitement des langues naturelles en vue d'applications informatiques** n'a pu avoir lieu que grâce à de généreuses contributions du Conseil de recherches en sciences humaines du Canada, du Département de langues et linguistique et de la Faculté des lettres de l'Université Laval. Nous tenons à remercier ces organismes.

L'organisation locale du colloque, sous la responsabilité conjointe du CIRB et de l'Association des étudiant(e)s diplômé(e)s inscrit(e)s en langues et linguistique (AEDILL), n'aurait pu suffire à la tâche sans le travail et le dévouement d'Esther Blais et d'Annie Bourret, chargées de l'organisation matérielle de la rencontre.

La préparation des ACTES, y compris les fastidieuses tâches de collecte, de traitement, de correction, de saisie des textes de présentation, était sous la responsabilité d'Esther Blais et de Bernard Guay.

Je veux enfin souligner le travail de madame Diane Tremblay qui a établi les premiers contacts entre le groupe RELAI et le LADL. Sans son esprit d'initiative, sa tenacité à nous convaincre de l'intérêt d'une telle rencontre, ce colloque n'aurait certainement pas eu lieu.

TABLE DES MATIÈRES

DISCOURS D'OUVERTURE

Lorné Laforge
Directeur du CIRB

- M. le président et directeur du département de langues et linguistique de l'Université Laval.
- M. le directeur du laboratoire d'automatique documentaire et linguistique de L'Université Paris VII
- Mme la présidente des étudiants diplômés inscrits en langues et linguistique
- Distingués invités d'honneur
- Chers collègues et participants inscrits à ce colloque

En guise de remarques préliminaires, je tiens d'abord à vous souhaiter la bienvenue et à exprimer au nom du CIRB l'immense satisfaction que nous ressentons à l'ouverture de ce colloque de voir ici réunis pour la première fois à l'Université Laval des chercheurs de très grande réputation pour communiquer les résultats de leur recherche sur *la description des langues naturelles en vue d'applications informatiques*. Nous sommes d'autant plus satisfaits de constater que cette entreprise a pris naissance sous le signe de la collaboration entre trois organismes, le LADL de Paris VII, le CERIL également de Paris VII et l'AEDILL de l'Université Laval et nous, c'est-à-dire le CIRB - Centre International de recherche en aménagement linguistique. La tenue de ce colloque a été rendue possible grâce à une subvention du Conseil de Recherche en sciences humaines du Canada, au soutien financier du projet RELAI et à l'appui de l'Université Laval, de la Faculté des lettres et du département de langues et linguistique. L'organisation du colloque est sous la responsabilité directe de M. Conrad Ouellon, directeur-adjoint du CIRB, appuyé par une équipe très efficace et dynamique. Nous aurons très souvent à lui rappeler cet évènement et à le féliciter pour ce qui s'annonce un succès éclatant.

Le thème du colloque qui nous rassemble ici aujourd'hui met l'accent sur la recherche fondamentale en linguistique, recherche préalable à toute application informatique. Cette recherche constitue sans qu'on en fasse toujours état, la toile de fond, le cadre théorique obligé des réalisations les plus spectaculaires de l'informatique moderne. Et les informaticiens le savent et commencent à l'admettre. Nous nous attendons donc à ce que les participants et les chercheurs-communicants nous permettent de mieux définir des concepts récents et imagés comme "les industries de la langue", qu'ils donnent des assises théoriques à l'informatique linguistique qu'on a souvent tendance à désigner dans certains milieux comme une recherche utilitaire et exclusivement appliquée, donc synonyme de recherche très peu universitaire.

Il faudrait rappeler que l'informatique linguistique peut être citée pour ses glorieux états de service et par l'intérêt qu'elle a généré auprès de chercheurs très prestigieux de l'époque. Mes souvenirs me ramènent à 1964 alors que j'assistais à un colloque sur la recherche en linguistique quantitative à l'Université de Strasbourg, et où on pouvait entendre les Moreau, Quémada, Potier, Coseriu, Herdan, Muller, Greimas et surtout Gougenheim fascinés par la traduction automatique, l'analyse par ordinateur du lexique et de la grammaire et par les travaux sur la concordance des textes. Au Canada, les travaux effectués à Ottawa et à l'Université de Montréal sur la traduction assistée par ordinateur, les banques de terminologie de l'OLF et du Secretariat d'État, l'EAO, l'analyse textuelle, la synthèse de la parole ont dominé les dernières décennies. Au niveau international, les spécialistes de l'informatique linguistique ont trouvé une tribune grâce à un organisme désigné par le sigle COLING, ou computational linguistics organisme regroupant surtout des intervenants anglophones.

Il appartient toujours aux aînés dont je suis de faire le pont entre le passé et le présent, d'assurer en quelque sorte une certaine continuité. Et dans cette continuité force nous est de constater le temps fort qui s'est installé dans notre milieu depuis quatre ou cinq ans en informatique linguistique. Serait-ce attribuable au fait que l'instrument s'est fait petit, souple et convivial, qu'il est devenu un bien de consommation accessible à tous et qu'il appartient à tous ceux qui ont les moyens de l'acquérir et de s'en servir? Serait-ce que les pouvoirs publics ont finalement perçu la portée de ce phénomène social et ont mis l'accent sur le développement d'outils utilitaires dans des domaines comme les communications, l'éducation, le commerce et l'industrie? Serait-ce également attribuable aux limites des artifices de l'informatique qui doit nécessairement utiliser le génie inventif et créateur des langues naturelles pour articuler l'intelligence artificielle et créer de nouvelles générations d'ordinateurs?

Toutes ces questions seraient pure rhétorique si l'on ne convenait pas que la conjoncture actuelle est extrêmement favorable à la conduite de recherches comme les nôtres et qu'il faut saisir l'occasion pour démontrer à tous les bailleurs de fonds la solidité, le sérieux et la grande rigueur de nos entreprises. En parcourant le programme du colloque nous ne pouvons nous empêcher de témoigner du très haut niveau des communications inscrites et de la qualité des communicants. Nous prédisons donc que ce colloque sera un événement marquant en informatique linguistique et qu'il sera cité à l'avenir comme un modèle à perpétuer. Est-ce trop ambitieux? Je ne le crois pas et vous me donnerez volontiers raison.

L'évolution que nous avons voulu imprimer à notre centre de recherche - le CIRB - depuis quelques années, se concrétise constamment et en particulier aujourd'hui par ce colloque. Dans une société comme la nôtre - je parle évidemment du contexte canadien et québécois - où les questions linguistiques ont toujours été viscérales, où les luttes pour résister ou pour conquérir du territoire, du pouvoir et de l'influence ont toujours été en apparence très civilisées, mais combien sournoises et soutenues, il nous est apparu nécessaire de changer notre orientation vers des recherches qui correspondraient davantage aux besoins de notre société, recherches permettant d'entrevoir l'objectivité des faits et non la subjectivité des opinions.

Voilà pourquoi le CIRB, sous le parapluie de l'aménagement linguistique et en particulier l'aménagement du corpus, a choisi d'entreprendre l'étude des problèmes demandant l'éclairage de la recherche scientifique pour que soient actualisés les objectifs et les moyens préconisés par les plans d'aménagement linguistique du territoire canadien et québécois. Par exemple, nous avons voulu définir notre programme scientifique en complémentarité et peut-être pour aller plus loin, au plan théorique, que l'Office de la langue française du Québec, organisme d'État voué à l'aménagement linguistique du Québec, en particulier, à l'aménagement du corpus, depuis plus de 20 ans. C'est donc par rapport à l'action de l'OLF et à tout ce qui est sous-jacent à cette action qu'un programme de recherche universitaire a pu être défini.

Le souci du CIRB dans sa programmation scientifique est de démontrer qu'elle génère un progrès intellectuel. La simple application pratique pourrait être le fait d'un organisme d'application, comme l'est l'Office, mais non l'objectif unique d'un organisme universitaire comme l'est un Centre de recherche comme le CIRB.

Dans cette perspective le CIRB favorise des recherches qui permettront de faire avancer nos connaissances sur l'informatique linguistique, sur l'enseignement et l'apprentissage des langues conduisant à la pratique de ces langues, sur l'enrichissement ou la modernisation lexicales des langues, sur la traduction, la terminologie et la néologie, sur la didactique des langues en un mot (ou deux), sur *les industries de la langue*.

Pour le CIRB, la langue ou les langues sont des richesses naturelles au même titre que les autres types de richesses naturelles, des biens collectifs partagés et en même temps des super-valeurs sociales. Par conséquent, elles doivent être aménagées, c'est-à-dire exploitées (les industries de la langue), diffusées (lexiques, dictionnaires, terminologies), banalisées ou démythi-

fiées (l'informatique linguistique), enseignées (la didactique des langues), être rendues à la communauté qui les utilise. Le CIRB veut s'appuyer sur des travaux de recherche fondamentale pour restituer aux langues toutes leurs fonctions et leurs valeurs. A ce titre le CIRB peut jouer un rôle de médiateur, non seulement auprès de la communauté universitaire, mais également auprès de tous les membres d'une communauté. Cette nouvelle approche ne peut être favorisée que par des travaux inter et multidisciplinaires.

La thématique des industries de la langue semble s'imposer aujourd'hui aux chercheurs du CIRB. Sa pertinence stratégique pour le monde francophone n'est plus à démontrer puisque déjà les deux premiers sommets l'ont identifiée parmi les axes prioritaires des actions communes de la francophonie. Récemment, une étude de Denis Monnier du Conseil de la langue française du Québec soulignait de façon explicite l'urgence de nous engager à fond dans cette voie et de relever un défi de taille en mobilisant toutes les ressources humaines et matérielles, en particulier celles des centres de recherche universitaires.

Le CIRB a déjà donné des signes qu'il veut bien relever ce défi. Ses réalisations à partir du projet RELAI et ses récentes publications sur ce sujet, publications qui seront lancées à la clôture du colloque, attestent qu'il s'est engagé par ses recherches à mettre son expertise linguistique diversifiée au service de la société dans le cadre d'un programme s'étendant sur les trois prochaines années. Il veut donc s'attaquer au problème ou à la tâche de rendre la langue française et même toute langue romane plus facilement traitable par ordinateur. Vos travaux nous en démontreront la possiblilité et fourniront les outils qui facilitent le travail sur l'ordinateur en français pour toutes les catégories de consommateurs d'informatique.

Ces brèves remarques liminaires ne sauraient passer sous silence la volonté du CIRB d'intégrer les étudiants de 2^{e} et 3^{e} cycles à nos travaux. Ce n'est pas la première fois que les étudiants diplômés participent de plein droit à des réunions savantes organisées par le Centre et nous desirons que cette participation devienne une des traditions du Centre. De cette façon, les recherches universitaires prennent vraiment leur sens puisque nous contribuons ainsi à la formation d'une nouvelle génération de chercheurs. Qui sait si nous n'avons pas réussi aujourd'hui, grâce à ce colloque, à réunir les Moreau, Quémada, Potier, Coseriu, Greimas, Herdan, Muller et Gougenheim d'aujourd'hui et de demain.

ANALYSE D'ADVERBES COMPLEXES

Antoinette Baldinger-Mrabti

1. TRAITEMENT DES NOMS DANS LE LEXIQUE-GRAMMAIRE ET NOTION D'ADVERBE COMPLEXE

Dans la théorie du lexique-grammaire de M. Gross 1975 (1981), les verbes supports (V_{sup}) servent à représenter les substantifs prédicatifs à l'intérieur du dictionnaire. Ainsi, le nom *enquête* correspond à l'entrée (J. Giry-Schneider 1978, 1987):

(1) N_0 *(=Jean) (fait + mène) une enquête sur* N_1 *(=: cette affaire)*

parallèlement à l'entrée du verbe *enquête*:

(2) N_0 *(=: Jean) enquête sur* N_1 *(=: cette affaire)*

Pour des noms donnés, les combinaisons verbe support - nom, illustrées par l'exemple (1), permettent de construire des représentations lexicales sous formes de phrases. Les rapports entre les noms considérés et les verbes sémantiques et morphologiquement apparentés (e. g. entre *enquête* et *enquêter*) permettent d'établir une relation de paraphrase entre beaucoup d'entrées du type de (1) (2) mais cette relation qui manque de généralité, lorsqu'on étudie sa reproductibilité sur le lexique, est secondaire.

Les verbes supports servent également à représenter les noms en position d'adverbes (M. Gross 1988) et on étudiera ici des propriétés syntaxiques d'adverbes encore assez mal connues qui mettent en jeu des verbes supports. Les adverbes que nous analyserons sont des groupes nominaux construits avec la préposition *dans* comme par exemple:

(3) *<Jean a disparu> dans des circonstances (accidentelles + inconnues)*

En grammaire traditionnelle le groupe *dans des circonstances (accidentelles + inconnues)* est un complément circonstantiel qui n'a pas reçu d'interprétation sémantique stable. Il n'est pas impossible qu'on l'ait négligé parce que les pronoms interrogatifs *où, quand, comment, pourquoi* lui étaient peu applicables:

- *Jean a disparu (où + quand + comment + pourquoi)*
- *dans des circonstances (accidentelles + inconnues)*

On remarque qu'il n'y a pas d'interdiction stricte sur la question *comment*. Toutefois le fait qu'il soit impossible de la pronominaliser par *ainsi*:

(2) = Jean a disparu ainsi

sera pour nous un critère pour l'exclure clairement de la classe des adverbes de manière.

L'adverbe en *circonstance* correspond à l'entrée

(4) *Que P (=: Que Jean ait disparu) (a eu lieu + s'est produit) dans des circonstances Adj (=: accidentelles + inconnues)*

La combinaison du verbe support d'occurrence d'événement (Harris 1976) *avoir lieu* ou *se produire* avec le nom *circonstance*, qui est son complément spécifique, permet d'attribuer une phrase à l'adverbe et de la faire dans le dictionnaire au même titre qu'un substantif prédicatif ou un verbe ordinaire. On sait que ce traitement unifie la représentation des groupes nominaux. Ces derniers, qu'ils soient adverbes ou prédicats, sont en effet régulièrement insérés dans des formes verbales composées avec des supports et on voit l'importance des V_{sup} pour l'établissement du dictionnaire. Par commodité, appelons *adverbes complexes* les groupes nominaux adverbiaux qui entrent dans les constructions du type de (3) puisque ces constructions, en apparence à un verbe conjugué, le verbe principal *V* d'une phrase *P* (e. g. *disparaître* dans *Jean à disparu*), nécessitent, pour être expliquées, qu'on ajoute un deuxième verbe, le support (e. g. le V_{sup} d'occurrence *avoir lieu* ou *se produire*), qui sous-tendra l'analyse élémentaire de l'adverbe et son placement à droite de *V* suivant un mécanisme transformationnel que nous détaillons plus loin.

C'est donc la présentation de quelques-uns des problèmes que soulève l'application de la méthode des verbes supports aux adverbes que nous allons aborder ici. Pour cela , nous rappellerons d'abord les types de contraintes distributionnelles qui peuvent s'observer dans les constructions adverbiales, afin de caractériser le degré de figement des adverbes complexes que nous discutons (§ 2). Nous examinerons ensuite quels sont les verbes qui sont candidats, dans ces constructions, pour supporter les adverbes en *dans* comme compléments spécifiques et nous rappellerons les mécanismes transformationnels des dérivations (§ 3). Existe-t-il des adverbes en *dans* à double portée (§ 4)? Théoriquement, l'adverbe peut complémenter trois sortes de verbes: des verbes ordinaires, des verbes supports variés, des verbes opérateurs. Effectuer leur tri est crucial pour l'analyse (§5). Enfin quel est exactement le degré de liberté des constituants de groupe nominal adverbial dans nos exemples (§6)?

2. ADVERBES LIBRES ET ADVERBES FIGÉS

Pour caractériser le degré de figement des adverbes complexes que nous décrivons, rappelons l'exemple

(3) *<Jean marche dans des circonstances (accidentelles + inconnues)*

et contrastons-le avec le bloc d'exemples suivant:

(4) *Jean marche dans la combine*
(5) *Jean est mort (dans son lit + sur le coup)*
(6) *<Jean a (parlé + répondu)> dans le droit fil de la (conversation + discussion)*
(7) *<Jean a (parlé + répondu)> (avec + dans) une langue (étrangère + inconnue)*
(8) *<Jean (travaille + voyage)> dans son genre.*
(9) *<Jean (travaille + voyage)> (dans l'insouciance la plus totale + insouciamment)*
(10) *<Jean travaille du chapeau> dans son genre*

Dans tous les cas, nous observons un complément prépositionnel en *dans*. Tout d'abord nous mettons à part les exemples (4) et (5) car ils ne comportent pas à proprement parler d'adverbes. En effet les groupes *dans la combine* et *dans son lit* ne sont pas séparables des verbes *marcher* ou *mourir* pour le calcul du sens et ne peuvent donc être analysés comme des ajouts à

droite d'une phrase simple *P* comme nous avons vu qu'il était possible de le faire pour (3) où *P(=: Jean a disparu)* existe indépendamment du groupe *dans des circonstances (accidentelles + inconnues)*. (4) et (5) sont des phrases figées au sujet humain près. On sait que les phrases figées s'analysent syntaxiquement de façon régulière. (4) et (5) ont un complément de verbe en *dans* plutôt qu'un adverbe. Soumettons ce complément à quelques tests syntaxiques. On remarque que le complément est questionné par *comment* dans (5) mais qu'il ne l'est pas dans (4). On remarque également que le complément ne s'impose pas de façon unique dans (5) puisqu'il peut commuter avec le groupe *sur le coup*. On en conclura que (5) est moins figé que (4).

Les exemples restants ont tous la même construction. La structure

<P=N_0 V W> Prép Dét N Modif

en donne une image simplifiée. Dans cette structure, nous attribuons une forme générale *Prép Dét N Modif* à l'adverbe: cette forme est celle d'un groupe nominal prépositionnel, appelé adverbe généralisé dans la terminologie de la grammaire transformationnelle (M. Gross 1988). Dans les exemples (6) et (7) le verbe principal *V (=: a parlé + répondu)* est moins libre que dans les exemples (8) et (9) où *V(=: travail + voyage)* est indifférencié par rapport à l'adverbe. Il appartient à une classe spécifique qui correspond à une interprétation sémantique de verbe de parole. Les adverbes correspondants peuvent être questionnés par *comment* et pronominalisés par *ainsi*: ce sont des adverbes de manière. On remarque que le degré de figement de l'adverbe est indépendant de celui de *P*: l'adverbe de (7) est beaucoup plus libre que celui de (6). Dans (7) le nom tête du groupe *N (=: langue)* a son sens ordinaire; la préposition n'est pas unique puisque *avec* commute avec *dans*; le modifieur adjectival *Adj (=: étrangère + inconnue)* est libre. Par contre dans (6) la séquence *dans le droit fil de* n'a pas un sens calculable à partir des mots pleins qui la composent *droit* et *fil*; le choix de la préposition est unique, le déterminant défini *le* est unique. On considérera que cette séquence occupe dans l'adverbe la position du déterminant *Dét* devant le nom libre *N (=: conversation + discussion)* et que *Dét* est figé.

Les exemples (8) et (9), comme on vient de le dire, ont un verbe libre, qui correspond ici à l'interprétation traditionnelle générale d'action. L'adverbe a, comme dans la paire précédente (6) - (7), un degré de figement indépendant de celui de *P*. C'est ainsi que *dans le genre* est figé et ne répond à aucune des questions traditionnelles de complément circonstanciel invoquées plus haut (§1). Par contre, l'adverbe de (9) est libre: il répond à la question *comment*, il est pronominalisé par *ainsi*; il donne lieu à la formation de l'adverbe en *-ment insouciamment* selon une procédure qui met la prépositon *dans* en parallèle avec la préposition *avec* qui sert à former l'adverbe synonyme *d'une manière insouciante*. En première approximation, ce sont les exemples (7) et (9) qui ressemblent le plus aux adverbes complexes en *dans* que nous étudions, c'est-à-dire des adverbes de manière couramment considérés comme libres. Rappelons toutefois que l'exemple (3) retenu n'est pas à strictement parler un adverbe de manière. A l'inverse, (10) est l'exemple le plus éloigné de (3) puisqu'il correspond à deux unités de sens, la phrase et l'adverbe, respectivement totalement figées; mais, on le remarque, combinables ensemble. L'étude de ce type d'exemple reste encore largement ouverte.

3. EXEMPLES DE VERBES SUPPORTS POUR LES ADVERBES COMPLEXES EN *DANS*

La propriété qui définit le mieux un verbe support est sa possibilité d'être réduit à zéro - et inversement sa possibilité d'être récupéré dans le même contexte - sous des conditions définies par des relations transformationnelles entre phrases. C'est ainsi que V_{sup} *faire observé au* §1 est réductible dans le cadre d'une relativation. Sa réduction est formalisée pas la règle [RédVsup] comme le montre l'exemple suivant:

(11) *<Paul a lu> l'enquête que Jean a faite sur cette affaire*
(11) [Réd$_V$sup]
= *<Paul a lu> l'enquête de Jean sur cette affaire*

De même, le V_{sup} *avoir lieu* est réductible dans le cadre d'une introduction coréférentielle de l'adverbe qu'on réalise par le biais d'un discours à deux phrases coordonnées. Sa réduction est formalisée par la règle [Pron V_{sup} z]. Cette règle s'applique après qu'un lien pronominal ait été formé qui explicite les relations de l'adverbe à la phrase *P*, comme le montre l'exemple suivant:

(12) *Jean a disparu, que Jean ait disparu a eu lieu dans des circonstances (accidentelles + inconnues)*
(12) [Pronomin]
= *Jean a disparu, cela a eu lieu dans des circonstances (accidentelles + inconnues)*
[Pron V_{sup} z.]
= (3) *Jean a disparu dans des circonstances (accidentelles + inconnues)*

Dans les deux cas de dérivation, la réduction du V_{sup} va de pair avec la réduction à zéro d'un pronom coréférent: relatif (e.g. *que*) dans la règle [Réd Vsup], démontratif (e.g. *cela*) ou personnel dans la règle [Pron Vsup z].

Quand un nom prédicatif complémente un *Vsup* (e,g, la combinaison verbe - nom *faire une enquête*) on a vu (§1) qu'il complète de la même façon des variantes de ce *Vsup* (e.g. la combinaison concurrente *mener une enquête*). Nous avons observé la même situation pour l'adverbe en *circonstance*. Il ne complémente pas seulement le *Vsup avoir lieu* mais aussi les variantes en liste ouverte *se produire, arriver, se faire* qui introduisent naturellement dans la langue les adverbes en *dans* du type de (3). L'analyse de l'adverbe complexe par ces *Vsup* possède une certaine généralité. Par exemple, nous la reproduisons facilement sur l'exemple suivant:

(13) *<On a arrêté Paul> dans un cadre arbitraire*

qui correspond au discours suivant:

On a arrêté Paul, qu'on ait arrêté Paul (a eu lieu + s'est produit + est arrivé + s'est fait) dans un cadre arbitraire

On voit que les *Vsup* d'occurrence de (3) sont de bons candidats pour analyser l'adverbe de (13).

Pour (13) comme pour (3), on observera que l'adverbe porte sur la phrase *P* puisque les *Vsup* ont une complétive *Que P (=: que Jean ait disparu + qu'on ait arrêté Paul)* pour sujet, complétive dont nous avons observé la pronominalisation en *cela*. On acceptera même assez bien la sélection du *Vsup être dans*:

Jean a disparu, c'est dans des circonstances (accidentelles + inconnues)
On a arrêté Paul, c'est dans un cadre arbitraire

La même portée s'observe quand on nominalise *P*:

La disparition de Jean a eu lieu dans des circonstances (accidentelles + inconnues)
L'arrestation de Paul a eu lieu dans un cadre arbitraire

Dans l'état actuel des recherches, la difficulté n'est pas de trouver des bons candidats pour supporter les adverbes comme compléments spécifiques mais de restreindre le champ des analyses possibles.

4. ADVERBES EN *DANS* À DOUBLE PORTÉE

On sait que les adverbes peuvent avoir plusieurs portées sémantiques. Cette situation est bien connue pour les adverbes de manière (Balibar-Mrabti 1987, Molinier 1985). Considérons la phrase.

(15) *Jean marche d'une manière rapide*

L'adverbe porte sur la phrase *P(=: Jean marche)* comme le montre le discours

(16) *Jean marche, sa manière de marcher est rapide*

mais il est possible également de poser le discours concurrent

(17) *Jean marche, il est rapide*

dans lequel le sujet du support *être* du prédicat adjectival *rapide*, attribué comme source à l'adverbe, n'est plus la phrase nominalisée par le pivot complexe *manière* mais le sujet N_0 *(=: Jean)* du verbe principal *V (=: marcher)* que nous pronominalisons par *il*. L'adverbe *d'une manière rapide* a donc une double portée, fondée sur l'observation des discours (16) et (17).

On remarque que pour (3) comme pour (13) il est impossible d'établir un lien coréférentiel, du type de celui que nous avons observé sur le discours (17), avec un des actants du verbe principal *V* de *P*, comme le montrent les interdictions suivantes:

Jean est dans des circonstances (accidentelles + inconnues)
Paul est dans un cadre arbitraire

Il n'existe pas, à propos des constructions adverbiales considérées (3) et (13), de proximité sémantique intuitive entre les noms *circonstance* et *cadre* et le *Vsup être dans*, si nous voulons lui donner un sujet humain, on ne pourrait pas non plus faire supporter ces noms par le *Vsup avoir*:

Jean a des circonstances (accidentelles + inconnues)
Paul a un cadre arbitraire

Cette deuxième possibilité en *avoir*, nous intéresse moins ici car elle ne conserve pas, pour la complémentation du verbe support, la préposition *dans* qui entre dans la composition des adverbes que nous cherchons à décrire. Néanmoins elle est un argument supplémentaire pour corroborer l'interprétation habituelle qui est attribuée à ces adverbes: celle d'être des adverbes de phrase et uniquement des adverbes de phrase.

On remarque dans l'examen des compatibilités entre verbes supports et groupes nominaux que le choix lexical de l'adjectif modifieur peut influencer autant celui du nom tête du groupe. Par exemple on peut dire

Jean a les circonstances atténuantes

Mais cette phrase correspond à un parallélisme connu entre le verbe support *avoir* et la préposition *avec*:

<Jean s'en est tiré> (avec + dans) les circonstances atténuantes

Ce parallélisme exclut ici le choix de la préposition *dans*.

Nous avons étudié ailleurs (article cité) une famille d'adverbes de manière dans lesquels le nom tête du groupe a un contenu sémantique presque vide et sert de pivot pour la montée (ou la descente) de l'adjectif qui s'observe comme modificateur dans le groupe. Soit par exemple

Jean marche (avec + d') un pas rapide

Le nom *pas* sert à former un complément interne *d'un pas rapide* pour le verbe principal *V (=: marcher)*. Il est le pivot de l'adverbation dans la phrase quasi synonyme (15). On pourrait considérer que les noms *circonstances, cadre*, de contenu sémantique beaucoup plus général que les adjectifs modifieurs qu'ils accueillent (e.g. *accidentelles, inconnues* ou bien *arbitraire*) fonctionnent eux aussi comme des pivots d'adverbation du modificateur, en combinaison avec certains verbes supports dont la sélection dépend autant des modificateurs que des noms. L'étude des combinaisons de verbes supports et des noms pivots d'adverbes est encore largement ouverte. Il existe des corrélations à approfondir, comme nous venons de le voir, entre ces combinaisons, certains adjectifs modificateurs et la préparation de groupe nominal adverbial. Notamment des verbes supports d'occurrence sont sélectionnés par certains noms qui vont de pair avec la préposition *dans*; le verbe support *avoir* est sélectionné par un nom pivot comme *pas* et va de pair avec la préposition *avec* qui entre dans la composition de certains adverbes de manière.

Existe-t-il des adverbes en *dans* dont la portée ne soit pas limitée à *P*? Considérons l'exemple suivant:

(18) *< Jean a compris cela> dans une vision (admirable + prémonitoire) des événements*

L'adverbe en *vision* qui s'y trouve porte sur la phrase *P (=: Jean a compris cela)* comme le montre le discours

Jean a compris cela , que Jean ait compris cela est contenu dans la vision (admirable + prémonitoire) qu'il a des événements

et sur la base de ce discours nous pouvons parler, comme pour les exemples précédents, d'adverbes complexes. Mais à côté de l'interdiction déjà observée pour les adverbes de (3) et de (13)

Jean est dans une vision (admirable + prémonitoire) des énénements

nous disposons de la phrase

(19) *Jean a une vision (admirable + prémonitoire) des événements*

La phrase (19) présente l'inconvénient, déjà souligné, de ne pas conserver la préposition *dans* de l'adverbe mais elle explicite la portée du nom *vision* sur le sujet *No (=: Jean)* de *P*. Nous avons donc ici un bon exemple d'adverbe en *dans* à double portée.

On aura remarqué l'impossibilité de construire une phrase en *Vsup avoir lieu*:

Que Jean ait compris cela a eu lieu dans la vision (admirable + prémonitoire) qu'il a des évémements

Par contre, nous utilisons une nouvelle forme pour supporter l'adverbe: le passif du verbe *contenir*. Cette nouvelle solution présente l'intérêt d'assimiler davantage que dans les cas prédicatifs le traitement des groupes nominaux adverbiaux à celui des groupes nominaux prédicatifs. En effet, l'adverbe n'est plus, comme pour (3) et (13), un complément spécifique, mais facultatif, du verbe. Il est , comme pour *faire*, ou *avoir*, un complément obligatoire: le complément d'agent du verbe. Cette situation rapproche le support *être contenu dans* du Vsup *être dans* déjà utilisé.

On remarque ainsi que (19) est la nominalisation de

Jean voit les événements d'une façon (admirable + prémonitoire)

On peut aussi mettre en évidence une relation causative entre (18) et

Que Jean ait une vision (admirable + prémonitoire) des événements a fait qu'il a compris cela

Cette direction d'interprétation n'existait pas pour (3) ni pour (13). La polysémie des adverbes, la variété des analyses syntaxiques qu'ils requièrent nous conduisent à les décrire individuellement. Ce statut les rapproche des expressions figées.

5. ADVERBES EN *DANS* ET TRI DES VERBES COMPLÉMENTÉS

Au §2 nous avons posé le problème du tri des constructions adverbiales en fonction de leur degré de figement. Nous conclurons maintenant notre approche syntaxique de quelques cas d'adverbes complexes en *dans* en examinant le problème du tri des verbes susceptibles d'être complémentés par l'adverbe. Quel statut syntaxique leur attribuer? Considérons pour cela le bloc d'exemples suivants:

(20) *Jean évoque le souvenir de son père dans une pensée pieuse*
(21) *Jean a agi dans la pleine jouissance de ses facultés*
(22) *Jean a évolué dans ses idées*
(23) *Jean a élevé Marie dans la religion*
(24) *Jean vit dans des conditions agréables*

Ces exemples sont disparates. Dans quel cas avons-nous affaire à un verbe ordinaire? Dans quel cas à un verbe support? Peut-on même parler de verbe opérateur appliqué à une phrase en verbe support (M. Gross 1981)?

L'exemple (20) nous ramène à l'exemple en adverbe complexe

(18) *Jean a compris cela dans une vision (admirable + prémonitoire) des événements*

dans lequel l'adverbe a une double portée puisqu'il est aisé de le relier au discours

Jean évoque le souvenir de son père, (c'est + il est plongé) dans une pensée pieuse

Dans le deuxième membre de ce discours, deux verbes sont en effet possibles pour supporter l'adverbe et ils ont respectivement pour sujet la phrase *P (=: Jean évoque le souvenir de son père)* que nous pronominalisons par *ce*; le sujet *No (=: Jean)* que nous pronominalisons par *il*. On aura remarqué que le verbe support compatible avec le sujet humain est une forme passive du verbe *plonger*. Cette particularité est un argument supplémentaire pour rapprocher (20) de (18) puisque dans les deux cas l'adverbe complémente un *Vsup* de forme passive. Mais dans (20) *être plongé dans* a un sujet humain tandis que dans (18), on l'a vu §4, *être contenu dans* avait au contraire un sujet phrastique.

Pouvons-nous parler de nominalisation à propos du nom tête *N (=: pensée)* de l'adverbe considéré? Nous ferons alors état de la paire

Jean pense à son père pieusement
= Jean a (une + des) pensée(s) pieuse(s) pour son père

Il semble que cette paire soit d'utilisation difficile ici car on ne voit pas comment justifier les différences de prépositions (e.g. *à son père* vs *pour son père*) d'ailleurs absentes de la forme adverbiale à analyser *dans une pensée pieuse.*

L'exemple (21) pose le problème des rapports entre les adverbations et les nominalisations à partir de données plus intéressantes. Comme pour (20) montrer qu'il s'agit d'un adverbe complexe complémentant un verbe ordinaire ne présente pas de difficulté particulière. On analysera le groupe *dans la pleine jouissance de ses facultés* au moyen du discours suivant:

Jean a agi, (c' + il) est dans la pleine jouissance de ses facultés

Ce discours nous montre que l'adverbe a, là encore, une double portée. On remarque surtout que le Vsup *être dans* présente la particularité de nominaliser la phrase

(25) *Jean jouit pleinement de ses facultés*

lorsqu'il est appliqué au sujet N_0 *(=: Jean)* de la phrase *P*. Autrement dit, nous disposons ici de la paire

(25) = *Jean est dans la pleine jouissance de ses facultés*

On connaît en français la complémentarité qui existe entre la préposition *dans* et la proposition *en*. Avec la préposition *dans* il faut un déterminant. Par contre la préposition *en* se construit sans déterminant comme le montrent par exemple les phrases

Jean est dans une colère noire
Jean est en colère

Nous rapprocherons la nominalisation que nous venons d'observer de la relation étudiée par D. de Négroni-Peyre 1978 sur une paire comme

(26) *Jean voyage*
= Jean est en voyage

Les formes verbales qui sont candidates pour l'analyse des adverbes complexes en *dans* sont donc variées. Aux verbes supports d'occurrence vus au §3 tels que *avoir lieu, se produire, arriver, se faire*, s'ajoutent des formes passives telles que *être contenu dans, être plongé dans*. Ce deuxième type de support présente l'intérêt de contraindre fortement l'adverbe en *dans* puisque celui-ci n'est plus introduit comme complément spécifique mais comme complément d'agent. On vient de voir qu'il faut ajouter à notre inventaire un troisième type de support: le *Vsup être dans* fonctionnant, à la manière du *Vsup être en*, comme support de nominalisation.

A propos des compléments en *dans* des exemples (22) à (24) nous ne pouvons plus effectuer d'analyse qui mette en jeu des discours. L'exemple (22) est la restructuration (A. Guillet, C. Leclère 1981) de la phrase

Les idées de Jean ont évolué

(22) est donc une phrase simple dans laquelle *évoluer* est un verbe principal *V* ordinaire. Le groupe *dans ses idées* complémente directement le verbe sans que son analyse passe par l'étude d'une construction adverbiale en verbes supports. Cet exemple retient notre attention dans la mesure où la transformation de restructuration donne lieu, entre autre, à des compléments en *dans*, dont le sens, on s'en aperçoit immédiatement, se rapproche de celui des adverbes que nous avons décrits. Dans le cadre de cet article, nous n'avons pas cherché à savoir si cette transformation pouvait jouer un rôle comparable à celui du passif, ou même des nominalisations, pour l'analyse d'adverbes complexes donnés.

Considérons maintenant l'exemple (23). Il pose un problème différent des exemples précédents, celui de savoir si nous devons analyser le verbe *élever* qu'il contient comme un verbe ordinaire ou comme un verbe opérateur. La décision dépend de l'acceptabilité d'une phrase en verbe support sur laquelle il puisse s'appliquer. Pouvons-nous accepter

Marie (est + vit) dans la religion

Si nous acceptons la phrase en *vivre, élever* sera considéré comme opérateur. On sait en effet que *vivre* a le statut d'une extension lexicale du verbe support *être Prép* pour *Prép =: dans* (L. Danies 1980, 1988) comme le montre l'exemple

(24) *Jean vit dans des conditions agréables*

Si nous n'acceptons pas *vivre* comme extension de *Vsup*, le caractère très contraint de la phrase (23) nous le fera considérer comme une phrase comparable à

(5) *Jean est mort dans son lit*

discutée au §2 donc comme une phrase figée avec un verbe ordinaire.

6. COMPOSITION DU GROUPE NOMINAL ADVERBIAL

Reprenons notre exemple initial

(3) *Jean a disparu dans des circonstances (accidentelles + inconnues)*

et détaillons les contraintes entre les constituants du groupe *Dé. N Modif: Jean a disparu dans (ces circonstances*

**** Aligner sous (ces circonstances

+ de telles circonstances
+ des circonstances (accidentelles + inconnues)
+ des circonstances des plus graves
+ des circonstances que tu connais
+ une circonstance à élucider
+ les circonstances (d'un accident

**** Aligner sous (d'un accident

+ de l'accident
+ que tu connais
+ les plus graves
+ habituelles

Jean a disparu dans (circonstances

***Aligner sous (circonstance

+ des circonstances
+ des circonstances (d'accident

*** Aligner sous (d'accident

+ d'un accident
+ de l'accident

*** Aligner sous (circonstance

+ une circonstance
+ la circonstance
+ les circonstances (accidentelles + inconnues)

Trois remarques s'imposent. Tout d'abord, le déterminant passe difficilement du pluriel au singulier donc la variation en nombre est presque interdite. Cette particularité est un indice de figement du groupe nominal. Ensuite, l'acceptabilité du déterminant pluriel dépend du choix des modifieurs. Ce phénomène correspond aux groupes à modifieur d'unicité décrits par M. Gross 1977. Enfin on notera la complexité de détail des contraintes énumérées. A ce propos, ajoutons l'observation que les modificateurs sont cumulables comme le montre la phrase

Jean a disparu dans les circonstances accidentelles que tu connais

Les paires suivantes:

- *Jean a disparu comment?*
- *En la circonstance*
- *Nous nous sommes abstenus comment?*
- *Dans les circonstances actuelles*

associées respectivement aux deux phrases attestées

Jean a disparu en la circonstance
Nous nous sommes abstenus dans les circonstances actuelles

montrent que l'adverbe change d'interprétation sémantique et de propriétés si on remplace *dans* par *en* ou si on choisit un adjectif approprié comme *actuelles* pour modifieur.

7. CONCLUSION

Les adverbes en *dans* que nous avons discutés sont appelés adverbes complexes quand leur caractérisation syntaxique met en jeu des discours à deux membres. Par cette méthode ils apparaissent naturellement dans la langue pour compléments des types de verbes supports (verbes d'occurrence d'événement, formes verbales passives, *Vsup être dans*). Considérés comme des formes libres, ils sont toutefois soumis à des contraintes nombreuses et imprévisibles a priori, quand on examine en détail la composition du groupe nominal introduit par la préposition et pour les décrire nous avons fait une liste de combinaisons autorisées. Si on rapproche ce constat de l'observation qu'à l'échelle de la phrase chaque construction adverbiale a présenté sa convergence propre de propriétés, nous en conclurons que la grammaire des adverbes libres, à travers les cas que nous avons traités, est peu différente de celle des formes figées.

Références

BALIBAR-MRABTI, Antoinette. 1987. *Règles formelles et règles rhétoriques sur un cas d'analyse d'adverbes*, **Linguisticae Investigationes** XI:2, Amsterdam: John Benjamin B. V.

DANLOS, Laurence. 1981. *Représentation d'information linguistiques: les constructions N être Prép X*, Thèse de 3e cycle, Université de Paris 7.

DANLOS, Laurence. 1988. *Les expressions construites avec le verbe support être Prép*, **Langages 90**, Paris: Larousse.

DE NEGRONI-PEYRE, Dominique. 1978. *Nominalisation par être en et réflexivation*, **Linguisticae Investigationes** II:I, Amsterdam: John Benjamins B.V.

GIRY-SCHNEIDER, Jacqueline. 1987. *Les prédicats nominaux en français*, Genève-Paris: Droz.

GROSS, Maurice. 1975. *Méthode en syntaxe*, Paris: Herman.

GROSS, Maurice. 1977. *Grammaire transformationnelle du français: syntaxe du nom*, Paris: Larousse.

GROSS, Maurice. 1981. *Les bases empiriques de la notion de prédicat sémantique*, **Langages 63**, Paris: Larousse.

GROSS, Maurice. 1988. *Grammaire transformationnelle du français: syntaxe de l'adverbe*, Paris: Cantilène.

GROSS, Maurice. 1988. *Les limites de la phrase figée*, **Langages 90**, Paris: Larousse.

HARRIS, Zellig. 1976. *Notes de cours de syntaxe*, Paris: Le Seuil.

MOLINIER, Christian. 1985. *Remarque sur une sous-classe d'adverbes en -ment orientés vers le sujet et les adjectifs sources*, **Linguisticae Investigationes** IX:2, Amsterdam: John Benjamins B.V.

APPORT DE L'ORDINATEUR À L'ANALYSE DES DONNÉES TEXTUELLES

Louis-Claude Paquin[1] et Jacques Beauchemin[2]
Université du Québec à Montréal

0. PRÉAMBULE

Ce texte[3] n'a pas pour but de presenter des outils ou des méthodes informatiques à ceux (chercheurs, gestionnaires, décideurs, etc.) dont la lecture et l'analyse du contenu des textes constituent la principale activité. Son objectif est plutôt d'exposer les besoins et les attentes de ces derniers à ceux (linguistes, informaticiens, etc.) qui les élaborent. Même si les outils et les méthodes informatiques pour la compréhension des textes n'ont cessé depuis les trente dernières années de se diversifier et de se perfectionner, tant sur le plan de la performance que de celui de la validité théorique, une insatisfaction persiste. Dans les pages qui suivent, au lieu de poser un diagnostic outil par outil, nous tentons de remonter les sources de cette insatisfaction.

Dès lors que l'on appréhende cet objet mouvant et volatile qu'est le texte, les problèmes se posent nombreux. Car, au-delà de la dimension proprement informatique, toute entreprise d'automatisation de la lecture repose sur ces questions à la fois élémentaires et extrêmememt complexes de savoir ce qu'est un texte et, plus fondamentalement encore, ce qu'est l'acte de la lecture. L'élaboration de même que l'utilisation d'outils informatiques dédiés à l'analyse de textes nous apparaît tributaire de la réponse à ces deux questions névralgiques.

Deux types d'outils d'analyse de textes se disputent la faveur des "travailleurs du texte"[4]. D'une part les analyseurs lexicographiques produisent des lexiques (listes de mots) et des concordances (liste de mots accompagnés d'un segment de leur contexte). D'autre part, les analyseurs morpho-syntaxiques associent aux phrases d'un texte les éléments d'une description structurale.

Ces deux types d'outils ont été associés plus ou moins exactement à deux méthodologies d'analyse des données textuelles qui, depuis toujours, sont tenues pour opposées: l'analyse quantitative où un maximum d'indices est pris en compte et l'analyse qualitative où seuls quelques indices jugés particulièrement significatifs sont considérés. Cette opposition méthodologique a été transposée sur le plan des familles d'outils informatiques. Les analyseurs lexicographiques sont utilisés pour produire des analyses quantitatives basées sur des calculs statistiques, alors qu'on attend des parseurs une description exhaustive permettant des analyses qualitatives.

---

[1] Chercheur au Centre d'Analyse de Textes par Ordinateur, Ph D en philologie médiévale, Édition critique d'un traité alchimique latin: Liber secretorum.

[2] Assistant de recherches au département de sociologie, rédige une thèse de doctorat en analyse du discours.

[3] Les auteurs participent depuis janvier 1988 à un projet de recherche, initié par Jules Duchastel, ayant pour objectif l'élaboration d'un Système d'Analyse de Contenu (des textes) Assisté par Ordinateur (SACAO) financé par le Fonds FCAR du Québec dans le cadre du programme «actions spontanées». Ils tiennent à souligner la précieuse contribution de Luc Dupuy avec lequel ils ont tenu des discussions enrichissantes.

[4] Le masculin est utilisé de façon générique et inclut la formulation féminine.

La pauvreté de certains résultats obtenus par des analyses lexicales imputables à une formalisation insuffisante des données textuelles a fait croire en la primauté du second type d'outils sur le premier. Un tel raisonnement repose sur une definition implicite suivant laquelle la langue naturelle correspond à un ensemble fini de règles circonscrivant un univers de «possibles». Or la supériorité présumée du « parsage » en analyse de texte est discutable pour peu que le texte soit considéré dans toutes ses dimensions et dans toutes ses manifestations. En effet, la description attendue des parseurs, bien qu'exhaustive, ne recouvre qu'un système du texte, celui qui régit l'enchaînement et la hiérarchisation des mots. Il en résulte que les autres dimensions (la référenciation, la thématisation, l'actantialité, l'intertextualité, etc.) restent à couvrir et que l'analyse doit être produite par d'autres moyens.

Face à la complexité de l'analyse des données textuelles, nous proposons de troquer l'automatisation de la lecture experte pour l'assistance à la lecture experte. Cela aura pour effet de privilégier la créativité du lecteur plutôt que l'exhaustivité mécanique d'une description ne recouvrant que partiellement ce qui est recherché dans les textes. Loin de rejeter l'un ou l'autre de ces types d'outils, nous proposons de les enrichir mutuellement en les intégrant dans un atelier « textuel » et surtout en calibrant la portée de leur intervention en fonction d'une méthodologie respectant les prémisses de celle qui avait cours avant l'utilisation de l'ordinateur.

1. LA LECTURE EXPERTE DES TEXTES

Notre groupe de recherche s'est constitué autour d'un besoin particulier en matière d'analyse de contenu de textes. Celui qu'expriment chercheurs, gestionnaires, décideurs, de tous horizons oeuvrant au sein d'organisations grandes productrices de textes. Leur rapport aux textes varie en fonction de leurs objectifs; accumulation de faits, d'événements ou de connaissances, interprétation, élaboration de stratégies, prise de décision, etc. Dans le mouvement sans cesse croissant de la technocratisation de la décision et de la gestion rationaliste de projets, les grands appareils, qu'ils soient privés ou publics, en sont venus à une production textuelle - faite de rapports, de directives, de projets ou de pré-projets - dont le volume grandissant a peu à peu rendu impossible leur exploitation véritable. Bref, ceux dont la lecture et l'analyse de texte constituent la principale activité, les travailleurs du texte, croulent sous la masse de documents qu'ils doivent analyser.

Mais qu'en est-il de cet objet texte?

Les mots « tissu » et « texte » ont une racine latine (*textus*) commune. Les réalités désignées se caractérisent par un enchevêtrement, dans un premier cas, de fils dans une trame et, dans le second, de systèmes dans l'espace discursif. Il n'y a ainsi de définition valable du texte que minimale: suite d'énoncés écrits en langue naturelle et enregistrés sur un support (papier ou magnétique). Pour le travailleur du texte, le texte est, au-delà de son apparence première, un objet stratifié qui ne se réduit pas plus à l'ensemble des mots qui le composent qu'aux relations réunissant ceux-ci en énoncés ou encore à un contenu pur et simple.

Le texte prend de multiples formes en fonction du projet communicationnel qui lui est assigné: études, rapports, directives, décrets, réponse en format libre à des questionnaires, retranscription d'entrevues, etc. Certes, le document se donne de prime abord comme contenu pur et simple. L'accès à ce contenu fait toutefois appel à un ensemble d'habiletés dont on sous-estime peut-être la complexité. Il nécessite bien sûr l'accomplissement de tâches qui, prises une à une, seraient informatisables: déchiffrer les caractères qui forment les mots, reconstituer l'enchaînement des mots en énoncés et la succession des énoncés en un contenu spécifique. Cependant cet ensemble de compétences s'avère insuffisant. Non seulement une connaissance minimale de

l'univers particulier du texte est-elle essentielle, mais encore le lecteur doit-il disposer d'un savoir renvoyant aux conventions sociales régissant l'énonciation et au champ de l'interdiscursivité constitutive du discours dans la société moderne. Arrêtons-nous un instant sur ces aspects fondamentaux de la discursivité.

Le texte, comme discours, déborde largement l'univers clos de la rationalité de son objet ou des catégories qu'il met en oeuvre. Il s'organise dans une économie de l'énonciation tout aussi porteuse de sens que les objets de la réalité qu'il désigne nommément au lecteur. Le texte connote ainsi les objets qu'il aborde tout autant qu'il les désigne. L'ironie, l'humour grinçant, la déférence, le discours d'autorité et combien d'autres dispositifs sont autant de procédés discursifs que le lecteur expert doit reconnaître et intégrer à son analyse globale du texte. Cette dimension constitutive du texte le pose en objet à « décoder » au-delà des règles proprement linguistiques qui le structurent.

Mais il y a plus. Le texte doit également être situé dans l'espace social qui le porte et dans les rapports de forces dans lesquels il s'insère. Le texte est toujours tissé de procédés ou de stratégies. Pourquoi en est-il ainsi? Pourquoi est-il davantage que ce qu'il dit explicitement? Parce que dans la société moderne, où les représentations du monde se sont affranchies du monolithisme et de la censure, l'espace dans lequel se meut le texte est celui d'un pluralisme où chaque discours dans un domaine donné coexiste avec un ensemble de représentations concurrentes. Dans un mouvement le plus souvent imperceptible à l'oeil nu, il converse avec quelque invisible interlocuteur, répond implicitement à ses détracteurs et appelle à sa rescousse ses alliés du moment.

L'interlocuteur absent ou invisible est celui qui hante le discours ou le regarde de l'extérieur mais qui d'une manière ou d'une autre le pose, par sa seule co-présence comme point de vue dans l'univers de tous les points de vue possibles. L'autre dans le discours c'est le rappel de la contingence d'une parole et donc de la volatilité de la vérité qu'elle prétend fonder. Cette modalité de la discursivité dans la société moderne, en vertu de laquelle la co-présence dans l'espace discursif de discours condamnés au dialogue permanent, a été saisie sous la notion d'interdiscursivité. Nous verrons maintenant quelle importance capitale revêt cette particularité du dicours pour l'analyse de texte et pour l'élaboration d'outils de support.

Mais qu'en est-il de la lecture?

Nous avons affirmé que le texte est polyphonique, traversé par les contraintes auxquelles le soumet l'espace pluraliste du discours dans lequel il se déplace et soumis à des modalités d'énonciation définies en société. Nous avons avancé qu'il est en cela deploiement de stratégies discursives. Le décodage des stratégies mises en oeuvre dans les textes - menées sur ses multiples registres (morphologique, syntaxique, rhétorique, etc.) - mobilise une expertise aussi vaste que variée. Or, malgré la complexité du processus discursif, un lecteur humain est en mesure, à un degré ou à un autre, de faire une lecture experte des textes qu'il aborde.

Cette capacité résulte du procès de la socialisation dans la foulée duquel se constitue une connaissance du monde extraordinairement ramifiée. La réalité, au-delà de ses manifestations empiriques, fait l'objet d'interprétations mobilisant tout autant les dimensions affective, culturelle qu'intellectuelle. En somme, lire un texte c'est tout à la fois prendre connaissance de l'information « brute » qu'il contient, considérer le dialoguisme que nous avons évoqué, s'y situer comme tiers et juger de la valeur de l'ensemble à partir de critères extrêmement complexes. C'est ce que nous appellerons la lecture experte. Mais cette expertise est paradoxale car elle relève d'un impensé qui fait en sorte que le lecteur est le plus souvent dans l'impossibilité d'énoncer les critères explicites qui le guident.

Or on ne peut renoncer au recours à l'ordinateur pour analyser les textes sous prétexte que les algorithmes qu'il peut mettre en oeuvre s'avèrent incapables dans un avenir prévisible de reproduire l'expertise humaine. Nous reconduirions alors le problème évoqué dès le début touchant la masse sans cesse croissante de textes en attente d'être analysés e la rigueur que ce travail nécessite. La solution nous semble résider dans la réconciliation des deux formes de lecture: il s'agit de mettre à la disposition du lecteur des instruments à l'aide desquels son expertise puisse être mise à profit, en même temps qu'il puisse lui garantir une capacité de lecture augmentée en termes de volume, de rigueur, bref de systématicité.

Rappelons d'abord le cadre à l'intérieur duquel s'est traditionnellement déployée la lecture experte. Nous verrons ainsi ce que nous pouvons retenir de cette méthode dans l'informatisation de l'analyse de contenu.

2. L'EXTRACTION ET L'ANALYSE PRÉ-INFORMATIQUE DES DONNÉES TEXTUELLES

La lecture effectuée par les travailleurs du texte n'a pas pour but d'épuiser les significations possibles d'un texte, mais d'en extraire des données en fonction d'intérêts qui leur sont propres. L'extraction s'effectue en deux temps: la sélection d'un segment porteur de données est d'abord opérée puis saisie, habituellement sous la forme de fiches. Les données extraites sont par la suite analysées. L'analyse prend la forme d'un classement des fiches recueillies pour réorganiser les données en sous-textes.

L'extraction des données textuelles nécessite d'abord la capacité de distinguer les contenus renvoyant au réel des éléments de discours. Il s'agit ensuite de ramener les formes différentes qui ont la même signification à une forme canonique. Parmi les contenus renvoyant au réel, les contenus pertinents sont sélectionnés. Cette sélection sera arbitraire si elle est fondée sur des critères souterrains, consistante si les critères découlent d'hypothèses explicites. Les fiches ont longtemps constitué une méthode privilégiée de rétention des données sélectionnées. Ses règles de rédaction, fort simples, (format fixe, conventions d'écriture, choix de mots-clés, mise en contexte de l'information, références, etc.) permettent de mener une analyse à grande échelle.

L'analyse consiste à isoler des régularités et des ruptures dans le matériel recueilli. Les fiches sont manipulées pour constituer des piles représentant des inventaires ou des configurations. Deux options méthodologiques sont possibles: l'analyse sera qualitative si peu de fiches considérées très représentatives sont retenues; elle sera quantitative si le plus de fiches possibles sont prises en compte. Cependant plus le nombre de fiches est élevé, plus il devient difficile d'être systématique, les régularités observées étant beaucoup plus le résultat d'une mise en forme de l'intuition que du calcul précis des unités retenues et de leur comportement.

Ce mode d'extraction des données textuelles laisse beau ip de place à l'improvisation. La motivation du lecteur à tendre la main pour prendre une fic ierge et la remplir tient tout à la fois de l'existence d'un seuil déclencheur conjoncturel (dont la règle qui le commande n'est pas clairement formulée) que de l'anticipation de l'importance d'un segment fondée sur l'expertise. Les difficultés reliées à la systématisation de l'extraction sont bien évidemment amplifiées si la tâche est confiée à une équipe de travail. Il est très difficile dans ce cas de s'assurer de l'uniformité de l'extraction tant l'expertise des lecteurs relève ultimement de dispositions intellectuelles et culturelles individuelles, au-delà de l'uniformité relative qu'a pu produire leur socialisation. De plus, il est impossible de valider l'exhaustivité, de vérifier si on a laissé passer de bonnes occurrences.

Nous voilà donc en face de deux caractéristiques principales de l'analyse pré-informatique des textes, caractéristiques inhérentes à l'acte de la lecture lui-même: l'analyse procède d'une lecture experte du texte en vertu, nous l'avons dit, de dispositions intellectuelles et culturelles

acquises, en même temps qu'elle est soumise à l'arbitraire d'un travail ignorant des règles souterraines qui le fo[illegible]ent. La lecture experte souffre donc d'un manque de rigueur rendant sa validation difficile. [illegible]r ailleurs, les procédés conventionnels d'analyse de textes interdisent à toutes fins pratiques le traitement de corpus de grande envergure typique des organisations. A la nécessité de systématiser la lecture s'ajoute donc celle de pouvoir appréhender de grands ensembles textuels. L'ordinateur nous apparaît être le seul outil susceptible de résoudre une part de ces problèmes.

3. L'INFORMATISATION DU REPÉRAGE ET DE L'ANALYSE DES DONNÉES TEXTUELLES

Les avantages d'une extraction des données textuelles basée sur la lecture humaine experte s'accompagnent donc d'inconvénients qu' il importe de palier. Elle n'est ni régulière ni systématique. De plus, il est impossible en cours d'analyse de changer les hypothèses sans avoir à reprendre la démarche à zéro, ce qui empêche une approche constructiviste de l'analyse. Dès l'apparition de l'ordinateur, on a tenté de le mettre à profit pour repérer et analyser les données textuelles en raison de sa rapidité d'exécution et de la régularité avec laquelle les tâches répétitives sont accomplies.

Les méthodologies de lecture des textes au moyen de l'ordinateur proposées aux travailleurs du texte tombent en deux catégories. La première est fondée sur la production et l'examen de listes ordonnées de mots, alors que la seconde tient compte de leur ordre dans le texte. Nous verrons pour chacune d'elles: leur présupposé, le type d'analyse produite, leurs limites et les améliorations qui ont été apportées et celles qui seraient souhaitables.

La lecture lexicale

En premier lieu, l'ordinateur a été considéré comme un outil de calcul; son recours a produit des analyses de textes strictement quantitatives. Le présupposé théorique est que l'ordre des mots n'influe pas sur la signification d'un texte; dans cette perspective, le texte est vu comme une population de mots. Dans un tel contexte, aucune hypothèse d'interprétation n'est nécessaire et un seul critère de repérage des formes significatives est appliqué: toute chaîne de caractères séparée par des « blancs ». Le repérage consiste à utiliser des algorithmes de tri pour produire des listes de mots ordonnées selon des critères alphabétiques ou leurs fréquences d'apparition (lexiques).

L'analyse des textes prend la forme de calculs statistiques décrivant la distribution des mots dans le texte en fonction de leur fréquence ou encore le texte est partitionné et les lexiques différents sont comparés pour établir la distance et la proximité des parties entre elles. Les analyses produites à partir d'une conception du texte exempte de connaissance, tant du système de la langue que du contenu des textes se sont avérées insatisfaisantes. Des améliorations ont été apportées dans plusieurs directions.

Les différentes désinences d'un même mot sont ramenées à une forme canonique (lemmatisation) afin que les fréquences prises en compte lors des calculs reflètent la distribution des mots et non pas leurs flexions. Cette mise à profit d'une connaissance linguistique minimale permet d'opérer une réduction dans le matériel et d'obtenir une plus grande précision. Les formes nominales et adjectivales sont ramenées au masculin singulier; par exemple, les formes *bons, bonne, bonnes* sont étiquetées *bon*. Toutes les formes conjuguées de tous les radicaux des verbes sont ramenées à la forme infinitive; par exemple, les formes *voulais, voudrions, voulu,* etc. sont étiquetées *vouloir*. Ce principe peut être étendu de la morphologie à la sémantique pour que

l'analyse de la distribution ne porte plus sur les unités lexicales mais sur les unités sémantiques et les formes nominales, adjectivales, verbales et adverbiales. Elles peuvent être ramenées à leur radical; par exemple aux formes *volonté, volontaire, vouloir, volontiers* et une même étiquette peut leur être accolée.

Un système de catégories issu d'hypothèses explicites quant à l'interprétation du texte est projeté sur le texte; les dénombrements sont par la suite effectués sur les catégories et non plus sur les mots. Ainsi, par exemple, tous les noms propres désignant des lieux de même que les adverbes de lieu peuvent être regroupés dans une catégorie étiquetée *espace*. Les catégories peuvent être inscrites dans une hiérarchie en vertu de critères théoriques. Une certaine connaissance du contenu du texte est ainsi introduite, ce qui force le lecteur à expliciter, non seulement les éléments textuels susceptibles d'être porteurs de sens, mais aussi d'arrêter les critères à partir desquels ceux-ci seront retenus et comptabilisés.

L'analyse portant sur la distribution des mots dans les sous-textes est complétée par le relevé du co-voisinage de mots tenus pour importants. Des concordances sont effectuées (mots-clés accompagnés de leur contexte) et, pour chacun des mots, un lexique est constitué sur l'ensemble des contextes rapportés. L'examen de la co-occurrence des mots permet de dépister des associations lexicales qui témoignent de la structuration de l'univers notionnel. Cette procédure permet un traitement statistique partiel de la mise en séquence des formes lexicales.

L'interactivité des dernières générations d'ordinateurs a favorisé la lecture plurielle. Les deux étapes consécutives de la lecture, repérage et analyse des données textuelles, peuvent être accomplies de façon cyclique. Il est devenu possible de relire plusieurs fois un texte selon de nouveaux réseaux d'hypothèses, dans la mesure où d'autres éléments pertinents à l'analyse sont identifiés et étiquetés. Sur la base de cette approche « construite » du texte, il deviendra possible, par exemple, de ramener de manière automatique et systématique des formes différentes qui ont la même signification.

Cependant, malgré les améliorations dont elle a fait l'objet, l'analyse lexicale souffre toujours d'importantes lacunes. La matière textuelle se retrouve disloquée au terme du processus informatique, de telle sorte que l'expertise ne peut intervenir que de manière rétrospective pour tenter de donner un sens aux résultats de l'analyse, certes précis et vérifiables, mais coupés du contexte de l'énonciation. Pour palier cet inconvénient, l'intérêt s'est déplacé vers l'utilisation des parseurs.

La lecture syntagmatique

Le projet d'informatiser la lecture humaine par la description grammaticale des phrases d'un texte a été formulé dès l'avènement des langages de programmation dédiés à la manipulation de structures symboliques, tels LISP. L'ordinateur n'est plus perçu strictement comme un puissant calculateur, mais comme un outil de modélisation sophistiqué, capable de gérer et d'accomplir des tâches réservées jusqu'alors au cerveau humain; d'où le terme «intelligence artificielle». Le présupposé qui fonde l'entreprise d'élaboration d'algorithmes de description (parseurs) des phrases est que l'appréhension d'un texte passe par la connaissance de la structure des phrases qui le composent. Il s'agit de segmenter les énoncés dans leurs constituants syntagmatiques, de les identifier et d'expliciter leurs rapports internes.

Il est très tôt apparu qu'il s'agissait d'une tâche très complexe. Le savoir-faire accumulé lors de l'élaboration de compilateurs (procédures visant à traduire des programmes écrits en langages source en instructions machine) ne s'est avéré que partiellement opérant puisque les

langues naturelles ne constituent pas des systèmes fermés, mais ouverts et que l'ambiguïté est présente à tous les niveaux. C'est ainsi que l'analyse des textes a été assujettie à une description linguistique des textes (voir fig. 1).

FIGURE 1:

Analyse linguistique mécaniste

⇩

niveau morphologique
reconnaissance du rôle des mots

⇩

niveau syntaxique
agencement des mots dans la proposition

⇩

niveau sémantique
articulation du développement formel du contenu

⇩

niveau pragmatique
modalités d'énonciation

⇩

Le texte est alors appréhendé comme une superposition de structures. Le niveau morphologique consiste en la reconnaissance du rôle des mots. Le niveau syntaxique proprement dit fait ressortir l'agencement des mots dans la phrase; d'abord l'assujettissement des mots à une tête pour constituer des groupes ou syntagmes, nominaux, prépositionnels ou verbaux; puis les rôles que tiennent les syntagmes dans les propositions; et enfin l'articulation formelle des propositions en phrases. Le niveau sémantique fait correspondre les mots ou syntagmes à des situations du monde: cas (agents, patients, instruments, etc.) rôles discursifs (thème et propos); référence (quantification, détermination, modulation, etc.); modalités (nécessités, possibilité, obligation, probabilité); temporalité. Le niveau pragmatique enfin s'intéresse aux modalités d'énonciation.

Si le modèle linguistique est le plus prometteur, son choix pose de nombreux problèmes. La formalisation des langues naturelles n'est que partielle, il reste des zones obscures non-négligeables telles l'anaphore, la coordination, les formulations incomplètes, etc. Si les théories

du fonctionnement de la langue foisonnent, toutes sont partielles et aucune ne fait l'unanimité. Par ailleurs, la théorie du parsage, c'est-à-dire la façon dont les algorithmes doivent être dessinés, est en devenir et se développe au rythme des tentatives de construction.

Voici un exemple du problème posé par le développement par tentatives. Une description syntagmatique des phrases d'un texte n'est possible que si une catégorisation morphologique des mots est effectuée. Pour effectuer cette opération avec efficacité, il faut mettre sur pied un dictionnaire où l'information morphologique est consignée en regard des mots. Toutefois, un tel dictionnaire ne peut une fois pour toutes être complété, car le type et le format de son contenu et le rapport qu'il devrait entretenir avec les procédures ne sont pas encore fixés; on ignore encore le niveau de sous-catégorisation nécessaire pour un fonctionnement optimal des procédures de parsage.

Les problèmes dont nous venons de faire état n'ont pas empêché le fait que des parseurs ont été construits et appliqués à de grands ensembles de textes. Ils produisent une description arborescente mettant en évidence les relations de dépendance contextuelle des mots. Ces informations permettent la constitution de lexiques de mots qualifiés par la syntaxe. Une analyse de type lexical peut donc être menée en tenant compte de propriétés sémantiques des énoncés. A titre d'illustration, deux exemples ont été retenus: la thématisation et la détermination nominale. Dans le premier cas, un lexique des mots qui occupent la première position de la phrase peut être constitué; il s'agit de ce dont on parle dans le texte. Dans le second cas, comme le déterminant et le déterminé sont distingués, il est possible de constituer pour chacun des mots déterminés un lexique de déterminants, il est aussi possible d'extraire du lexique global les mots qui ne concourent pas directement à la thématique du texte. De même, on pourra dans les deux cas produire pour chacun des mots des indices de thématisation et de détermination.

Ces tentatives d'utiliser la description syntagmatique dans des analyses de données textuelles connaissent un succès mitigé. Comme nous l'avons souligné le fonctionnement des parseurs n'est pas conforme, la plupart du temps, aux principes acceptés par les linguistes, ayant été en grande partie « bricolés » par accumulation d'heuristiques. Il en résulte que leur fiabilité est douteuse, et que leur architecture est difficile à rectifier. Les programmes informatiques qui les mettent en oeuvre étant complexes et écrits dans des langages évolués mais non-performants, les temps de réponse sont longs, ce qui rend le traitement de grandes masses lourd et leur coût parfois prohibitif. En raison de l'aspect normatif des règles constituant le savoir-faire des parseurs, la description produite ne convient qu'aux expressions bien formées. Quant à la description structurelle produite, les règles de son interprétation demeurent à produire.

Par ailleurs, comme les « travailleurs du texte » sont absents des équipes qui élaborent les parseurs, les préoccupations des linguistes priment. Ceux-ci ont tendance à entretenir un rapport réflexif à l'outil et à considérer le parseur comme un banc d'essai pour valider des hypothèses théoriques sur quelques phrases choisies. L'exhaustivité et la complexité sont les caractéristiques recherchées alors que la complétude et la couverture importent peu.

Les contributions à la théorie du parsage étant trop nombreuses et pointues pour être exposées ici dans le détail, seules les tendances générales sont évoquées. A l'instar des systèmes experts qui séparent le savoir exprimé sous forme de règles d'inférence du moteur qui les invoque, le savoir linguistique est de plus en plus tenu à part du mécanisme informatique qui le met en oeuvre. Il est exprimé de façon modulaire et lisible de telle sorte qu'il puisse aisément être relu et révisé. Dans la foulée du courant de l'informatique de l'usager final, des progiciels simples et conviviaux pour la génération d'analyseurs ont été développés afin que les linguistes puissent, à la suite d'un léger entraînement, participer directement à l'élaboration de parseurs.

En dernière analyse, il nous semble que les parseurs, tributaires de la linguistique computationnelle, pour les besoins de l'analyse des données textuelles, font trop et trop peu à la fois. Le niveau de complexité et d'exhaustivité de la description syntaxique visé, mais difficilement atteignable dans un avenir prévisible, n'est pas nécessaire. En effet, l'analyse

cherche des indices en termes de régularités ou de ruptures textuelles, elle indique des tendances et caractérise des ensembles d'énoncés pris globalement. Ainsi les parseurs conviennent à l'étude raffinée de l'énonciation, mais négligent les macro-structures textuelles qui dénotent l'anatomie du texte, la stratégie discursive qui y est mise en oeuvre.

Pour une lecture experte assistée par ordinateur

Face à l'ampleur des problèmes énoncés plus haut, nous esquissons quelques pistes qui nous semblent en mesure d'arrimer la production d'outil aux besoins des « travailleurs du texte ». Sur le plan théorique, le modèle linguistique qui s'avère trop centré sur la langue devrait être assujetti à un modèle textuel qui reste à formaliser. Les propositions pour une morphologie discursive (A. Lecomte et J.-M. Marandin), développées dans le cadre de travaux en analyse du discours, nous semblent constituer un point de départ prometteur.

L'analyse morphologique du discours repose en grande partie sur l'hypothèse selon laquelle les énoncés d'un discours se présentent comme des formes d'objets-noyaux aux configurations régulières. Analyser la morphologie d'un discours revient à construire un modèle général du texte en repertoriant à travers les strates du discours la manifestation des objets de schématisation et, au-delà des limites strictes de la phrase, en reconstituant les itinéraires sémantiques que ces objets empruntent. Les schématisation sont des opérations qui structurent des objets cognitifs et les articulent dans l'espace d'un savoir (référenciation). Ces opérations sont toujours tributaires de circonstances spécifiques, soit la pratique sociale qui en détermine les conditions de possibilité.

En plus du système de relations de dépendance contextuelle, les objets de schématisation sont inscrits dans un système de relations de transformation d'objets, de relations méta-fonctionnelles (l'introduction d'un texte, d'un auteur, ...), etc. Les objets d'une schématisation sont récurrents, étant constamment repris et reformulés par les interlocuteurs tout au long du processus discursif. Le processus par lequel les unités sémantico-cognitives faisant référence au réel sont stabilisées à l'intérieur de formes linguistiques pour constituer des schématisations, est appelé ancrage. Les ancrages nominaux matérialisent les objets en décrivant leurs propriét . Les ancrages verbaux fournissent les éléments de la dynamique des objets: leurs relations.

Ce type d'analyse du discours exploite la particularité du langage naturel d'être à lui-même son propre métalangage, c'est-à-dire qu'il sert à la fois à représenter la réalité et à représenter la représentation de la réalité. Ceci justifie une lecture par extraction et échantillonnage de segments de texte (en termes techniques on parle de « thématisation par spécification »), tenus pour représentation canonique des enjeux importants du discours. Ces segments, articulés les uns aux autres, forment un nouveau texte se donnant comme résultat de l'acte d'interprétation. Une grammaire discursive du texte analysé est en quelque sorte mise au point progressivement.

L'automatisation de la lecture des textes nous apparaît être un objectif impossible à atteindre dans un avenir prévisible. C'est pourquoi nous proposons de remplacer cet objectif mécaniste pour fournir une assistance au lecteur expert afin d'accroître l'efficience du processus en termes de consistance et de rapidité. Qui plus est, une description exhaustive mais statique des textes générés de façon déterministe, même si elle était sans faille, ne serait que partiellement utile pour réaliser l'analyse de grandes surfaces textuelles. Dans une telle perspective, l'investigation des régularités et des ruptures textuelles se fait par accumulation d'indices de plusieurs natures, telle l'agglomération d'items lexicaux en certains points stratégiques du texte, les procédés stylistiques, etc.

Une approche interactive à l'analyse de textes où la dimension heuristique prime nous semble préférable. L'analyse prend alors la forme d'une démarche cyclique composee d'autant de boucles extraction / validation que jugées nécessaires; les résultats obtenus guidant la suite des

opérations. La gouverne (contrôle) des opérations est donc laissée à l'expert lecteur. En somme, à la moulinette requérant une confiance aveugle, nous préférons la calculette où les manipulations répétées, libres et variées augmentent la créativité de l'utilisateur.

Pour correspondre aux caractéristiques exposées précédemment, l'architecture informatique souhaitable prend la forme d'un atelier « textuel » où dans un univers intégré coexistent un analyseur lexicographique et des sous-parties de parseurs, notamment pour décrire les séquences nominales et rattacher celles-ci aux séquences verbales (voir fig. 2). Au lieu d'une description arborescente de chacune des phrases qui s'avère lourde et difficile à valoriser, les résultats que nous visons prennent la forme de topographies: des inventaires, des classifications ou encore des partitions du texte selon des critères internes. Nous introduisons le terme topographie car la catégorie d'espace nous apparaît importante pour décrire les relations qu'entretiennent les objets de schématisation.

FIGURE 2:

Assistance à la lecture experte

catégorisation
morphologique
conceptuelle

Analyse

des constituants
blocage des locutions
distribution lexicale

des relations
séquences nominales
détermination
séquences verbales
cas

topographies
inventaires, classifications, partitions

En attendant que les parseurs pouvant être appliqués à n'importe quel texte en produisant des descriptions linguistiquement fiables soient disponibles, nous préconisons une solution mixte qui consiste à faire des analyses lexicographiques qui tiennent minimalement compte de la distribution positionnelle des mots dans les phrases. Nous expérimentons présentement l'application informatique des principes de la morphologie textuelle par l'extension du calcul de co-occurrence basé sur une catégorisation morphologique. Ceci nous permet d'ores et déjà d'assister le dépistage et le blocage des locutions, c'est-à-dire les unités sémantiques, appelées termes, composées de plusieurs mots qui, pris séparément, ont chacun une signification (par ex.: *traitement de texte*). Cette opération apporte une rigueur accrue à l'analyse des constituants.

En conclusion, il nous apparaît essentiel de ne pas plier la méthode d'analyse des textes pré-existante aux impératifs techniques de l'ordinateur, de refuser que leur langage d'exploitation parasite la méthodologie. De même, la discussion sur la primauté d'un type d'outil sur l'autre doit être modifiée en faveur de l'enrichissement mutuel de leur apport. La portée de leur intervention doit être calibrée en fonction de la méthodologie.

BILAN D'UN AN D'OBSERVATION ET D'ACTION EN INDUSTRIES DE LA LANGUE AU PLAN FRANCOPHONE

André Abbou
Observatoire français des industries de la langue

Sans vouloir accorder une importance démesurée aux questions de forme, il n'est pas dépourvu d'intérêt de relever que la désignation même de l'objet dont nous allons nous entretenir au cours de ce colloque, à savoir le traitement de la langue naturelle en vue d'applications industrielles, est sujet à des appellations diverses: linguistique informatique, informatique linguistique, langue et informatique, industries de la langue. Si les trois premières appellations se contentent d'emboîter, de façon plus ou moins précise et claire, les théories, les procédures et les hypothèses de traitement automatique des composantes de la langue, il en va autrement de la quatrième.

Elle est, tout le monde le sait, d'usage récent. Juin 1983, c'était hier ou presque. Au cours du colloque COFORMA, nous avions risqué, eu égard aux premiers produits apparus sur le marché et aux besoins dont on commençait à percevoir la nature et l'étendue, la dénomination d'industries de la langue. Elle était commode. Elle laissait le champ libre à toutes les innovations et à toutes les évolutions de ce terrain à peine reconnu. Mais elle était elliptique et cavalière. Elle posait comme acquis que la langue avait généré une industrie, comme le cuir ou la locomotion automobile l'avaient fait à des époques antérieures. On pouvait cependant trouver à l'expression une double justification. Le terme de langue signale que celle-ci, en tant que matériau et véhicule, est entrée dans un processus de transformation destiné à l'introduire dans des systèmes complexes d'information et dans des outils évolués, appelés à remodeler de façon importante et durable, l'organisation de secteurs d'activité aussi vitaux pour une société que la bureautique, la domotique, la productique, la conception et la fabrication assistées par ordinateur, l'archivage et la documentation, les éditions imprimées et audiovisuelles, l'électronique, la médecine et la rééduction des nandicaps moteurs, la traduction, etc.

Le terme d'industrie souligne que les services et le produit ainsi conçus donnent lieu à l'exercice d'activités conjointes et standardisées, telles que la transformation de matériaux, la conception, l'élaboration et la confection des systèmes ou de dispositifs d'assistance nouveaux. Les industries de la langue concernent donc les techniques, produits, activités et services qui s'appuient sur un traitement de la langue naturelle.

Les explications ne sont survenues, comme il se doit, qu'une fois le terme et le concept lancés. Ils étaient, on le distingue encore plus nettemment maintenant, prématurés. Car il a fallu, entre décembre 1986 et juillet 1987, s'éloigner des balbutiements, des survols cavaliers et erronés et des évaluations fantaisistes. De façon analogue aux symbolistes qui voulaient reprendre à la musique "leur bien", il a fallu scruter les projets et les produits industriels français ou étrangers, impliquant de près ou de loin un traitement de la langue naturelle, les programmes de la Communauté économique européenne (Esprit I et II, Eureka notamment), pour y rechercher la présence éventuelle des composants "industries de la langue", en décrire la nature, les fonctions, les utilisations, etc. De même, il a fallu tenter de décrire, à un moment donné, l'état de l'offre et l'état prévisionnel de la demande, les marchés confirmés et ceux projetés par les instituts de consultants internationaux.

Aujourd'hui, l'expression paraît plus communément admise. Le périodique *Language Technology* s'est explicitement donné comme sous-titre "Magazine des industries de la langue", le colloque de l'INRIA en décembre 1987, a choisi, pour désigner l'objet de son travail, le vocable

d'industrie de la langue, et la CEE, depuis mars 1988, parle couramment, pour annoncer ses réunions et ses programmes, d'industries de la langue. L'expression semble donc désormais consacrée.

Cette démarche paraît avoir été la bonne car nous avons pu ainsi mettre en place les réflexions et les actions de Réseau francophone des industries de la langue.

1. LE RÉSEAU FRANCOPHONE DES INDUSTRIES DE LA LANGUE

■ *PREMIER EXERCICE 1987* (Suivi du Sommet de Paris)

Il a donc connu, ce n'est pas un secret, bien des vicissitudes. Le premier Sommet des Chefs d'État et de gouvernement des pays ayant en commun l'usage de la langue française, qui s'est tenu en février 1986 à Paris et qui a créé le Réseau des industries de la langue, avait suivi les recommandations du rapport introducteur, validé un certain nombre de propositions sensées mais inapplicables. Elles réclamaient annuellement un montant de 88MFF, soit, à elles seules, la totalité du budget des cinq réseaux mis en place par le Sommet. Autant dire que ces propositions se coupaient par là de toute possibilité de mise en oeuvre et manquaient de sousbassement puisque l'*Étude d'opportunité, de faisabilité et de mise en marché* n'avait été ni entreprise ni envisagée.

Il a fallu, dans le cadre que vous connaissez, à savoir la mise à disposition des Réseaux de 30% en 1987 du budget "Intervention" de l'ACCT, tenter d'opérer. Le Réseau des industries de la langue, à l'instar des autres réseaux, reçut une enveloppe de 3,2 MFF, dont à peine 1 MFF put être utilisé, puisque l'ACCT avait, au moment où on lui annonçait cette réquisition, déjà engagé ses crédits.

Il n'empêche que quatre réunions purent se tenir avant juillet 1987, qu'on y débattit des questions préalables, des secteurs d'intervention à privilégier, d'études de faisabilité pour des projets importants, que les *Études d'opportunité, de faisabilité et de mise en marché* furent conduites en France[1] et en Belgique, que des séminaires purent se tenir dont certains prolongèrent leurs effets jusqu'en mars 1988 (séminaire international de TAO réuni à Paris, Séminaire francophone sur les questions de formation en industries de la langue). Fin juin 1987, le programme à présenter au Sommet de Québec était rédigé et avalisé par les instances préparatoires, le budget d'opérations estimé à 15 MMF et les dossiers à soumettre aux Chefs d'État et de gouvernement en cours de constitution.

Le Sommet de Québec se tint, valida les propositions et enregistra les promesses de versement des États au fonds multilatéral industries de la langue. La France, qui avait promis 7,5 MFF, en versa finalement 4,6 MFF en 1988 - faute d'un concours ministériel classé comme multilatéral, mais bilatéral en réalité - et le Québec s'acquitta de 900 000 FF, le Canada de 200 000 FF. Il manquait donc 9,3 MFF.

■ *DEUXIÈME EXERCICE* (Suivi du Sommet de Québec)

Il débuta par le renfort de Bernard Quémada, directeur de l'INALF (Institut national de la langue française) promu responsable de Réseau, dont le rédacteur de ces lignes, à qui avait échu le rôle de délégué permanent du Réseau et de responsable-adjoint du Réseau en 1987, devint l'adjoint.

---

[1] ABBOU, LEFAUCHEUR, M[illegible]t, Les industries de la langue: applications industrielles du traitement de la langue par les machines, Volu[illegible]t 2, DAICADIF, septembre 1987.

Ces deux animateurs tentèrent dans un premier temps d'obtenir des crédits complémentaires, notamment des pays industrialisés ayant peu ou pas du tout versé de contribution. La notion des lignes budgétaires disponibles ("aide au développement" par exemple) se révéla un obstacle insurmontable.

Le Réseau tint cependant trois réunions en 1988 (Bruxelles en mars, Montréal en juin, Rabat en novembre), organisa son travail, incita à la mise en place des Observatoires des industries de la langue dans les pays ou régions concernés. Les observatoires chargés de recenser, trier, et interpréter toutes les informations, d'informer et de conseiller les comités d'experts nationaux, constituent et constitueront de plus en plus la colonne vertébrale de toute politique nationale et internationale au plan des industries de la langue. La preuve en est que la CEE, quand elle a envisagé de lancer un programme "industries de la langue" en 1990, a lancé un appel d'offres destiné à constituer, sous son patronage, un observatoire européen des industries de la langue. Ainsi, s'est mis en place en France, et officiellement, l'Observatoire français des industries de la langue dès le 4 janvier 1988. Ainsi se sont constitués ou sont en cours de constitution des observatoires québécois, canadien, belge, suisse et africain.

Avec des moyens réduits, Le Réseau - après appels d'offre et présentation par les Comités nationaux - a agréé 20 à 25 projets, auxquels il a accordé des dotations - récupérables sous forme de fourniture d'un nombre variable d'exemplaires des produits agréés et financés à hauteur maximale de 40% - dans des domaines variés, en rapport avec les trois axes d'action retenus.

A) Recherche-développement industriel

domaines intéressés:

- TAO (bilan de l'offre et analyse des systèmes)
- EAO
- Communication parlée (bases de données des sons du français avec variétés socio-géographiques)
- Bureautique (détecteur contextuel de fautes d'orthographes - pour traitement de texte)
- Aide à l'analyse et à l'interprétation des textes (systèmes de dépouillement terminologique par ordinateur et dispositif informatisé de sondage des flux d'information)
- Analyse et évaluation de l'offre en outils et services dérivés du traitement de la langue naturelle (Répertoire des outils et services)
- Création d'un réseau des observatoires nationaux des industries de la langue (concertation, outils et procédures de travail)

B) Néologie - terminologie

Inventaire actualisable des travaux terminologiques, productions terminologiques et néologiques, concertation et action du Réseau international de néologie-terminologie.

C) Formation - perfectionnement

Inventaire des centres de formation et des programmes disponibles, sélection des programmes et des centres, appel à candidature, sélection et affectation des étudiants en cours de spécialisation, dotation en bourses et en frais de formation.

Nous n'évoquons que pour mémoire la préparation du programme 1989 du Réseau, la préparation des propositions des programmes et des budgets à soumettre au Troisième Sommet (Dakar, mai 1989).

Voilà un bilan précis de l'action du Réseau au plan francophone.

2. PROBLÈMES EN SUSPENS E. QUESTIONS POSÉES

Le travail à conduire est immense, on le perçoit. Mais les politiques nationales et multilatérales des pays francophones butent sur des difficultés administratives et financières.

a) difficultés administratives:

Les industries de la langue sont un domaine qui transcende des découpages théoriques (disciplines) et les sphères d'action des ministères en cause. La Recherche et la technologie, l'Industries, le Commerce intérieur et extérieur, la Francophonie, la politique de la langue (maintien "un potentiel technologique important permettant de conserver l'usage du français comme . d'information et de traitement de l'information), l'Education et la Formation professionnelle. Ces ministères n'ont pas tous les mêmes optiques. Il est parfois difficile aux services de ces ministères de comprendre que toute politique qui vise, à court terme, à privilégier le "business" sur la constitution de potentiels scientifiques et technologiques, va à l'échec. Les erreurs répétées en matière de filière électronique et de filière informatique au plan français le prouvent. L'échec de THOMSON et de MATRA en micro-informatique est encore dans toutes les mémoires. L'impulsion pour briser ces autarcies ne peut venir que des programmes internationaux (communauté européenne, communauté francophone).

Au plan francophone, dans la mesure où le Réseau et les Observatoires recoupent produits, utilisations, technologies, effets culturels, sociaux et professionnels, ils assurent une veille technologique et linguistique. 80 à 85% des produits et services conçus au plan européen concernent principalement ou ne concernent que la langue anglaise.

b) difficultés financières

Les budgets d'intervention sont faibles. Les degrés d'information et les préocupations divergentes ne permettent pas à une volonté politique de s'exprimer clairement. Tout le monde sait que les industries de la langue intéressent l'avenir et l'existence de la communauté ayant en commun l'usage du français. A croire, ou laisser croire, que cette communauté n'a que des problèmes d'auto-suffisance - très importants d'ailleurs - et à ne mobiliser que des budgets en relation avec l'aide au développement, c'est se tromper de perspective, d'objectif et de vocation de la communauté de ces pays.

Faute de crédits d'intervention, faute de volonté politique clairement démontrée, faute de marchés organisés et viables, les industriels courent au plus pressé. La survie ou le développement de leurs activités leur commandent de s'adapter et de suivre l'évolution des marchés, de produire ce qui est vendable et exportable, donc ce qui tourne le dos à une recherche-développement et à des investissements onéreux pour la création d'une industrie de la langue, francophone ou multilingue. Si les industriels manquent d'esprit de risque et paraissent frileux en ce domaine, c'est que l'environnement les y porte.

CONCLUSION

La création et l'organisation d'une industrie de la langue à usage francophone et multilingue paraissent encore sujettes à caution. Tout dépendra de la clairvoyance, de la levée des barrières administratives et de la volonté politique qui s'exprimera.

Pour ma part, il y a deux mots que je n'emploierai désormais qu'après mûre réflexion. Ce sont ceux d'enjeu et de défi. Ces mots, ballotés et galvaudés, prennent l'allure de tartes à la crème et vont à l'encontre des clarifications souhaitées. Les défis sont ces actes de foi et de volonté qu'on se donne à soi-même, compte tenu d'une conscience aiguë des risques et des conséquences d'une situation défavorable. Les enjeux procèdent de l'existence d'objectifs clairs, de projets non ambigus, et du souci constant de détecter derrière les situations imposées et les comportements conditionnés, les effets et les perspectives de leurs conséquences.

Mais à la base, il faut une conscience claire et de ce que l'on est, et de ce que l'on veut, et de ce qu'il est encore possible d'entreprendre. Après viennent les questions de cohérence et de moyens d'action.

"Celui qui regarde longtemps ses rêves finit par ressembler à son ombre" rapportait MALRAUX, citant un proverbe asiatique. Il serait regrettable que la communauté des pays ayant en commun l'usage de la langue française, et par voie de conséquence la langue française elle-même, se perdent dans les volutes du songe.

LA CARTOGRAPHIE LEXICOGRAPHIQUE DES AVIS OFFICIELS

Jean-Claude Boulanger
Université Laval

1. INTRODUCTION

L'un des aspects récents et peu explorés de la lexicographie française contemporaine concerne les rapports que les dictionnaires généraux monolingues (DGM) entretiennent avec les avis officiels de recommandation et de normalisation issus des travaux des commissions ministérielles de terminologie françaises et québecoises. De prime abord, les avis sont destinés à des groupes de spécialistes appelés à manipuler des vocabulaires spécifiques au cours de leurs activités professionnelles. Dès le moment où les lexicographes généralistes les prennent en charge, le public destinataire des DGM les consulte; à tout le moins, il les a sous les yeux lorsqu'il ouvre un dictionnaire.

Trois concepts doivent être circonscrits pour saisir pleinement le cheminement des unités lexicales cautionnées qui, au sortir des officines d'État, sont récupérées ou non par les dictionnaires de langue. Il s'agit des concepts de «politique linguistique», d'«aménagement linguistique» et d'«avis officiel».

Une politique linguistique est une décision d'ordre législatif qui concrétise l'intérêt de l'État pour le domaine de la langue. Cet intérêt est largement répandu dans le monde contemporain. Une étude récente montre que plus de la moitié des États souverains du monde sont intervenus dans le champ langagier et plus particulièrement dans le secteur de l'affichage (voir Leclerc 1988). L'aménagement linguistique est un processus d'intervention étatique volontaire en vue de planifier et de modeler le changement linguistique dans une société. L'un des objectifs de l'aménagement est de façonner la langue elle-même, soit en la décrivant, soit en l'enrichissant du point de vue lexical. Les manifestations de la standardisation et de la description de la langue trouvent leur accomplissement dans le dictionnaire et la grammaire. L'intervention dans les langues de spécialité (LSP) est dite *aménagement terminologique*. Enfin l'avis officiel est un document de nature institutionnelle, émanant d'une autorité mandatée pour intervenir dans la langue, et portant à la connaissance du public et des usagers les décisions prises à l'égard d'un terme, d'un groupe de termes, d'un plus vaste ensemble d'unités, un dictionnaire terminologique par exemple, etc. Avant d'être relayés par divers canaux médiatiques et de parvenir au catalogue des mots du dictionnaire, les avis des commissions ministérielles françaises ou québécoises sont d'abord publiés dans les organes parlementaires de chaque État (*Journal officiel* en France et *Gazette officielle du Québec*). Outre l'ensemble des décisions, le terme *avis officiel* désigne chacune des unités qui est l'objet d'une sanction de normalisation ou de recommandation.

L'activité étatique et gouvernementale déployée autour de ces trois différents axes interventionnistes a des répercussions de plus en plus visibles et tangibles dans les DGM. Depuis une quinzaine d'années, en effet, les répertoires lexicaux prennent une importance accrue en tant que courroie de transmission des décisions ministérielles à caractère technolectal. De fait, l'une des missions du dictionnaire consiste à banaliser l'usage des termes entérinés par des autorités. Avec comme conséquence, que le dictionnaire peut contribuer à installer ou à maintenir dans

l'usage des formes élues par le groupe socioprofessionnel responsable de l'intervention (ex. *logiciel, matériel, listage, didacticiel, sortant*). Ceci ne signifie nullement que des formes jugées comme étant répréhensibles ou à remplacer n'ont plus de vigueur ou d'adeptes dans certaines circonstances du discours, y inclus le discours lexicographique lui-même. Voir par exemple la définition de *donnée* (Annexe 6.1) dans le *GRLF* qui maintient la forme *digitale* alors qu'à l'article *digital* du même dictionnaire il est dit: «Rem. On recommande officiellement l'adj. *numérique* pour remplacer cet anglicisme, qui crée en français des confusions avec 1. *digital* [...].»

Le dictionnaire a depuis belle lurette la responsabilité de véhiculer le bon usage et une certaine vision de la norme. Depuis l'origine de la lexicographie française, les rédacteurs de dictionnaires sont sans cesse à l'écoute de ceux qui font la langue, qu'il s'agisse de personnes, de groupes ou d'institutions. Cela fait partie de la nature du dictionnaire, de sa vocation, de son aspect captif. Quant aux commissions ministérielles de terminologie, elles constituent des académies modernes dont les travaux méritent un coup d'oeil circonstancié afin de faire le point sur leur impact dans le public général.

La normalisation organisée, institutionnalisée, planifiée remonte à un peu plus d'une quinzaine d'années. La mise en activité des mécanismes français et québécois d'interventions ministérielles date respectivement de 1972 pour la France et de 1978 pour le Québec. Les premiers décrets de normalisation figurent au *Journal officiel* français du 18 janvier 1973 tandis que les premiers avis de l'Office de la langue française sont diffusés par la *Gazette officielle du Québec* le 26 mai 1979.

À partir de 1975, les officialismes sont introduits dans les dictionnaires de langue. Depuis leur nombre augmente régulièrement. L'exemple des répertoires pionniers (*Lexis* et *Petit Robert*) gagne maintenant toutes les entreprises lexicographiques et toutes les catégories de dictionnaires de langue: les dictionnaires pour les enfants (*Dictionnaire CEC jeunesse*), les dictionnaires pour les collégiens (*Micro-Robert*) et les grands dictionnaires (*GRLF*). Les formes lexicales estampillées, sanctionnées viennent manifestement perturber la macrostructure et plus visiblement encore la microstructure des DGM. Elles offrent aux lexicographes un nouveau réservoir dans lequel ils peuvent puiser des entrées nouvelles tout comme elles requièrent un traitement adéquat dans l'article, au même titre que l'étymologie, le réseau analogique, les citations, quand il y a lieu.

2. LES DISCOURS DICTIONNAIRIQUES

Les attitudes des lexicographes envers les avis officiels se répercutent dans deux genres de discours dans les dictionnaires: le discours prélexicographique et le discours lexicographique.

2.1 Le discours prélexicographique ou commercial est celui qui se trouve en ouverture des dictionnaires. Il s'agit des préfaces, introductions, présentations, etc. En principe, c'est le lieu où le lexicographe définit ou explique la position de son équipe de rédaction à l'égard des différents types de mots qu'il traite (régionalismes, néologismes, emprunts, notamment les anglicismes, avis officiels, etc.) et les critères de choix qui président à la sélection. Un examen attentif de ces discours dans sept dictionnaires courants, tous parus depuis la fin des années 70, dessine un portrait assez juste de la situation. Des sept ouvrages suivants: *GRLF, PR, LEXIS, PLI89, DHLF, DFH* et *DFP*, seuls les quatre premiers se prononcent sur les termes recommandés. J'illustre par les textes du *PLI89* et du *PR* (1986) l'opinion de deux équipes de lexicographes français.

Dans *PLI89*, il est stipulé: «Les recommandations de l'Académie française ont été mentionnées chaque fois que l'état d'avancement des travaux du Dictionnaire nous l'a permis. [...] Les recommandations officielles en matière de terminologie ont été mentionnées chaque fois qu'elles existaient» (p. 6). Cette dernière remarque peut s'interpréter de deux manières: les termes recommandés ont tous leur place dans la nomenclature tout en étant affublés de la marque d'officialisation; les termes recommandés ont leur place dans la nomenclature avec ou sans indice d'officialisation. Dans les faits, aucune des explications ne prévaut, car sur 36 termes du corpus, 31 seulement sont retenus dont 5 sont marqués (voir le tableau 1).

Dans le *PR*, il est précisé que: «Le Petit Robert signale les «recommandations officielles» françaises (recomm. offic.), soit sous l'emprunt, soit, lorsqu'elles semblent effectivement en usage, à l'ordre alphabétique. *Bulldozer* malgré l'existence d'un remplaçant officiel *bouteur*, demeure dans l'usage; *matériel* et *logiciel* concurrencent heureusement *hardware* et *software*, que la description ne peut, par ailleurs, négliger. Les termes approuvés par arrêtés ministériels -- à partir des arrêtés du 12 janvier 1973 -- ont été mentionnés dans le dictionnaire dans la mesure où ils remplaçaient un anglicisme figurant à la nomenclature, et quand leur emploi était effectif, ou probable dans les années à venir. La publication exhaustive et commentée des termes officiellement approuvés relèverait d'une autre perspective, ouvertement normative, que nous n'avons jamais adoptée» (1986, p. XVIII-XIX).

Quant au seul dictionnaire québécois qui aurait pu se prononcer sur le sujet, le *DFP*, il demeure muet. Sur un total de 21 pages imprimées formant quatre textes différents, il n'est nulle part question des décisions officielles de l'Office de la langue française, ni de ses trois énoncés de politique linguistique portant respectivement sur l'emprunt de formes linguistiques étrangères, sur les québécismes et sur les titres et fonctions au féminin. Pourtant, à l'intérieur des articles du *DFP*, il est fréquemment question des avis linguistiques et terminologiques de l'OLF (ex. 1: *crédit*, sens 7: «Unité de valeur dans l'enseignement universitaire et collégial. [...] REM.: L'OLF recommande d'employer plutôt *unité*; ex. 2: *académique*, sens 3: *Année académique*: temps qui s'écoule entre le début et la fin des classes, des cours. REM. L'OLF recommande de remplacer ce terme par *année scolaire* ou *année universitaire*, selon le cas). [Pour un examen plus poussé des discours introductifs, on se reportera à Boulanger 1988c.]

2.2 Le discours lexicographique n'a pas la souplesse du discours d'introduction. Il est plus rigide, plus codé puisqu'il constitue une armature sur laquelle se greffe l'information à transmettre. La grille de synthèse s'est développée et fixée au fil des siècles permettant une présentation des données suivant un ordonnancement bien précis des rubriques. L'ajout d'une information nouvelle, comme celle qui concerne les recommandations ministérielles, peut donc perturber la physionomie séculaire de l'article.

Il est utile de s'arrêter sur quelques brefs constats pour illustrer ce phénomène. Ils procèdent du plus général au plus particulier, sans être exhaustifs.

1. Tous les dictionnaires de langue récents incorporent un nombre plus ou moins élevé d'avis ou ils marquent du sceau d'officialisation des unités déjà traitées dans les articles.

2. Aucun répertoire ne catalogue ou n'identifie l'ensemble des unités scrutées par les commissions de terminologie, peu importe l'arrêté en cause.

3. Un discours codé, c'est-à-dire une terminologie particulière s'est créée pour rendre lexicographiquement compte de l'interventionnisme étatique dans les lexiques spécialisés. À titre d'exemple, voici une série de verbes relevés dans les articles de quelques dictionnaires: *recommander, remplacer, préconiser, conseiller, proposer.* Ces termes prennent une coloration sémantique axée sur la terminologie lexicographique déjà disponible, comme c'est le cas des unités *normaliser, franciser, traduire*, également repérées dans les microstructures. Cette terminologie est produite par les lexicographes afin de pouvoir discourir sur le phénomène d'officialisation des termes. La plupart du temps, elle renvoie explicitement à l'autorité normative tout en dégageant le rédacteur de la responsabilité de l'intervention.

4. Le discours reflétant l'officialisation d'une unité lexicale niche à peu près dans n'importe quelle rubrique microstructurale. Contrairement aux autres rubriques au contenu et à la place fixée à l'avance dans chaque article, l'indicatif de l'officialisation d'un terme apparaît au petit bonheur pour le moment. Aucun dictionnaire n'a établi ou proposé de politique cohérente à ce sujet. De fait, la notation des officialismes introduit une nouvelle marque prescriptive dans la tradition lexicographique française. Même s'il n'est pas récursif dans chaque article, pour des raisons évidentes, l'indice en question est néanmoins très présent et il joue un rôle suffisamment déterminant pour qu'on songe à lui attribuer une dénomination qui le personnalise et qui confirme sa place et son utilité au sein du vocabulaire lexicographique. Sur le modèle des autres dénominations, je suggère d'appeler *officialisation* ou *label* ce nouvel élément du discours lexicographique codé. La rubrique ou la marque d'officialisation ou de label rejoint ainsi des congénères comme la datation, la définition, la citation.

3. LE TRAITEMENT MICROSTRUCTURAL

Rien ne distinguant réellement les officialismes retenus en entrée des autres formes-vedettes, c'est le contenu de l'article qu'il convient de scruter pour recueillir les indications idoines. Afin d'illustrer le processus, j'ai constitué un minicorpus de termes extraits de l'*Arrêté du 22 décembre 1981 relatif à l'enrichissement du vocabulaire de l'informatique* (voir *DNO*, 1984, p. 401-406). Cet arrêté répertorie 54 entrées: 34 formes simples (dont 1 est accompagnée d'une variante (*visu* ou *visuel*) et 1 autre d'un synonyme (*tirage* ou *fac-sim*)), 1 entrée à trois volets morphologiques (*bi-, tri-, multiprocesseur*) et 19 syntagmes terminologiques. J'ai choisi d'examiner les 36 termes simples qui demeurent après l'addition des co-entrées et l'élimination de la forme gigogne. Le tableau qui suit (tableau 1) montre la distribution et le traitement des termes dans trois dictionnaires publiés par des éditeurs différents: le *GRLF*, le *DFP* et le *PLI89*. Le terme a été marqué d'un [+] lorsque la forme et le sens de l'entrée renvoyaient à l'informatique; le signe [-] indique donc que le dictionnaire ne consigne pas le sens informatique de l'entrée.

Des 36 termes examinés, le *GRLF* en retient 33, le *PLI* 31 et le *DFP* 25, ce qui illustre bien l'importance du vocabulaire de l'informatique dans les DGM. Cette terminologie se banalise de plus en plus; elle rejoint l'ensemble des usagers, ce que les dictionnaires ne peuvent ignorer (voir Boulanger, 1988b). Parmi les unités traitées, 23 sont reconnues par les trois dictionnaires. Seul le terme *fac-sim* est laissé de côté par l'ensemble des répertoires. Parmi les 36 entrées, le *GRLF* en officialise 12, dont 2 indirectement puisqu'il signale l'équivalent anglais sans faire allusion à la recommandation française ou à la solution de remplacement comme dans les autres cas (voir Annexe 6.2). Pour le lecteur non averti, il est quasi impossible de déchiffrer le message derrière la référence cachée. Le *PLI* marque 4 termes tandis que le *DFP* en étiquette 3. *Logiciel. matériel* et *numérique* sont les seuls à faire l'unanimité du point de vue du label.

TABLEAU 1:

Statut des avis

Dictionnaires	GRLF			DFP			PLI		
Termes	Entrée	Officialisation	Localisation	Entrée	Officialisation	Localisation	Entrée	Officialisation	Localisation
autonome	+	-	RA-AI	-	∅		-	∅	
bureautique	+	-		+	-		+	-	
compatibilité	+	-		-	∅		-	∅	
disquette	+	-		+	-		+	-	
donnée	+	-	RA-AI	+	-		+	-	
fac-sim	-	∅		-	∅		-	∅	
incrément	+	-		+	-		+	-	
infographie	+	-		-	∅		+	-	
information	-	∅		+	-		+	-	
informatique	+	-		+	-		+	-	
instruction	+	-		+	-		+	-	
interactif	+	-		-	∅		+	-	
interface	+	+	RI	+	-		+	-	
listage	+	+	CD	+	-		+	+	DM
lister	+	-		+	-		+	-	
logiciel	+	+	RI	+	+	RI	+	+	RI
matériel	+	+	RI	+	+	RI	+	+	RI
mémoire	+	-		+	-		+	-	
microprocesseur	+	-		+	-		+	-	
multiprogrammation	+	-		+	-		+	-	
multitraitement	+	+	RI	-	∅		+	-	
numérique	+	+	CD	+	+	RI	+	+	RA
photostyle	+	+	CD	-	∅		+	-	
portabilité	+	-		-	∅		-	∅	
processeur	+	-		-	∅		+	-	
progiciel	+	-		+	-		+	-	
robotique	+	-		+	-		+	-	

Dictionnaires	GRLF			DFP			PLI		
Termes	Entrée	Officia-lisation	Locali-sation	Entrée	Officia-lisation	Locali-sation	Entrée	Officia-lisation	Locali-sation
serveur	+	+	DR	+	-		+	-	
téléinforma-tique	+	-		+	-		+	-	
télématique	+	-		+	-		+	-	
télétraitement	+	-		+	-		+	-	
terminal	+	-		+	-		+	-	
tirage	+	+	CD	-	ø		-	ø	
visu	-	ø		-	ø		+	-	
visuel	+	+	CD	+	-		+	-	
visualiser	+	-		+	-		+	-	
36	33	12		25	3		31	4	

Légende: + --> le mot possède une entrée et un sens informatique dans le dictionnaire; le mot est labellisé.
+ --> le mot ne possède pas de sens informatique; le mot n'est pas labellisé.
ø --> ne s'applique pas

Abréviations: AI --> allusion indirecte
CD --> commentaire dans la définition
DM --> définition métalinguistique
DR --> définition référencée
RA --> rubrique analogique
RI --> rubrique indépendante

Les rubriques utilisées pour véhiculer le message officiel sont la définition, l'officialisation et le réseau analogique. Dans le minicorpus, on compte 19 indications de ce genre, réparties comme suit (voir le tableau 2):

TABLEAU 2:

Distribution des indications officielles

RUBRIQUE	PROCÉDÉS	EXEMPLES
DÉFINITION	- déf. métalinguistique [1]	LISTAGE (PLI)
	- déf. référencée [1]	SERVEUR (GRLF)
	- déf. commentée [5]	LISTAGE, NUMÉRIQUE, PHOTOSTYLE, TIRAGE, VISUEL, (GRLF)
LABEL	marqueurs	
	• REM [3]	INTERFACE, LOGICIEL, MATÉRIEL (GRLF)
	• parenthèses [4]	LOGICIEL, MATÉRIEL, NUMÉRIQUE (DFP); MATÉRIEL (PLI);
	• ∅ [2]	LOGICIEL (PLI); MATÉRIEL (GRLF)
ANALOGIE	identification de la forme étrangère [3]	AUTONOME, DONNÉE(GRLF); NUMÉRIQUE (PLI)

Le corpus restreint n'a permis de repérer l'information officielle que dans les trois rubriques mentionnées. D'autres recherches menées par ailleurs montrent que l'indicatif ministériel peut apparaître dans la parenthèse étymologique (ex. *remue-méninges* (*GRLF*), en entrée-renvoi, dans l'exemple, etc. (voir Boulanger 1988c)). La définition et le label demeurent pour le moment les rubriques privilégiées.

La répartition proposée prouve que le traitement n'est pas systématique, tant s'en faut. En fait, des trois termes communs à tous les répertoires, deux ont le même traitement partout (*logiciel* et *materiel* ont une rubrique indépendante (RI)) et un a trois traitements différents (*numérique*: commentaire dans la définition (CD), rubrique indépendante (RI) et renvoi analogique (RA)).

J'ai examiné aussi dans quelles mesures la définition officielle avait des chances de poursuivre sa carrière dans le dictionnaire de langue. Elles sont bien minces comme l'illustre les constatations suivantes faites à partir du *GRLF* (voir l'Annexe 6.1 où toutes les définitions des termes officialisés sont données):

- La définition constitue une information métalinguistique (ex. *listage* (*PLI*)).
- La définition est modifiée au point qu'elle se détache totalement de la source officielle (ex. *autonome, visuel* (*GRLF*)).
- La définition est modérément retouchée de façon à être adaptée au public-cible du DGM (ex. *interface, listage* (*GRLF*); *logiciel,* (*DFP*); *numérique* (*DFP*)).

- La définition est très légèrement modifiée; elle ne s'écarte pas beaucoup de l'énoncé officiel (ex. *logiciel, matériel (GRLF)*; *logiciel (PLI)*; *matériel (DFP)*).
- La définition officielle est tronquée, c'est-à-dire qu'un segment est abondonné (ex. *tirage*: la portion retranchée [... *présentée sur une visu*] contient un terme (*visu*) non traité dans le dictionnaire (*GRLF*)).
- La définition officielle est citée intégralement et accompagnée de sa référence ministérielle (ex. *serveur (GRLF)*).

Le traitement microstructural des arrêtés de terminologie en est encore à sa phase exploratoire. Seule l'écume de la surface a été remuée. Il faut encore se pencher sur les secrets des profondeurs. Notamment sur les critères de sélection des avis par les lexicographes. La plupart des méthodologies de la recherche lexicographique étant antérieures à 1975, il n'est guère étonnant qu'elles ne fassent aucune allusion au sujet (voir Boulanger 1988a et 1988c).

Une rapide analyse fournit quelques critères de surface:

- L'usage réel et non pas artificiel du terme (ex. *didacticiel/fac-sim*).
- La concurrence avec l'emprunt (ex. de *secours/back up*).
- La nouveauté conceptuelle (ex. *infographie/ographeur, virus, vaccin*).
- Le degré de technicité (ex. *codet, tableur*).
- La provenance ou l'emploi géographique (ex. *bogue* (n.f.), *spoule* (n.m.)).
- La synonymie de normalisation (ex. *visu/visuel, fac-sim/tirage*).
- Le statu quo lexicographique (ex. *bit, disquette, information*).

Les critères doivent être considérés dans leur ensemble car il est rare que chacun fonctionne indépendamment d'un ou de plusieurs autres. Ainsi *tutoriel* qui correspond à un emprunt sous la forme du calque (anglais *tutorial*) et qui identifie un concept relativement récent et dont le degré de technicité est élevé.

4. CONCLUSION

L'intervention étatique dans le domaine de la langue ramène à la mémoire le concept de «norme». Or les rapports entre la norme et le dictionnaire sont loin d'être clairs et de faire l'unanimité (cf. Rey 1972 et 1983). Malgré leur volonté de se cantonner dans les limites de la description, de l'observation, les dictionnaires français d'aujourd'hui, comme ceux d'hier d'ailleurs, endossent volontairement ou non, la responsabilité d'une prescription partielle du lexique synchronique français. Les dictionnaires offrent à l'utilisateur un répertoire de mots choisis, acceptés d'emblée et fixés, l'absence d'un mot est vue comme le signe d'une condamnation implicite par le lexicographe.

Le dictionnairiste est perçu comme un mediateur entre la société et les gens ordinaires. À travers son anonymat, il devient le garant de la norme et de la connaissance lexicale, ce qui entraîne que ce qu'il entérine en tant que responsable d'un dictionnaire est le fait linguistique décrit, à l'exclusion des autres. Le DGM réglemente et régente tout à la fois puisqu'il impose au public une image concertée du lexique.

Le lexicographe, on l'a vu devient en outre un intermédiaire entre le pouvoir étatique et les utilisateurs de répertoires en consignant les décisions officielles qu'il filtre plus ou moins. Il marque les avis grâce à une série d'intervention et d'étiquettes, introduisant ainsi un renforcement de la norme par le simple fait qu'il identifie l'autorité interventionniste. Simultanément, il crée

une distance entre lui et les instances décisionnelles. L'attribution d'un label officiel a pour effet de distinguer la norme sociale ordinaire, qui est rattachée à la description, de la nouvelle norme institutionnelle qui est rattachée à la prescription d'origine législative. Le consommateur n'a plus qu'à se soumettre à l'usage ministériel ou à le rejeter. La consignation fréquente de l'emprunt ou de la forme à remplacer laisse le choix au locuteur. Si, désormais, les lexicographes rendent compte des avis, ils ne songent nullement à se substituer aux autorités désignées. C'est ce qui explique qu'ils «labellisent» le plus souvent les officialismes et qu'ils se permettent à l'occasion des commentaires microstructuraux personnels ou des critiques à l'égard des suggestions. Ainsi dans le *GRLF*, à l'entrée *bouteur*: «REM. Ce mot n'est pas attesté à notre connaissance dans l'usage spontanée». L'intrusion du *je* (ici *notre*) dans le discours lexicographique est plutôt exceptionnel au sein des microstructures contemporaines. Il est même à remarquer d'une manière toute particulière. De fait, si ce n'était de son étiquette officielle, *bouteur* serait toujours dans l'antichambre ou le purgatoire des fichiers en attendant une problématique naturalisation lexicographique. Il est manifeste que le commentaire du lexicographe signifie que la consignation de *bouteur* est contraire à l'usage puisque ce n'est pas un terme observé et dont la vitalité est démontrée. Si l'on se fie aux différents traitements qu'il reçoit dans plusieurs DGM consultés, *bouteur* serait une tentative de francisation infructueuse, une forme artificielle, un mot-éprouvette.

Les lexicographes accueillent les officialismes mais pas à n'importe quel prix. Ils n'assurent pas le gîte et le couvert à tous. Comme le souligne l'un d'eux, ils enregistrent «les condamnations et recommandations officielles en matière de termes techniques jugés indésirables: l'intrusion de la norme prend ici figure officielle ministérielle -- et les dictionnaires ne peuvent refuser cette manifestation évaluative et prescriptive, alors même qu'ils se veulent descriptifs» (Rey, 1983, p. 546). Malgré cela, le lexicographe conserve toujours la prérogative de sélectionner les unités qu'il veut retenir. La subjectivité est en concordance avec l'idéologie qu'il prône lui-même et avec celle qui est façonnée par l'institution dictionnairique qui l'emploie. Celle-ci a des objectifs de rentabilité économique qui ne sont pas toujours en synergie avec l'efficacité didactique et scientifique des DGM. Au sens le plus noble, la fabrication de dictionnaires est l'un des plus importants maillons des industries de la langue et cela depuis des décennies, longtemps avant que l'on reconnaisse l'existence de ce concept sous la forte poussée expansionniste des outils informatiques et que l'on fonde la "linguismatique".

Bibliographie

5. BIBLIOGRAPHIE

5.1 Linguistique

Boulanger (Jean-Claude), 1988a, Quelques considérations sur le statut de la «res lexicographica» en linguistique, dans *Dictionaries*, (Texte dactylographié, 29 p., à paraître).

Boulanger (Jean-Claude), 1988b, Remarques sur l'aménagement du statut du français en informatique, dans Annie Bourret et Marie-Claude L'Homme (éd.), *Les industries de la langue: au confluent de la linguistique et de l'informatique*, (sous la direction de Pierre Auger, avec la collaboration de Carole Verreault), coll. "K", n° 9, Québec, Centre international de recherche sur le bilinguisme, Université Laval, p. 61-69.

Boulanger (Jean-Claude), 1988c, Lexicographie et politique langagière: l'exemple français des avis officiels, dans *Encyclopédie internationale de lexicographie*, Berlin-New York, Walter de Gruyter (Texte dactylographié, 31 p., à paraître).

Leclerc (Jacques), 1988, *Les langues de l'affichage dans le monde*, (Texte dactylographié, 300 p., à paraître).

Rey (Alain), 1972, Usages, jugements et prescriptions linguistiques, dans *Langue française*, n° 16, décembre, p. 4-28.

Rey (Alain), 1983, Norme et dictionnaires (domaine du français), dans *La norme linguistique*, Textes colligés et présentés par Édith Bédard et Jacques Maurais, coll. "L'ordre des mots", Paris/Québec, Le Robert/Conseil de la langue française, p. 541-569.

5.2 Dictionnaires

Abenaim (Raymonde), Boulanger (Jean-Claude), Shiaty (A.E.) et Vaugeois (Denis), 1986, *Dictionnaire CEC jeunesse*, Nouvelle édition, Montréal, Les Éditions CEC, 1200 p.

Arrêté du 22 décembre 1981 relatif à l'enrichissement du vocabulaire de l'informatique (Journal officiel. N.C. du 17 janvier 1982), 1984, dans *Dictionnaire des néologismes officiels. Tous les mots nouveaux*, Paris, Franterm, p. 401-406.

Dictionnaire du français, 1987, Paris, Hachette, 1782 p.

Dictionnaire du français plus. À l'usage des francophones d'Amérique, 1988, Édition établie sous la responsabilité de A.E. Shiaty, avec la collaboration de Pierre Auger et de Normand Beauchemin, Rédacteur principal: Claude Poirier, Montréal, Centre éducatif et culturel inc., XXIV + 1857 p.

Dictionnaire Hachette de la langue française, 1980, Sous la direction de Françoise Guérard, Paris, Encyclopédies Hachette, 1817 p.

Larousse de la langue française. Lexis, 1979, Sous la direction de Jean Dubois, 2e édition, illustré, Paris, Librairie Larousse, XVI + 2111 p.

Le Micro-Robert. Langue française plus noms propres, chronologie, cartes, 1988, Rédaction dirigée par Alain Rey, Paris, Dictionnaires Le Robert, XXVII + 1992 p.

Petit Larousse illustré 1989, 1988, Paris, Librairie Larousse, 1680 p. + Atlas.

Robert (Paul), *Le Grand Robert de la langue française. Dictionnaire alphabétique et analogique de la langue française*, 1985, 2e édition entièrement revue et augmentée par A. Rey, Paris, Dictionnaires Le Robert, 9 vol., LVIII p. + p.v.

Robert (Paul), *Le Petit Robert 1. Dictionnaire alphabétique et analogique de la langue française*, 1986, Nouvelle édition revue, corrigée et mise à jour, Rédaction dirigée par Alain Rey et Josette Rey-Debove, Paris, S.N.L. - Dictionnaire Le Robert, XXXI + 2175 p.

Annexe

6. ANNEXE

6.1 Définitions

autonome:	GRLF -->	Qui n'est pas connecté à un calculateur central, qui est indépendant des autres éléments du système.
	ARRÊTÉ -->	Se dit d'un matériel lorsqu'il fonctionne indépendamment de tout autre [...].
donnée:	GRLF -->	Représentation conventionnelle d'une information (fait, notion, ordre d'exécution) sous une forme (analogique ou digitale) permettant d'en faire le traitement automatique.
	ARRÊTÉ -->	Représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement [...].
interface:	GRLF -->	Jonction entre deux éléments d'un système informatique (connexion physique ou connexion de programmation).
	ARRÊTÉ -->	Jonction entre deux matériels ou logiciels leur permettant d'échanger des informations par l'adoption de règles communes, physiques ou logiques.
listage:	GRLF -->	Document qui reproduit une liste (souvent produit par l'imprimante d'un ordinateur; [...]).
	PLI -->	Recomm. off. pour listing.
	ARRÊTÉ -->	Document en continu produit par une imprimante d'ordinateur.
logiciel	GRLF -->	Ensemble des programmes, procédés et règles, éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de l'information.
	DFP -->	Ensemble des règles et des programmes relatifs au fonctionnement d'un ordinateur, par oppos. à matériel*.
	PLI -->	Ensemble de programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de l'information.
	ARRÊTÉ -->	Ensemble des programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de données [...].
matériel:	GRLF -->	Ensemble des éléments employés pour le traitement automatique de l'information.
	DFP -->	Ensemble des éléments physiques employés pour le traitement de l'information, par oppos. à logiciel.
	PLI -->	Ensemble des éléments physiques d'un système informatique.
	ARRÊTÉ -->	Ensemble des éléments physiques employés pour le traitement des données [...].
multitraitement:	GRLF -->	Traitement simultané de plusieurs programmes (par un ordinateur).
	ARRÊTÉ -->	Mode de fonctionnement d'u . ordinateur selon lequel plusieurs processeurs ayant accès à des mémoires communes peuvent opérer en parallèle sur des programmes différents.
numérique:	GRLF -->	Se dit de la représentation de données d'information ou de grandeurs physiques au moyen de caractères, chiffres, systèmes, dispositifs ou procédés employant un mode de représentation discrète.
	DFP -->	Qui utilise des nombres, des grandeurs discrètes (opposé à analogiques).

	PLI -->	a. Se dit de la représentation d'informations ou de grandeurs physiques au moyen de caractères, tels que des chiffres, ou au moyen de signaux à valeurs discrètes. [...] b. Se dit des systèmes, dispositifs ou procédés employant ce mode de représentation discrète. par opp. à analogique.
	ARRÊTÉ -->	Se dit, par opposition à analogique, de la représentation de données ou de grandeurs physiques au moyen de caractères -- des chiffres généralement -- et aussi des systèmes, dispositifs ou procédés employant ce mode de représentation discrète [...].
photostyle:	GRLF -->	Dispositif permettant d'introduire dans la mémoire d'un ordinateur une information (coordonnées ponctuelles) sur un écran de visualisation [...].
	ARRÊTÉ -->	Dispositif d'entrée que l'opérateur pointe directement sur l'écran d'une visu [...].
serveur:	GRLF -->	«Organisme exploitant un système informatique permettant à un demandeur la consultation et l'utilisation directes d'une ou plusieurs banques de données» (Journ. off., 17 janv. 1982).
	ARRÊTÉ -->	Organisme exploitant un système informatique permettant à un demandeur la consultation et l'utilisation directes d'une ou plusieurs banques de données.
tirage:	GRLF -->	Document graphique résultant du transfert sur un support permanent d'une image [...].
	ARRÊTÉ -->	Document graphique résultant du transfert sur un support permanent d'une image présentée sur une visu [...].
visuel:	GRLF -->	Dispositif d'affichage, d'inscription sur un écran ou une console à tube cathodique. -- Par ext. L'écran, la console [...].
	ARRÊTÉ -->	Appareil permettant la présentation visuelle et non permanente d'informations [...].

6.2 Officialisation: énoncés et marqueurs

autonome:	GRLF -->	syn.: non connecté (angl. off-line). [RA-AI]
donnée:	GRLF -->	(pour traduire l'angl. data). [RA-AI]
interface:	GRLF -->	REM. Dans ce sens, le mot est admis (Journ. off., 13 janv. 1974) ainsi que jonction. [RI]
listage:	GRLF -->	recomm. off. pour franciser l'anglic. listing, n.m. [...]. [CD]
	F -->	Recomm. off. pour listing. [DM]
logiciel:	GRLF -->	REM. L'administration recommande ce terme pour traduire l'anglais software*. [RI]
	DFP -->	(Mot recommandé pour remplacer software.) [RI]
	PLI -->	Recomm. off. pour software. [RI]
matériel:	GRLF -->	Recomm. off. pour hardware [.]. [RI]
	DFP -->	(Équivalent français recommandé pour remplacer hardware.) [RI]
	PLI -->	(Recomm. off. pour hardware.) [RI]
multitraitement	GRLF --	REM. Équivalent proposé pour remplacer l'anglicisme multiprocessing. [RI]
numérique:	GRLF -->	(recomm. off. pour remplacer digital*) [CD]

	DFP -->	(Terme officiellement recommandé pour remplacer digital.) [RI]
	PLI -->	Syn. (anglic. déconseillé): digital. [...] Syn.: digital. [RA]
photostyle:	GRLF -->	(créé pour rendre l'anglais light pen; recomm. off.) [CD]
serveur:	GRLF -->	«Organisme exploitant un système informatique permettant à un demandeur la consultation et l'utilisation directes d'une ou plusieurs banques de données» (Journ. off., 17 janv. 1982). [DR]
tirage:	GRLF -->	(recomm. off. pour l'angl. hard copy). [CD]
visuel:	GRLF -->	(trad. offic. de l'angl. display). [CD]

CODIFICATION "PHONOGRAPHIQUE" DE L'ESPAGNOL

Silvia Faitelson-Weiser
Université Laval

Au cours des dernières années, et avec le généreux appui du Conseil de Recherches en Sciences Sociales du Canada, nous avons travaillé à l'élaboration d'un Dictionnaire Inverse de l'Espagnol élaboré à l'aide de l'ordinateur et contenant les quelque 181 000 mots qui constituent les lexiques de 16 ouvrages lexicographiques de l'espagnol[1]

Lors du déroulement de ce travail, nous nous sommes heurtée, dès le début, à un problème de notation. En effet, même si l'espagnol est probablement la langue romane qui a le taux le plus élevé de correspondance entre son système phonologique et son code orthographique, cette correspondance est loin d'être absolue, même au niveau phonologique. Évidemment, nous aurions pu opter pour l'emploi d'une transcription phonétique plus ou moins large. Toutefois, cette solution nous est vite apparue peu économique, et celà pour deux raisons principales:

a) les cas problèmes n'impliquaient que quelques phonèmes et,

b) afin de faciliter la consultation de notre dictionnaire, élaboré à partir de sources écrites, nous voulions que dans le produit fini, les mots apparaissent comme ils s'écrivent couramment en espagnol.

Le but de cette communication est d'exposer les problèmes qui se sont posés et la manière dont nous les avons résolus.

Notre objectif était, nous le rappelons, de permettre à l'ordinateur de classer les mots du corpus d'une façon rigoureusement phonologique, indépendamment des divers graphèmes utilisés pour les transcrire et, en même temps, de présenter ces mots avec leur orthographe courante en espagnol.

Afin d'atteindre cet objectif il a fallu, dans un premier temps, identifier les cas posant problème. Ceci a été relativement vite fait et nous avons trouvé trois groupes de cas problèmes impliquant des consonnes et un impliquant des voyelles, comme suit:

1. PHONÈMES CONSONANTIQUES NORMALEMENT TRANSCRITS EN ESPAGNOL PAR DES "GRAPHÈMES COMPLEXES"

En premier lieu, il y avait les six phonèmes consonantiques transcrits en espagnol par des "graphèmes complexes", c'est-à-dire par plus d'un graphème. Parmi eux, cinq sont des phonèmes

---

[1] Faitelson-Weiser, S (1987): DIASLE: Dictionnaire inverse et analyse statistique de la langue espagnole - Diccionario inverso y analisis estadistico de la lengua espanola. - Reverse Dictionary and Statistical Analysis of the Spanish Language. Les Presses de l'Université Laval, Québec.

assez courants en espagnol, soit l'affriquée /ʧ/, transcrite par ch, comme dans chico, la vibrante longue /r/, représentée par rr, comme dans perro, la palatale latérale /ʎ/, transcrite par ll, comme dans le mot calle, la vélaire sourde /k/, qui peut parfois être représentée par qu (guiar). A ces cinq phonèmes courants de l'espagnol il fallait ajouter le phonème /ʃ/, présent notamment dans des emprunts récents et transcrit par sh, comme dans le mot flash.

Phonèmes		Graphèmes	Exemples		
/ʧ/	=	ch	chico	/'ʧiko/	(petit)
/r/	=	rr	perro	/'pero/	(chien)
/ʎ/	=	ll	calle	/'kaʎe/	(rue)
/ʃ/	=	sh	flash	/'flaʃ/	(flash)
/k/	=	qu	querer	/ke'rer/	(vouloir)
/g/	=	gu.	guiar	/gi'ar/	(guider)

2. GRAPHÈMES CONSONANTIQUES REPRÉSENTANT PLUS D'UN PHONÈME

Un autre cas problème était celui des graphèmes consonantiques servant à transcrire plus d'un phonème. C'est le cas des graphèmes c et g, qui représentent deux phonèmes différents, selon qu'ils sont suivis d'une voyelle antérieure (i ou e) ou non. En effet, suivi de ces voyelles, le graphème c transcrit le phonème /θ/ (par exemple, cinco) et dans les autres cas, le phonème /k/ (caza). Quant au graphème g, devant e ou i, il se lit /χ/ (general), et dans les autres cas, /g/, comme dans gato.

Graphèmes		Phonèmes	Exemples		
c	=	/θ/ (devant /e/ ou /i/)	cinco	/'θinko/	(cinq
	=	/k/ (dans les autres cas)	caza	/'kaθa/	(chasse)
g	=	/χ/ (devant /e/ ou /i/)	general	/χene'ral/	(général)
	=	/g/ (dans les autres cas)	gato	/'gato/	(chat)

3. PHONÈMES CONSONANTIQUES REPRÉSENTÉS PAR PLUS D'UN GRAPHÈME

Or s'il y avait seulement deux cas de graphèmes simples servant à transcrire plus d'un phonème, nous avions aussi à nous occuper de cinq cas de phonèmes qui peuvent être représentés en espagnol par plus d'un graphème. En effet, dans cette langue, les phonèmes /θ/, /χ/, /k/, /g/ et /b/ peuvent être transcrits par plus d'un graphème: le phonème /θ/ est tantôt représenté par c (cinco), tantôt par c (encore caza), tantôt par k (dans les mots d'origine étrangère, comme kilometro), et aussi par qu (querer); le phonème /g/, tantôt par g (gato), tantôt par gu (guiar) et, finalement, le phonème /b/, est représenté tantôt par b (beber), tantôt par v (vivir).

Phonèmes		Graphèmes	Exemples		
/θ/	=	c	cinco	/'θinko/	(cinq)
	=	z	caza	/'kaθa/	(chasse)
/χ/	=	g	general	/χene'ral/	(général)
	=	j	junta	/'χunta/	(réunion)
/k/	=	c	caza	/'kaθa/	(chasse)
	=	k	kilómetro	/ki'lometro/	(kilomètre)
	=	qu	querer	/ke'rer/	(vouloir)
/g/	=	g	gato	/'gato/	(chat)
	=	gu	guiar	/gi'ar/	(guider)
/b/	=	b	beber	/be'ber/	(boire)
	=	v	vivir	/bi'bir/	(vivre)

Dans tous les cas dont nous venons de parler, il fallait faire quelque chose si nous voulions, comme c'était le cas, que la machine puisse trier ensemble les phonèmes, et non simplement les graphèmes.

Le cas le plus simple à régler était celui du phonème /b/, dans lequel il suffisait de demander à l'ordinateur de ne pas discriminer lors du tri entre les deux signes, et c'est ainsi que nous avons procédé (au tableau 1, on peut voir que les deux, ensemble, occupent le rang 27).

Les autres cas étaient un peu plus compliqués et nous avons opté pour la création d'un ensemble de "caractères intermédiaires" qui sont ceux qui figurent au tableau 1 sous la rubrique "caractère d'entrée". Ces caractères permettaient d'établir une relation univoque - un graphème = un phonème et seulement un phonème - entre tous nos signes et les phonèmes qu'ils représentent. Cette relation d'univocité était indispensable pour que l'ordinateur puisse trier nos mots selon leurs caractéristiques phonologiques et présenter ensuite, une fois les tris opérés, par une simple commande de conversion, les mots avec leur orthographe courante.

Ainsi, pour les fins du tri, nos six "graphèmes complexes" ont été entrés comme des majuscules simples et un rang de tri particulier leur a été attribué: le double ll a été entré comme L majuscule (comme dans cALe) et classé au rang 9, alors que le l minuscule, représentant le phonème latéral non palatal /l/, était classé au rang 10; le rr, entrécomme R majuscule (pERo), était classé au rang 12, après le r minuscule simple (phonème /ɾ/), classé au rang 11; le groupe ch, entré comme C majuscule (CIco), obtenait le rang 13, différent du rang 33 du c minuscule; le groupe sh, quand il représentait le phonème palatal /ʃ/, a été entré comme S majuscule (fIAS), et classé au rang 14, avant le s minuscule, servant à transcrire le phonème /s/, et situé au rang 16; quant aux ensembles gu et qu, ils ont été entrés respectivement comme G majuscule (GIAr) et Q majuscule (QerEr), et triés, ainsi que nous le verrons tout de suite, sous les rangs 32 et 33 avec d'autres signes. Une fois les tris effectués, les majuscules ont été reconverties en signes complexes.

Quant aux cas des graphèmes simples présentant des relations non univoques avec les phonèmes qu'ils transcrivaient, et vice-versa, nous avons adopté une solution semblable. Nous avons d'abord, au moyen de majuscules employées comme caractères d'entrée, désambiguisé la relation et ensuite, nous avons, comme nous l'avons fait pour les graphèmes b et v, demandé à l'ordinateur d'attribuer un même ranglors du tri aux divers signes - majuscules ou minuscules - correspondant au même phonème. Ainsi, les c minuscules, qui représentaient le phonème /θ/, ont été entrés comme des Z majuscules (ZIuco) et ces Z majuscules ont été triés au rang 17, avec les z minuscules (cAza), transcrivant le même phonème. Les groupes gu, dont nous avons déjà

parlé, qui avaient été entrés comme des G majuscules, ont été triés au rang 32, avec les g minuscules, servant également à transcrire le phonème /g/. La même procédure a été suivie dans le cas des groupes qu, transcrits par Q majuscules et triés au rang 33, avec les c minuscules, et les k, puisque les trois représentant le phonème /k/. Finalement, les g suivis de e ou de i, et transcrivant le phonème /χ/, ont été entrés comme J majuscules (JenerAl) et triés au rang 35, avec les j minuscules, représentant le même phonème (jUnta). Comme dans le cas des graphèmes complexes, une fois les tris effectués, les majuscules sont redevenues les minuscules qu'elles étaient avant leur conversion en "caractères d'entrée", tel qu'indiqué sous la rubrique "caractère d'édition".

TABLEAU 1:

Solution adoptée pour les consonnes

Rang de tri	Valeur phonologique	Caractère d'entrée	Exemple	Caractère d'édition	Exemple
9.	/ʎ/	L	cALe	ll	calle
12.	/r/	R	pERo	rr	perro
13.	/ʧ/	C	Cico	ch	chico
14.	/ʃ/	S	flAS	sh	flash
17.	/θ/	z	cAza	z	caza
		Z	ZInco	c	cinco
27.	/b/	b	bebEr	b	beber
	/v/	v	vivIr	v	vivir
32.	/g/	g	gAto	g	gato
		G	GUAr	gu	guiar
33.	/k/	c	cAza	c	caza
		k	kilómetro	k	kilómetro
		Q	QerEr	qu	querer
35.	/χ/	j	jUnta	j	junta
		J	JenerAl	g	general

4. LES VOYELLES

Pour ce qui est des voyelles, il n'y a en espagnol qu'un vrai cas de relation non-univoque phonème/graphème: il s'agit du phonème vocalique /u/, normalement représenté par le graphème u mais qui, situé entre un g et une voyelle antérieure (e ou i) est transcrit comme ü tréma. Le cas a été traité comme les cas d'ambiguité consonantique: les ü ont été entrés avec des W majuscules et triés au rang 31, avec les autres u minuscules.

Toutefois, nous voulions, dans notre Dictionnaire, et contrairement à la pratique courante des dictionnaires de l'espagnol, tenir compte de la différence entre les voyelles toniques et les

voyelles atones, qui nous semblait très importante pour un Dictionnaire Inverse. Cette exigence a tout de suite introduit un nouveau problème d'ambiguité. En effet, en espagnol, on peut marquer l'accentuation d'intensité d'un mot en mettant un "accent aigu" sur la voyelle qui porte cette accentuation, mais les voyelles toniques ne portent pas toujours cette marque, ce qui donne la situation présentée dans le tableau qui suit:

Phonèmes		Graphèmes	Exemples		
/a/ atone	=	a	casa	/'kasa/	(maison)
/'a/ tonique	=	a	casa	/'kasa/	(maison)
	=	â	mamá carácter	/ma'ma/ /ka'rakter/	(maman), (caractère)
/e/ atone	=	e	nene	/'nene/	(bébé)
/'e/ tonique	=	e	nene	/'nene/	(bébé)
	=	é	bebé débil	/be'be/ /'debil/	(bébé) (faible)
/i/ atone	=	i	difícil	/di'fiθil/	(difficile)
/'i/ tonique	=	i	rico	/'riko/	(riche)
	=	í	difícil	/di'fiθil/	(difficile)
/o/ atone	=	o	mano	/'mano/	(main)
/'o/ tonique	=	o	hablo	/'ablo/	(je parle)
	=	ó	habló	/a'blo/	(il a parlé)
/u/ atone	=	u	publico	/pu'bliko/	(je publie)
	=	ü	agüero	/a'gwero/	(augure)
/'u/ tonique	=	u	cutis	/'kutis/	(peau)
	=	ú	público	/'publiko/	(public)

Tenant donc, ainsi que nous l'avons dit, à séparer les voyelles toniques, nous avons fait appel, une fois de plus, à nos "caractères d'entrée". Cette fois-ci, nous avons entré en majuscules les voyelles toniques qui ne portaient pas d'accent écrit. Ainsi, au Tableau II on constate que le premier /a/ du mot **casa**, qui est la voyelle tonique, est devenu un **A** majuscule, ainsi que le /e/ tonique du mot **nene**, le /i/ tonique de **rico**, le /o/ tonique de **monte** et le /u/ tonique de **cutis**. Ces conversions effectuées, nous avons attribué, lors du tri, un seul et même rang à chaque voyelle tonique, qu'elle soit représentée par une voyelle accentuée ou par une voyelle majuscule (rang 3 pour **á** accentué et **A** majuscule, rang 5 pour **é** accentué et **E** majuscule, rang 7 pour **í** accentué et **I** majuscule, rang 28 pour **ó** accentué et **O** majuscule, et rang 30 pour **ú** accentué et **U** majuscule). Un rang différent a été attribué aux voyelles atones, transcrites toujours, sauf pour le **ü** dont nous avons déjà parlé, par des minuscules (rang 4 au **a** atone, rang 6 au **e** atone, rang 8 au **i** atone, rang 29 au **o** atone, et rang 31 au **u** atone, représenté par **u** minuscule ou par au **W** majuscule). Comme dans les autres cas, après les tris, les majuscules ont été remplacées par les "caractères d'édition" correspondants.

TABLEAU 2:

Solution adoptée pour les voyelles

Rang de tri	Valeur phonologique	Caractère d'entrée	Exemple	Caractère d'édition	Exemple
3.	/'a/ (tonique)	á	mamá carácter	á	mamá carácter
		A	cAsa	a	casa
4.	/a/ (atone)	a	cAsa	a	casa
5.	/'e/ (tonique)	é	bebé débil	é	bebé débil
		E	nEne	e	nene
6.	/e/ (atone)	e	nEne	e	nene
7.	/'i/ (tonique)	í	difíZil	í	difícil
		I	rico	i	rico
8.	/i/ (atone)	i	difíZil	i	difícil
28.	/'o/ (tonique)	ó	habló	ó	habló
		O	mOnte	o	monte
29.	/o/ (atone)	o	hAblo	o	hablo
30.	/'u/ (tonique)	ú	público	ú	público
		U	cUtis	u	cutis
31.	/u/ (atone)	u	publico	u	publico
		W	agWEro	ü	agüero

Cette manière de faire qui posait - nous a-t-on assuré - très peu de problèmes du point de vue de la programmation, et qui était relativement facile à appliquer lors de la saisie des données, nous a permis d'atteindre nos objectifs. En effet, dans notre Dictionnaire Inverse et Analyse statistique de la langue espagnole, tous les mots sont classés selon un alphabet phonographique qui tient compte de leurs traits phonologiques tout en les présentant avec leur orthographe courante.

Auteure **Catherine Péquégnat**
Laboratoire de Génie informatique, Grenoble

Titre **Conception en DELPHIA-PROLOG d'une interface simple et efficace pour l'interrogation de bases de données en français - Une application industrielle**

RÉSUMÉ

Le but est de construire dans des temps raisonnables :

1) des interfaces relativement simples;
2) faites pour des SGBD dont le langage d'interrogation est de type SQL en fonction d'heuristiques générales et comportant un maximum de composants réutilisables d'une application à l'autre.

Ceci nous a conduit à la définition d'un noyau du système dit statique, dont la conception et la réalisation sont présentées brièvement dans cette communication. Par ailleurs, et de façon tout à fait expérimentale, nous avons voulu ouvrir ce type d'interfaces à certaines procédures d'évaluation déductive de requêtes, possibles du fait de l'environnement PROLOG du système.

Il s'agit aussi d'expliciter nos heuristiques dans la perspective de définir et de réaliser un système dynamique, i.e. un système intégrant des composants permettant l'acquisition contrôlée semi-automatique de l'information lors du passage d'une application à l'autre.

Auteure **Agnès Tutin**
Université de Montréal

Titre **Les constructions libres de forme Nom + Nom**

RÉSUMÉ

A côté de formes figées dont les constituants sont soudés, il existe en français des constructions libres Nom + Nom (notées N1 N2) précédées ou non d'un déterminant.

A une structure de surface unique N1 N2 correspondent en fait plusieurs phénomènes qu'on peut délimiter à l'aide de critères syntaxiques :

•**Le N2 adjectival** *(Ex :* une visite éclair, une note limite*) est un nom qui a acquis un statut adjectival autonome. Il peut généralement apparaître en position d'adjectif attribut et subir une modification de son degré d'intensité.*

•**Les N2 apposés** *se construisent par simple juxtaposition (Ex :* une femme médecin, un objet symbole*) et sont paraphrasables par une phrase à verbe être.*

•**Les compléments de nom construits sans préposition** *rapprochables de compléments de nom prépositionnels (Ex :* le réseau banlieue = le réseau de la banlieue, un fichier matières = un fichier par matières*).*

•**Les juxtapositions par coordination** *(Ex :* une dérivation-composition, un teinturier-blanchisseur*) sont un phénomène plus lexical que syntaxique.*

•**Les N1 prépositionnels**, *non précédés de déterminant, introduisent directement les noms et sont parfois rapprochables de groupes prépositionnels développés (Ex :* côté études = du côté des études*).*

Auteure **Andrée Borillo**
Université Toulouse Le Mirail

Titre **De quelques procédés de caractérisation des noms d'action en français**

RÉSUMÉ

On connaît la difficulté de dégager en français les différentes catégories de noms, en particulier la difficulté qu'il y a à caractériser les noms d'action. Plusieurs critères ont été proposés que l'on examinera brièvement et dont on montrera les insuffisances.

On propose d'examiner les noms dans un cadre aspectuo-temporel et d'établir différentiellement des traits de catégorisation exactement comme on le fait pour les verbes (ex. traits duratif, terminatif, ponctuel de la classification vendlerienne); ceci :

- *en les faisant entrer, comme argument, dans des constructions verbales considérées comme prototypes de l'expression prédicative de durées,* **durer Quant Ntps, mettre Quant Ntps, prendre Quant Ntps, passer Quant Ntps.**
- *en les combinant, toujours dans un rôle d'argument, avec des verbes ayant par nature la fonction d'auxiliaires aspectuels dans la spécification des phases de déroulement d'une situation:* **commencer N, se mettre à N, être en cours de N, cesser N, achever N.**
- *en les examinant dans leur fonction de noms prédicatifs, avec des verbes support dotés d'un sémantisme explicite d'action ou d'état, ex.* **procéder à N, opérer N, subir N, être dans N (en N, à N).**
- *également, en les confrontant avec des adjectifs ou des adverbes manifestant des propriétés sémantiques d'action ou d'état bien établies.*

Ainsi, on arrive à dégager quelques critères opératoires permettant de faire la distinction entre différentes catégories de noms; en tout premier lieu, la distinction entre noms d'état et noms d'action, puis parmi ces derniers, la mise à jour de sous-classes distinctes sur la base de la nature agentive du sujet logique (par ex. action volontaire vs action non volontaire).

Auteur **Jacques Ladouceur**
Université Laval

Titre **Le découpage automatique de textes en unités lexicales**

RÉSUMÉ

Le découpage d'un texte en unité lexicales est une opération fondamentale dans un grand nombre de processus de traitement automatique des langues naturelles. Son automatisation pourtant soulève un certain nombre de difficultés.

D'une part, un texte (au sens large) contient de l'information textuelle (suites de mots) et paratextuelle (numéros de pages, références, etc) Comment faire pour qu'un système de découpage ne confonde pas ces deux types d'informations? D'autre part, le découpage automatique de textes est très souvent une opération qui s'insère dans un processus plus large de traitement de la langue. On découpe un texte en vue de pouvoir le traiter. Par sa nature donc, un système de découpage devrait être en mesure de produire des résultats qui s'adaptent à un grande variété de systèmes de traitement automatique de la langue. Finalement, un bon système de découpage devrait être en mesure de reconnaître les unité lexicales graphiquement complexes. Un mot comme **pomme de terre** *devrait être reconnu comme étant un seul mot et non pas trois.*

Nous présentons deux logiciels, DAT et SYREX, qui se veulent une solution partielle aux trois problèmes que nous avons soulevés. Le premier logiciel découpe un texte en mots graphiquement simples. Le second reconnaît les unités lexicales graphiquement complexes.

Auteur **Gaston Gross**
Laboratoire de Linguistique Informatique
Université Paris XIII

Titre **Degré de figement des composés N de N**

RÉSUMÉ

Le recensement systématique des noms composés de type N de N pose de difficiles problèmes de délimitation de ce qui est ou n'est pas figé. Les nombreux travaux sur ce point de la tradition grammaticale ne nous sont pas d'une grande aide puisque les définitions qu'ils proposent ne sont que des variations autour du thème de "l'idée unique", critère sémantique dont il est très facile de démontrer qu'il n'est valable que pour les composés figés, c'est-à-dire à peu près 15% des composés. Le but de cette communication est de mettre en evidence les propriétés syntaxiques du second substantif (détermination propre, compatibilité de cette détermination avec celle du premier substantif, rupture distributionnelle, pronominalisation par en, remplacement par le possessif ou par un adjectif de relation, etc.) et de montrer ainsi que la composition correspond à des degrés de figement différents, représentant un nombre de classes très élevé.

On en tirera comme conclusion que les descriptions qui ont été proposées dans la littérature grammaticale ne rendent pas compte de l'importance de la notion de figement car elle réduisent le phénomène à un cas particulier seulement, à partir duquel on forge un critère qui masque la complexité du phénomène.

LES DICTIONNAIRES ÉLECTRONIQUES DELAS ET DELAC

Blandine Courtois, et Max Silberztein
Laboratoire d'Analyse Documentaire et Linguistique, Paris

INTRODUCTION

Le système DELA des dictionnaires électroniques élaborés au Laboratoire d'Automatique Documentaire et Linguistique a pour but la description et l'analyse de la langue française en vue des traitements sur ordinateurs. Par système de dictionnaires, nous entendons une base de données linguistiques, et les programmes permettant de les traiter.

Une description systématique de la langue implique la représentation de la syntaxe. A cet égard, de nombreux travaux ont été réalisés au LADL, en particulier sur la syntaxe des verbes, largement décrite sur les tables du lexique-grammaire (J. P. BOONS, A. Guillet, C. Leclère, M. Gross, 1976, 1979, 1982, 1988). Toutefois une description linguistique complète nécessite également la construction de lexiques contenant l'ensemble des mots de la langue, avec toutes leurs variations de formes. C'est l'objectif du système DELA.

Sur le plan formel, l'emploi du séparateur comme le blanc entre les mots fait que les unités de texte se répartissent en **mots simples et mots composés.**

Les mots simples sont des séquences contiguës de lettres, comprises entre deux séparateurs, telles que *table, mangerions* et *donc.*

Les mots composés sont des séquences comportant au moins un séparateur, par exemple le blanc (*pomme de terre*), le trait d'union (*face-à-face*), l'apostrophe (*aujourd'hui*), ou une combinaison de séparateurs (*c'est-à-dire*).

En conséquence, nous distinguons dans le système DELA deux types de données et de lexiques. Les unités simples sont rassemblées dans le dictionnaire de mots simples appelé DELAS. Les unités composées sont dans le dictionnaire de mots composés DELAC.

Dans le DELAS, chaque mot simple est accompagné du code de sa classification morphologique. Cette classification est systematique, et sert de base à l'exécution d'une procédure automatique de génération de formes fléchies. L'ensemble des formes construites à partir des mots du DELAS constitue le dictionnaire DELAF.

En parallèle avec la classification morphologique, une représentation phonémique a été élaborée, et son application à l'ensemble des mots du DELAS a permis de construire les dictionnaires phonétiques DELAP (E. Laporte, 1988).

Depuis les debuts de la traduction automatique, de nombreux systèmes d'analyse morphologique ont été proposés. C'est un exercice de choix pour étudiants en informatique. Mais c'est la première fois qu'un lexique de taille réaliste est constitué avec une description systématique de la morphologie. A ce jour, l'ensemble des mots simples du DELAS comporte plus de 70 000 mots, et celui des mots composes du DELAC plus de 80 000 entrées.

CARACTÉRIQUES DES DICTIONNAIRES ÉLECTRONIQUES

Entre les dictionnaires électroniques et les dictionnaires usuels, il existe des différences structurelles profondes. Sans reprendre la discussion relative aux différences entre les uns et les autres (M. Gross, 1988), indiquons cependant quelques propriétés et contraintes des dictionnaires électroniques.

1) Descriptions systématiques

Dans les dictionnaires usuels, certaines informations et règles ne sont pas notées parce que supposées connues. Par exemple, dans le Petit Larousse illustré 1986, la règle de formation du pluriel des mots en *al* est implicitement la suivante: le pluriel des mots en *al* est en *aux*. D'où l'absence d'indication de la flexion:

un journal ---des journaux

et l'impossibilité de trouver le mot *journaux* dans le PLI.

Une telle présentation implicite des mots et des règles est inadaptée aux dictionnaires électroniques. Ceux-ci doivent contenir des représentations systématiques, aussi bien des mots que des règles, et ces dernières doivent être définies sans ambiguïtés pour toute entrée lexicale. En conséquence un dictionnaire morphologique comportera, pour tout nom et adjectif, une description de la mise au féminin et au pluriel, et pour tout verbe une description de sa conjugaison.

2) Données formelles

Dans les lexiques du LADL, les mots sont considérés sous leur aspect formel, les données retenues étant strictement syntaxiques et morphologiques. Les données d'ordre culturel ou étymologique ne sont pas prises en compte, parce qu'elles ne sont pas significatives pour l'analyse synchronique de la langue à laquelle visent les dictionnaires électroniques. En outre, aucune information sémantique n'est introduite, du fait qu'il n'existe pas de système de description applicable à tous les éléments de la langue.

3) Accès normalisés

La recherche de mots dans un dictionnaire usuel présuppose un niveau de connaissance du lecteur, et table sur sa faculté d'interprétation.

D'une part, une même graphie peut apparaître dans plusieurs entrées lexicales, en nombre variable selon les dictionnaires, le choix de l'entrée utile étant laissé au lecteur. D'autre part, pour atteindre un mot composé ou une expression figée, l'accès se fait par l'un des mots constituants, choisi sur des critères lexicographiques non définis. Par exemple dans le PLI, l'expression *angle de tir* se trouve dans l'entrée *tir*, le *tir à blanc* se trouve dans l'entrée *blanc*. Par contre, *blanc d'oeuf* ne se trouve pas dans l'entrée *oeuf*, mais dans l'entrée *blanc*. Les procédures d'accès aux mots dans les dictionnaires sont donc variables. Au contraire, dans les dictionnaires électroniques, ces procédures sont nécessairement normalisées et identiques, quels que soient les mots recherchés.

4) Extensivité

Outre les aspects systématiques et normalisés des dictionnaires électroniques, il faut signaler aussi leur objectif d'une couverture lexicale aussi étendue que possible. Le but de cette large couverture est de permettre des applications nombreuses et diversifiées, par exemple:

- la reconnaissance de mots dans des textes, quel que soit le domaine traité,
- l'analyse syntaxique, prenant en compte tous les emplois des mots, ainsi que toutes les constructions possibles de phrases,
- l'élaboration de dictionnaires spécifiques: homographes, lexiques triés en ordre inverse, lexiques par parties du discours,
- la vérification orthographique automatique,
- l'analyse et l'étude statistique des mots eux-mêmes, de leur fréquence, leur structure et leurs mécanismes de formation,
- et toute la gamme des jeux basés sur la combinaison des lettres dans les mots, tels les anagrammes, mots croisés ou scrabble.

5) Cohérence

Des impératifs de cohérence des données sont à respecter lors de la conception et de la réalisation de dictionnaires électroniques. Du fait de la notation systématique des informations associées à chaque mot, le format des entrées est homogène, et chaque entrée a une structure interne cohérente.

Cependant, afin d'obtenir une plus grande homogénéité des lexiques, nous avons distingué au LADL trois ensembles contenant des unités de texte différentes:

1) les mots simples sous leur forme canonique
2) les formes fléchies
3) les mots composés

Ces trois ensembles seront donc présentés séparément dans la suite de cet exposé.

Le lexique DELAS

DELAS est le Dictionnaire Électronique du LADL pour les mots Simples du français. Nous en donnerons d'abord une vue globale en décrivant la structure des données, avec des exemples concrets d'entrées lexicales. Ensuite, nous examinerons séparément chacun des éléments constitutifs d'une entrée: mot, classification grammaticale, code morphologique.

1. Entrées du DELAS

La structure d'une entrée se décompose en deux parties:

- un mot simple, noté sous sa forme canonique,

- des informations grammaticales et morphologiques associées, se présentant sous forme d'au moins un code morphologique. Celui-ci se compose pour les mots variables de deux éléments:

1- un symbole de partie du discours, (N pour les noms, V pour les verbes,...)
2- un numéro de code morphologique, lequel renvoie à une classe formelle de variations morphologiques.

S'il s'agit de mots invariables, ce deuxième élément est omis.

Les différents exemples donnés ci-dessous illustrent la structure des entrées du DELAS:

table..N21
soigneux,.A63
grandir..V18
admirablement,.ADV

Dans ces exemples, les codes *.N21*, *.A63*, *.V18*, *.ADV*, représentent respectivement:

- *N21* un nom féminin singulier, formant le pluriel en ajoutant s à la fin du mot,
- *A63* un adjectif de formes identiques au masculin singulier et pluriel, et prenant les terminaisons respectives *e* et *es* au féminin singulier et pluriel,
- *V18* un verbe régulier du deuxième groupe,
- *ADV* un adverbe invariable.

En juin 1988, le dictionnaire DELAS comportait plus de 64 000 entrées, qui sont toutes de graphies différentes. Ceci entraîne le rassemblement des mots de même orthographe dans des entrées communes, ce qui minimise le nombre des entrées. En fait, l'ensemble du DELAS représente un corpus de près de 72 000 mots.

2. ÉLÉMENTS CONSTITUTIFS DES ENTRÉES

2.1. Mots du D E L A S

2.1.1. *Formes canoniques*

Nous sommes habitués dans les dictionnaires usuels à la représentation des mots sous leur forme canonique. Celle-ci a été adoptée dans le lexique DELAS. Donc:

- les noms masculins et les adjectifs sont mis sous la forme du masculin singulier,
- les noms et adjectifs exclusivement féminins sous la forme du féminin singulier,
- les verbes sous la forme infinitive.

2.1.2 *Mots simples*

Par définition, le DELAS ne contient que des mots simples, c'est-à dire ne comportant aucun séparateur tel que le blanc, le trait d'union ou l'apostrophe. Les mots composés comme,

par exemple, *arc-en-ciel* ou *ver de terre*, sont recensés ailleurs, dans les tables du dictionnaire de mots composés DELAC. En outre, les mots sont exclusivement en minuscules, accentuées ou non. Les mots avec des majuscules, noms propres et sigles, sont reportés dans des listes annexes.

2.1.3. *Mots venant de contextes identifiés*

Certaines entrées du DELAS sont à noter parce que les mots qui y sont consignés n'ont pas d'existence autonome en tant que mots isolés. Ce sont des mots qui n'apparaissent que dans des contextes bien identifiés. Il s'agit:

- soit d'unités issues de mots composés comme *tohu, bohu, ex, libris, pick, up, prud, homal, homie, ping, pong, check, list, week, end,...*
- soit de parties de locutions, conjonctions, prépositions, ou adverbes composés: *parce, jusque, tandis, afin, aujourd, hui, ad, hoc, ipso, facto, catimini, cahin, caha....*
- soit de préfixes servant à former des mots composés, tels que *anti, cardio, hyper, sous....*
- soit de préfixes dérivés de noms propres, comme *anglo, italo, stalino,...*
- soit de mots élidés: *c', d', j', l', m', n', s', t', qu', jusqu', lorsqu', puisqu', quoiqu', quelqu', presqu', entr'*. Ces derniers sont mis sans apostrophe dans le DELAS, puisque le séparateur est par définition exclu.

D'un point de vue formel, tous ces éléments sont des mots simples ordinaires, c'est pourquoi ils sont consignés dans la version actuelle du DELAS. Toutefois les préfixes ne sont pas recensés de façon exhaustive, du fait de leur présence dans une liste à part.

2.1.4. *Mots de même graphie*

Dans les entrées du DELAS, aucune distinction n'est faite entre deux mots de sens différents, mais de même orthographe et ayant des formes fléchies identiques. Dans ce cas, un seul code morphologique est mis. Par exemple, le mot botte, qui a trois entrées différentes dans le PLI, est noté par l'entrée simple:

botte,.N21

N21 etant un nom féminin, prenant un s au pluriel, qu'il s'agisse d'une botte de foin, d'une botte d'escrime ou d'une botte de caoutchouc. De même, le verbe *voler* se conjugue de la même façon quel que soit son sens, et ne donne lieu qu'à une entrée:

voler,.V3

dans laquelle V3 est le code morphologique des verbes réguliers du premier groupe.

2.2. La classification grammaticale

2.2.1. *Classement selon les parties du discours*

Sur le plan grammatical, les mots sont répartis en neuf catégories, ou parties du discours (*Le Bon Usage*, M. Grevisse). On distingue les noms, les adjectifs, les verbes, les adverbes, les

pronoms, les articles, les prépositions, les conjonctions et les interjections. La classification selon les parties du discours étant traditionnellement utilisée, nous l'avons intégrée dans le dictionnaire morphologique DELAS.

Les codes grammaticaux sont les suivants:

.A (adjectif), .N (nom), .V (verbe), .ADV (adverbe), .CONC (conjonction de coordination), .CONS (conjonction de subordination), .PREP (préposition), .INTE (interjection), .DETE (déterminant), .PRON (pronom), et .XINC (inclassable).

2.2.2. *Entrées à un code*

Les mots qui appartiennent à une catégorie grammaticale unique et bien définie donnent lieu à des entrées simples.

Exemples: *colloque..N1*
linguiste..N31
discursif..A38
cordialement..ADV
ni..CONC
dans..PREP

2.2.3. *Entrées à plusieurs codes*

Les entrées du DELAS étant toutes de graphies différentes, il en résulte qu'une même entrée lexicale regroupe souvent plusieurs mots grammaticalement différents. En conséquence, cette entrée contient plusieurs codes associés. Par exemple, le mot *sur* étant d'une part adjectif, d'autre part préposition, comporte deux codes:

sur..A32.PREP

De telles entrées sont dites homographes. Nous considérons aussi comme homographes les noms humains, spécialement les noms de profession, fréquemment employés comme adjectifs, par exemple:

boulanger, dans un apprenti boulanger
assistant, dans un maître assistant

Un double codage nom et adjectif (.N et .A) est alors nécessaire pour ce type d'items:

boulanger..N42.A42
assistant..N32.A32

2.3. La codification morphologique

Le système de codification morphologique du DELAS sépare deux types de donnée:

- d'une part les noms et adjectifs,
- d'autre part les verbes.

2.3.1. *Codification des noms et adjectifs*

En français, les variations de formes des mots ne portent que sur leur terminaison. Pour les noms et les adjectifs, une suite de quatre terminaisons suffit à représenter toutes les variantes en genre et nombre. Par exemple la suite:

[f, ve, fs, ves]

décrit les quatre formes fléchies *oisif, oisive, oisifs, oisives*, de l'adjectif *oisif*. Cette suite, appelée flexion, peut s'appliquer à tous les adjectifs terminés en *if*. Elle détermine donc une classe d'équivalence morphologique, ou classe flexionnelle, à laquelle se rattachent tous ces adjectifs.

Pour coder la morphologie de tous les noms et adjectifs français, nous avons dressé la liste de toutes les classes flexionnelles existantes. Chaque classe est numérotée, le numéro servant de référence pour coder les entrées du DELAS. Par exemple, la flexion mentionnée plus haut correspondant à la classe 38, on a dans le DELAS:

intempestif..A38
craintif..N38.A38

On voit sur cet exemple que les noms et les adjectifs sont rattachés aux mêmes classes flexionnelles. Pour les noms de genre unique, nous avons maintenu la description générale des flexions avec quatre composantes. Simplement, le signe - a été introduit pour désigner des formes manquantes, ainsi:

[l, -, ux, -]

represente la flexion de la classe 4 regroupant des mots masculins tels que:

cheval..N4 *canal..N4*

Au total, 80 classes flexionnelles ont été répertoriées pour l'ensemble des noms et adjectifs. La liste détaillée des flexions est donnée en annexe 1. Elle a eté organisée de façon à regrouper d'abord les mots exclusivement masculins (classes 0 à 19), puis les mots exclusivement féminins (classes 20 à 29), enfin les mots à double genre (classes 30 à 80).

2.3.2. *Codification des verbes*

Alors que le code morphologique des noms et adjectifs renvoie à une flexion, celui des verbes renvoie à une conjugaison-modèle, identifiée par son numéro. Par exemple, on trouve dans le DELAS l'entrée verbale:

construire..V91

spécifiant que ce verbe se conjugue comme le verbe-modèle de numéro 91: *cuire*.

Déjà au siècle dernier, une classification systématique avait été élaborée par M. Bescherelle Aîné pour décrire la conjugaison française. Depuis, plusieurs systèmes de classement des verbes

ont été proposés, certains basés sur les variations phonétiques des formes conjuguées. Dans le DELAS, seuls les critères orthographiques ont été pris en compte pour établir les 96 conjugaisons-modèles auxquelles se réfère le codage des verbes.

Le DELAS étant simplement un dictionnaire morphologique, les caractères syntaxiques de construction des verbes ne s'y trouvent pas. Ils sont décrits ailleurs dans les tables syntaxiques du lexique-grammaire du LADL. Cependant, certains verbes présentent dans leur conjugaison des particularités provenant de leur construction, qui nécessitent l'emploi de marqueurs spécifiques. Ce sont les cas suivants:

1) verbes à participe passé invariable, c'est-à-dire verbes intransitifs, qui ne peuvent pas se conjuguer avec l'auxiliaire *être*. Alors le marqueur U est utilisé pour signaler une forme unique pour le participe passé, par exemple:

circuler,.V3U.

2) verbes qui, pour diverses raisons, ne se conjuguent qu'à la troisième personne. Ainsi sont les verbes dits impersonnels pour lesquels le sujet est obligatoirement *il*. Ceux-ci sont codés avec le marqueur I, comme:

neiger,.V5I *(il neige)*

3) verbes défectifs: la défectivité est un phénomène irrégulier, différent selon chaque verbe concerné. Elle est donc indiquée par le marqueur général D spécifiant l'existence de formes manquantes dans la conjugaison. Ainsi est le verbe:

frire,.V90D

Le marqueur D n'est pas accompagné de la liste des temps et personnes manquants. Ces informations sont reportées dans un fichier externe, qui peut être exploité par le programme de génération des formes verbales de façon à éviter la production des formes inexistantes.

3. STATISTIQUES SUR LE LEXIQUE DELAS

Le lexique DELAS est divisé en 26 fichiers, où les mots sont répartis en fonction de leur lettre initiale, comme dans un dictionnaire classique, et triés alphabétiquement. Les statistiques sont évaluées d'une part sur chaque fichier isolé, d'autre part sur la totalité du lexique.

Une page en annexe donne la répartition des mots, par lettre initiale et au total, dans chacune des catégories grammaticales suivantes: adjectifs, noms, verbes, adverbes. Les autres catégories de mots sont comptabilisées ensemble. Les totaux, apparaissant en bas de page, montrent que la version du DELAS de 1988 comporte près de 10 000 verbes, 17 000 adjectifs, 42 000 noms et 2800 adverbes.

Le lexique DELAF

DELAF est l'ensemble exhaustif et ordonné de toutes les formes fléchies des mots du DELAS. Celles-ci sont construites à l'aide d'une procédure automatique de génération des formes. Nous devons donc considérer celle-ci, avant d'aborder la constitution des entrees du DELAF.

1. LA GÉNÉRATION DES FORMES FLÉCHIES

Dans le système de codage morphologique du DELAS, on distingue les noms et adjectifs d'une part, et les verbes d'autre part. La génération automatique des formes fléchies se subdivise suivant ce principe en deux types de traitements:

1) le premier engendrant les formes fléchies des noms et des adjectifs,

2) le second servant à conjuguer les verbes, et à construire l'ensemble des formes verbales.

1.1. Génération des formes nominales et adjectivales

Sur le plan formel, les formes nominales et adjectivales sont composées de deux éléments:

- un radical R, indépendant du genre et du nombre,
- une terminaison T, variable en genre et en nombre.

Nous avons déjà vu que la terminaison T est un vecteur à quatre composantes, équivalant à une flexion numérotée et formalisée comme suit:

$$F_n = (tms.\ tfs,\ tmp,\ tfp)$$

où n = numéro de la flexation,
tms = terminaison du masculin singulier,
tfs = terminaison du féminin singulier,
tmp = terminaison du masculin pluriel.
tfp = terminaison du féminin plur

Dans le programme de génération automati[illegible] mots fléchis sont donc calculés par une expression de la forme:

$$R.F_n$$

où R désigne le radical commun à toutes les formes fléchies d'un même mot.

Suivant les cas, un nom ou un adjectif engendre de une à quatre formes fléchies, comme le montrent les exemples ci-dessous:

bois..N2	-->	*bois, ms + mp*
balle,.N21	-->	*balle, fs balles, fp*
pâle,.A31	-->	*pâle, ms + fs pâles, mp +fp*
pieux,.A63	-->	*pieux, ms +mp pieuse, fs pieuses, fp*
acteur..N36	-->	*acteur, ms actrice, fs acteurs, mp actrices, fp*

1.2. Génération des formes verbales

Celle-ci est basée sur un programme de conjugaison écrit au LADL, et exploitant les 96 verbes-modèles utilisés lors du codage morphologique des verbes du DELAS. Pour le programme,

deux fichiers auxiliaires sont nécessaires: le premier, dit fichier de terminaisons, contient la liste de toutes les terminaisons existantes à tous les temps de la conjugaison; le second, dit fichier de conjugaisons, contient la description de la conjugaison des 96 verbes-modèles sélectionnés. Ces deux fichiers permettent au programme le calcul automatique des formes conjuguées des verbes du DELAS, à partir de leur numéro de code morphologique.

Seuls les temps simples de la conjugaison donnent lieu à des formes conjuguées simples, donc seules ces dernières sont prises en considération pour construire le dictionnaire DELAF. Un verbe régulier du premier groupe, par exemple *inventer*, donne lieu à 39 formes simples de graphies différentes, dont

- 6 formes impersonnelles:
  invent-[er, ant, é, és, ée, ées].

- et 33 formes personnelles:
  invent-[e, es, ons, ez, ent] = 5 formes
  invent-[ais, ait, ions, iez, aient] = 5 formes
  invent-[erai, eras, era, erons, erez, eront] = 6 formes
  invent-[erais, erait, erions, eriez, eraient] = 5 formes
  invent-[ai, as, a, âmes, âtes, èrent] = 6 formes
  invent-[asse, asses, ât, assions, assiez, assent] = 6 formes

Sur le plan théorique, les verbes sont conjugables à toutes les personnes de tous les modes-temps. Cependant, dans l'usage courant, un certain nombre de formes verbales sont rarement employées, comme celles de l'imparfait du subjonctif. Malgré tout, ces formes existent, elles sont donc engendrées par le programme de conjugaison automatique du LADL. L'ensemble des formes résultantes est un ensemble maximal. Le sous-ensemble des formes effectivement rencontrées dans les textes est fonction à la fois du style des auteurs, et du domaine d'application.

2. LA CONSTITUTION DES ENTRÉES DU D E L A F

Une entrée du DELAF est construite en deux parties: 1) une forme canonique ou fléchie, 2) son identification en genre et nombre pour les formes nominales et adjectivales, ou son identification en mode, temps, personne et nombre pour les formes verbales. Les enregistrements du DELAF, obtenus informatiquement, se présentent comme il apparait sur les exemples ci-dessous:

table,.N21:Nfs
tables,-1.N21:Nms
soigneux,.A63:Ams:Amp
soigneuse,-2x.A63:Afs
soigneuses,-3x.A63:Afp
grandirais,-3.V18:C1s:C2s

dans lesquels:

:Nfs et :Nfp spécifient des formes nominales au féminin singulier et féminin pluriel,

:Ams :Amp :Afs :Afp désignent des formes adjectivales, respectivement masculin singulier, masculin pluriel, féminin singulier, féminin pluriel,

:C1s :C2s représentent le conditionnel présent, 1re et 2e personnes du singulier.

Les expressions précédées du signe - permettent de retrouver la forme canonique. Par exemple: *soigneuses,-3x* signifie qu'en enlevant trois caractères en fin de mot et en ajoutant x, on obtient le mot de base *soigneux*.

Par rapport au nombre de mots simples contenus dans le DELAS, le nombre des formes engendrées et stockées dans le DELAF est multiplié au plus par:

2 pour un nom de genre unique,
4 pour un adjectif ou un nom à double genre,
39 pour un verbe régulier du primier groupe.

Ces proportions sont maximales, car beaucoup de noms et adjectifs sont invariables en genre et ne donnent que deux formes. Néanmoins le dictionnaire DELAF, engendré à partir de la version du DELAS de 64 000 entrées, contient de l'ordre de 530 000 formes. En machine, le volume est près de 10 fois plus important que celui du DELAS. Ce phénomène d'expansion est d'autant plus marqué que les entrées du DELAS sont factorisées, c'est-à-dire qu'elles peuvent contenir, en une entrée unique, plusieurs mots de même graphie. Au contraire, les entrées du DELAF ne sont que partiellement factorisées. En effet, les formes identiques ne sont regroupées que si elles proviennent du même mot grammatical. C'est par exemple le cas des deux formes de l'adjectif *soigneux* au masculin singulier et pluriel.

La différence de présentation et de volume entre les deux dictionnaires DELAS et DELAF ne doit pas faire oublier leur unité et leur cohérence. A la base, les informations sont essentiellement les mots du DELAS avec leur morphologie. Que toutes les formes soient déployées de façon exhaustive comme dans le DELAF, ou qu'elles soient implicitement présentes grâce au code du DELAS, l'ensemble formel représenté reste toujours le même: celui des mots simples de notre langue.

Le lexique DELAC

Le Dictionnaire Électronique du LADL pour les mots composés (le DELAC) comporte à ce jour plus de 80 000 mots composés. Le DELAS et le DELAC sont des dictionnaires électroniques morphologiques, c'est-à-dire que leurs entrées sont associées à une catégorie grammaticale et a un code flexionnel. Cette description permet en particulier de reconnaître automatiquement les mots dans les textes. On peut ainsi reconnaître que dans la phrase:

Les pieds noirs sont venues

l'occurrence *pieds noirs* représente le nom composé *pied noir* (qui peut être féminin) au pluriel, *sont* représente la troisième personne du pluriel du présent de l'indicatif du verbe *être*, et *venues* représente le participe passé féminin pluriel du verbe *venir*.

Blandine Courtois recense les mots simples; Maurice Gross recense les adjectifs et adverbes composés; Gaston Gross, René Jung, Michel Mathieu-Colas et Robert Vivès recensent les noms composés. Ces derniers constituent la majorité des mots composés, et sont classés selon leur structure. Nous avons actuellement les classes suivantes:

-NA (Nom Adjectif): *carte bleue, pied noir,*
-NDN (Nom de Nom): *pomme de terre, coup de force.*
-AN (Adjectif Nom): *beau frère, blanc-bec,*
-NN (Nom Nom): *homme grenouille, chien-loup,*
-NAN (Nom à Nom): *pelle à gâteau, tarte à la crème,*
-VN (Verbe Nom): *trompe l'oeil, gratte-papier,*
-PN (Préposition Nom): *en cas, arrière garde,*

Il existe de nombreuses autres classes de noms composés ([Michel Mathieu 1988j), mais qui représentent nettement moins de noms composés. Les classes NA et NDN sont les plus importantes en nombre.

Classe	Nombre de mots
NA	44 985
NDN	20 899
AN	1 324
NN	2 376
NAN	2 542
PN	604
VN	1 021

Notre travail consiste à créer à partir de ces listes brutes le dictionnaire électronique DELAC, utilisable par des programmes de traitement automatique. Il s'agit donc d'attribuer des codes de flexion aux différentes catégories et d'introduire une classification morpho-syntaxique de ces termes. La différence entre le DELAC et les listes données par des lexicographes ne concerne pas uniquement la forme des données; elle concerne la **qualité** de l'ensemble des données, et se traduit par un **ajout d'informations**. Cette qualité est difficilement mesurable, mais ce n'est pas pour cela qu'elle est facile à obtenir:

- plus de 10 % des entrées lexicales recensées par les lexicographes ne sont pas cohérentes au sens formel, or les programmes ne peuvent traiter que les bases de données *ne contenant aucune erreur*; il est donc indispensable de reconnaître les données génératrices de bruit.

- les informations données par les lexicographes concernant le nombre, le genre, la flexion en nombre et en genre des noms composés font souvent défaut, et ne sont pas toujours cohérentes; or, pour être utilisable par des procédures d'analyse automatiques, le DELAC doit décrire sans aucune erreur le comportement flexionnel des noms composés; il est donc indispensable de détecter et de corriger les informations erronées, et d'ajouter les informations manquantes.

Toutes les procédures présentées sont fondées sur un outil (automate non déterministe) qui permet de reconnaître dans des textes ou dans des listes des séquences comprenant des mots explicites, des formes lemmatisées, ou des parties du discours. Il est possible par exemple de reconnaître toutes les lignes qui débutent par la séquence *grand/*, suivie par un nom féminin, suivi par la séquence *:une*. Cette séquence est entrée ainsi:

%grand/<N-f>:une

Cette séquence reconnait par exemple les lignes:

grand/voile:une
grand/rue:une
grand/mère:une

<apercevoir> représente toutes les formes du verbe *apercevoir*; la séquence *j'<V-1s>* reconnait toutes les séquences de *j'* suivi d'un verbe à la première personne du singulier, etc.

La construction d'un dictionnaire électronique à partir de listes brutes est une opération relativement complexe; ceci est dû à au moins trois raisons:

- le nombre d'entrées traitées est de l'ordre de plusieurs dizaines de milliers; ceci entraîne que les traitements à effectuer sont longs et lourds;

- de nombreuses procédures doivent être mises en oeuvre et être reliées entre elles; le nombre de fichiers intermédiaires générés est de l'ordre de plusieurs centaines;

- certaines procédures ne peuvent pas être automatiques.

Nous présentons en annexe l'organigramme des procédures effectuées. Les fichiers générés sont représentés par des cercles. Les cercles numérotés correspondent à des problèmes discutés dans ce rapport. Par exemple, le cercle numéroté 0 correspond à une liste brute. Les procédures sont représentées par des rectangles, marqués « A » (procédure automatique) ou « M » (procédure manuelle). Certaines procédures génèrent plusieurs fichiers. Par exemple, la vérification orthographique d'un fichier de noms composés génère deux fichiers: le fichier des noms composés sans faute d'orthographe détectée, et le fichier des noms composés qui contiennent une faute supposée. Dans ce cas, un des fichiers résultats est marqué « O » (oui: sans erreur).

1. LE FORMAT DES LISTES

Le format des listes de noms composés dépend de la classe considérée. Nous le donnons ici:

Classe	Format	Exemple
NA	<MD>/<MD>:<D>	*carte/bleue:une*
	<MD>-/<MD>:<D>	*Pays-/bas:les*
NDN	<MD>/de//<MD>:<D>	*pomme/de//terre:une*
	<MD>/de/<MD>/<MD>:<D>	*accident/de/la/route:un*
AN	<MD>/<MD>:<D>	*grand/père:un*
	<MD>-/<MD>:<D>	*basse-/cour:une*
	<MD>'/<MD>:<D>	*grand'/mère:une*
NN	<MD>/<MD>:<D>	*homme/grenouille:un*
	<MD>-/<MD>:<D>	*abri-/bus:un*
NAN	<MD>/à//<MD>:<D>	*manche/à//balai:un*
	<MD>/à/<MD>/<MD>:<D>	*tarte/à/la/crème:une*
NàV	<MD>/à/<MD>:<D>	*sauc/à/manger:une*
VN	<MD>-/<MD>:<D>	*garde-/manger:un*
PN	<MD>-/<MD>:<D>	*arrière-/garde:une*

Le symbole <MD> représente tous les mots simples que l'on trouve dans le DELAF. Le symbole <D> représente une séquence parmi les neuf suivantes:

E, le, la, les, un, une, des, de le, de la

E représente le déterminant vide, que l'on met pour certains noms propres, par exemple: *François/premier:E*. Le caractère « / » est un séparateur de zones, utilisé dans des programmes d'extraction et de tri. Toutes les entrées d'une classe déterminée ont le même nombre de zones, et donc de séparateurs « / ». Par exemple, les classes NDN et NAN ont quatre zones, les classes NA et AN ont deux zones, etc. Le caractère « : » sépare l'entrée lexicale des informations

associées. Il y a un caractère « : » par entrée, quelle que soit la classe de noms composés. La seule information associée aux entrées est le déterminant choisi parmi les neuf séquences correspondant à <D>.

Certains noms composés s'écrivent avec un trait d'union; on place celui-ci avant le séparateur de zones. Remarquons en particulier que tous les noms composés VN ou PN ont un trait d'union.

Le premier traitement est un programme qui vérifie le format des listes données. A partir de la liste « brute » (fichier 0 en annexe), ce programme d'acquisition des données engendre deux fichiers: la sous-liste des entrées bien formées, et la sous-listes des entrées rejetées. Cette dernière peut représenter plus de 5 % de la liste brute. Les entrées rejetées correspondent aux trois types:

1) Les fautes de frappe concernant la ponctuation, comme par exemple un mauvais usage du caractère « / », un mauvais nombre de zones, etc.; les entrées correspondantes peuvent représenter plus de 2 % des entrées. Par exemple, sur une liste de 2 376 NDN, il y a 61 fautes de ce type (2,5 %). La correction de ces fautes ne pose pas de problème, mais elle nécessite bien entendu un traitement spécifique (détecter et corriger les fautes, puis fusionner la liste corrigée avec la liste globale).

2) Les entrées dont le déterminant n'est pas valide. On trouve parfois l'article contracté *:du* au lieu de sa forme explicite *:de le*, *:de les* au lieu de *:des*, ainsi que des fautes de frappe dans les déterminants. Ces fautes touchent plus de 2 % des entrees (par exemple, 457/20946 NDN). De même que pour les fautes de ponctuation, il faut détecter et corriger ces fautes, puis fusionner la liste corrigée avec la liste globale.

Nous corrigeons les entrées qui possèdent une ou plusieurs fautes de ponctuation ou de déterminant.

3) Les mots composés dont la structure ne correspond pas au schéma de définition (fichier 1 dans l'organigramme). Par exemple, les entrées:

abri/anti-atomique:un (NA)
art/arabo-musulman:le (NA)
non-prolifération/nucléaire:une (NA)
sous-marin/atomique:un (NA)

sont de notre point de vue mal classées, car on y trouve des mots composés à l'intérieur d'une zone; les entrées:

bernard/l'ermite:un (NN)
bernard/l'hermite:un (NN)

de la liste NN ne sont pas cohérentes vis-à-vis des definitions formelles à la base des traitements: ces entrées obéissent à la définition formelle *Nom/Déterminant/Nom*. et non pas *Nom/Nom*. L'entrée:

traveller's/chèque:un (NN)

pose le problème général du traitement de l'apostrophe. Nous traitons toutes les séquences qui comportent une apostrophe, comme:

aujourd'hui, *entr'acte*, *entr'aider*, *grand'mère*, *grand'rue*, *Levi's*, *O'Connors*, *prud'homme*, *traveiler's chèque*, etc.

ces mots composés sont associés à une classe et à une procédure spécifiques. Les noms composés du type Adjectif/Nom qui ont une apostrophe entre l'adjectif et le nom (*grand'rue*) sont décrits dans la liste AN.

4) Les mots composes dont une partie simple ne figure pas dans le dictionnaire DELAF. Ces mots concernent des fautes de frappe ou d'orthographe d'usage, des noms propres (**Alexandre** *le grand*, l'**Afrique** *équatoriale*, etc.), et aussi des mots communs valides qui manquent dans le DELAF. Les entrées de ce type peuvent représenter jusqu'à 4 % de la liste brute. Par exemple, sur la liste de 21 573 noms composés NDN, 627 ont été rejetés lors de la vérification orthographique (2,9 %). Afin de récupérer les noms propres, nous extrayons de la liste des 627 mots rejetés les mots rejetés en majuscule. On obtient alors deux listes:

a) la liste de mots composés qui comportent un mot en majuscule non trouvés dans le dictionnaire (fichier 2 dans l'organigramme): l'**Europe** *occidentale*.

b) la liste des mots composés qui comportent un mot en minuscule non trouvés dans le dictionnaire DELAS (fichier 3 dans l'organigramme):

*acide/***parabutoxyphénylacéthydroxamique***:de le*

Les premiers mots sont pour la plupart des noms propres, mais certains peuvent être erronés (*l'Archipel des Galapos*); les seconds sont soit des mots fautifs, soit des mots valides qu'il faudra ajouter dans le dictionnaire des mots simples.

2. LA STRUCTURE DES ENTRÉES

La simple verification de l'orthographe et du format des entrées ne suffit pas. Les listes doivent aussi être cohérentes du point de vue de la structure de leurs entrées:

Classe	Format	Exemple
NA	<N>/<A>:<D>	*carte/bleue:une*
	<N>-/<A>:<D>	*Pays-/bas:les*
	<N>/<V-ant>:<D>	*agent/neutralisant:un*
	<N>-/<V-ant>:<D>	*cerf-/volant:un*
	<N>/<V-pp>:<D>	*crime/organisé:le*
	<N>-/<V-pp>:<D>	*cou-/nu:un*
NDN	<N>/de//<N>:<D>	*pomme/de//terre:une*
	<N>/de/<DET>/<N>:<D>	*accident/de/la/route:un*
AN	<A>/<N>:<D>	*grand/père:un*
	<A>-/<N>:<D>	*basse-/cour:une*
	<A>'/<N>:<D>	*grand'/mère:une*
NN	<N>/<N>:<D>	*homme/grenouille:un*
	<N>-/<N>:<D>	*abri-/bus:un*
NAN	<N>/à//<N>:<D>	*manche/à//balai:un*
	<N>/à/<DET>/<N>:<D>	*tarte/à/la/crème:une*
NàV	<N>/à/<V>:<D>	*salle/à/manger:une*
VN	<V>-/<N>:<D>	*garde-/manger:un*
PN	<PREP>-/<N>:<D>	*arrière-/garde:une*

avec:

<A> représente un adjectif (*bleue*),
<DET> représente un déterminant (*la*),
<N> représente un nom (*carte*),
<PREP> représente une préposition (*arrière*),
<V> représente un verbe (*manger*),
<V-ant> représente un verbe au participe présent (*volant*),
<V-pp> représente un verbe au participe passé (*volé*),

Remarque

Nous avons placé les noms composés des types:

<N>/(<A>+<Vant>+<V-pp>)

dans la liste NA. De façon analogue, on pourrait ranger les noms composés:

(<A>+<V-ant>+<V-pp>)/<N>

dans la liste AN. Actuellement, cela n'a pas d'importance car nous n'avons pas de nom composé *<V-ant>/<N>*. Les seuls noms composés pouvant être considérés comme *<V-pp>/<N>* sont:

Sacré/Coeur:le
Sacré/Collège:le

car *sacré* est décrit dans le DELAS comme adjectif et aussi comme participe passé.

Ces contraintes sont la conséquence du fait que le dictionnaire DELAC fait partie du système de dictionnaires électroniques du LADL qui doit être cohérent au niveau morphologique. En particulier, *<A>* représente les mots codés comme adjectifs **dans le DELAS**.

Nous effectuons donc une deuxième passe de vérification, cette fois plus précise, ce qui permet de rejeter des noms composés qui correspondent soit à une faute dans le DELAS, soit à une faute de classement du nom composé (fichier 4 dans l'organigramme). Par exemple, nous avons trouvé dans une liste NA les entrées suivantes:

argot/polytechnicien:le
cellule/tueuse:une
championne/amateur:une
beaujolais/primeur:de le
prévisions/météo:des

Ces noms composés sont mal classés par rapport au dictionnaire DELAS, dans lequel les mots *polytechnicien*, *tueuse*, *amateur*, *primeur* et *météo* sont décrits exclusivement comme des noms. *championne/amateur:une* devra probablement être classé comme NN; par contre, *tueuse* et *météo* seront codés comme adjectifs dans la prochaine version du DELAS. De même, le mot *blindé* figurait dans le DELAS (version 4) comme participe passé du verbe *blinder*, mais non comme nom. Le nom composé NA *blindé/léger:un* avait alors été rejeté à tort.

3. LE DÉTERMINANT ET LA DESCRIPTION FLEXIONNELLE

Pour chaque entrée, les lexicographes choisissent un déterminant parmi les neuf:

E, le, la, les, un, une, des, de le, de la

Le déterminant « zéro » (E) n'a été utilisé jusqu'à présent que dans deux NA; le déterminant un est de loin le plus fréquent.

Déterminant	AN	NDN	VN	PN	NAN	NA
un	446	10 278	984	319	1 233	15 996
une	229	7 096	21	250	860	16 011
le	262	1 038	5	16	73	3 764
la	208	862	1	11	162	4 986
de le	32	549	8	5	52	1 156
de la	12	442			53	457
E						2
les	112	222			15	1 254
des	23	412	2	3	94	1 359
TOTAL	1 324	20 899	1 021	604	2 542	44 985

Du point de vue du nombre et du genre des noms composés, le déterminant a plusieurs fonctions qu'il est important de préciser:

3.1. Le nombre

Le déterminant décrit le nombre du nom composé:

- *E, le, la, un, une, de le,* et *de la* représentent un nom composé singulier,
- *les* et *des* représentent un nom composé pluriel.

De plus, il représente la flexion singulier->pluriel:

- *un* et *une* désignent un nom composé qui accepte le pluriel,
- *E, le, la, de le, de la, des* et *les* désignent des noms composés invariables en nombre.

Le nombre des noms composés n'est pas toujours facile à déterminer. Par exemple, pourquoi le nom composé:

fous/de//Dieu:les (NDN)

a-t-il été entré au pluriel (déterminant *les*), et non pas au singulier *fou/de//dieu:un?* Le déterminant les signifie-t-il que les phrases:

Luc est un fou de dieu.
Un fou de dieu s'est suicidé hier par le feu.

sont inacceptables en français?

Par ailleurs, on trouve des noms donnés de façon indépendante au singulier et au pluriel. Parfois, c'est à dessein qu'on a dédoublé les entrées. Par exemple:

- *un aide social* est un homme:

  L'aide social viendra visiter la maison demain.

- *une aide sociale* est soit une femme (féminin de *un aide social*), soit une contribution:

  L'aide sociale viendra visiter la maison demain.
  Luc reçoit une aide sociale de 1 000 F par mois.

- *les aides sociales* peut désigner soit le pluriel du nom ambigu *une aide sociale*, soit éventuellement l'organisme, comme dans la phrase:

  Luc envoie sa fiche de paie aux aides sociales.

mais bien souvent, le dédoublement résulte d'un manque de vérification. Certaines listes ont été construites indépendamment par plusieurs lexicographes. Il n'y a pas moyen de savoir si les dédoublements observés sont significatifs ou non, puisque les noms composés n'ont pas d'information de type distributionnelle ou sémantique associée (comme par exemple, **humain** ou **non humain**).

La variabilité des noms composés ne peut pas être décrite dans une liste indépendamment de considérations syntaxiques. Il faudrait plonger le nom dans une phrase simple. Nous reprenons l'exemple de Gross 1988: le nom composé *planche/à//voile* (NAN) peut ou non se mettre au pluriel, cela dépend. Dans la phrase:

Luc construit une planche à voile.

c'est un nom concret, qui peut se mettre au pluriel (*Luc construit des planches à voiles*); par contre, dans la phrase:

Luc fait de la planche à voile.

il ne peut pas être mis au pluriel. Il faudra donc étudier pour chaque nom la possibilité qu'il y ait un verbe support associé (ici **faire** *de la planche à voile*), et les contraintes dues au verbe support autorisé, puisqu'elles peuvent varier. En attendant que cette étude soit faite, le nombre attaché aux noms composés est approximatif.

Comme le nombre d'un nom est une notion purement lexicale, il est indispensable que le DELAC et le DELAS soient cohérents sur le plan strictement morphologique. Nous avons donc vérifié que lorsque le déterminant est singulier (*E, un, le, de le, une, la, de la*), le nom composé est constitué de mots singuliers (procédure analogue avec le pluriel).

Cette procédure nous a permis de trouver la seule erreur: *coupe/sombres:des* (fichier 7). Il est remarquable que la détection d'une seule erreur (sur plusieurs dizaines de milliers d'entrées!) nécessite une procédure spécifique relativement coûteuse: on doit vérifier le nombre de chaque

mot dans le dictionnaire. Nous n'avons pas trouvé de cas exceptionnel, c'est-à-dire de nom composé dont le nombre ne correspondrait pas à celui de ses constituants (comme *une deux chevaux*).

3.2. Le genre

D'une part, le déterminant ne décrit pas le genre des noms composés dans les trois cas suivants:

- *E* est masculin ou féminin:

  Action/directe:E (NA), *François/premier:E* (NA),

- *un* désigne un nom composé masculin, qui peut avoir ou non le double genre:

  chef/de/service:un (NDN) (double genre)
  chef/de//inculpation:un (NDN) (masculin uniquement)

- *les* et *des* désignent un nom composé masculin ou féminin:

  bains/douche:des (NN)
  avaries-/frais:des (NN)

D'autre part, le déterminant ne décrit pas toujours la flexion masculin->féminin:

- *un* désigne un nom masculin variable ou non en genre:

  appui/politique:un (NA), *cousin/germain:un* (NA)

- *les* et *des*, quand ils s'appliquent à des noms masculins, peuvent désigner des noms variables ou non en genre.

Pour les noms associés aux déterminants *un, de le* et *le* d'une part, *une, de la* et *la* d'autre part, nous avons vérifié que le genre des parties simples du nom composé correspond bien à celui du déterminant. Bien sûr, nous avons trouvé des erreurs (27 erreurs pour les NA, 36 pour les NDN, 2 pour les AN), mais aussi des cas exceptionnels, comme:

grosse/légume:une (AN)
peau/rouge:un (NA)

Nous avons vu que le déterminant ne décrit pas toujours le genre et le comportement flexionnel en genre des noms composés. Nous devons donc rétablir le genre des noms composés pluriel, et décrire la flexion en genre des noms composés en *un, les* ou en *des*.

4. LE GENRE DES NOMS COMPOSÉS PLURIEL

Nous traitons la liste des noms composés au pluriel (liste 12 en annexe). Nous distinguons les classes NA, AN, NDN, NAN, et NàN d'une part, des classes VN, et PN d'autre part. Pour les premiers, il est possible d'utiliser le genre d'une partie simple « caractéristique » du nom composé afin de déterminer le genre du nom composé:

Classe	Caractéristique
AN	A et N
NA	N et A
NDN	le premier N
NAN	le premier N
NàV	N

Pour les autres classes, il n'est pas possible de déduire le genre du nom composé à partir des parties simples: nous avons donc dû entrer « manuellement » le genre des noms composés VN et PN au pluriel. Par ailleurs, le genre des deux noms d'un NN n'est pas toujours le même:

animaux-/machines:des
avaries-/frais:des
public-/relations:des
...

remarquons que le genre du nom composé peut être celui du premier nom (*animaux-/machines*, *avaries-/frais* sont féminins), ou du deuxième nom (*public-/relations*); les NN pluriel sont peu nombreux (36), nous avons donc mis le genre manuellement.

Nous donnons ici les deux procédures pour mettre le genre des noms composés. Le premier exemple concerne la classe NDN, pour laquelle on se fonde sur le premier nom. Cette routine est appliquée telle quelle pour la classe NAN. Le second exemple concerne la liste NA, pour laquelle deux mots apportent des informations. Cette routine est aussi appliquée à la liste AN.

4.1. Les NDN pluriel

A partir de la liste des 659 noms composés NDN obligatoirement au pluriel, on génère deux fichiers:

- le fichier *FEMININ* des noms composés dont le premier nom est obligatoirement féminin (272 noms composés), par exemple:

aigreurs/de//estomac:des
bottes/de//chasse:des

- le fichier *AUTRES* des noms composés dont le premier nom est masculin (mais il peut être aussi au féminin), soit 387 noms composés.

Cette classification utilise le programme *extra* qui compare la contrainte *%<N-m>* (les lignes qui commencent par un nom masculin) avec les lignes de la liste des noms composés au pluriel. *AUTRES* vérifie la contrainte, *FEMININ* non. La commande utilisée est:

extra PLURIEL % <N-m> AUTRES FEMININ

Les noms composés de *AUTRES* peuvent être soit masculins, soit masculins et féminins. Nous les séparons en utilisant la même procédure, cette fois avec la contrainte *%<N-f>*:

extra AUTRES %<N-f> MANUEL MASCULIN

qui génère le fichier *MANUEL* des 46 noms composés dont le premier nom est soit à double genre, par exemple:

enfants/de/les/écoles:les

soit ambigu, par exemple:

manoeuvres/de//abordage:des

et le fichier *MASCULIN* des noms composés dont le premier nom est exclusivement masculin (341 noms composés), par exemple:

barons/de/la/presse:les
cheveux/de//ange:des

Le genre féminin est automatiquement mis pour les 272 noms du fichier *FEMININ*; le genre masculin est automatiquement mis pour les 341 noms du fichier *MASCULIN*; le genre des 46 noms du fichier *MANUEL* est mis « à la main ». Après avoir introduit le genre dans ces trois fichiers, nous fusionnons la liste des noms au singulier enrichie automatiquement du genre, avec les trois fichiers *FEMININ*, *MASCULIN* et *MANUEL*.

4.2. Les NA pluriel

On utilise cette fois le genre du nom et de l'adjectif. Le nom présente trois possibilités: soit masculin (m), soit féminin (f), soit les deux (mf); l'adjectif aussi; nous avons donc les neuf possibilités suivantes:

Nom	Adjectif	Exemple
m	m	*abonnés/absents:les*
m	f	*?*
m	mf	*accidents/domestiques:les*
f	m	*?*
f	f	*activités/culturelles:des*
f	mf	*archives/judiciaires:des*
mf	m	*Artistes/Associés:les*
mf	f	*manches/courtes:des*
mf	mf	*livres/historiques:les*

Il nous faut séparer les entrées de façon à avoir les listes:

- noms composés féminins: f-f, f-mf, mf-f
- noms composés masculins: m-m, m-mf, mf-m
- noms ambigus (que l'on devra traiter manuellement): mf-mf
- erreurs que l'on doit rejeter, ou cas exceptionnels: m-f, f-m.

Comme on le voit en annexe, les programmes *extra* génèrent les 7 fichiers suivants:

Fichier	Nombre	Type	Exemple
13	59	mf-mf	livres/historiques:les
14	54	mf-m	Artistes/Associés:les
15	34	mf-f	manches/courtes:des
16	1010	m-m,m-mf	abonnés/absents:les
17	2	m-f	Indes/occidentales:les
18	1439	f-mf,f-f	archives/judiciaires:des
19	4	f-m	Bouffes/parisiens:les

Le fichier 13 comporte les noms composés pour lesquels il n'est pas possible de déterminer le genre à partir des composants; les fichiers 14 et 16 comportent des noms composés masculins; les fichiers 15 et 18 comportent des noms composés féminins; les fichiers 17 et 19 sont:

Liste 17:

Indes/orientales:les
Indes/occidentales:les

Le mot *inde* (sans majuscule) est codé dans le DELAS comme un nom masculin (nom de couleur, qui signifie bleu indigo). Nous n'avons actuellement pas de dictionnaire de noms propres, et nous ne reconnaissons donc pas les noms propres qui sont homographes avec un mot commun du DELAS (par exemple Pierre). Nous ne pouvons pas nous fier exclusivement en la présence d'une majuscule pour reconnaître les noms propres, car il existe de nombreux noms propres composés dont une ou plusieurs parties simples sont écrites en majuscules, sans pour autant être elles-mêmes des noms propres: *les* **Artistes Associés**, *la* **Banque Nationale** *de Paris*, *une* **Carte** *Visa*, *le* **Secours** *catholique*, etc.

Liste 19:

Bouffes/parisiens:les
glycérides/partiels:les
langues/altaïques:les
lettres/royaux:les

Bouffes peut être un nom propre masculin (même phénomène qu'avec *Inde*). *glycérides-/partiels:les* a été entré comme faute parce que dans la version 4 du DELAS, *glycérides* n'était pas entré comme nom masculin. *altaïques* a été entré dans le DELAS version 4 comme masculin exclusivement, ce qui doit être corrigé. L'entrée *lettres/royaux:les* est une faute.

5. LE GENRE DES NOMS COMPOSÉS DE DÉTERMINANT UN

Les noms composés au singulier qui ont le déterminant *un* doivent être aussi étudiés du point de vue du genre. Nous distinguons trois types de noms composés en *un*:

1) les noms qui ne peuvent être qu'uniquement masculins à cause de leurs composants: *terrain de jeu* ne peut pas être féminin puisque *terrain* n'a pas de féminin. Cette liste correspond à la liste 20 en annexe.

Cette description est limitée par le fait que notre outil est purement morphologique: dans la phrase *une pied noir est venue*, le nom composé est féminin, même si *pied* n'a pas de féminin.

Cette possibilité n'est pas décrite dans la liste NA, où n'apparait que l'entrée *pied/noir:un*. Ce phénomène doit être étudié généralement pour tous les noms humains; l'étude a été entreprise pour les NA par Gui on 1988.

2) les noms à double genre: *enfant pauvre* peut être masculin ou féminin, puisque *enfant* et *pauvre* ont le double genre;

3) les noms qui ont une flexion en genre: *cousin germain* donne *co. ine germaine*.

Afin de détecter les noms composés du deuxième type, nous extrayons de la liste des noms associés au déterminant *:un* les noms composés dont le nom et l'adjectif peuvent être aussi féminins. Par exemple, pour les NA, nous appelons la commande:

extra na.un %<N-f>/(<A-f> + <V-ppf>) DOUBLE MASCULIN

	NA	NDN	AN	NAN
un	15 996	10 278	446	1 233
DOUBLE	774	4	56	56
double genre	160	456	2	7

Le fichier *DOUBLE* (21 en annexe) contient des noms composés dont le nom (pour les NDN, NAN), ou le nom et l'adjectif (pour les NA et AN) sont à double genre. Nous étudions systématiquement les noms de ce fichier, et dédoublons les noms composés à « vrai » double genre. Ce n'est pas parce qu'un nom composé a structurellement le double genre qu'il l'a globalement. Considérons les entrées:

sale/tour:un (AN)
enfant/de/la/Assistance:un (NDN)

sale/tour est obligatoirement au masculin; il est dans la liste DOUBLE parce que *sale* est un adjectif masculin ou féminin, et *tour* est ambigu (*une tour* ou *un tour*). Par contre, *enfant de l'Assistance* est rangé dans la liste 23, puis est dédoublé.

Cette procédure ne peut pas être automatique, même si les noms simples sont marqués **humain** ou **non humain**: par exemple, seule une connaissance des noms composés suivants permet de savoir si l'on a affaire à un nom humain à double genre:

agent/de//conservation (NDN)
agent/de//maîtrise (NDN)
chef/de//inculpation (NDN)
chef/de//cabinet (NDN)

Les listes de noms composés des classes NDN et NAN dont le déterminant est *un* sont traitées de façon analogue:

extra nan.un %<N-f> double masc

En utilisant ces procédures, nous avons pu ajouter des noms composés féminins « cachés ». Afin de trouver les noms composés du troisième type, nous devrons utiliser le code flexionnel du nom et de l'adjectif du DELAS. Les noms et adjectifs qui ont un féminin ont un code morphologique supérieur à 30.

6. LES VARIATIONS GRAPHIQUES

L'orthographe des mots composés est loin d'être définie ([Mathieu-Colas 1988]). Les noms composés acceptent des variations orthographiques:

appuie-main (VN) *appui-main* (NN)

l'usage du trait d'union est souvent fluctuant:

moyen âge (AN) *moyen-âge* (AN)

l'usage des lettres majuscules aussi:

Moyen Age (AN)

Lorsqu'un nom composé accepte plusieurs graphies, nous le dédoublons. Ce dédoublement a le désavantage de traiter les variantes de la même entrée comme des entrées indépendantes. Rétablir le lien de synonymie entre les variantes nécessitera l'utilisation d'un outil de gestion de ces liens (base de données relationnelles du type DB2/SQL) et pourra donner lieu à des études plus générales:

voyage présidentiel (NA) = *voyage du président* (NDN)

7. LA MAINTENANCE DU D E L A C

Les listes de noms composés sont en cours de construction. Elles représentent néanmoins plus de 80 000 noms composés, ce qui est un nombre important pour le mini-ordinateur VAX-730. A titre d'exemple, un traitement comme l'extraction selon la catégorie grammaticale de mots parmi 20 000 noms composés représente plus de 4 heures de traitement en mono-utilisateur. Or, à chaque ajout de nom composé, toutes les procédures décrites doivent être appliquées. Il n'est donc pas question de réeffectuer ces procédures pour l'ensemble des noms d'une classe à chaque étape du recensement. Nous avons donc établi une procédure de maintenance dans laquelle nous distinguons d'une part les ajouts de nouveaux mots, d'autre part les suppressions de mots incorrects, et les modifications apportées à la liste.

Les lexicographes disposent de la liste générale des noms déjà recensés sur support papier. Lorsqu'ils veulent ajouter un mot à la liste générale, ils écrivent les mots à ajouter sur une disquette PC; lorsqu'ils veulent supprimer ou corriger un mot, ils annotent la liste sur papier. Nous transférons les données de la disquette d'ajout sur le mini-ordinateur, et nous effectuons toutes les procédures de vérification, de classement et d'ajout d'information sur la liste d'ajouts. Ensuite, nous fusionnons la liste d'ajout avec la liste générale. Par ailleurs, nous lisons la liste des suppressions et des modifications en parcourant « manuellement » la liste générale sous éditeur, et nous y entrons les modifications au fur et à mesure.

CONCLUSION

Nous avons présenté ici deux dictionnaires électroniques, le DELAS et le DELAC. Nous avons montré que la nécessité de gérer un nombre important de données parfaitement cohérentes entraîne d'effectuer des procédures nombreuses, ce qui n'est pas nécessaire pour les dictionnaires

classiques. Ces procédures coûteuses en temps et en place mémoire n'ont pas un effet toujours perceptible pour le lexicographe qui fournit la liste brute; néanmoins, elles sont indispensables, et constituent le problème principal de la construction d'un dictionnaire électronique.

Nous avons vu que les problèmes soulevés ne sont pas tous solubles automatiquement: plusieurs procédures ne peuvent être que manuelles, ce qui exclut l'objectif « zéro erreur ». Nous avons tout fait pour que ces procédures s'appliquent à des listes les plus petites possibles.

Le point de vue purement morphologique a comme conséquence de ne décrire qu'une seule fois des entrées syntaxiquement dédoublées (comme les deux verbes *voler*) dans le DELAS. Par ailleurs, les procédures utilisées pour vérifier les listes de noms composés sont fondées sur l'utilisation des informations du DELAS, ce qui présente certains inconvénients (comme par exemple la description du féminin du nom *pied noir*). Ces problèmes seront résolus lorsque les entrées du DELAS et du DELAC seront associées à des propriétés syntaxiques.

Bibliographie

BOONS Jean-Paul, GUILLET Alain, LECLÈRE Christian, 1976. *La structure des phrases simples en français: Les verbes intransitifs.* Droz, Genève.

COURTOIS Blandine, 1987. DELAS: *Dictionnaire Electronique du LADL pour les mots Simples du français.* Rapport technique du LADL, Université Paris 7.

GREVISSE Maurice and GOOSSE André, 1986. *Le bon usage,* douzième édition. Editions Duculot, Paris-Gembloux.

GROSS Gaston, 1986. *Typologie des noms composés.* Rapport A.T.P. Nouvelles recherches sur le langage, Paris XIII, Villetaneuse.

GROSS Gaston, JUNG René and MATHIEU-COLAS Michel, 1987. *Noms composés.* Rapport n° 5 du Programme de Recherches Coordonnées « Informatique Linguistique », Université Paris 7.

GROSS Maurice, 1986. *Les adjectifs composés du français.* Rapport n° 3 du Programme de Recherches Coordonnées « Informatique Linguistique », CNRS, Paris.

GROSS Maurice, 1977, 1982, 1989. *Grammaire transformationnelle du français: 1 Syntaxe du verbe. 2 Syntaxe du nom. 3 Syntaxe de l'adverbe.* Cantilène, Paris.

LAPORTE Eric, 1988a. *Méthodes algorithmiques et lexicales de phonétisation de textes. Applications au français.* Thèse de doctorat en informatique, LADL, Université Paris 7.

LEEMAN Danièle, 1988. *Echantillons des adjonctions au DELAS d'adjectifs en -able.* Rapport du Programme de Recherches Coordonnées « Informatique Linguistique », LADL, Université Paris 7.

MATHIEU-COLAS Michel, 1987. *Variations graphiques de mots composés.* Rapport n° 4 du Programme de Recherches Coordonnées « Informatique Linguistique », CNRS, Paris.

Annexe 1

Répartition des classes flexionnelles des noms/adjectifs

Groupe	Noms/adjectifs	Numéros des classes	Relations entre les formes
I	masculins	0 à 19	fs = fp = 0
II	féminins	20 à 29	ms = mp = 0
III	masc.et fém. plur.en 's'	30 à 59	mp = ms + 's' fp = fs + 's'
IV	masc.et fém. masc.invar.	60 à 69	fs = fp + 's' mp = ms
V	flexions résiduelles	70 à 80	

ms = masculin singulier, mp = masculin pluriel,
fs = féminin singulier, fp = féminin pluriel.

Annexe 2

Groupe I : Classes flexionnelles des noms et adjectifs exclusivement masculins.		
Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
0 1	*,-,*,- .,-,s,-	oeil,.N0S yeux,.N0P mot,.N1 violat,.A1
2 3	.,-,.,- .,-,x,-	bois,.N2 arsénieux,.A2 bateau,.N3 chou,.N3
4 5	l,-,ux,- il,-,ux,-	journal,.N4 ciel,.N4 corail,.N5
6 7	.,-,s,s us,-,i,-	orgue,.N6 naevus,.N7
8 9	um,-,a,- homme,-,shommes,-	quantum,.N8 bonhomme,.N9
10 11	man,-,men,- y,-,ies,-	recordman,.N10 lobby,.N1
12 13	.,-,es,- o,-,i,-	box,.N1 tempo,.N1
14 15	.,-,im,- .,-,m,-	kibboutz,.N14 sefardi,.N15
16	e,-,i-	nuraghe,.N16

* classe avec renvoi singulier-pluriel
- formes qui n'existent pas
. terminaison non spécifique
S = singulier seulement, P = pluriel seulement

Annexe 3

Groupe II : Classes flexionnelles des noms et adjectifs exclusivement féminins.		
Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
20	-,*,-,*	madame,.N20S mesdames,.N0P
21	-,-,-,s	maison,.N21 épinière,.A21
22	-,-,-,-	croix,.N22 caudines,.A22P
23	-,-,-,x	eau,.N23
24	-,y,-,ies	lady,.N24
25	-,man,-,men	recordwoman,.N25
* classe avec renvoi singulier-pluriel - formes qui n'existent pas . terminaison non spécifique S = singulier seulement, P = pluriel seulement		

Annexe 4

Groupe III : Classes flexionnelles des noms et adjectifs des deux genres, avec pluriels en 's'.		
Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
30	*,*,*s,*s	empereur,.N30M impératrice,.N30P
31	.,.,s,s	apte,.A31 artiste,.N31
32	.,e,es	têtu,.A32 veinard,.N32
33	.,se,s,ses	andalou,.N32.A32
34	.,te,us,tes	favori,.N34.A34
35	eur,euse,eurs,euses	rageur,.A35 voleur,.N35
36	eur,rice,eurs,rices	ambassadeur,.N36
37	eur,eresse,eurs,eresses	vengeur,.N37.A37
38	f,ve,fs,ves	oisif,.A38 veuf,.N38
39	.,sse,s,sses	traître,.N39
40	l,lle,ls,lles	cruel,.N40.A40
41	n,nne,ns,nnes	bon,.A41 citoyen,.N41
42	er,ère,ers,ères	léger,.A42 ouvrier,.N42
43	et,ète,ets,ètes	discret,.A43 préfet,.N43
44	ef,ève,efs,èves	bref,.A44
45	ec,èche ecs,èches	sec,.A45
46	c,que,cs,ques	laïc,.N46.A46
47	c,che,cs,ches	blanc,.A47
48	c,chesse,cs,chesses	duc,.N48
49	g,gue,gs,gues	oblong,.A49
50	.,sque,s,sques	maure,.N50.A50
51	gu,guë,gus,guës	ambigu,.A51
52	n,gne,ns,gnes	malin,.N52.A52
53	ou,olle,ous,olles	foufou,.N53
54	er,euse,ers,euses	streaker,.N54
55	.,ine,s,ines	feuillant,.N55.A55
56	.,esse,s,esses	clown,.N56
57	o,a,os,as	aficionado,.N57
58	ète,étesse,ètes,étesses	poète, N58
59	ec,èque,ecs,èques	grec,.N59.A59

* classe avec renvoi masculin M ---> feminin F
. terminaison non spécifique

Annexe 5

Groupe IV : Classes flexionnelles des noms et adjectifs des deux genres, avec masculin invariable.

Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
60	*,*,*,*s	tiers,.A60M tierce,.A60F
61	.,e,.,es	niais,.N61 obtus,.A61
62	.,se,.,ses	métis,.N62.A62 épais,.A62
63	x,se,x,ses	jaloux,.N63.A63 fameux,.A63
64	x,sse,ux,sses	faux,.A64
65	x,ce,x,ces	doux,.A65
66	ux/il,ille,ux,illes	vieux,.N66.A66
67	ès,esse,ès,esses	profès,.N67 exprès,.A67

* classe avec renvoi masculin M --> féminin F
. terminaison non spécifique

Annexe 6

Groupe V : Classes flexionnelles des noms et adjectifs des deux genres, groupe résiduel.

Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
70	*,*,*x,*s	hébreu,.A70M hébraïque,.A70F
71	l,le,x,les	aïeul,.N71
72	au,lle,aux,lles	agneau,.N72
73	au/l,lle,aux,lles	beau,.A73
74	el/elle, aux, elles	séquentiel, .A74
75	t,te,s,tes	tout,.A75
76	al,ale,aux,ales	provincial,.N76.A76 légal,.A76
77	.,.,x,x	gâteau,.A77 (adj.)
78	um,a,a,a	maximum,.A78
79	us,a,a,a	valgus,.A79

* classe avec renvoi masculin M ---> féminin F
. terminaison non spécifique

Groupe VI : Classe flexionnelle des noms et adjectifs invariables aux deux genres.

Numéro de classe	Flexion ms,fs,mp,fp	Exemples de mots du dictionnaire DELAS
80	.....	sable,.A80 (couleur) quatorze,.A80P

. terminaison non spécifique

Annexe «2»

Verbes-modèles pour les classes de conjugaisons

Classe Numéro	Verbe modèle
1	avoir
..	être

3	aimer
4	placer
5	manger
6	peser
7	céder
8	jeter
9	appeler
10	assiéger
11	dépecer
12	rapiécer

13	broyer
14	payer
15	envoyer
16	aller

18	finir
19	haïr

20	acquérir
21	assaillir
22	bouillir
23	couvrir
24	cueillir
25	fleurir
26	endormir
27	fuir
28	sentir
29	servir
30	vêtir
31	courir
32	mourir
33	tenir
34	ouir
35	faillir
36	gésir
37	chauvir

Classe Numéro	Verbe modèle
40	asseoir
41	devoir
42	mouvoir
43	pourvoir
44	pouvoir
45	prévoir
46	recevoir
47	savoir
48	surseoir
49	valoir
50	voir
51	vouloir
53	déchoir
55	falloir
56	pleuvoir
57	seoir
58	prévaloir
59	promouvoir

60	vaincre
61	absoudre
62	coudre
63	moudre
64	peindre
65	résoudre
66	prendre
67	rendre

68	battre
69	connaître
70	croître
71	mettre
72	naître
73	accroître

74	suivre
75	vivre
76	foutre
77	rompre
78	fiche

Classe Numéro	Verbe modèle
80	faire
81	traire
82	plaire
83	boire
84	croire
85	enclore
86	conclure
87	inclure
88	taire

89	suffire
90	confire
91	cuire
92	écrire
93	dire
94	lire
95	rire
96	maudire
97	nuire
98	circoncire

Codes-lettres additionnels

I = verbe impersonnel,

U = unique forme pour le participe passé,

D = défectif,

R = verbe n'existant pas sans pronom réfléchi

Exemple DELAS

f,.N2
fa,.N2
fable,.N21
fabliau,.N3
fablier,.N1
fabricant,.N32
fabricateur,.N36
fabrication,.N21
fabricien,.N41
fabriquer,.V3
fabrique,.N21
fabriquer,.V3
fabulateur,.N36.A36
fabulation,.N21
fabuler,.V3
fabuleusement,.ADVE
fabuleux,.A63
fabuliste,.N31
fac,.N21
façade,.N21
face,.N21
facer,.V4
facétie,.N21
facétieusement,.ADVE
facétieux,.A63
facettage,.N1
facette,.N21
facetter,.V3
fâché,.A32
fâcher,.V3
fâcherie,.N21
fâcheusement,.ADVE
fâcheux,.N63.A63
facho,.N31.A31
facial,.A76
faciès,.N2
faci.e,.A31
facilement,.ADVE
facilitation,.N21
facilité,.N21
faciliter,.V3
façon,.N21
faconde,.N21
façonnage,.N1
façonné,.A32
façonnement,.N1
façonner,.V3
façonneur,.N35
façonnier,.N42.A42
factage,.N1
factal,.A32
facteur,.N1.N36
factice,.N1.A31
facticement,.ADVE
facticité,.N21
factieusement,.ADVE
factieux,.N63.A63
faction,.N21
factionnaire,.N1
factionner,.V3
factitif,.N1.A38
factitivement,.ADVE
factorage,.N1
factorerie,.N21
factoriel, A40
factorielle,.N21
factoring,.N1
factorisation,.N21
factoriser,.V3
factotum,.N1
factrice,.N21
factuel,.A40
factuellement,.ADVE
factum,.N1
facturation,.N21
facture,.N21
facturer,.V3
facturier,.N1.N42
facule,.N21
facultaire,.A31
facultatif,.A38
facultativement,.ADVE
faculté,.N21
fade,.N31.A80
fadaise,.N21
fadasse,.A31
fadassement,.ADVE
fadasserie,.N21
fade,.A31
fadé,.A32
fadement,.ADVE
fader,.V3
fadeur,.N21
fading,.N1
fado,.N1
faena,.N21
fafiner,.V3
fafiot,.N1
fagale,.N21
fagne,.N21
fagot,.N1
fagotage,.N1
fagoter,.V3
fagotier,.N42
fagotin,.N1
fagoue,.N21
fahrenheit,.N2.A80
faiblage,.N1
faiblard,.N32.A32
faiblardement,.ADVE
faible,.N1.N31.A31
faiblement, ADVE
faiblesse,.N21
faiblir,.V18u
faiencage,.N1
faience,.N21
faiencé,.A32
faiencerie,.N21
faiencier,.N42
faignant,.N32.A32
faignanter,.V3
faille,.N21
failé,.A32
failler,.V31
failli,.N32.A32
faillibilité,.N21
faillible,.A31
faillir,.V35u
faillite,.N21
faim,.N21
faine,.N21
faîne,.N21
fainéant,.N32.A32
fainéanter,.V3u
fainéantise,.N21
fair,.A80
faire,.N25.V80
faisabilité,.N21
faisa[illegible],.A31
faisa[illegible],.N1.N32
faisandage,.N1
faisandé,.A32
faisandeau,.N3
faisander,.V3
faisanderie,.N21
faisane,.N21.A21
faisanneau,.N3
faisceau,.N3
faiseur,.N35
faisselle,.N21
fait,.N1.A32
faîtage,.N1
faîte,.N1
faîteau,.N3
faîtière,.N21.A21
faitout,.N1
faix,.N2
fakir,.N31
fakirisme,.N1
falaise, N21
falarique,.N21
falbala,.N1
falbalasser,.V3
falciforme,.A31
falconidé,.N1
faldistoire,.N1
falerne,.N1
fallacieusement,.ADVE
fallacieux,.A63
falloir,.V55u
falot,.N1.A32
falourde,.N21
falquer,.V3
falsifiabilité,.N21
falsifiable,.A31
falsificateur,.N36
falsification,.N21
falsifier,.V3
faluche,.N21
fatum,.N1

Exemple DELAF

f,.N2:Nms:Nmp
fa,.N2:Nms:Nmp
fable,.N21:Nfs
fables,·1.N21:Nfp
fabliau,.N3:Nms
fabliaux,·1.N3:Nmp
fablier,.N1:Nms
fabliers,·1.N1:Nmp
fabricant,.N32:Nms
fabricante,·1.N32:Nfs
fabricantes,·2.N32:Nfp
fabricants,·1.N32:Nmp
fabricateur,.N36:Nms
fabricateurs,·1.N36:Nmp
fabrication,.N21:Nfs
fabrications,·1.N21:Nfp
fabricatrice,·4eur.N36:Nfs
fabricatrices,·5eur.N36:Nfp
fabricien,.N41:Nms
fabricienne,·2.N41:Nfs
fabriciennes,·3.N41:Nfp
fabriciens,·1.N41:Nmp
fabricota,·1er.V3:J3s
fabricotai,·2er.V3:J1s
fabricotaient,·5er.V3:I3p
fabricotais,·3er.V3:I1s:I2s
fabricotait,·3er.V3:I3s
fabricotâmes,·4er.V3:J1p
fabricotant,·3er.V3:G00
fabricotas,·2er.V3:J2s
fabricotasse,·4er.V3:T1s
fabricotassent,·6er.V3:T3p
fabricotasses,·5er.V3:T2s
fabricotassiez,·6er.V3:T2p
fabricotassions,·7er.V3:T1p
fabricotât,·2er.V3:T3s
fabricotâtes,·4er.V3:J2p
fabricote,·0r.V3:P1s3s:S1s3s:Y2s
fabricoté,·1er.V3:Kms
fabricotée,·2er.V3:Kfs
fabricotées,·3er.V3:Kfp
fabricotent,·2r.V3:P3p:S3p
fabricoter,.V3:W00
fabricotera,·1.V3:F3s
fabricoterai,·2.V3:F1s
fabricoteraient,·5.V3:C3p
fabricoterais,·3.V3:C1s:C2s
fabricoterait,·3.V3:C3s
fabricoteras,·2.V3:F2s
fabricoterent,·5er.V3:J3p
fabricoterez,·2.V3:F2p
fabricoteriez,·3.V3:C2p
fabricoterions,·4.V3:C1p
fabricoterons,·3.V3:F1p
fabricoteront,·3.V3:F3p
fabricotes,·1r.V3:P2s:S2s
fabricotes,·2er.V3:Kmp
fabricotez,·1r.V3:P2p:Y2p
fabricotiez,·3er.V3:I2p:S2p
fabricotions,·4er.V3:I1p:S1p
fabricotons,·3er.V3:P1p:Y1p
fabrique,·1er.V3:J3s
fabriquei,·2er.V3:J1s
fabriquaient,·5er.V3:I3p
fabriquais,·3er,V3:I1s:I2s
fabriquait,·3er.V3:I3s
fabriquâmes,·4er.V3:J1p
fabriquant,·3er.V3:G00
fabriquas,·2er.V3:J2s
fabriquasse,·4er.V3:T1s
fabriquassent,·6er.V3:T3p
fabriquasses,·5er.V3:T2s
fabriquassiez,·6er.V3:T2p
fabriquassions,·7er,V3:T1p
fabriquât,·2er,V3:T3s
fabriquâtes,·4er,V3:J2p
fabrique,.N21:Nfs
fabrique,·0r.V3:P1s3s:S1s3s:Y2s
fabriqué,·1er.V3:Kms
fabriquée,·2er.V3:Kfs
fabriquées,·3er.V3:Kfp
fabriquent,·2r.V3:P3p:S3p
fabriquer,.V3:W00
fabriquera,·1.V3:F3s
fabriquerai,·2.V3:F1s
fabriqueraient,·5.V3:C3p
fabriquerais,·3.V3:C1s:C2s
fabriquerait,·3.V3:C3s
fabriqueras,·2.V3:F2s
fabriquèrent,·5er.V3:J3p
fabriquerez,·2.V3:F2p
fabriqueriez,·3.V3:C2p
fabriquerions,·4.V3:C1p
fabriquerons,·3.V3:F1p
fabriqueront,·3.V3:F3p
fabriques,·1.N21:Nfp
fabriques,·1r.V3:P2s:S2s
fabriqués,·2er.V3:Kmp
fabriquez,·1r.V3:P2p:Y2p
fabriquiez,·3er.V3:I2p:S2p
fabriquions,·4er.V3:I1p:S1p
fabriquons,·3er.V3:P1p:Y1p
fabula,·1er.V3:J3s
fabuler,·2er.V3:J1s
fabulaient,·5er.V3:I3p
fabulais,·3er.V3:I1s:I2s
fabulait,·3er.V3:I3s
fabulâmes,·4er.V3:J1p
fabulant,·3er.V3:G00
fabulas,·2er.V3:J2s
fabulasse,·4er.V3:T1s
fabulassent,·6er.V3:T3p
fabulasses,·5er.V3:T2s
fabulassiez,·6er.V3:T2p
fabulassions,·7er.V3:T1p
fabulat,·2er.V3:T3s
fabulâtes,·4er.V3:J2p
fabulateur,.N36:Kms
fabulateur,.A36:Ams
fabulateurs,·1.N36:Kmp
fabulateurs,·1.A36:Amp
fabulation,.N21:Nfs
fabulations,·1.N21:Nfp
fabulatrice,·4eur.N36:Nfs
fabulatrice,·4eur,A36:Afs
fabulatrices,·5eur.N36:Nfp
fabulatrices,·5eur,A36:Afp
fabule,·[illegible]3:P1s3s:S1s3s:Y2s
fabulé,·1er.V3:Kms
fabulée,·2er.V3:Kfs
fabulées,·3er.V3:Kfp
fabulent,·2r.V3:P3p:S3p
fabuler,.V3:W00
fabulera,·1.V3:F3s
fabulerai,·2.V3:F1s
fabuleraient,·5.V3:C3p
fabulerais,·3.V3:C1s:C2s
fabulerait,·3.V3:C3s
fabuleras,·2.V3:F2s
fabulèrent,·5er,V3:J3p
fabulerez,·2.V3:F2p
fabuleriez,·3,V3:C2p
fabulerions,·4.V3:C1p
fabulerons,·3.V3:F1p
fabuleront,·3.V3:F3p
fabules,·1r.V3:P2s:S2s
fabulés,·2er.V3:Kmp
fabuleuse,·2x.A63:Afs
fabuleusement,.Adv
fabuleuses,·3x.A63:Afp
fabuleux,.A63:Ams:Amp
fabulez,·1r.V3:P2p:Y2p
fabuliez,·3er.V3:I2p:S2p
fabulions,·4er.V3:I1p:S1p
fabuliste,.N31:Nms:Nfs
fabulistes,·1.N31:Nmp:Nfp
fabulons,·3er.V3:P1p:Y1p
fac,.N21:Nfs
face,·2cer.V4:J3s
facede,.N21:Nfs
facades,·1.N21:Nfp
facai,·3cer.V4:J1s
facaient,·6cer.V4:I3p
facais,·4cer.V4:I1s:I2s
facait,·4cer.V4:I3s
facames,·5cer.V4:J1p
facant,·4cer.V4:G00
facas,·3cer.V4:J2s
facasse,·5cer.V4:T1s
facassent,·7cer.V4:T3p
facasses,·6cer.V4:T2s
facassiez,·7cer.V4:T2p
facassions,·8cer.V4:T1p
facat,·3cer.V4:T3s
facates,·5cer.V4:J2p
face,.N21.Nfs
face,·0r.V4:P1s3s:S1s3s:Y2s
facé,·1er.V4:Kms
facée,·2er.V4.Kfs
facées,·3er.V4:Kfp

Schéma 5

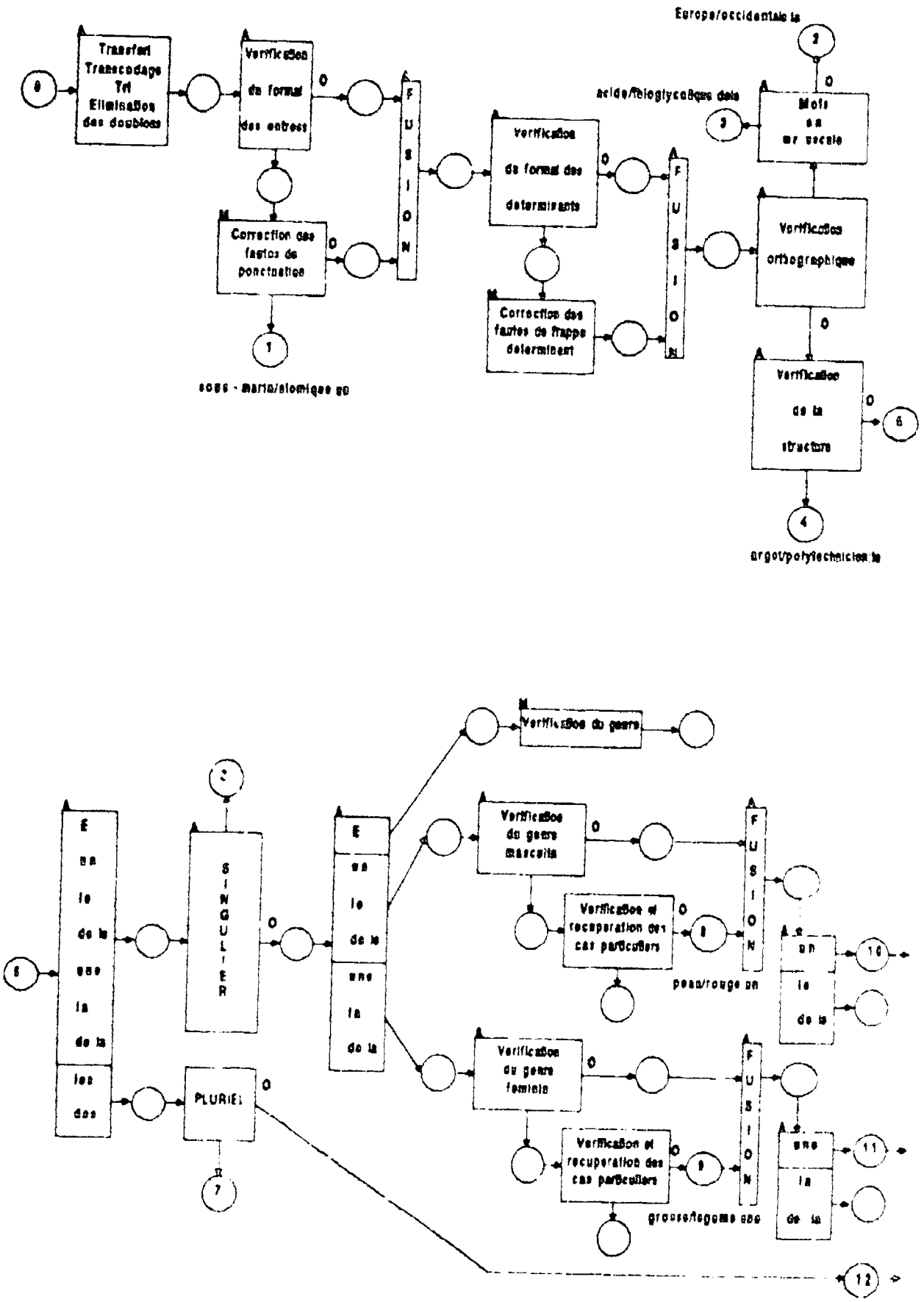
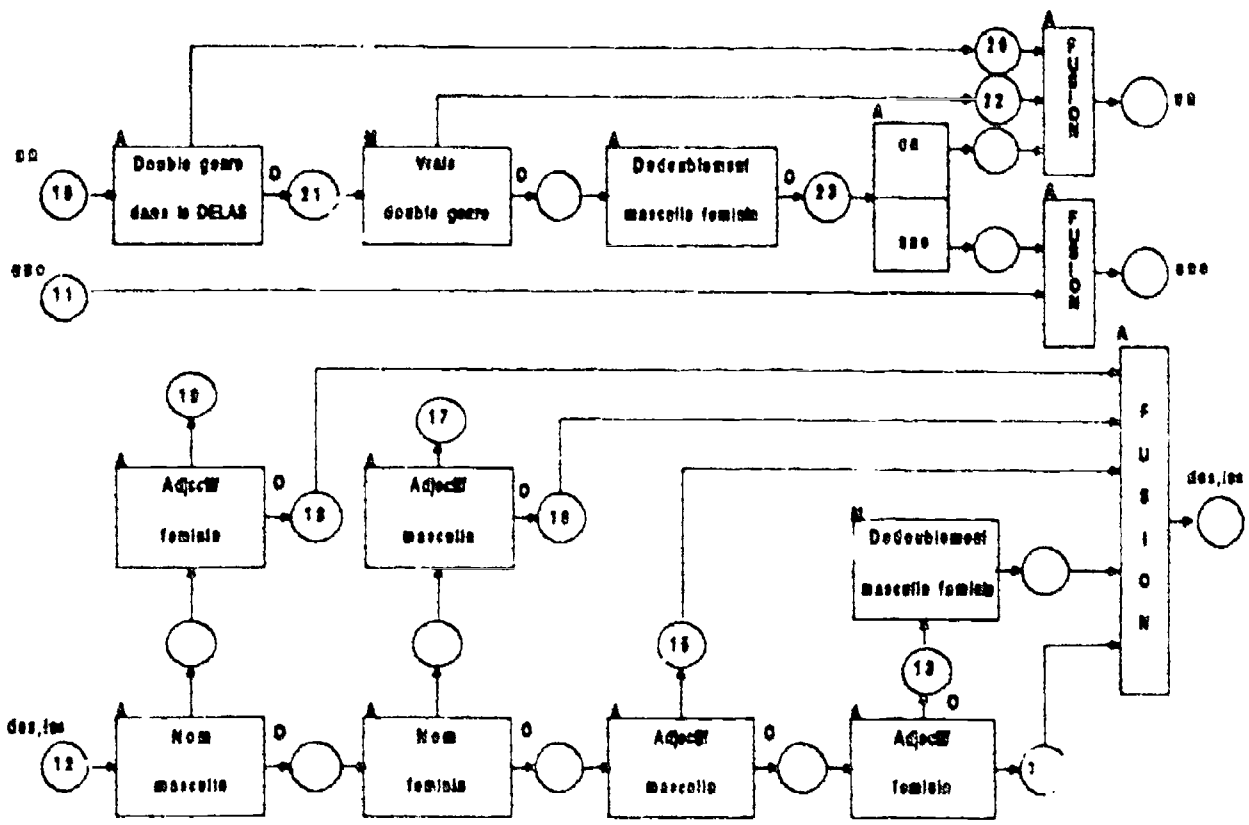

Transfert
Transcodage
Tri
Eliminations
des doubles
Verification
de format
des entrees
Correction des
fautes de
ponctuation
FUSION
Verification
de format des
determinants
Correction des
fautes de frappe
determinant
FUSION
Europe/occidentale la
acide/thioglycolique dala
Mots
en
majuscule
Verification
orthographique
Verification
de la
structure
argot/polytechnicien la
scos - marin/biologie ap
Verification du genre
Verification
du genre
masculin
Verification et
recuperation des
cas particuliers
peau/rouge an
Verification
du genre
feminin
Verification et
recuperation des
cas particuliers
grosse/legume abs
SINGULIER
PLURIEL

Schéma 6



FICHIER 4: Mauvaise structure

abbaye/bénédictine:une:fs:NA
abeille/charpentière:une:fs:NA
abeille/maçonne:une:fs:NA
acier/triplex:de&le:ms:NA
acte/illicite:un:ms:NA
adiabatique/saturée:la:fs:NA
adiabatique/sèche:la:fs:NA
affecté/spécial:un:ms:NA
agate/chrysoprase:une:fs:NA
agénésie/folliculaire:la:fs:NA
agent/aérateur:un:ms:NA
aigue-/marine:une:fs:NA
aigue/marine:une:fs:NA
air/fossile:de&le:ms:NA
alphabet/gotique:le:ms:NA
alphabet/morse:le:ms:NA
amanite/oronge:une:fs:NA
amicale/bouliste:une:fs:NA
amplitude/occase:une:fs:NA
âne/cornard:un:ms:NA
annélide/hirudinée:une:fs:NA
annélide/oligochète:un:ms:NA
annelide/oligochète:une:fs:NA
annélide/polychète:une:fs:NA
antenne/paraboloïde:une:fs:NA
appareil/décompresseur:un:ms:NA
araignée/chasseuse:une:fs:NA
arase/régulière:la:fs:NA
arbre/arsin:un:ms:NA
arbre/fossile:un:ms:NA
argile/figuline:de&le:fs:NA
argot/polytechnicien:le:ms:NA
artère/carotide:une:fs:NA
association/écologiste:une:fs:NA
Atlantique/Nord:le:ms:NA
atlantique/occidental:le:ms:NA
Atlantique/Sud:le:ms:NA
atome/donneur:un:ms:NA
atome/marqueur:un:ms:NA
auto/sacramental:un:ms:NA
avenir/commercial:un:ms:NA
avenir/économique:un:ms:NA
avenir/immédiat:le:ms:NA
avenir/politique:un:ms:NA
avenir/professionnel:un:ms:NA
avion/monoplan:un:ms:NA
avion/traiteur:un:ms:NA
beau/parleur:un:ms:NA
beaujolais/primeur:de&le:ms:NA
bec/pelleteur:un:ms:NA
blé/mâteil:de&le:ms:NA
blindé/léger:un:ms:NA
bombe/aérosol:une:fs:NA
bordeaux/primeur:de&le:ms:NA
botanique/fossile:la:fs:NA

DES ÉLÉMENTS D'UN ATELIER DE GÉNIE LINGUISTIQUE

H. Habrias, J.-F. Hue, J-H. Jayez, P. Legrand, Y. Simon, D. Vellard
LIANA Nantes France

0 INTRODUCTION:

L'analyse et la compréhension de textes en langue française, en vue d'applications diverses (traduction automatique, construction d'interfaces utilisateur conviviales, système de résolution de problèmes posés en langue naturelle) sont une tâche complexe qui nécessite un environnement informatique spécialisé; un tel environnement est appelé **atelier de génie linguistique**.

Un tel atelier se compose des élément suivants:

- un générateur d'analyseurs syntaxico-sémantiques.
- un système de gestion de bases de données linguistiques.
- un système expert en résolution de problèmes posés en langue naturelle.
- un illustrateur graphique.

Nous présentons dans cet article deux réalisations du LIANA: un générateur d'analyseurs syntaxico-sémantiques pour des grammaires lexicales réalisé par Jean-François Hüe, et une interface, construite par Patrick Legrand, pour constituer les dictionnaires électroniques qui utilisent les représentations de Maurice Gross [11].

1 UN GÉNÉRATEUR D'ANALYSEURS SYNTAXICO-SÉMANTIQUES

1.1 Introduction

Dans la nature, le sens s'exprime à partir des formes et de leur langage. La syntaxe de ces formes peut être fort complexe et très irrégulière. Heureusement, il est un domaine, celui de nos langues indo-européennes, où cette irrégularité et cette complexité sont fortements atténuées. Trois principes sont utilisés à cet effet: l'usage d'un nombre restreint de symboles graphiques, la linéarité de la disposition de ces symboles pour former un texte, l'existence de règles de grammaire en nombre peu élevé et relativement stable au cours du temps.

Cependant l'immense variété de ce qui doit être décrit par ces langages: les objets, les actions, les synchronisations a pour conséquence de rendre difficile l'analyse syntaxico-sémantique des textes en langue naturelle.

Une science particulière, la **linguistique**, s'est développée dans le but d'étudier les langues humaines sous tous leurs aspects. Une sous-branche très riche de cette science générale est la **théorie des langages** [1], [2], [3], [4], [5] dont une des ambitions est la réalisation pratique des analyseurs syntaxico-sémantiques que demande la construction des compilateurs des langues artificielles de l'informatique. Cette même théorie des langages, combinée avec d'autres, est à la base des réalisations pratiques des systèmes automatiques de compréhension des langues naturelles [6].

Un grand nombre d'analyseurs syntaxico-sémantiques pour textes en langue naturelle suivent avec rigueur dans leur analyse le principe de la présentation linéaire de nos langues parce que leur conception découle directement de la définition des **grammaires lexicales** dans laquelle certains terminaux joueront un rôle particulier; leur ensemble sera nommé lexique.

1.2 Résumé des idées qui inspirent ce travail.

*** A un langage dont la grammaire peut être complexe on associe une grammaire lexicale dont les grammaires associées sont plus simples (Algébriques et déterministes [7]).

*** L'analyse ne se fera pas linéairement sur le texte d'entrée de l'analyseur mais à partir des symboles particuliers de lexique; une analyse linéaire normale pourra se produire autour d'eux. Une généralisation possible, qui sera étudiée ultérieurement, est de ne plus s'intéresser à de simples terminaux mais à des groupes de terminaux; on parlera alors de **grammaires polylexicales**.

***Ces grammaires lexicales peuvent être définies à partir de grammaires elles-mêmes lexicales ce qui permet une analyse syntaxico-sémantique par couches successives, en profondeur.

1.3 Définition d'une grammaire lexicale.

On appelle grammaire lexicale un ensemble (L, G, F) où:

L est un ensemble de symboles appelé lexique.

G est une grammaire dont l'ensemble des terminaux est égal à L.

F est une fonction de L vers E*E où E est un ensemble de grammaires.

F: L-->E*E
1-->(Gg(1), Gd(1))

avec Gg(1) grammaire gauche de 1 et Gd(1) grammaire droite de 1.

Exemple.

Soit le langage U= {x/x=a b c n N*} *pour lequel il n'existe pas de grammaire algébrique* [7].

Définissons une grammaire lexicale GL1=(L1,G1,F1) qui l'engendre exactement:

L1={b}

G1=($V1_N$,$V1_T$,A1,P1) avec
$V1_N$={A1}
$V1_T$=L1
P1={A1-->bA1,A1-->nil}

F1:L1-->E1*E1
b-->(G1g(b),G1$_d$(b)) avec

G1$_g$(b)={G1$_g$N(b),G1$_g$T(b),A1$_g$(b),P1$_g$(b)}

où

G1$_g$N(b)={A1$_g$(b)}
G1$_g$T(b)={a}
P1$_g$(b)={A1$_g$(b)-->a}

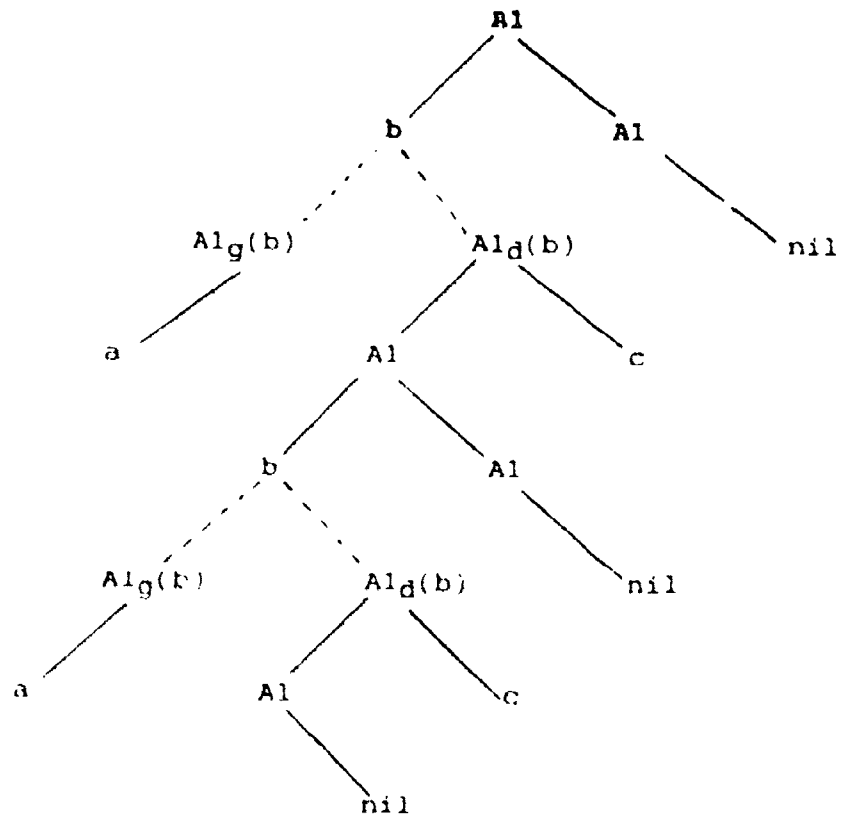
G1$_d$(b)={G1$_d$N(b),G1$_d$T(b),A1$_d$(b),P1$_d$(b)}

où

G1$_d$N(b)={A1$_d$(b),GL1}
G1$_d$T(b)={c}
P1$_d$(b)={A1$_d$(b)-->GL1c}

Comme on peut le remarquer les grammaires ci-dessus sont toutes algébriques et déterministes et il y a un appel récursif de la grammaire lexicale dans la grammaire gauche de 'b'. *Ainsi un langage pour lequel il n'existe pas de grammaire algébrique est engendré par une grammaire lexicale qui fait appel à deux grammaires algébriques déterministes; ce résultat est en soit remarquable.*

ARBRE SYNTAXIQUE ASSOCIÉ À 'aabbcc'



le trait pointille est utilise pour representer les derivations de la grammaire gauche et de la grammaire droite d'un symbole du lexique.

ANALYSE SYNTAXIQUE DE 'aabbcc'

Recherche du premier 'b' dans une lecture gauche droite, un symbole souligné n'est plus considéré dans la suite de l'analyse syntaxique.

aabbcc
dérivation à gauche

aabbcc
dérivation à droite

recherche du second 'b'

aabbcc
dérivation à gauche

aabbcc
dérivation à droite

aabbcc (il n'y a plus de b)
dérivation à droite

aabbcc

Le texte est reconnu comme syntaxiquement correct vis a vis de GL1.

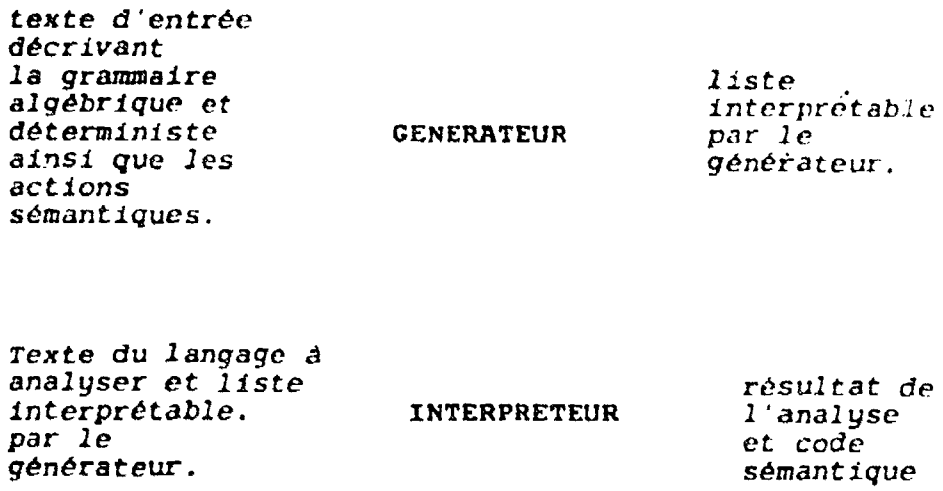
1.4 Construction des analyseurs associés aux grammaires lexicales

Nous utilisons pour réaliser un analyseur d'une grammaire algébrique déterministe $G=(V_N,V_T,A,P)$ une interprétation par les fonctions Booléennes [8]; l'analyseur associé à la grammaire sera nommé analyseur **Booléen** [9].

Il va s'agir de construire, en s'appuyant sur les symboles du texte à analyser et sur les regles de production de la grammaire une suite de fonctions Booléennes $(P_k)_{k=1}$ $P_k:V_T^{k}\rightarrow\{0,1\}$ dont on évalue le produit au fur et à mesure de l'analyse. Si à la fin du texte le produit égale à 1 alors le texte est syntaxiquement correct sinon il est syntaxiquement incorrect. Pour un texte syntaxiquement incorrect l'analyse s'arrête des la rencontre d'un symbole pour lequel on ne peut trouver, en appliquant les règles de production de la grammaire, une évaluation vraie.

Suivant ce principe nous avons construit un **générateur d'analyseurs Booléens** pour grammaires algébriques déterministes qui par extension est capable de générer des analyseurs pour grammaires lexicales. En effet si pour toute grammaire algébrique déterministe nous pouvons faire correspondre un analyseur Booléen, pour une grammaire lexicale dont le lexique comporte n symboles, nous avons au plus 2n+1 analyseur Booléens (d'après la définition donnée au paragraphe 3).

ARCHITECTURE DU GÉNÉRATEUR D'ANALYSEURS BOOLÉENS:



La liste interprétable est rangée dans le dictionnaire des analyseurs pour grammaire algébrique et déterministe.

ARCHITECTURE DU GÉNÉRATEUR D'ANALYSEURS POUR GRAMMAIRES LEXICALES.

Un analyseur pour grammaire lexicales est composé de plusieurs analyseurs Booléens secondaires (deux par symboles de lexique) et par un analyseur Booléen principal qui va être chargé de l'analyse du langage lié au lexique, et des appels successifs des analyseurs Booléens secondaires.

Le générateur d'analyseurs pour grammaires lexicales est construit à partir d'un générateur d'analyseur Booléens:

L'interpréteur pour analyseur de grammaires lexicales est construit à partir d'un interpréteur d'analyseurs Booléens:

1.5 Vers un atelier de génie linguistique

Le génie linguistique recouvre la mise en oeuvre des méthodes issues de la théorie des langages dans le but de réaliser des applications industrielles en :

- Compréhension des langues naturelles.
- Traduction des langues naturelles.
- Résolution de problèmes posés en langue naturelle.

Un atelier de génie linguistique doit contenir les outils logiciels nécessaires à la réalisation de ces produits. Ces outils sont:

- Générateur d'analyseurs syntaxico-sémantiques.
- Système de gestion de base de données linguistiques.
- Générateur de systèmes experts en linguistique.
- Générateur de représentations graphiques.

2. UN OUTIL D'ÉDITION DE LEXIQUE

Tout d'abord on donnera brièvement un ensemble d'arguments qui mettent en lumière le pourquoi d'un tel logiciel. Ces arguments ne sont ni originaux, ni ici démontrés ou illustrés, mais ils situent bien une école de pensée.

Il faut différencier les dictionnaires (traductionnels) enregistrés sur support magnétique dans un but d'édition, et les dictionnaires dits électroniques, conçus, organisés à des fins d'utilisation par des programmes réalisant des traitements automatiques de la langue.

On peut songer à utiliser les premiers (dictionnaires traditionnels) pour constituer les seconds. Cependant, cela présente quelques difficultés:

1) techniques. Il n'est pas évident de décoder les supports magnétiques des dictionnaires traditionnels.

2) L'information stockée dans les dictionnaires traditionnels n'est pas organisée de façon suffisament rigoureuse (non par rapport à un absolu, mais par rapport à ce que l'on veut en faire).

3) Mais surtout l'information manque. Elle est incomplète, parcellaire.

Devant cet état de fait, on est amené à concevoir et à constituer des dictionnaires, pleinement utiles aux traitements automatiques, différents dans leur forme et par la qualité et la cohérence des informations qu'ils renferment, et dont la référence est le lexique-grammaire du L.A.D.L.

Le logiciel présenté ici, se veut une aide à la constitution de tels dictionnaires. Il permet d'étudier, par des méthodes similaires à celle du lexique-grammaire, toutes propriétés linguistiques souhaitées sur n'importe quel ensemble d'unités linguistiques.

Le logiciel

Fondamentalement, il sert à faire de l'édition (entrer en machine des données) et aussi un peu de sélection.

Aspect général des tables

Ce sont des tableaux à doubles entrées, en lignes et colonnes (cf. figure 4). Au croisement d'une ligne et d'une colonne est stockée de l'information (codée avec peu de valeurs en général). Chaque ligne ou entrée horizontale peut correspondre à un item lexical; chaque colonne, à une propriété que possèdent ou non les items lexicaux.

On est amené à éditer trois ensembles de données. Deux sont des listes, les entrées horizontales et verticales, le troisième est un tableau bidimensionnel, la matrice elle-même. On dispose de deux éditeurs: un pour les listes, l'autre pour la matrice.

***ÉDITION DES LISTES** *(cf. figures 3 et 3 bis)*

Type des listes pouvant être éditées:

Simple ou double

Simple: une liste de verbes, de noms...

.
.
lister
localiser
maintenir
maîtriser
.
.

Double: chaque élément de la liste est un enregistrement de deux champs.

Le premier champ,le champ principal et celui qui est manipulé en priorité lors de la matrice.

Cela peut être par exemple un code dénotant une structure syntaxique NO V N1.

Le second champ étant une information complémentaire. utilisée à d'autres fins. Dans notre exemple cela peut être l'explication du code (phrase élémentaire: *Sujet Verbe Complément*)

Ordonnée ou séquentielle

Soit un des champs est ordonné suivant l'ordre lexicographique, soit il n'y a aucun ordre à priori (séquentiel).

Indexation

Si l'un des champs ou la liste n'est pas ordonné, on peut cependant construire un index qui permet virtuellement de l'ordonner.

Duplication autorisée ou non

Que la liste soit ordonnée ou indexée, on peut indiquer si une entrée quelconque peut on non être dupliquée.

Voilà l'ensemble des impératifs qui ont régi la construction de l'éditeur de listes. Au début de la construction d'une nouvelle liste se déroule une phase dite de configuration qui permet de déterminer l'organisation future de la liste, l'éditeur tient compte ensuite au cours des manipulations de la liste de cette organisation.

Quant à l'édition proprement dite, on y retrouve les fonctions classiques de l'édition: création, suppression, modification, déplacement d'entrées (pour les listes séquentielles), impression, enregistrement des données, et bien sûr on visualise et on se déplace aisément dans la liste.

Des modifications apportées à une liste peuvent avoir des répercussions sur la matrice qui lui est liée.

A:\t6

F1 -édition des entrées horizontales.

F2 -édition des entrées verticales.

F3 -édition de la matrice.

F4 -matrice sélectionnée.

F10 -Fin

Votre choix: F1

A:\t6.E00

CONFIGURATION

liste simple ordonnée sur le champ principal duplication non autorisée.

---

abandonner

abolir

abroger

accepter

accueillir Advm

adjurer

admettre

adopter

affecter

afficher

1er champ:

CREER

A:\t6.MAT SÉLECTIONNÉE

F1 -édition des entrées horizontales.

F2 -édition des entrées verticales.

F3 -édition de la matrice.

F4 -matrice sélectionnée.

F10 -Fin

Votre choix: F2

aucune ligne n'a été sélectionnée.

A:\t6.E01

CONFIGURATION

liste double séquentielle.

NO h:sujet, substantif humain (Paul)

NO-r:sujet, groupe nominal non restreint P

NO-rle faitQuP:sujet, groupe nominal non restreint,complétive:le fait QU P.

NO-rV1C:sujet, groupe nominal non restreint, infinitive dont le sujet est le le .

NOV:Sous-structure, les compléments peuvent être facultatifs

NOVcontreN h:distribution sujet verbe 'contre' substantif humain

NOVaprèsN h:distribution Sujet Verbe 'après, substantif humain

N1QuPind:complément direct, complètive à l'indicatif

N1VOC:complément direct, complètive Qu Pind restructurée en une infinitive dont

N1auxVOC:complément direct, complètive Qu Pind restructurée en une infinitive a

1er champ:

2ième champ:

CRÉER AVANT

A:\t6.MAT SÉLECTIONNÉE

F1 -édition des entrées horizontales.

F2 -édition des entrées verticales.

F3 -édition de la matrice.

F4 -matrice sélectionnée.

F10 -Fin

Votre choix: F3

aucune ligne n'a été sélectionnée.

A:\t6.MAT

	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	q	q	è
	O	O	O	O	O	O	O	l	l	i	l	l	l	l	l	u	u	t
		-	-	-	V	V	V	Q	V	a	d	V	V	è	A	e	e	r
	h	r	r	r		c	a	u	O	u	e	l	l	t	d	N	N	e
			l	V		o	p	P	C	x	V	i	a	r	j	O	O	A
			e	l		n	r	i		V	O	n	n	e		V	V	d
			f	C		t	è	n		O	C	f	t	A		A	è	j
			a			r	s	d		C		C	C	d		d	t	Q
			i			e	N							j		j	r	u
			t			N											e	P
			Q				h										A	
			u			h											d	
			p														j	

abandonner	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
abolir	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
accepter	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
accueillir Advm	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
adjurer	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
admettre	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

CONCLUSION

Le traitement automatique des documents naturels ne peut se faire efficacement que si l'on ne dispose d'un environnement logiciel spécialisé et suffisamment convivial afin que des non informaticiens linguistes puissent l'utiliser aisément. **Un prototype d'atelier de génie linguistique: l'atelier L1 est développé en Smalltalk [10] dans le cadre du LIANA** pour satisfaire les besoins de divers projets: Le projet Franco Canadien de Systèmes Intelligents en Langue Française, le **PRC** informatique et langue naturelle.

La fabrication d'un protot pe de compilateur en langue naturelle à l'aide de l'atelier L1 doit illustrer la puissance et l'efficacité d'un tel environnement.

Bibliographie

AHO, Alfred V. & ULLMAN, Jeffrey D.: *The Theory of Parsing Translation and Compiling*, Prentice-Hall, New Jersey 1972.

CHOMSKY, N.: *Structures syntaxiques*, Seuil, Paris, 1969.

DUCROT et TORODOV: *Dictionnaire encyclopédique des sciences du language*, Seuil, Paris,1972.

GINSBURG, S.: *The mathematical Theory of context-free languages*, MacGraw-Hill, New-Jersey, 1966.

LAURIERE, J.-L.: *Résolution de problèmes par l'Homme et la machine*, Eyrolles, Paris, 1986.

JAYEZ, J.-H.: *Compréhension automatique du langage naturel*, Masson, Paris, 1982.

AUTEBERT, J.-M.. *Langages algébriques*, Masson, Paris, 1987.

KLEENE, S.-C.: *Logique mathématique*, A Colin, Paris, 1971.

HUE, J.-F.: *Conception d'un générateur*, DEA, Toulouse, 1984.

GOLDBERG, A.: *Smalltalk-80*, Addison Wesley, New Jersey, 1984.

GROSS, M.: *Grammaire transformationnelle du Français*, LADL, Paris 1986.

ÉTUDE DU DEGRÉ DE DIFFICULTÉ DE TEXTES RELATIFS À L'INFORMATIQUE

Jocelyne Bourque
Université Laval

0. INTRODUCTION

Cette étude a pour objectif le développement d'une méthode de mesure du degré de difficulté de textes relatifs à l'informatique. Les mesures ainsi obtenues devraient guider le lecteur éventuel dans le choix de textes adaptés à sa competence.

Le principe général de la méthode est de mesurer la difficulté en fonction du vocabulaire contenu dans le texte, en tenant compte du contexte. Il s'agit donc d'évaluer chaque terme selon divers critères, tant lexicaux que contextuels. Une mesure globale pour le texte à l'étude peut ensuite être obtenue en fonction des cotes de difficulté de l'ensemble de son vocabulaire. Nous présentons ici les différentes variables, lexicales et contextuelles, qui interviennent dans l'évaluation du degré de difficulté des termes propres à l'informatique.

1. DESCRIPTION DES VARIABLES DE MESURE

1.1 Variables lexicales

1.1.1 *Catégorie grammaticale*

La catégorie grammaticale a déjà été avancée comme un facteur influençant la difficulté dans le domaine des langues secondes (Mackey, 1965). On y mentionnait que les verbes étaient plus difficiles à apprendre que les noms ou substantifs. D'autre part, des recherches en terminologie (Guilbert, 1981) ont appuyé l'hypothèse voulant que la catégorie des substantifs présente moins de difficulté que les autres catégories grammaticales.

Bien que rien n'indique la présence d'une hiérarchie des difficultés parmi les catégories autres que les substantifs, nous avons tout de même distingué plusieurs valeurs pour la variable catégorie grammaticale. Il sera possible d'évaluer la pertinence de cette distinction au cours d'une analyse subséquente. Les cinq valeurs retenues sont : substantifs, verbes, adjectifs, adverbes, autres.

La majorité des termes ne posent aucun problème pour l'identification de la catégorie grammaticale. Toutefois, quelques cas méritent une attention particulière.

Le premier d'entre eux touche les noms de programmes, de commandes ou de fichiers, considérés normalement comme des substantifs. Si "command.com", et "exe2bin" sont immédiatement identifiés comme des substantifs, il n'en va pas de même pour un cas comme le programme "debug". S'agit-il d'un verbe ou d'un substantif ? Son comportement en contexte peut nous renseigner sur sa nature, particulièrement lorsqu'il n'est pas précédé du mot programme qui confère inévitablement à l'expression une valeur de substantif.

Dans tous les exemples retrouvés dans le texte, "debug" se comporte comme un substantif en tenant le rôle de sujet d'un verbe. C'est donc la valeur substantif et non verbe qui sera

attribuée au terme "debug" lors de la saisie de la variable catégorie grammaticale. Il en ira de même des autres cas bâtis sur le même modèle ("select", "write", "replace"...).

Un autre indice contextuel vient appuyer cette décision. La plupart des portions de textes qui présentent un programme ou une commande débutent par le nom de cette commande (en anglais) suivi d'une périphrase en français expliquant sa signification. Par exemple : "RESTORE - Restauration d'un disque dur". Chacun des noms de commande est expliqué au moyen d'un substantif ce qui renforce l'idée que ces termes ont vraiment une telle valeur dans le corpus étudié.

Le cas de certains termes tels que "ax" et "bp" pose également certains problèmes. Leur absence totale de signification à première vue empêche une identification immédiate de la catégorie grammaticale. Encore une fois, la référence au contexte s'avère nécessaire. D'après celui-ci, les termes recherchés sont des noms de registres. Les registres se définissent comme des zones de mémoire ou emplacements et on peut leur trouver une certaine similitude avec des lieux géographiques. Dans cette optique, les noms de registres, tout comme les noms de lieux, doivent figurer dans la catégorie des substantifs.

Aux problèmes mentionnés plus haut s'ajoute celui des nombreuses abréviations dont regorge le texte à l'étude. "Ko", "md" et "alt", abréviations respectives de "kilo octet", "mkdir" (elle-même abréviation de "make directory") et "alternate" n'en sont que quelques exemples. Pour chaque abréviation, la prise de décision quant à la catégorie grammaticale a débuté par une recherche du mot abrégé et c'est la catégorie de ce dernier qui a été affectée à son abréviation. "Ko" et "md" (nom de commande) se sont donc vus attribuer la catégorie substantif tandis que "alt" prenait place dans le groupe des verbes.

1.1.2 *Longueur*

Les études portant sur la longueur du mot (Henderson, 1982, McNeil, 1987) n'ont pas démontré d'influence de ce facteur sur la difficulté de compréhension. Cependant, la longueur était alors évaluée en fonction du nombre de lettres ou du nombre de syllabes constituant le mot. Il serait intéressant d'évaluer la longueur, non pas d'après le nombre de lettres ou de syllabes, mais d'après le nombre de composants (mots) qui forment un terme. La longueur, toujours égale à 1 pour les termes simples, prendrait une valeur supérieure dans le cas des syntagmes (ex.: longueur = 1 pour "disquette", 2 pour "disque dur", 3 pour "unité grande capacité").

La longueur est très facile à déterminer pour les trois exemples précédents : il suffit de compter le nombre de constituants. Cependant, de nombreux syntagmes contiennent des mots grammaticaux tels qu'articles, prépositions et conjonctions. Ces éléments, bien qu'ils entraînent parfois une certaine complexité sur le plan syntaxique, ne posent pas de problème quand on parle de difficulté du vocabulaire car ils portent très peu d'information de nature sémantique. Nous avons donc résolu de ne pas en tenir compte dans la mesure de la longueur des syntagmes et de compter uniquement les éléments porteurs de sens (ex.: longueur = 2 pour "impression écran" et pour "impression d'écran").

Bien que de longueur égale, ces deux expressions peuvent sembler différentes du point de vue de la difficulté. "Impression écran" paraît peut-être moins clair que "impression d'écran" et cette différence n'est pas reflétée dans la longueur telle que nous la mesurons. Ce phénomène est tout à fait normal puisque la difficulté amenée par "impression écran" ne tient pas à la longueur du syntagme, mais à la façon dont il est formé. C'est la construction anglaise du syntagme qui diminue la clarté et ce fait est noté au moyen de la variable mode de formation qui sera expliquée plus loin.

On ne sait pas de façon certaine si la longueur du terme, basée sur le nombre de constituants, a une influence sur la difficulté de compréhension ni, si influence il y a, dans quel sens elle se manifeste. Trois hypothèses peuvent être soulevées à ce sujet.

La première hypothèse veut que la longueur n'ait aucune influence sur la difficulté et qu'un syntagme long puisse être aussi facile ou aussi difficile qu'un terme simple.

Selon la deuxième hypothèse, plus un terme est long, plus il est difficile car il comporte un plus grand nombre de mots susceptibles d'apporter des difficultés. Par exemple, "commande" par rapport à "commande externe" peut sembler plus facile à cause de la présence, dans le deuxième terme, de l'élément "externe" dont la signification présente certaines difficultés.

Enfin, la troisième hypothèse suppose que plus un terme est long, plus il est facile, car, à la manière d'un contexte, les éléments supplémentaires peuvent apporter des éclaircissements sur le sens d'un élément inconnu. Ainsi, lorsqu'on compare les termes "tête" et "tête de lecture-écriture de l'unité de disquette" et que le sens de l'élément "tête" n'est pas connu, le syntagme long est plus explicite que le terme simple. En effet, il indique où se trouve la tête en question et ce à quoi elle sert.

De ces trois hypothèses une seule devra être retenue dans la formule finale du calcul de la difficulté. C'est dans une étape subséquente de la recherche et en utilisant un logiciel de statistiques que le choix pourra s'effectuer.

1.1.3 *Degré d'abstraction*

Tel que l'ont avancé certaines études dans le domaine de l'enseignement des langues secondes (Mackey, 1965), l'influence du degré d'abstraction sur la difficulté de compréhension est assez facilement concevable. On peut, avec raison, présumer que les termes concrets s'apprennent et se comprennent plus facilement que les termes abstraits.

L'assignation de la valeur concrète ou abstraite aux différents termes du corpus s'effectue assez facilement dans certains cas. "clavier" et "imprimante" se classent sans problème dans le groupe des termes concrets, tandis que "suppression" et "originale", désignant respectivement une action et une qualité, appartiennent indubitablement aux termes abstraits.

Un problème se pose avec certains termes qui ne se situent pas clairement dans la classe abstraite ou concrète. Par exemple, le terme "octet", en tant qu'unité de mesure de la capacité mémoire, paraît plutôt abstrait, alors qu'il penche davantage vers les termes concrets lorsqu'on l'envisage dans le sens d'une portion physique d'un disque.

Cette incertitude oblige la création d'une autre classe, située entre les groupes concret et abstrait. Cette classe, que nous qualifierons de mi-concrète mi-abstraite, pourra recueillir les cas qui ne peuvent se placer résolument dans les deux autres groupes.

Malgré cet ajout, il demeure que les frontières ne sont pas tout à fait fermées d'une classe à l'autre et que le choix d'une valeur pour le degré d'abstraction peut parfois s'avérer complexe. L'intuition ne suffisant pas toujours à trancher la question, l'emploi de critères un peu plus rigoureux devient nécessaire pour délimiter clairement les trois groupes.

La classe concrète regroupe les termes qui représentent des objets qu'on peut toucher ou pointer. Les termes abstraits sont ceux qui ne peuvent avoir de représentation physique sans le recours à un objet de référence. L'exemple suivant permet de mieux comprendre ce critère:

Une "copie" (action de copier) ne peut être représentée visuellement sans un item à copier. C'est cet item (fichier, disquette ou répertoire), qui constitue l'objet de référence grâce auquel "copie" devient représentable.

La troisième classe, celle des termes mi-concrets mi-abstraits, peut plus difficilement être délimitée par des critères absolus. Nous dirons simplement qu'elle regroupe tous les termes qui n'entrent dans aucune des deux autres catégories.

1.1.4 *Mode de formation*

En dépit de plusieurs recherches en ce sens, nous n'avons malheureusement pas pu mettre la main sur une étude comparative des divers modes de formation des termes en fonction du degré de difficulté. C'est donc le corpus lui-même qui a servi de base à la définition de ce critère. En effet, pour déterminer les différentes valeurs possibles de la variable, nous avons élaboré une liste de tous les modes de formation du vocabulaire spécialisé rencontrés dans le corpus à l'étude.

Cette liste tente de présenter un ordre intuitivement croissant de la difficulté de compréhension, mais les indices sont trop peu nombreux pour permettre un classement sûr des valeurs de la variable. Comme c'était le cas pour la variable longueur, une analyse devra être effectuée ultérieurement pour s'assurer de la position respective des différentes valeurs. Celles-ci sont au nombre de neuf et correspondent aux neuf modes de formation recensés dans le corpus.

Le premier groupe réfère aux termes qui sont puisés dans un dictionnaire de français général et employés avec leur sens courant. Les exemples suivants en font partie : "ordinateur", "effacer", "imprimer".

Les termes qui sont formés par analogie proche se retrouvent dans la deuxième classe. L'analogie proche signifie que la forme du terme est la même que celle d'un mot général, mais que son sens s'éloigne légèrement du sens courant. Ainsi, un "fichier", en informatique, a pour mode de formation l'analogie proche : sa forme est identique à celle du mot général et ils ont en commum plusieurs caractéristiques sémantiques.

Il arrive que la forme d'un terme spécialisé soit celle d'un terme général, mais que leurs sens soient assez éloignés. Nous parlons dans ce cas d'une analogie éloignée, laquelle constitue le troisième mode de formation. Par exemple, "configuration", "partition" et "unité" appartiennent à ce groupe.

Les abréviations sont regroupées sous le quatrième mode de formation. Notre corpus en propose une assez bonne sélection, dont les exemples suivants : "car" (abréviation de caractère), "impec" (abréviation de impression écran) et "con" (abréviation de console).

Plusieurs des termes spécialisés reconnus dans notre étude sont empruntés totalement ou en partie à une langue étrangère, en l'occurrence l'anglais. Le cinquième mode de formation les rassemble sous l'étiquette emprunts. Ce sont surtout des noms de commandes, comme en font foi les exemples qui suivent : "copy", "erase", "tree", "end", "select".

Dans quelques cas, des termes tout à fait nouveaux se sont ajoutés au vocabulaire de l'informatique. Ces néologismes, formés spécifiquement pour désigner une notion nouvellement apparue, constituent le sixième mode de formation. Nous en avons exclu les emprunts pour en faire une classe à part, bien que ceux-ci soient parfois considérés comme un type de néologisme. La raison est que l'utilisation d'emprunts, étrangers mais peut-être connus, n'offre pas nécessairement la même difficulté que l'apparition d'une forme entièrement nouvelle pour le lecteur. "Formatage", "octet" et "disquette" sont les exemples les plus connus dans le groupe des néologismes de forme.

Le septième mode de formation est celui des acronymes. C'est un groupe très restreint ne contenant que quelques termes et des syntagmes formés à partir de ces derniers (exemples : "dos", "basic", "ascii", "bit").

Le groupe suivant n'est pas davantage productif dans notre corpus. Il renferme les noms propres et dérivés de noms propres, tels que "Pascal" et "Microsoft".

Enfin, les éléments correspondant au neuvième mode de formation portent le nom de symboles. Il s'agit d'assemblages apparemment arbitraires de caractères, sans rapport visible avec le sens. Le corpus en compte plusieurs exemples, entre autres : "xt", "wtvqq" et "nv".

L'attribution de l'une ou l'autre des neuf valeurs à la variable mode de formation pour les termes à analyser pose rapidement un problème : il est fréquent de rencontrer dans un même terme une combinaison de deux ou plusieurs modes de formation. Par exemple, "del", abréviation de "delete", correspond aux quatrième et cinquième mode de formation. Ce phénomène de combinaison, tout de même assez rare chez les termes simples, devient presque universel lorsque le terme est composé de plusieurs éléments. Il faut donc établir une règle de conduite qui soit applicable à tous les cas pour éviter de faire des choix arbitraires.

Nous avons pensé que le lecteur, placé devant un terme comportant plusieurs éléments ayant des modes de formation différents, saisira facilement l'élément dont le mode de formation est d'un degré de difficulté peu élevé. Par contre, la compréhension du terme entier sera rendue plus difficile par la présence de l'élément de difficulté élevée sur lequel le lecteur butera sans doute. C'est donc le mode de formation porteur de la plus grande difficulté qui détermine la difficulté d'un syntagme à ce niveau.

Par conséquent, l'attribution d'une valeur à la variable mode de formation, en présence d'une combinaison de modes, équivaut à l'attribution de la valeur la plus élevée rencontrée dans le terme. Exemples :

> "Wo", abréviation du mot anglais "word", ayant un sens particulier en informatique, provient d'une combinaison des modes de formation 4 (abréviation) et 5 (emprunt). C'est cette dernière valeur qui sera attribuée au terme "wo".
>
> "Impression en écho" contient d'abord l'élément "impression" ayant la valeur 1 (sens courant) et ensuite l'élément "écho" portant la valeur 3 (analogie éloignée). "Impression en écho" se verra donc attribuer la valeur 3 pour la variable mode de formation.

1.1.5 *Degré de spécificité*

Le vocabulaire scientifique et technique a déjà été catégorisé selon le degré de spécificité (Descamps et Phal, 1968). Il en est ressorti trois groupes de termes :

1- le vocabulaire scientifique
2- le vocabulaire semi-spécifique
3- le vocabulaire technique

Le vocabulaire scientifique, tel que décrit par Descamps et Phal, comprend les mots de sens très général communs à plusieurs spécialités au niveau fondamental. Ceux-ci peuvent être le point de départ de lexies complexes (ex.: coefficient). "Coefficient d'absorption" appartient au vocabulaire semi-spécifique tandis que "coefficient d'absorption totale linéaire" représente la classe du vocabulaire technique.

Les exemples présentés dans l'étude de Descamps et Phal laissent supposer que c'est la longueur du terme qui détermine le degré de spécificité. Effectivement dans ce cas précis, "coefficient d'absorption" est plus spécifique que "coefficient" car il y a spécification au moyen d'un terme propre à un domaine ("absorption"). "Coefficient" passe donc de la classe du vocabulaire scientifique général à celle du vocabulaire semi-spécifique par l'adjonction d'un terme appartenant à cette deuxième classe.

Mais, lorsque le mot spécifié est de type général et que celui qui le spécifie est également membre de la première classe, le terme composé obtenu n'appartient pas à la classe du vocabulaire technique ni même du vocabulaire semi-spécifique. Il demeure, au contraire, parmi les termes scientifiques généraux.

Prenons, par exemple, le cas de "touches numériques" dont les deux éléments font partie du vocabulaire scientifique général. Le terme entier demeure général et applicable à différents domaines. La longueur du terme n'a donc pas permis de déterminer le degré de spécificité. Il en va de même pour des termes simples qui ne sont pas assez généraux pour se situer dans le premier groupe. Nous pensons aux termes tels que "bit", "octet", "formatage", etc.

Voici les critères qui nous ont paru, mieux que la longueur, indiquer le degré de spécificité :

Dans le cas d'un terme simple, si ce terme est général et commun à plusieurs sciences, on lui attribue le degré de spécificité 1, correspondant à la classe du vocabulaire scientifique général (ex.: "copie"). Si, par contre, le terme est spécifique à un domaine, en l'occurrence l'informatique, on considère qu'il fait partie du vocabulaire semi-spécifique, dont le degré de spécificité est 2 (ex.: "disquette").

Pour un terme composé, il faut d'abord distinguer l'élément spécifié du ou des éléments qui le spécifient. Cela fait, trois situations peuvent se présenter : les deux premières se produisent lorsque le mot spécifié relève du vocabulaire scientifique général. Dans le premier cas, l'élément qui spécifie appartient également à cette catégorie. Nous avons déjà donné l'exemple de "touches numériques", où le terme composé avec deux éléments de degré 1 porte lui aussi le degré de spécificité 1. L'autre cas est celui où le mot qui spécifie appartient au vocabulaire semi-spécifique tandis que l'autre est général (ex.: "copie de disquette"). Le terme composé se verra alors attribuer le degré 2, correspondant au vocabulaire semi-spécifique.

Enfin, la troisième situation prend en compte les mots spécifiés qui appartiennent à un domaine spécialisé. Les mots composés à partir de tels éléments portent le degré de spécificité 3, correspondant au vocabulaire technique de Descamps et Phal, car il y a spécification d'un terme déjà spécialisé (exemple : "disquette simple face"). Le nombre d'éléments dans un terme de degré 3 n'est pas limité. En voici un exemple :

> "Valeur de déplacement hexadécimale". La partie "valeur de déplacement" se compose d'un élément de degré 1 ("valeur") spécifié par un élément de degré 2 ("déplacement" au sens utilisé spécifiquement en informatique). Cette partie constitue donc un élément de degré 2, qui grimpe au degré suivant lorsque spécifiée par l'élément "hexadécimale".

1.2 Variables contextuelles

1.2.1 *Définition*

La présence d'une définition dans le contexte entourant un terme diminue sans contredit la difficulté de compréhension. L'évaluation de cette variable consisterait donc, pour chaque occurrence, à dire s'il y a ou non présence d'une définition.

Cependant, lorsqu'une occurrence d'un terme est définie, la compréhension de toutes les occurrences suivantes est facilitée, et non pas seulement celle de cette occurrence. Donc, lorsqu'un terme a été défini quelque part dans le texte, il faut considérer comme définies toutes les occurrences suivantes.

1.2.2 *Illustration*

Un illustration peut, dans bien des cas, éclairer le lecteur sur le sens du terme illustré. Cette variable, comme la variable définition, peut prendre deux valeurs, selon qu'il y a présence ou non de l'élément contextuel explicatif, en l'occurrence une illustration.

De même que pour la définition, la présence d'une illustration accompagnant une occurrence d'un terme facilitera la compréhension des occurrences suivantes. Celles-ci seront donc considérées comme illustrées, même si l'illustration n'est pas répétée.

1.2.3 *Synonyme*

La présence d'un synonyme peut aider à la compréhension d'un terme, quoique ce ne soit pas toujours le cas. En fait, un synonyme agit parfois comme une définition. Dans ce cas, il facilite la compréhension et peut être considéré comme un critère d'évaluation de la difficulté.

Par exemple, dans notre corpus, "disque dur" a pour synonyme "disque fixe". Le premier terme est le plus utilisé mais le second est plus facilement compréhensible. Le fait de donner ce synonyme à "disque dur" facilite sa compréhension. "Disque fixe" est donc considéré comme un synonyme à valeur de définition. Les synonymes "première disquette" et "disquette originale" agissent de la même façon auprès de "disquette source".

Lorsqu'une occurrence d'un terme est accompagnée d'un synonyme à valeur de définition, la présence de ce synonyme sera enregistrée pour cette occurrence et toutes les occurrences suivantes.

1.2.4 *Exemple*

La présentation d'un exemple dans un texte aide à la compréhension du terme concerné dans le contexte spécifique de cette occurrence. Des occurrences différentes peuvent nécessiter des exemples différents selon le contexte.

Ainsi, la commande "copy" s'utilise de différentes façons et des exemples distincts accompagnent certaines occurrences de ce mot. D'autres types d'utilisation de cette commande ne sont pas présentés avec un exemple.

Par conséquent, la présence d'un exemple auprès d'une occurrence n'entraînera pas l'enregistrement de cet exemple pour d'autres occurrences du terme considéré.

2. CONCLUSION

Ces différentes variables appliquées aux termes du corpus étudié, feront l'objet d'une analyse de type régression "stepwise" qui permettra de déterminer leur influence relative sur la difficulté de compréhension.

Quels que soient les résultats de cette analyse, il demeure certain que la difficulté de compréhension du vocabulaire est un facteur important dans la compréhension générale d'un texte et qu'elle mérite qu'on s'y attarde.

Bibliographie

DESCAMPS, J.L. et PHAL, A. *La recherche linguistique au service de l'enseignement des langues de spécialité.* Dans *Le français dans le monde*, Hachette et Larousse, Paris, 1968, vol. 8, no 61, pp.12-19.

GUILBERT, Louis. *La relation entre l'aspect terminologique et l'aspect linguistique du mot.* Dans *Textes choisis de terminologie*, G. Rondeau et H. Felber, rédacteurs, Groupe interdisciplinaire de recherche scientifique et appliquée en terminologie Girsterm, Université Laval, Québec, 1981, 334 p.

HENDERSON, Leslie. *Orthography and Word Recognition in Reading.* Academic Press, London, 1982, 397 p.

MACKEY, William Francis. *Language teaching Analysis.* Longman Group Ltd, London, 1965. 562 p.

McNEIL, Davic. *Psycholinguistics, a new approach.* Harper & Row Publishers, New York, 1987, 290 p.

Auteur **André Dugas**
Université du Québec à Montréal

Titre **L'évaluation de la productivité lexicale et les dictionnaires électroniques**

RÉSUMÉ

Le phénomène de la productivité lexicale est peu étudié. D'une part, les lexicographes doivent se contenter, d'un dictionnaire à l'autre, de réunir les unités lexicales attestées sans pouvoir en assurer une certaine "homogénéité". Par exemple, la dernière édition du Grand Robert a l'entrée SURDIMENSIONNER mais n'a pas SOUS(-)DIMENSIONNER. D'autre part, l'examen par les linguistes descriptivistes d'unités lexicales non attestées n'est pas courant.

L'étude du matériel virtuel disponible présente un grand intérêt dans le cadre de la constitution puis de la consultation de dictionnaires électroniques. Ce type de dictionnaire dont l'une des caractéristiques est de prétendre à la plus grande exhaustivité suppose une redéfinition des rapports à la norme et entraîne l'examen inédit de nombreux paradigmes lexicaux.

En prenant comme base de données la classe verbale des entrées du DELAS (un dictionnaire électronique de LADL), nous proposons dans cet exposé un inventaire des processus productifs de la formation des unités lexicales de cette classe. Nous verrons ensuite des modèles fondamentaux pour la formation de nouvelles unités bien formées mais non attestées dans les ouvrages lexicographiques connus.

FORMALISATION ET VARIATION LINGUISTIQUE: Le français du Québec*

Jacques Labelle
UQAM

SOMMAIRE:

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■

1. INTRODUCTION

Dans la perspective d'un traitement automatique, les phrases suivantes (1-4) devraient être reconnues comme françaises:

(1) *Max n'arrête pas de barber Pierre*
(2) *Luc accote Simon aux échecs*
(3) *Paul a pris une brosse toute la nuit*
(4) *Guy ne se laisse pas piler sur les pieds*

aussi bien que (5-8):

(5) *Le spectacle barbe Pierre*
(6) *Luc ne me remet pas*
(7) *Paul en a pris une, de cuite*
(8) *Guy se fait un sang d'encre*

---

*La recherche sur le "Lexique-Grammaire du français du Québec" bénéficie d'une subvention du Conseil de Recherche en Sciences Humaines du Canada (n° 410-85-1382)

L'analyse de ces séquences donne des résultats différents, sur le plan de l'acceptabilité, selon que les tests sont effectués par un linguiste parisien ou québécois. Les phrases (1-4) pourront générer des "*" ou "?*"; le linguiste québécois accolera le symbole "*" aux séquences (5) et (7) et cherchera une analyse pour (6) et (8) en se demandant, par exemple, si *sang d'encre* n'est pas un nom composé, un mot technique. Il est bien connu que le travail du linguiste s'appuie sur l'intuition, sur sa connaissance implicite de la langue. Comment pourrait-il, autrement, vérifier des équivalences ou des différences de sens? C'est cet aspect du travail d'analyse (indispensable pour l'identification et la caractérisation des phrases élémentaires) qui entre en jeu lorsque nous établissons, par exemple, des définitions (ou mieux, des approximations) sémantiques; notons entre guillemets ces "approximations sémantiques" pour les exemples (1-8):

(1') "Max n'arrête pas de provoquer Pierre"
(2') "Luc est aussi bon que Simon aux échecs"
(3') "Paul a pris un coup toute la nuit"
(4') "Guy ne se laisse pas marcher sur les pieds"
(5') "Le spectacle ennuie Pierre"
(6') "Luc ne me replace pas (dans sa mémoire)"
(7') "Paul a pris un coup"
(8') "Guy se fait du souci"

Il faut comprendre que les équivalents sémantiques sont éventuellement très nombreux et qu'aucun critère formel ne permet de choisir la meilleure équivalence; c'est le problème de la synonymie. Il est entendu que cette perception intuitive de sens ou mieux, de différence sémantique entre deux phrases simples, doit être prolongée par au moins une différence observable, formelle; nous reviendrons sur ce point en 2.

L'expérience qui à consiste vérifier l'acceptabilité d'une phrase doit donc être faite par un linguiste de la variété à décrire. On a eu trop souvent tendance à négliger cette importante condition expérimentale. L'astérisque devant une phrase devrait être précisé par une indication de la variété où ont été faites les vérifications. Dans le cas qui nous occupe ici, nous utiliserons le symbole Q pour français du Québec et F pour français normé. Ainsi, des phrases comme (2) et (5) ne seront pas affublées de symboles "*" ou "?*", mais marquées selon la variété, comme:

(2a) (Q) *Luc accote Simon aux échecs*

(5a) (F) *Le spectacle barbe Pierre*

Ainsi, la formalisation en vue de l'établissement d'un lexique-grammaire (LG) du français et de son traitement automatique en est de beaucoup facilitée et améliorée. Le LG s'appuie sur le principe de l'exhaustivité (M. Gross, 1975), c'est-dire sur la description de l'ensemble des phrases élémentaires du français (et des mécanismes de mise en relation des phrases); ce LG doit, par conséquent, inclure le français du Québec[1] (ainsi que d'autres variétés de français), ce qui implique un niveau de description tout aussi complet et détaillé pour ce "français" que pour le "français" métropolitain. Nous montrerons ici que ce travail n'est pas simple et qu'il exige une technique d'autant plus précise que l'objet à décrire (Q) n'est pas habituellement perçu comme distinct de l'objet (F) traditionnellement étudié.

[1] Au département de linguistique de l'université du Québec à Montréal, le Groupe de Recherche sur la Formalisation Linguistique (GRFL) se consacre actuellement à l'élaboration de descriptions linguistiques systématiques du français du Québec, utilisables par un ordinateur. La perspective de ces travaux est celle du LADL (Laboratoire d'Automatique Documentaire et Linguistique, CNRS, France).

Au département de linguistique de l'université du Québec à Montréal, le Groupe de Recherche sur la Formalisation Linguistique (GRFL) se consacre actuellement à l'élaboration de descriptions linguistiques systématiques du français du Québec, utilisables par un ordinateur. La perpective de ces travaux est celle du LADL (Laboratoire d'Automatique Documentaire et Linguistique, CNRS, France).

2. DESCRIPTION DE PHRASES SIMPLES

Les dictionnaires d'usage courant donnent pour chaque mot un certain nombre de "sens" servant à distinguer les divers emplois de ce mot; mais ces distinctions, faites essentiellement sur une base intuitive et non systématique, sont destinées à l'usager ayant déjà une bonne connaissance de la langue et ne sont pas utilisables, comme telles, dans un système automatique (cf. M. Gross, conférence d'ouverture). La procédure du LG consiste à séparer (sur une base extensive) et formaliser les phrases simples, c'est-dire les unités lexico-syntaxiques nucléaires.

Nos travaux à Montréal donnent priorité à l'examen des éléments lexicaux opérateurs de phrases, en particulier les verbes à complétive(s) (cf. A. Blanger 1987, 1988) ou sans complétive, les constructions à verbes supports (cf. F. Caviola et L. Grou 1988) et les expressions figées (cf. J. Labelle 1988a, 1988b). Nous illustrerons ici ce programme à partir de verbes et d'expressions figées (EF).

2.1 Les verbes

Pour réaliser une bonne couverture lexicale, nous utilisons plusieurs moyens:

- le dépouillement direct de corpus existants en français du Québec (par ex. le corpus Bibeau-Dugas 1964);
- le dépouillement de dictionnaires ou glossaires du français du Québec (par exemple, le Glossaire de 1937 et le dictionnaire Plus de 1988);
- la cueillette d'éléments lexicaux provenant des membres du groupe.

Il faut souligner que nous avons bénéficié d'une liste de verbes du TLFQ,[2] laquelle a beaucoup contribué à compléter nos listes.

Nous avons vu, à propos des exemples (1-8), que tous les verbes F doivent être examinés en détail, ainsi que les entrées lexicalement distinctes comme (Q)*achaler, canter, enfarger*, etc., puisque leurs emplois syntaxiques risquent fort de différer. Nous n'avons pas une idée précise de l'ampleur de ce phénomène, encore peu exploré. Prenons l'exemple du verbe *accoter* qui a deux entrées dans le *Petit Robert* (dont une difficilement interprétable, parce que sans exemple) et sept dans le dictionnaire *Plus*. Nous reprenons les diffrents emplois de ce verbe en les motivant de la façon suivante: chaque entrée (=emploi) doit correspondre à une interprétation distincte (intuition sémantique) et être marquée formellement:

[2]Nous tenons à remercier l'équipe du Trésor de la Langue Française du Québec, en particulier M. Claude Poirier, directeur du projet, pour son aimable collaboration.

(1) *N0 V N1 Loc N2 =:*
(Q) *Luc accote l'échelle (contre + sur + ...) le mur*
= *L'échelle accote (contre + sur + ...) le mur*

(2) *N0 V N1 Prép N2 =:*
(Q) *Luc accote Simon aux échecs*
= *Luc et Simon s'accotent aux échecs*

où *Loc ≠ contre, près de, ...*, où le verbe signifie approximativement "être aussi bon que, rivaliser, ..." et où N2 est un abstrait, contrairement à (1); notons que cet emploi est symétrique:

(Q) *Luc s'accote avec Simon aux échecs*
= *Simon s'accote avec Luc aux échecs*

(3) *N0 V N1 =:*
(Q) *Le piquet accote la porte brise*
= *La porte est accote (avec + par) un piquet*

(4) *N0 V =:*
(Q) *La porte accote (deux places + sur le seuil)*

différent de (2), parce qu'elle ne peut en être dérivée par relation de neutralité et à cause de l'interprétation "la porte frotte".

(5) *N0 V N1 (=Nhum) =:*
(Q) *Les comédiens ont accoté Clémence*

cet emploi (5) se distingue de (2) par le sens ("donner son appui à quelqu'un, encourager quelqu'un") et par l'impossibilité de l'emploi symétrique en *se V:*

(Q) **Les comédiens se sont accotés avec Clémence*
**Les comédiens et Clémence se sont accotés*

(6) *N0 se V avec N1 =:*
(Q) *Jo s'est accotée avec un gars de Québec*
("Jo vit en concubinage avec un gars de Québec")

où *N0* et *N1* sont des noms "humains", où le complément *Prép N2 =: dans ce domaine, sur ce point* est interdit et où l'interprétation est différente de (2):

(Q) *Jo s'est accotée avec un gars de Québec sur ce point

L'emploi (6) se démarque formellement de (2) par l'emploi adjectival suivant:

N0 être V-é (E + avec N1) =:
(6') (Q) *Jo est accotée (E + avec un gars de Québec)*
(2') (Q) **Luc est accoté avec Simon aux échecs*

Nous donnons en annexe un extrait d'une table de verbes transitifs à complément "humain" (32H-Q) qui illustre le cas de figure (5), c'est-à-dire la construction transitive à complément humain.

Certains verbes, F et Q, de champs sémantiques voisins présentent des difficultés de passage d'une variété à l'autre. C'est le cas de la phrase:

Max barbe Pierre

bien formée, en F et en Q; toutefois, en français du Québec, contrairement à F, l'interprétation est active et synonyme de "provoquer". Cette intuition sémantique n'est toutefois pas suffisante pour marquer une séparation nette des deux emplois; il convient d'appuyer formellement cette hypothèse. Il ressort de la paire suivante:

(Q) **Le fait d'aller au concert barbe Pierre*
(F) *Le fait d'aller au concert barbe Pierre*

que la même forme verbale, *barber*, a des comportements syntaxiques distincts en F et en Q et qu'il s'agit bien là d'un fait observable. Il en est ainsi de beaucoup de verbes comme: *caler, gêner, niaiser, planter, replacer, ...*, dans des phrases comme:

(F) **Luc a calé Paul lors de l'entrevue*
(Q) *Luc a calé Paul lors de l'entrevue*

(F) *La chaise gêne Paul*
(Q) **La chaise gêne Paul*

(F) **Max me niaise*
(Q) *Max me niaise*

où les emplois de verbes ne permettent pas, dans l'une ou l'autre des variétés de français, de conclure à la bonne formation de la séquence. D'autre part, on pourra, dans beaucoup de cas, établir des équivalences de formes dans des constructions comme:

Luc ne me ((F)remettait + (Q)replaçait) pas

où l'on remarque, dans les deux cas, l'emploi métaphorique et la possibilité d'un complément abstrait (ou psychologique) du type *dans sa tête, dans sa mémoire*.

2.2 Une classe de verbes

La table 32H-Q[3] illustre quelques propriétés de constructions. Cette classe de verbes a un caractère résiduel et ne constitue pas un objet syntaxique spécialement intéressant; mais, à cause de sa simplicité, elle est particulièrement désignée pour mettre en relief quelques variantes.

[3]Le code de la table de constructions est celui du LADL et le suffixe, ajouté par nous, indique la variété linguistique.

D'une part, les grandes caractéristiques de la table 32H-F se retrouvent en Q: comme le complément direct, le sujet est très généralement de type "humain", ce qui donne à la classe un caractère de symétrie distributionnelle. Elle contient également deux types de compléments qui ont une distribution peu remarquable par rapport à F: ce sont ceux qui correspondent à la question *de combien*? et le complément *sur ce point*:

(Q) *Le marchand a fourré Luc de deux piastres*
(Q) *Pierre a niaisé Luc sur ce point*

D'autre part, la classe 32H-Q se distingue de F par la construction dite "neutre" et la suffixation adjectivale. Dans Boons, Guillet et Leclère (1976), il n'est signalé qu'un seul emploi neutre:

(F) *Des voyous ont dérouillé Paul*
Paul a dérouillé

De notre côté, nous avons remarqué que plus de 10% des verbes actuellement recensés dans cette classe acceptent la relation de neutralité:

(Q) *Le professeur a coulé Max*
Max a coulé

(Q) *Paul a planté Max*
Max a planté

Cette situation va dans le sens d'une intuition partagée par quelques linguistes français ayant observé cette tendance, en français du Québec, à privilégier la construction neutre (Jean-Paul Boons fut le premier à nous le faire remarquer il y a plusieurs années). Ce phénomène n'est pas limité à la classe syntaxique 32H. Nous en avons donné un exemple à propos du verbe *accoter*, dans son emploi à deux compléments, dont un locatif:

(Q) *Luc accote l'échelle (contre + sur + ...) le mur*
= *L'échelle accote (contre + sur + ...) le mur*

Les tables de constructions F nous servent de point de départ et de comparaison; mais nous ajoutons, au besoin, des propriétés du français du Québec. C'est le cas, par exemple, de la dérivation adjectivale en -eux:

N0 être V-eux =:
(Q) *Luc est baveux*
(Q) *Luc est lâcheux*

ayant une relation avec:

(Q) *Luc bave (le monde + les gens)*
(Q) *Luc lâche (le monde + les gens)*

Dans certains cas, le complément peut être du type "non humain":

(Q) *Luc lâche (les . ires + ses entreprises)*

Le complément direct, régulièrement effacé dans la structure adjectivale en *-eux*, est "générique" et correspond à l'aspect "avoir l'habitude de" comme dans la sous-structure verbale correspondante:

(Q) *Luc boude*

pouvant recevoir les interprétations:

(Q) *Luc a l'habitude de bouder*

ou:

(Q) *Luc boude Marie*

Il faut noter que le suffixe *-eux* n'est pas automatiquement commutable avec *-eur* et que, malgré une certaine productivité, il n'est pas prédictible:

(Q) *Luc est (bosseux + bosseur)*
(Q) *Luc est (barbeux + *barbeur)*
(Q) *Luc est (*crouseux + crouseur)*

L'adjectif *niaiseux* n'est pas relié à la structure 32H, ce qui explique le signe "-" dans la colonne *V-eux:*

(Q) *Luc est niaiseux*
("Luc est stupide, maladroit, ...")

= *Luc niaise*
("Luc perd son temps")

≠ *Luc niaise Paul*
("Luc agace Paul")

2.3 Un exemple de forme composée: CP1-Q

Depuis quelques années, nous avons établi des inventaires de constructions figées; mais, dans ce domaine, les listes étant particulièrement pauvres, il nous a fallu fournir des efforts particuliers dans la cueillette des données. Dans un premier temps, nous disposions d'environ 10 000 EF, de type verbal en grande majorité, que nous avons classées sur le modèle de M. Gross (1982) et que nous avons ramenées à environ 4 000 après étude comparative des lexiques F et Q, en 1986-87.[4] Mais il reste, comme pour les verbes, à revoir systmatiquement toutes les EF du LADL et à examiner leurs interprétations et propriétés de constructions: les résultats sont tout aussi imprévisibles que pour les verbes, puisque l'EF est en fait une entrée lexicale composée, appartenant des catégories syntaxiques variées: verbes, adjectifs, adverbes, noms (G. Gross 1988), ...; rappelons l'exemple *avoir du chien*, dans:

(Q) *Ce petit enfant a du chien*

[4]De nouvelles EF continuent régulièrement à s'ajouter à la banque de données qui en contient environ 5 000 actuellement.

qui produit un effet de surprise chez les locuteurs parisiens pour la bonne raison qu'il signifie (F)"avoir du sex-appeal"; cette EF a fait l'objet d'une comparaison détaillée des propriétés de constructions et a donné le résultat suivant, imprévisible au départ: la distribution des propriétés syntaxiques, à première vue identique en F et en Q, s'avère différente et donne des bases formelles aux deux interprétations:

(Q) *Luc a eu le chien de lui répondre cela*
(F) **Luc a eu le chien de lui répon cela*

Les structures dérivées d'EF peuvent éclairer certains emplois de verbes et vice-versa. Ainsi:

(Q) *Paul ambitionne sur le pain béni*
("Paul exagère")

a la sous-structure:

(Q) *Paul ambitionne*

de même sens. Le verbe simple *ambitionner* a un emploi intransitif tout à fait équivalent, comme dans l'exemple suivant, à l'impératif:

(Q) *Paul, ambitionne pas!*

qui semble plus près de l'EF que du verbe (Q) *ambitionner sur:*

(Q) *Paul ambitionne sur Marie*

au sens de "Paul abuse de Marie". Plusieurs exemples de sous-structures sont considérées dans ce sens (voir annexe) et permettent de mettre en lumière le lien étroit qui lie structures figées et libres.

La classe CP1-Q mise en annexe se caractérise par le fait que les phrases élémentaires qu'elle décrit sont du type:

N0 V Prép C1 =:
Jo s'enfarge dans les fleurs du tapis
("Jo se perd dans les détails")

3. COMPARAISONS ET TRANSFERTS

Ces études de lexique-grammaire du français du Québec conduisent à des comparaisons systématiques F-Q et à des applications à l'ordinateur comme la documentation automatique ou les transferts automatiques F-Q (L. Danlos, 1988). Dans un cas ou dans l'autre, il est indispensable de disposer d'une information linguistique formalisée dans le détail et tenant compte des nombreuses différences variationnelles, malheureusement trop souvent laissées pour compte. Pour procéder à des transferts automatiques F-Q, il faudra disposer d'informations complètes sur les phrases élémentaires, figées ou libres, et de mécanismes de mise en équivalence.

3.1 Équivalences lexicales

Dans le cas des équivalences lexicales, on considérera que les phrases élémentaires F1 et Q1 sont en relation d'équivalence parce qu'elles offrent un profil de propriétés syntaxiques identiques et qu'elles ont été jugées sémantiquement semblables, c'est-dire pouvant apparaître dans les mêmes contextes:

F1: *Luc a pris une pelle*
= Q1: *Luc a pris une fouille*

alors que F2 et Q2 ne pourront être dans une telle relation, malgré leur identité lexicale:

F2: *Luc a du chien*
≠ Q2: *Luc a du chien*

Le glossaire d'équivalences des EF devra représenter formellement les éléments lexicaux, en fonction de leur appartenance à telle classe de constructions (et donc un profil précis de propriétés syntaxiques) et fournir les indications lexicales permettant de passer de l'une à l'autre variété.

3.2 Équivalences syntaxiques

Rappelons une règle syntaxique qu'il est nécessaire de compléter par une règle de transfert F-Q: le détachement.

Cette règle transformationnelle (M. Gross 1968) détache le groupe nominal, ou une partie de ce groupe, après pronominalisation.

Luc a coulé (Max + le projet de Max)
[détachement]
= *Luc # il a coulé (Max + le projet de Max)*
[détachement]
= *Luc l'a coulé # (Max + le projet de Max)*

Il n'y a aucune différence F-Q à noter, qu'il s'agisse de *N0* ou de *N1*. Mais certains déterminants, indéfinis et numéraux, entraînent des comportements différents; avec le *ppv en*, Q efface la *Prép* =: *de* alors que F la conserve. Comparons, dans les même conditions transformationnelles de [détachement], des phrases construites sur des verbes lexicalement équivalents:

(Q) *Luc a barbé (un + plusieurs) élève (s)*
[détachement]
= (Q) *Luc en a barbé (un + plusieurs) # (E + *de) élève(s)*

(F) *Luc a chiné (un + plusieurs) élève(s)*
[détachement]
= (F) *Luc en a chiné (un + plusieurs) # (*E + de) élève(s)*

Ce phénomène de détachement, en français du Québec, n'exclut pas la liaison la "pause-intonation" entre le déterminant au masculin singulier et le nom détaché:

(Q) *Max en a barbé un # élève*

où *élève* est phonétiquement relié à *un* par le [n] de liaison. Le détachement est alors phonétiquement marqué par l'allongement de la voyelle précédant ce [n]. Signalons que le détachement sans *de* doit être distingué d'autres cas, comme celui des adjectifs:

(Q) *Luc a plusieurs foulards: des noirs et des blancs*
*Luc en a plusieurs # (*noirs + de noirs)*

Mais si l'ensemble des *foulards* sont noirs, alors le groupe nominal se trouve être: *plusieurs foulards noirs* et seul le détachement sans *de* est autorisé:

(Q) *Luc a plusieurs foulards noirs*
= *Luc en a plusieurs # (noirs + *de noirs)*

En d'autres termes, dans l'exemple précédent, l'ensemble des "foulards" est identique à l'ensemble des "foulards noirs" et le *de* marquant la partition n'est pas admis. Notons qu'avec *beaucoup de*, on n'effacera pas le *de* qui a sa source dans:

Luc a beaucoup de foulards noirs
*Luc en a beaucoup # (*noirs + de noirs + de foulards noirs)*

3.3 Équivalences non formelles

Dans le cas des EF, il y a possibilité de recourir à un lexique de phrases simples où l'on aura mis en relation non formelle des équivalents sémantiques formellement éloignés l'un de l'autre comme:

(Q)	(F)
Il y a rien là!	*Ca mange pas de pain!*
Luc a frappé un noeud	*Luc est tombé sur un os*
Max s'enfarge dans les fleurs du tapis	*Max se perd dans les détails*
Luc a le feu au passage	*Luc est en colère*
Jo a déjà vu neiger	*Jo n'est pas tombé de la dernière pluie*
etc.	

c'est alors une solution complémentaire, pratique, qui fait appel à la "traduction" des expressions, coup par coup.

Il y a une autre situation: celle où les phrases simples ne peuvent pas recevoir d'équivalent, même sémantique, parce qu'il y a lacune totale due aux différences de cultures. Autre problème de traduction. Doit-on mettre en correspondance une expression comme *partie de sucre* avec une

périphérase (F ou FQ) ou emprunter l'EF comme telle, comme on le ferait avec une langue étrangère? La première solution est probablement la meilleure... Mais cette question déborde les cadres et objectifs de notre propos.

Terminons par une indication, bien imparfaite, sur la situation comparative comme elle nous apparaît en ce début de recherche. La comparaison des tables F et Q a donné, dans l'état actuel des travaux, quelques résultats chiffrés partiels que nous préciserons dans la poursuite des travaux et qui n'ont encore qu'une valeur très approximative. Dans les tables de verbes et d'EF examinées jusqu'à présent, l'intersection FQ (c'est-à-dire le fond linguistique commun à F et à Q), en termes d'équivalences lexico-syntaxiques, représente près de 80% des phrases examinées jusqu'ici. Est-ce une tendance significative? Nous n'avons, à l'heure actuelle, aucun moyen d'en décider.

4. CONCLUSION

Nous avons voulu illustrer ici une méthode de description formalisée qui, dans le cas du français du Québec, peut revêtir une grande importance, puisqu'elle précède et conditionne le traitement automatique qui s'en suivra.

Le fait d'introduire le lexique-grammaire du français du Québec dans une importante banque de données formalisées du français, et cela dès le départ, signifie, à notre avis, prendre le train quand il passe (ou ne pas rater le coche); cela signifie participer activement au virage technologique qui s'amorce dans notre discipline.

L'élaboration d'un lexique-grammaire du français du Québec, en vue de son intégration progressive au LG général du français, vise aussi une meilleure connaissance du français et du concept même de langue et de variété de langue.

Références

BÉLANGER, Alain (1987) "Contribution au lexique-grammaire des verbes psychologiques en français du Québec", dans *Lexique-grammaire du français du Québec*, rapport technique n° 18, LADL, CNRS, Paris.

------ (1988) "Syntaxe comparée des verbes à construction complétive", *Verbes et expressions figées en français du Québec*, Rapport de Recherche n° 4, UQAM, Montréal.

BIBEAU, Gilles et A. DUGAS, (1964) *Corpus Bibeau-Dugas*, document non publié réalisé en 1963-64, en dépôt à l'UQAM, Montréal.

BOONS, Jean-Paul, Alain GUILLET, Christian LECLÈRE (1976) *La structure des phrases simples en français: 2- Classes de constructions transitives*, Rapport de recherche n° 6, LADL, Paris.

CAVIOLA, Francine et Lyette GROU (1988) "Quelques observations sur les nominalisations avec le verbe-support *faire*", Rapport de Recherche n° 4, GRFL, UQAM, Montréal.

DANLOS, Laurence (1988) "Interaction des décisions dans un système de génération automatique de textes", Communication présentée au colloque international sur la grammaire et le lexique comparés des langues romanes, déc. 1988, Université Laval, Québec.

DUBOIS, Jean et coll. (1979) *Larousse de la langue franaise, Lexis*, Larousse, Paris.

GROSS, Gaston (1988) "Un projet d'étude systématique des noms composés du français" dans *Verbes et expressions figées en français du Québec*, Rapport technique #4, GRFL, UQAM, Montréal.

GROSS, Maurice, (1968) *Grammaire transformationnelle du français: syntaxe du verbe*, Larousse, Paris.

------ (1975) *Méthodes en syntaxe*, Hermann, Paris.

------ (1982) "Une classification des phrases figées du français", *Revue Québécoise de Linguistique*, Vol.11, No 2, UQAM, Montréal.

LABELLE, Jacques (1988a) "Lexiques-grammaires comparés: formes verbales figées en français du Québec", *Langages*, n° 90, Larousse Paris.

------ (1988b) "Note sur les expressions figées à un complément" *Linguistica Communicatio*, Université Sidi Mohamed Ben Abdellah, Fès, Maroc.

ROBERT, Paul et coll. (1967) *Dictionnaire alphabétique et analogique de la langue française. Le Petit Robert*, Société du Nouveau Littré, Le Robert, Paris.

SHIATY, A.E. et coll. (1988) *Dictionnaire du français PLUS*, Centre éducatif et culturel, Montréal.

TABLE 32H-Q

(Extrait)

N 0 = N hum	N 0 = N - hum	N 0 = V - n		N 0 V	N 1 V	N 0 V N 1 de combien	N 0 V N hum sur ce point	N 1 = V - n	N 1 = N - hum, concret	N 1 = N - hum, abstrait	N 0 être V - eux
+	-	-	ABRILLER	-	-	-	+	-	+	+	-
+	-	-	ACCOTER	-	-	-	+	-	-	+	-
+	-	-	BAQUER	-	-	+	+	-	-	-	-
+	-	-	BARBER	-	-	-	+	-	-	-	+
+	-	-	BARDASSER	+	-	-	+	-	+	-	+
+	-	-	BAVER	-	-	-	+	-	-	-	+
+	-	-	BLÂMER	-	-	-	+	-	-	+	-
+	-	+	BOSSER	+	-	-	+	-	-	-	+
+	-	-	BOUDER	+	-	-	+	-	-	+	+
+	-	-	BOUQUER	-	-	-	-	-	-	-	-
+	-	-	BOURASSER	+	-	-	+	-	-	-	+
+	-	-	BOURRER	-	-	-	+	-	-	-	-
+	-		BRIFFER	-	-	-	+	-	-	-	-
+	-	-	CANTER	-	+	-	-	-	-	-	-
+	-	-	CLAIRER	-	-	-	-	-	-	-	-
+	-	-	CÔXER	-	-	-	+	-	-	-	-
+	-	+	CÔXER	+	-	-	+	-	-	-	-
+	+	-	COULER	-	+	-	+	-	-	+	-
+	-	-	CROSSER	-	-	+	+	-	-	-	+
+	-	-	CROUSER	+	-		-	-	-	-	-
+	-	-	ENFARGER	-	-	-	-	-	-	-	-
+	-	-	ESPÊRER	+	-	-	-	-	-	-	-
+	-	-	ÊTAMPER	-	-	-	-	-	-	-	-
+	-	-	FOURRER	+	+	-	-	-	-	-	-
+	+	-	FOURRER	-	-	+	+	-	-	-	-
+	-	-	LÂCHER	-	-	-	-	-	-	-	+
+	-	+	MINOUCHER	-	-	-	-	-	-	-	+
-	+	-	MOPPER	-	-	-	-	-	-	-	-
+	-	-	NIAISER	-	-	-	+	-	-	-	-
+	+	-	PAQUETER	-	+	-	-	-	-	-	-
+	-	+	PEIGNER	-	-	-	-	+	+	-	-
+	-	-	PLACER	-	-	-	-	-	-	-	
+	-	-	PLANTER	-	+	-	-	-	-	-	-
+	-	-	REPLACER	-	-	-	-	-	-	-	-
+	-	-	TROSTER	-	-	-	+	-	+	+	-

TABLE CP1-Q

(Extrait)

N O = : N hum	N O = : N - hum	N O V	N 1 = : N O pc
+	- agir	- à Poss-O	- façon
+	- agir	- en	- sous-main
+	- aller	- à le	- balai
+	- aller	- à les	- bécosses
+	- aller	- en	- cour
+	- ambitionner	+ sur le	- pain bénit
-	+ arriver	+ dans le	- meilleures familles
+	- arriver	- en	- ville
+	- s'asseoir	+ en	- sauvage
+	- s'asseoir	- sur Poss-O	- steak
+	- s'atteler	- à le	- ouvrage
+	+ branler	+ dans le	- manche
+	+ changer	+ de	- poil
+	- chier	- dans Poss-O	- culottes
+	- chier	- sur le	- bacul
+	- coucher	- sur la	- corde à linge
+	- courir	- après le	- trouble
+	- courir	- après Poss-O	- souffle
+	- se cracher	- dans les	+ mains
+	- crever	- de	- faim
+	- crever	+ de	- chaleur
+	- crier	+ à le	- meurtre
+	- croupir	- dans Poss-O	- crasse
+	- se débrouiller	+ avec les	- moyens du bord
+	- se déguiser	- en	- coup de vent
+	- s'endormir	+ sur la	- job
+	- s'enfarger	- dans les	- fleurs de le tapis
+	- s'enfarger	- dans Poss-O	- mots
+	- s'enfarger	- dans Poss-O	- ombrage
+	- s'enfarger	- dans Poss-O	- histoire
+	- s'enfarger	- dans Poss-O	- mensonges
+	- s'enfarger	- dans un	- grain de sable
+	- s'ennuyer	- de Poss-O	- mère
+	- se fendre	- en	- quatre
+	- fesser	- dans le	- tas
-	+ finir	- par une	- basse-messe
+	- flâner	+ sur la	- job
+	- foncer	+ dans le	- tas
+	- grimper	- dans les	- rideaux

FRANA: LOGICIEL FRANÇAIS DE GÉNÉRATION DE TEXTES

Chantal Contant
Université de Montréal

Le système FRANA est un logiciel qui génère, de façon entièrement automatique, des rapports portant sur l'activité boursière de New York. Dans Contant (1985) et Contant (1988), nous décrivons les caractéristiques du sous-langage boursier que nous avons observées par l'étude d'un corpus composé de 119 rapports publiés dans des journaux francophones. Dans ce présent article, nous nous attarderons plutôt à décrire le système de génération en soi.

L'outil utilisé pour la programmation de FRANA est le langage de production OPS-5 (Forgy 1981). Par la suite, Sylvie Giroux, Evelyne Millien et Michel Boyer de l'Université de Montréal l'ont réécrit en PROLOG, améliorant ainsi la vitesse de rédaction en passant de quelques minutes à quelques secondes pour un texte de 3 paragraphes.

Le traitement linguistique dans FRANA se fait à un macro-niveau, c'est-à-dire qu'on combine des syntagmes (ex: SN + Verbe + SP) pour constituer des propositions qui, à leur tour, se combinent en phrases. Une phrase peut contenir 1 à 3 propositions, chaque proposition reflétant un « message à exprimer ». Malgré ce traitement au niveau syntagmatique, certains accords demeurent nécessaires: l'accord du verbe avec son sujet, l'accord des participes passés, l'accord des adjectifs attributs, et celui des adjectifs possessifs qui sont à l'intérieur du complément mais qui doivent s'accorder avec le sujet.

Le système FRANA a eté élaboré à la suite des travaux de Karen Kukich (1983) qui a conçu le système initial ANA, qui génère des rapports boursiers anglais. ANA est constitué de quatre modules indépendants mais séquentiels. Le premier module est un **générateur de faits** écrit en langage C qui prend pour entrée les données numériques de la bourse de New-York et qui les transforme en une série de faits sous forme de banque traitable par OPS-5

Le deuxième module est écrit en OPS-5 et prend pour entrée la banque de données sortant du module 1 et décide **quoi dire** à partir de ces faits. C'est la **sémantique** du système. A l'aide de ses 142 règles de productions (de type SI -> ALORS), le module 2 produit une dizaine de **messages** à exprimer.

Ces messages sont ensuite envoyés au module 3. Ce module s'occupe de l'**organisation du discours** et choisit l'ordre dans lequel les messages seront exprimés.

Enfin un quatrième module, le module **linguistique**, procède aux choix lexicaux (ou plutôt syntagmatiques) et aux choix des formes syntaxiques. Il fait ensuite les accords et imprime le texte. Le module 2 a donc décidé QUOI dire, le module 3 QUAND le dire et le module 4 COMMENT le dire. Le module 4 de K. Kukich contient 110 règles de production et 450 entrées syntagmatiques anglaises.

Etant donnée la modularité de ce système initial anglais, nous avons décidé de réaliser un module linguistique français pouvant se substituer au module 4 de ANA, donnant ainsi naissance à un générateur de rapports boursiers français: FRANA. Puisque les deux systèmes ont une partie

commune (les trois premiers modules), le contenu sémantique des rapports boursiers français et anglais est identique et, de ce fait, les textes générés véhiculent la même information dans les deux langues. Mais ces textes sont indépendants au niveau des choix syntaxiques et lexicaux. Il ne s'agit donc pas de traduction mais de génération bilingue. Le module linguistique français contient 143 règles, 371 entrées syntagmatiques et une table de conjugaison de 75 verbes.

Voyons d'abord un exemple de message sémantique qui sert d'entrée au quatrième module:

```
(make message ^pry 2 ^repdate 04/20 ^top GENMKT
^subtop MKTSTAT ^subjclass MKT ^dir up ^deg great
^tim close ^sco broad)
```

qui pourrait s'exprimer sous la forme: « le marché des valeurs boursières a clôturé en forte hausse ».

Voyons maintenant un exemple de rapport boursier rédigé par FRANA, puis nous expliquerons quelques étapes qui nous permettent de passer de l'entrée du module 4 (messages) à la sortie (texte).

Rapport boursier

Jeudi, le 21 avril 1983

le marché des actions est demeuré soutenu tout au long de la journée hier, à Wall Street, où les titres ont fermé sur une forte hausse. L'activité a été fébrile.

après avoir enregistré son plus gros gain en fin d'après-midi, l'indice Dow Jones des industrielles a clôturé avec une avance de 16,93 points, à 1191.47. l'indice des transports s'est établi à 531.53, en hausse de 6.1? points et celui des services publics a inscrit un gain de 1.44 points, à 128.05.

le volume a été de 110.2 millions d'actions échangées au regard de 91.2 millions le jour précédent, alors que les titres à la hausse enterraient ceux à la baisse par 1175 contre 499.

Les principales étapes de la rédaction d'une proposition sont les suivantes: choisir un prédicat (verbe + complément) associé à un message à exprimer; choisir la forme syntaxique à utiliser; choisir le sujet du prédicat; procéder à la conjugaison du verbe et aux accords morphologiques s'il y a lieu; écrire la proposition. Autour de ces étapes gravite un bon nombre d'éléments de contrôle qui vérifient la pertinence de mettre une virgule ou un point, de changer de paragraphe, d'alterner entre phrases longu s et phrases courtes, d'insérer un syntagme adverbial, etc...

La première étape consiste donc à choisir un prédicat associé à un message à exprimer. Voici un exemple d'entrée syntagmatique prédicative.

```
(make phraselex ^ptype pred ^top DOW ^subtop DOWPT ^subjclass
      DOW
^classespec DOW ^vardeg x ^subsubtop FIRST ^len 15 ^rand 10
^dir up ^deg great ^tim first
^verbe enregistrer ^sppre !
^predrem une avance de plus de <x> points en tout début de séance)
(make sppré ^clé 1 ^terme en hausse de <x> points à l'ouverture
      ^len 9)
```

On remarque dans cette entrée plusieurs attributs sémantiques correspondant à la signification de ce prédicat (topique, direction, degré...). Ce sont ces attributs qui doivent correspondre avec ceux du message sémantique à exprimer. Lorsque différentes expressions synonymes sont disponibles, un choix est fait au hasard mais la probabilité pour une entrée d'être choisie dépend aussi de sa fréquence d'utilisation (basée sur l'étude de notre corpus).

L'étape suivante permet de choisir la forme syntaxique à utiliser. Celle-ci dépend entre autres de l'état du système (ex: est-il au début d'une phrase ou a-t-il déjà rédigé une première proposition?). Par défaut, tout prédicat peut s'exprimer en une phrase simple (proposition indépendante). Mais il existe d'autres formes syntaxiques disponibles. Parfois, un verbe et son complément peuvent être remplacés par un syntagme prépositionnel ou adjectival. Dans l'entrée prédicative ci-haut, on constate que l'utilisation d'un syntagme prépositionnel « en hausse de x points à l'ouverture, l'indice... » pourrait remplacer la proposition indépendante « l'indice a enregistré une avance de plus de x points en tout début de séance » si tel était le choix syntaxique du système. Voici la liste des choix syntaxiques possibles dans FRANA:

Variantes avec le verbe

1- Phrase simple (indépendante ou principale)
ex: le marché a clôturé à la baisse

2- Proposition coordonnée (et - mais)
ex: ... et (mais) le marché a clôturé à la baisse.

3- Subordonnée (alors que - tandis que)
ex: ... alors que (tandis que) le marché clôturait à la baisse.

4- Complément de temps
a) antéposé ex: après avoir connu une baisse, le marché...
b) postposé ex: ..., avant de connaître une forte baisse en après-midi

5- Infinitive avec POUR
ex: ... pour clôturer en hausse marquée

6- Relative avec OU
ex: à la Bourse de New York, où l'activité a été modérée.

Variantes sans le verbe

7- Syntagmes prépositionnels
a) antéposé ex: en baisse à l'ouverture, ...
b) postposé ex: ... à l'issue d'une séance mouvementée.

8- Nominalisation avec préposition
ex: après un grand mouvement de baisse initial, ...

9- Epithète détachée
ex: faible à l'ouverture, ...

Autres

10- Adverbes

ex: Au total, ...
En fin de journée, ...
... hier à la Bourse de New York.

Nous voici maintenant à l'étape qui consiste à choisir un sujet pour accompagner notre prédicat. Si par exemple l'entrée prédicative demande la classe sujet (subjclass) DOW, alors il faut choisir parmi les entrées syntagmatiques sujets celle qui est de la classe sujet DOW. S'il y a plusieurs entrées disponibles, deux facteurs déterminent le choix du système: la fréquence d'utilisation observée dans notre corpus ainsi que le niveau d'hyponymie. Plus on avance dans la rédaction du texte, moins on a à utiliser des termes spécifiques concernant le sujet du discours. Voici un exemple d'entrée sujet:

(make phraselex ^ptype subj ^top GENMKT ^subjclass MKT
^classespec MAR ^subjterm le marché new-yorkais
^subjnumber sing ^subjgenre masc
^rand 10 ^subjhypolev 3 ^len 6 ^usage 0)

Les entrées sujets sont des syntagmes nominaux et contiennent donc des informations relatives au genre et au nombre de ceux-ci. Ces informations sont utiles pour conjuguer le verbe et pour accorder certains mots à l'intérieur du reste du prédicat (predrem) lorsque nécessaire. La conjugaison du verbe se fait à l'aide d'une table de conjugaison en fonction du choix syntaxique qui détermine le temps du verbe (passé composé, imparfait, infinitif présent ou passé), et en fonction du nombre du sujet et également de son genre lorsqu'il y a présence d'un participe passé conjugué avec le verbe « être » (ex: s'est ou se sont redressé-e-s). Lorsque le verbe de l'entrée prédicative est un verbe d'état, certains mots attributs doivent être accordés avec le sujet. Pour réaliser cet accord nous avons dû ajouter, dans le module français, des variables morphologiques. En anglais, un tel problème d'accord ne se posait pas. Voici un exemple d'utilisation d'une variable morphologique:

Prédicat: ^verbe demeurer
^predrem <mot> tout au long de la journée
^choixms irrégulier ^choixmp irréguliers
^choixfs irrégulière ^choixfp irrégulières

Lors de l'impression, la variable <mot> sera remplacée par la valeur appropriée, en fonction du genre et du nombre du sujet.

Revenons au choix du sujet. Lorsque la forme syntaxique « complément de temps antéposé » ou « épithète détachée » est choisie, le sujet n'apparaît pas tout de suite en surface. Il ne sera exprimé que dans la proposition suivante.

Ex: Après être demeuré(-e-s) en baisse toute la matinée, l'indice (la Bourse, les titres...)
Ex: Faible(s) à l'ouverture, l'indice (les titres)...

Pourtant, il faut faire dès maintenant le choix du syntagme nominal sujet afin d'accorder le participe passé ou l'adjectif épithète. Il faut d'abord s'assurer que ce syntagme nominal est

compatible avec la classe sujet de la prochaine proposition puisqu'il sera également sujet de cette seconde proposition. De plus, il faut conserver cette entrée sujet en mémoire car elle ne sera imprimée que dans la seconde proposition.

Inversement, lorsque FRANA choisit la forme syntaxique «infinitive avec POUR» ou «complément de temps postposé», il faut s'assurer que le sujet de la proposition précédente est identique au sujet du prédicat actuel et il faut même récupérer l'information sur le genre et le nombre de cette entrée sujet (déjà imprimée) si on veut bien accorder les variables morphologiques. C'est ce que FRANA fait avec succès. Ex: «l'indice a..., avant d'inscrire *son* meilleur gain».

Nous conclurons en disant que ANA et FRANA sont des logiciels efficaces et entièrement automatisés, ne nécessitant pas de révision humaine. Les rapports produits sont linguistiquement bien formés et décrivent de façon cohérente des faits réels. Les structures syntaxiques sont adéquates et les termes employés sont justes, reflétant par leur fréquence le style des textes rédigés manuellement.

Ce rapport a été subventionné en partie par le CRSHC et par le Fonds F.C.A.R. Merci à Karen Kukich, Richard Kittredge, Guy Lapalme et Michel Boyer pour leur support respectif.

Bibliographie

CONTANT, Chantal (1988) « Génération automatique de rapports boursiers français et anglais » dans *Revue québécoise de linguistique*, vol. 17, no 1, Montréal, p. 197-222.

------ (1985) *Génération automatique de texte: Application au sous-langage boursier français*, mémoire de maîtrise, Université de Montréal.

CONTANT, Chantal et M.-H. GAUTHIER (1983) *Manipulation du corpus, grammaires de textes, paraphrases*, projet de recherche sur les sous-langages, Département de linguistique et philologie, Université de Montréal.

DANLOS, Laurence (1985) *Génération automatique de textes en langues naturelles*, Masson.

Journal Le DEVOIR, rapport boursier dans les pages économiques du 18 octobre 1983 au 10 décembre 1983 (Bourses de New York, Toronto et Montréal).

FORGY, C.L. (1981) *OPS-5 User's manuel*, Department of Computer Science, Carnegie-Mellon University.

KITTREDGE, Richard (1982) "Variation and Homogeneity of Sublanguages" dans *Sublanguage: Studies of Language in Restricted Semantic Domains*, Walter de Gruyter, p. 107-137.

KUKICH, Karen (1983) *Knowledged-based Report Generation: A Knowledge-Engeneering Approach to Natural Language Generation*, thèse de Ph.D., Department of Information Science, University of Pittsburg.

HÉTÉROGÉNÉITÉ ET INTRICATION DANS LES ÉNONCÉS CONSÉQUENCES POUR LE PARSAGE*

Mazurdin Jean-Marie
SLD, INaLF (CNRS)

0. INTRODUCTION.

Un colloque consacré à « la description des langues naturelles en vue d'applications informatiques » adresse une question spécifique à la théorie et à la pratique du parsage syntaxique: le problème de la couverture des parseurs. Ce problème émerge, par ailleurs, sous le fait d'une pression externe: la demande de parseurs robustes liée au (projet de) développement d'interface grand public en langue naturelle, ou de divers types de traitement de données textuelles. Or, la théorie et la pratique du parsage se sont développées « dans un cocon »: les parseurs de référence peuvent faire s'appuyer sur des théories syntaxiques sophistiquées, mais ils ne s'attaquent généralement qu'à quelques aspects de la langue ou des énoncés, et sont appliqués en miroir sur un corpus de phrases généralement tirées des articles de linguistique théorique. Ou bien, ils sont développés sur un langage restreint dans le cadre d'une application particulière.[1] Il est clair, dans l'état actuel du domaine du parsage syntaxique, que l'accroissement de la couverture n'est pas un problème simplement quantitatif: accroître le nombre de règles ou le nombre d'entrées lexicales dans le dictionnaire associé au parseur. C'est un problème théorique pour la linguistique et pour le parsage.

J'admets la définition suivante de parsage syntaxique (je me limite au parsage syntaxique d'énoncés écrits): *parser* consiste à reconnaître dans une suite de mots (la séquence d'entrée) des dispositions d'entités. Ces dispositions sont représentées afin d'exhiber les propriétés structurales, positionnelles et interprétatives des entités reconnues et de leurs relations. Le nombre, la syntaxe et le vocabulaire des représentations appariées aux suites d'entrée (arborescence ou non, catégorie monadique ou composée, utilisation d'indice ou non, ...) varient selon les théories syntaxiques. Il est généralement admis maintenant, en linguistique et en intelligence artificielle, que les propriétés représentées feraient la voie à l'interprétation de la séquence d'entrée; ou à minima, que les représentations syntaxiques seront manipulables par des procédures intelligentes de traitement de données en langue naturelle. Un parseur syntaxique particulier dispose, donc, en entrée de deux données: (i) une suite de mots et (ii) un savoir syntaxique. Il fournit en sortie une ou plusieurs représentations. La couverture d'un parseur sera définie en fonction de son savoir syntaxique: plus ce savoir sera important en compréhension et en extension, plus nombreuses seront les entités et les dispositions d'entités qu'il pourra parser et plus expressives seront les représentations qu'il en fournira.

On reconnait dans des langues comme le français ou l'anglais la place centrale d'un type d'entité: le syntagme. La donnée est simple: plusieurs unités lexicales forment une unité (un tout) pour une règle ou pour un processus. Par exemple, dans la phrase *La fille de la concierge est*

---

*Ce travail s'inscrit dans le cadre du contrat de coopération Franco québécois « Conception et applications d'un analyseur lexico-syntaxique du français (ALSF) » (Centre d'ATO, UQAM et INaLF/LISH, CNRS). Je remercie la Commission Permanente Franco-québécoise pour le soutien financier qu'elle apporte à ce projet.

[1] A la notable exception des grammaires en chaînes, même si Salkoff (1973, 1979) limite son entreprise au discours scientifique.

venue », le sujet de la phrase (S) (du syntagme verbal (SV) ou du verbe, peu importe ici) est *la fille de la concierge* et non le nom *fille* pris isolément (quel que soit la définition ou la place que l'on donne à la notion de fonction et l'assignation des fonctions). La linguistique contemporaine s'est développée sur une hypothèse (issue de la description distributionnelle): syntagmes et dispositions de syntagmes sont hiérarchiquement organisés, et sur une représentation de cette hypothèse sous forme de configurations arborescentes étiquetées de catégories. Sur cette hypothèse s'est élaborée une mise en forme de la syntaxe permettant de générer entités et configurations par des règles formellement identiques: des règles syntagmatiques. Or, ce modèle, qui semblait acquis, est en débat dans les développements actuels des différentes théories syntaxiques: la syntaxe d'inspiration chomskyenne tend à l'élimination des règles syntagmatiques, alors que la grammaire syntagmatique généralisée (GSG) revient au noyau de départ (pour me limiter à ces deux théories).

Les parseurs pouvaient être dessinés assez simplement dans la configuration de départ. Ils disposent en entrée de deux données:

- (i) une suite de mots considérés sous l'angle de leur catégorie grammaticale,
- (ii) une grammaire syntagmatique, généralement augmentée afin de pouvoir traiter la structure de surface (en particulier, les écarts entre structure de surface et structure profonde pour les parseurs se référant aux théories transformationnelles).[2] Un parseur peut, alors, être défini comme un interprète qui applique des règles syntagmatiques sur les suites d'entrée pour les convertir en une représentation arborescente étiquetée, et le parsage comme l'ensemble des algorithmes ou des techniques réglant cette application.

Je propose trois problèmes auxquels un parseur, exposé au tout-venant des productions langagières, sera immanquablement confronté. Au travers de leur description (nécessairement partielle, je les considère comme prototypiques), je montrerai comment leur prise en compte peut amener à modifier le dessin général de parseur exposé en introduction. On verra qu'ils posent à chaque fois des problèmes de représentation à la théorie syntaxique; plus précisément dans cet exposé à la représentation en termes de règles syntagmatiques. J'ai privilégié les règles syntagmatiques dans le format X-barre parce qu'étant les plus contraintes, elles permettent de poser les questions plus radicalement.[3]

1. INTRICATION LEXICALE/SYNTAGMATIQUE

Un parseur à couverture maximale va devoir reconnaître et représenter des entités subsyntagmatiques: des suites de mots traitées par les règles syntagmatiques comme des unités. Ce sont des unités lexicales polylexicales; a priori, elles devraient appartenir aussi bien aux catégories majeures qu'aux catégories mineures.

---

[2] De fait, les parseurs sont construits sur l'intuition des linguistes et que Milner énonce ainsi à propos de « l'inutilité des transformations »: « L'intuition fondamentale n'a pas été remise en cause: on suppose toujours qu'entre la configuration immédiatement observable et la structure proprement linguistique, il peut y avoir discrépance, mais on ajoute que la configuration observable contient toujours les indices suffisants qui permettent de restituer la structure » (Milner, 1985a :50).

[3] Je me focalise artificiellement sur les règles dans cet exposé; je renvoie à Morin (sd) pour une défense et illustration des catégories composées en théorie du parsage.

1.1. Exemples d'unités polylexicales mineures

1. a. **De** l'eau suintait de la fontaine.
   a'. L'eau suintait de la fontaine.

2. b. Il a fait **à peine** de linguistique.
   b'. Il a fait **beaucoup** de linguistique.

3. c. Il fume **dans les** dix cigarettes par jour.
   c'. Il fume **environ** dix cigarettes par jour.

Un parseur doit pouvoir traiter les suites soulignées en (a, b et c) comme des unités structuralement ou fonctionnellement équivalentes aux unités simples soulignées de (a', b' et c'): des unités réalisant la position de spécifieur en (1) et (2); une unité "modifiant" le numéral en (3)[4].

On ne peut pas toujours effectuer indépendamment de l'analyse syntaxique la reconnaissance et le traitement de ces unités: il faut, en effet, leur assigner une identité catégorielle qui ne découle pas de la composition des traits des unités qui les composent: ni **à** ni peine pris isolément ne peuvent réaliser la position de spécifieur dans un SN. Si on peut imaginer traiter l'unité **à peine** dans une phase de pré-parsage (dans une procédure chargée de reconnaître des unités polylexicales et de leur assigner les traits qui les identifient), il n'en va pas de même pour **de le** ou **dans les**: il faudrait singulièrement augmenter la phase de pré-parsage pour qu'elle ne traite pas, en (1), la suite *de la* dans *de la fontaine* comme de *l'* dans *de l'eau*. La reconnaissance de ces unités est inséparable de l'analyse syntaxique: le partitif (par exemple) n'apparaît que dans un groupe nominal régi directement par un verbe, une préposition ou en position sujet (moins fréquemment).[4] Les parseurs syntaxiques apparaissent, donc, condamnés à inclure des opérations lexicales dans le parsage des unités syntagmatiques et ces opérations ne sont pas concevables comme l'application de règles générales. Elles requièrent un dictionnaire où sont données ces unités.

1.2. Les unités polylexicales majeures

Soit les exemples suivants:

(4). a. La **bibliothèque usager** est au fond du couloir.
   b. On a réparé le **moteur diesel à quatre temps** responsable de la panne.
   c. On a réparé le **moteur** responsable de la panne.
   d. Je n'ai pas trouvé une **bibliothèque** potable dans cette fac.

Admettons un parseur disposant de règles syntagmatiques comme (5) ci-dessous. Leur formulation précise importe peu ici.

(5) n2 ---> spécifieur n1 complément(s).
   n1 ---> ... n0 ...

---

[4] Certains cribles de numéraux sont, de plus, discontinus: Il fume de dix à vingt / **entre** dix **et** vingt cigarettes par jour (Gross, 1975). On notera, en passant et plus généralement, que la représentation X-barre classique ne dit rien des suites « crible + quantifiant » du type illustré par (3) ou comme: Il a fait assez peu de linguistique, ...

Ce parseur analysera les suites *le moteur responsable de la panne* et *une bibliothèque potable* comme des groupes nominaux, mais ne réussira pas à analyser correctement comme un (et un seul) n2 les suites *le moteur diesel à quatre temps responsable de la panne* en (4.a) et *la bibliothèque usager* en (4.b). La description du groupe nominal encapsulée dans les règles de (5) est insuffisante pour traiter ces tours.

1.2.1. *Première description*

On peut faire une première hypothèse. De la même manière qu'il existe des unités polylexicales mineures, il existe des unités polylexicales majeures: des noms composés. On admet que *moteur diesel à quatre temps* et *bibliothèque usager* forment de telles unités. Les règles (5) ne sont pas remises en question, si elles peuvent voir ces suites comme ne comptant que pour une seule entité et qu'elles les traitent comme la réalisation d'une tête lexicale de n2 à l'instar des unités simples (*moteur* en (4.c) ou *bibliothèque* en (4.d)).

Cette hypothèse revient à admettre que n0 n'est pas une catégorie terminale: elle peut dominer une combinaison d'items analysable en plusieurs catégories. Cette modification implique, plus généralement, qu'un groupe nominal peut « avoir plus de niveaux » que ceux qui sont stipulés dans les règles de (5). Ou bien, que l'analyse hiérarchisée du groupe nominal français n'est pas adéquate.

Le parsage des suites soulignées de (4.a,b) poserait un problème identique au parsage des unités polylexicales de(1)-(3): on doit rendre le parseur capable de reconnaître et de représenter des entités syntagmatiques et des entités sub-syntagmatiques (lexicales). Dans les deux cas, le recours à un dictionnaire listant les unités est incontournable. On notera, par ailleurs, que le pré-traitement de ces entités (l'image de celui qui est envisageable pour une unité comme *à peine*) est presque toujours impossible.

1.2.2. *Discussion de la première hypothèse*

Considérons le projet de lister les unités polylexicales majeures. Si le projet de répertorier les unités polylexicales mineures est possible (elles forment une liste quasiment fermée), on s'attend à ce que la liste des unités polylexicales majeures soit extrêmement longue et ouverte par créativité lexicale (au même titre que la liste des unités simples de même catégorie). Ce que confirment les recensements terminologiques.

A considérer ces recensements, on constate deux phénomènes:

- les mots composés « ressemblent » aux syntagmes: ils semblent formés sur le même répertoire de formes que les syntagmes nominaux « normaux ».[5] Le nom *réacteur atomique à neutrons rapides* ne se distingue pas formellement du n1 libre *fille blonde à tresses vertes*.

- le jugement catégorisant un fragment de groupe nominal comme un mot composé est fluctuant et dépendant des univers d'emploi. Je renvoie sur ce point à G. Gross 1988. La catégorisation d'une suite de mots comme formant un nom composé implique des considérations prag-

[5] De fait, il faut distinguer les noms composés qui ressemblent à des groupes nominaux et ceux qui présentent une composition différente: un sans-culotte, un chez-soi, un rendez-vous, etc (voir Gross G 1988: 61). Ces derniers doivent, sans doute, être listés.

matiques, plus que syntaxiques, portant sur les dénominations dans une sphère d'activité ou un univers de discours donné (univers de discours qui peut s'étendre aux frontières de la langue dans le cas d'unités comme *pomme de terre* ou *chemin de fer*).

1.2.3. *Deuxième hypothèse*

Admettons le principe formulé par Milner (1985b :15): « si des multiplicités linguistiques apparaissent comme des unités, c'est que des processus bien définis dans la grammaire les traitent comme telles. Autrement dit, toute proposition de la grammaire, toute règle, toute opération est en droit de définir un type d'unité ». Selon ce principe, une suite de mots peut former une unité pour les processus référentiels (un nom composé) tout en relevant pour leur composition d'un autre principe de groupement (les règles « normales » de composition du groupe nominal).

Il n'en demeure pas moins que les règles de (5) ne permettent pas d'analyser les suites de (4.a-b). Il faut ici étendre le paradigme des groupes nominaux considérés pour la description du groupe nominal.

Soit les groupes nominaux:

(6) a. Le **président Mitterand** inaugure une bibliothèque.
b. Paul a attrapé un **papillon toto vulgaris**.
c. Le **mot chaise** n'a pas cinq lettres.
d. **L'agence de presse officielle britannique Tartempion** a rapporté que ....

Les GN soulignés en (6) manifestent un même principe d'organisation que ceux de (4). Je l'appelle parataxe: adjonction sans discontinuité. On peut représenter ces structures comme (7) ci-dessous:

(7) [n0 N N] : [n0 moteur diesel]
[nmax. Nmax. Nmax.]: [nmax [n2 le président] [n? Mitterand]]

En effet, la parataxe ne semble pas s'opérer au même niveau: au niveau lexical pour *moteur diesel* et au niveau maximum pour le *président Mitterand*.[6]

Sans que cela constitue une argumentation, on remarque que si on analyse *moteur diesel* et *président Mitterand* comme une parataxe au niveau maximum, on ne peut pas analyser *moteur diesel responsable de la panne* en (4.b), sinon admettre le résultat absurde où *diesel* serait pris comme tête d'un SN régissant le SA *responsable de la panne*. A l'inverse, si on analyse *président Mitterand* et *moteur diesel* comme une parataxe au niveau minimum, on ne peut plus analyser le groupe nominal souligné de (6.d).

Admettons l'hypothèse de la parataxe. On peut classer les groupes nominaux examinés dans ce paragraphe comme des structures exceptionnelles, des tours périphériques (il n'en demeure pas moins qu'un parseur à couverture maximale doit être capable de les traiter). Ce jugement ne peut être maintenu que s'ils sont isolés. Or, il est, à cet égard, remarquable que le groupe nominal soit par excellence le lieu de la parataxe (ce trait le distingue des autres constituants majeurs de S). Ce qu'illustrent les quelques exemples suivants:

[6] Il faut mettre à part le cas de la mention (6.c): elle semble appartenir au système du nom propre.

8. a. **La belle jeune fille.**
   b. Un tricot **de laine à encolure de soie de ma grand-mère.**
   c. Rien **d'autre d'intéressant.**

On notera seulement quelques propriétés des éléments soulignés en (8):

- ils ont la même identité catégorielle (adjectif, nom, SN, SA,..)
- ils ne relèvent pas directement du système de rection (voir ci-dessous) de la tête lexicale du syntagme,
- ils se différencient selon plusieurs paramètres: matériel en (8.b) [différence de préposition: **en** *laine*, **à** *encolure*], classe syntaxique en (8.c) [en admettant que (8.c) est un SN. Voir Huot 1981 sur ce point], classe sémantique en (8.a), statut référentiel en (8.b) [différence entre *de laine à encolure ...* vs *de ma grand-mère*] ou (6) ci-dessus [GN « normal » vs GN régi par un nom propre].

Les faits sont complexes et ils n'ont pas été, à ma connaissance, décrits formellement.[7] Je ne l'entreprends pas ici. On retiendra:

- que les structures internes du GN sont ouvertes à des phénomènes de parataxe (plus formellement d'adjonction) qu'il faut explorer avant de maintenir une analyse hiérarchisée de ce constituant. Sur ce point, une comparaison s'impose avec les autres structures plates dans la langue: les groupes coordonnés.[8] Plus généralement, il s'agit de l'intrication d'une organisation hiérarchique et des développements horizontaux qui s'y greffent. Une représentation cohérente doit être donnée de ces deux modes de structuration.
- que les éléments en parataxe sont soumis à une contrainte de distinction. Il reste à établir le rapport entre cette contrainte (portant sur des éléments qui ne sont le support d'aucune fonction) et les autres principes de distinction posés par ailleurs: disjonction référentielle et non-redondance fonctionnelle (Milner, 1981). Si on admet que la distinctivité est une des propriétés des unités linguistiques, nous tenons là un indice que les tours considérés dans ce paragraphe forment bien, au regard de ce principe, des unités.

On notera que la deuxième hypothèse est cohérente avec la description des phénomènes de polylexicalité affectant d'autres constituants, par exemple le groupe verbal. On y retrouve la même situation: des structures régulières et des interprétations particulières.[9] Je ne peux développer ici la critique de l'interprétation sémantique dans les formes de la compositionalité. Le pro-

---

[7] Petite note: le développement de la linguistique contemporaine relève plus de l'éclatement en factions ou en chapelles que de la communauté d'échanges des résultats et des hypothèses. Résultat: une situation d'éclatement bibliographique, où il est très difficile d'accéder aux travaux qui ne sont pas cités dans les bibliographies des ouvrages se situant en dehors de son horizon intellectuel, institutionnel ou théorique.

[8] On peut coordonner à tous les niveaux de n2. Coordination sous n0: l'**académicien et ministre** a encore frappé; coordination sous n1: beaucoup **d'étudiants et de professeurs** (Milner 1978), coordination sous n2: **les femmes et les enfants**. Il semble qu'il n'y ait pas de parataxe au niveau n1. Par ailleurs, la parataxe au niveau maximum semble limitée au nom propre, aux appellatifs et, sous réserve, à la mention.

[9] Ce que confirme l'étude de Gross (1988: 22): « les phrases figées (...) s'analysent pratiquement toutes de façon systématiquement régulière. Les règles qu'elles subissent sont exactement les règles de la syntaxe des phrases libres et ce, aussi bien pour leurs parties libres que pour leur parties figées». Voir également Greciano 1983.

blème demeure, néanmoins, dans la perspective d'un traitement sémantique des représentations syntaxiques, d'imaginer les formes de représentation et de stockage des éléments constitutifs des contenus « figés » ou métaphoriques convoqués dans l'interprétation de certaines combinaisons de signifiants.

Enfin, la seconde hypothèse implique que l'on distingue la polylexicalité des unités mineures et des unités majeures. La polylexicalité mineure demande la confection de répertoire et des opérations de parsage particulières (appariement de la chaîne de mots parsés et de la suite stockée dans la base lexicale). La polylexicalité majeure est d'abord le symptôme d'un défaut de description et de formalisation du GN.

La voie de recherche esquissée dans la deuxime hypothèse déplace le problème de la complexité du parsage d'un constituant comme le groupe nominal français: ce n'est pas tant la composition nominale qui est problématique que la syntaxe même du groupe nominal. Les noms composés ne feraient qu'exploiter les virtualités structurales du groupe nominal. Le problème de la couverture requiert donc bien que l'on augmente la couverture de la théorie syntaxique: empiriquement et formellement.

2. INTRICATION SYNTAGMATIQUE

Les grammaires syntagmatiques captent l'organisation de plusieurs unités lexicales en groupe: les syntagmes. A propos de cette organisation, elles font deux hypothèses particulières (qui ne découlent pas nécessairement de l'intuition de départ):

- l'organisation des groupes est exprimable en termes de composition catégorielle: une phrase est constituée d'un SN et d'un SV, un SN est constitué d'un déterminant et d'un nom, etc.,
- les catégories constitutives d'un groupe sont hiérarchisées: n0 est inclus dans n1, n1 est inclus dans n2, etc.[10]

On a vu, au paragraphe précédent, que ces différentes hypothèses, représentées de façon compacte par les règles syntagmatiques, devaient être disjointes et soumises à l'épreuve de la description de tours qui ne sont généralement pas considerés.

Les parseurs font une assomption supplémentaire: les règles syntagmatiques décrivent les dispositions de surface des mots dans la chaîne parlée/écrite. Le rapport entre les configurations syntagmatiques et leur projection dans la chaîne n'est pourtant pas direct: deux faits s'y opposent. Le premier est bien connu: l'ordre des constituants est variable (ce qui a entraîné l'augmentation du formalisme des grammaires syntagmatiques: transformation, chaîne dans les grammaires d'inspiration chomskyenne, méta-règle et distinction entre règle de dominance et règle de précédence dans la GSG). Je laisse ce point de côté.

Le second l'est moins: un groupe peut être discontinu par insertion d'autres groupes en son intérieur.

---

[10] Je reprends ici Milner 1985b. Il y a ici un problème de fond: est-ce que n2 est inclus dans S de la même manière que n1 est inclus dans n2? On peut en douter. On sait, par ailleurs, que l'on peut douter de la catégorie « groupe verbal » (voir sur ce point, Gross 1975, Milner 1985b, Rouveret-Vergnaud 1980). D'où la valse-hésitation à propos de la définition de la catégorie S dans la linguistique d'inspiration chomskyenne. L'approche projective (introduite au paragraphe 3) n'implique pas que le principe de cohésion dans les différentes entités syntagmatiques soit identique; mais il faut, alors, donner une définition autre que structurale à la catégorie « tête lexicale » d'un syntagme.

2.1. Exemples d'insertion intra-syntagmatique

En restant dans un cadre syntagmatique strict esquissé plus haut, l'insertion est possible à l'intérieur d'un groupe:

(9). a. Il a, **chose exceptionnelle**, revu Marie.
a'. Ils ont, **les uns et les autres**, fait de la linguistique.
b. La destruction, **illégale comme le tribunal l'a montré**, des pièces n'a pas été jugée.

Aux frontières entre groupes, c'est-à-dire dans l'intérieur d'un groupe enchâssant:

(10) a. Paul, **aux policiers**, a répondu non.
a'. Paul, **dès le lendemain**, a répondu non aux policiers.
b. Il opte **sur le champ** pour la liberté.
c. La destruction des pièces, **événement inadmissible**, n'a pas été jugée.

Aux frontières gauche et droite de S:

(11) .a. **A Paris, Marie, le directeur**, elle ne le rencontre pas.
b. Marie ne l'a pas rencontré à Paris, **le directeur**.
c. Il a dit que, **Paul, à Paris**, il ne le voyait pas.
d. La pomme que je lui ai donnée, **à Paul**, était empoisonnée.

Il faut admettre que les portions de chaînes syntagmatiques immunes à de telles insertions sont rares. Pour le français (et sans souci d'exhaustivité): clitique--verbe *(*il, aux policiers, dit .... ; *il le, aux policiers, dénonce.)*, déterminant--adjectif--nom *(*la, comme il dit, fille ; *la belle, vraiment très belle, fille.)* et peut-être, préposition--sn *(??il est venu dans, chose exceptionnelle, sa voiture. ; *il l'a donné à, dit-il, la fille.)*[11]

L'insertion intra-syntagmatique est donc un phénomène généralisé.

Admettons un parseur disposant d'une grammaire syntagmatique. Il se trouve face aux tours (9-11) dans une situation analogue à celle qui était introduite au premier paragraphe: il doit parser autre chose que ce qui est prévu par la description de la catégorie qu'il est en train d'appliquer à la suite de mots en entrée.

Or, rien dans le format des grammaires syntagmatiques ne permet de prévoir cette situation (la différence des entités que je qualifiais de sub-syntagmatiques au paragraphe 1): la solution ne peut être que technique. C'est une telle solution technique que propose Marcus avec ses « attention-shifting rules »: le parseur sursoit à une analyse pour en mener une autre,

---

[11] Ces insertions peuvent avoir lieu à l'intérieur de suites analysables comme des formants polylexicaux, comme le montrent les exemples de Piot 1988:

(iii). a. ... **de telle sorte**, disait-il à Marie, **que** tu ne partes pas ....
b. ... **avant**, disait-il à Marie, **que** tu ne partes ....
c. .. **dès lors**, bien sûr, **que** tu es d'accord avec lui....

L'analyse des suites de telle sorte que, avant que, etc comme des formants polylexicaux (les locutions conjonctives de la grammaire traditionnelle) n'est peut-être pas la meilleure. On peut les analyser, en suivant Emonds 1985, comme la réalisation régulière de la combinaison: « préposition S' ». Cette analyse implique que les prépositions ici en cause sous-catégorisent leurs compléments.

mais cette possibilité est limitée, dans l'implémentation de Parsifal, aux groupes nominaux et à une seule portion de la chaîne syntagmatique: le GV (précisément, et pour l'anglais, la combinaison « aux-verbe »).[12] Si l'on veut « armer » un parseur syntagmatique pour le rendre apte à parser correctement les énoncés (9-11), la technique du « détournement d'attention » devrait être généralisée sans que rien dans la théorie syntaxique ne vienne l'étayer: elle se trouve donc être ad hoc vis-à-vis de cette théorie.

2.2. Quelques éléments de description

Les grammaires syntagmatiques captent l'organisation des groupes en phrase (ou dans la phrase): les groupes qui y sont reconnus entrent dans les processus syntaxiques ou sémantiques qui ont pour domaine la phrase.

Certes, et c'est un point important, les groupes insérés sont des groupes régulièrement formés, mais ils ne relèvent pas de la phrase pour leur position de réalisation ou pour leur interprétation. Ils sont insérés « en surplus de S » (Cadiot-Fradin, 1988) et sont en relation paratactique avec un élément de S ou S en son entier.[13] Ils sont le support de différents processus interprétatifs que l'on désigne habituellement sous les chefs de: apposition, thématisation, repérage, support, reprise ("tail function" dans Dik 1981).

Je laisse de côté les différents traitements proposés en grammaire générative (je renvoie à Fradin (en prép.) pour leur critique). La description de ces ajouts montre qu'ils ne sont pas des phénomènes « sauvages »: ils sont contraints dans leur forme et leur interprétation. Deux points émergent de la description:

Les ajouts sont contraints par l'organisation de S. Par exemple, la position entre auxiliaire et verbe n'est pas ouverte à tout GN:

12. a. Paul a, le traître, revu Marie dès le lendemain.
    a'. Marie a, cette garce, revu Paul dès le lendemain.
    b. * Marie a, le voleur, revu dès le lendemain.
    b'. * Marie l'a, le voleur, revu dès le lendemain.
    c. Marie l'a revu dès le lendemain, le voleur.
    c'. Le voleur, Marie l'a revu dès le lendemain.

Elle n'est pas ouverte à un GN régi par le verbe (b) ou à un GN lié à un clitique; de plus, il semble devoir être interprétable comme un nom de qualité (Milner, 1978). Il est possible de décrire ainsi les contraintes pesant sur chacune des positions de la chaine permettant la réalisation d'un ajout à S.

Leur interprétation dépend de leur position dans la chaîne; cette position est caractérisée par deux paramètres: droite ou gauche et contiguïté/proximité d'un élément de S. Je ne peux pas entrer dans le détail de la description ici. Qu'il suffise ici de rappeler qu'un GN à la droite

---

[12] Marcus 1980: PARSIFAL est, à ma connaissance, le seul parseur à base syntagmatique qui pose ce problème. Il est, bien sûr, pleinement reconnu dans les parseurs en chaînes (Salkoff 1979).

[13] Seuf (8a): Paul, **aux policiers**, a répondu non , si on analyse aux policiers comme le complément sous-catégorisé de répondre. La description de ce tour pose des problèmes complexes. Je m'appuie dans ce paragraphe sur la description de Fradin 1988. Illustrant le problème de l'insertion, je n'opère pas de distinctions sur ces ajouts.

de la frontière de S ne peut pas être traité par les processus de thématisation ou qu'un GN à la droite de V ne peut être traité que comme une reprise (ou une apposition au GN contigu s'il n'est pas lié par une anaphore).[14]

La description de ces tours met en jeu deux types de structures: les structures relevant de la phrase et des structures relevant d'un système multiforme qu'on peut appeler énoncé. Elles sont dans la chaîne parlée/écrite entremêlées. On peut faire l'hypothèse que cet entremêlement est dû à la contrainte de proximité/contiguïté avec le terme avec lequel les éléments « en surplus » sont en relation ou les termes distingués de S (le sujet ou le verbe). Sous cette contrainte, les syntagmes ne sont pas des domaines fermés: ils sont disruptables. Il est à noter que les représentations arborescentes de « la structure de surface » ne peuvent pas représenter ces insertions.

2.3. Le problème pour un parseur

Un parseur exposé au tout-venant des énoncés sera affronté à des tours tels que (9-12). Pour le parsage (et pour un parseur particulier), la question peut être posée brutalement: va-t-on se décharger sur l'interpréteur du traitement de ces faits, ou bien va-t-on chercher à les représenter dans une syntaxe du français?

Je doute qu'on puisse les représenter par des opérations de déplacement, et la saisie des contraintes de réalisation dans la chaîne ne peut pas être spécifiée localement (cf. les règles de précédence linéaire de la GSG). En d'autres termes, il y a une lacune conceptuelle dans les théories syntaxiques actuelles.[15]

Il faut traiter de ces structures dans une syntaxe du français, pour deux raisons:

- elles sont contraintes syntaxiquement et ces contraintes interviennent dans leur interprétation,

- les processus interprétatifs dont elles sont le support ne sont pas purement pragmatiques (sans mettre en doute que des contenus pragmatiques sont mis en jeu dans ces processus). La théorie du parsage syntaxique a, aussi, l'ambition de représenter les aspects des énoncés qui feraient l'interprétation. Il est clair que ces tours sont déterminants pour la mise en place des rapports entre le contexte d'occurrence et l'interprétation du noyau phrastique (prédication et interprétation événementielle). Un parseur sémantique qui ne les prendrait pas en compte ne saisirait qu'une petite partie de ce qu'est interpréter un énoncé.

3. RÈGLES SYNTAGMATIQUES ET LEXIQUE

La remise en cause des règles syntagmatiques n'est pas effectuée à propos des constructions que j'ai introduites aux paragraphes précédents. Elle est liée au traitement de la souscatégorisation stricte des entités lexicales majeures (en particulier les verbes). Je développe brièvement quel est l'enjeu pour le parsage.

---

[14] Pour une étude de cas éclairante, voir la description de Franckel 1988 consacrée à l'interprétation du gérondif français, selon qu'il se trouve à gauche ou à droite du verbe tensé de S.

[15] Les descriptions de la grammaire fonctionnelle (Dik 1981) ou des grammaires en chaîne ne sont pas locales: elles prennent en compte la chaîne en son entier, mais elles souffrent, par ailleurs, d'un manque de précision qui demande une réélaboration.

3.1. La sous-catégorisation stricte

3.1.1. Dans l'approche d'origine, les règles syntagmatiques sont regroupées: elles forment un composant autonome (la base) dans le modèle des grammaires génératives d'avant Gouvernement et Liage (GB), et un module autonome dans la GSG et GB. Dans le modèle classique de la grammaire générative, les règles mettent en place progressivement la structure en constituants de S en « allant de S vers les items lexicaux ». J'ai déjà noté que l'on pouvait douter de l'hypothèse selon laquelle il y a « solution de continuité » entre les différents constituants de S et S.

Ce développement s'opère de façon autonome par rapport aux items lexicaux: ils viennent se ranger dans les structures ainsi développées lors de la réécriture des symboles pré-terminaux: ils **s'insèrent** dans la structure syntagmatique (voir la règle d'insertion lexicale dans *Aspects* par exemple).

C'est le modèle de base pour les parseurs. L'interprète parcourt les règles "en allant des items lexicaux vers la catégorie racine (S) ou de la catégorie racine vers les items lexicaux". Les items lexicaux y sont vus sous l'angle de leur identité catégorielle, identité catégorielle qui fait l'objet d'un test validant l'application de la règle ou son échec. On peut dire que de tels parseurs parsent non pas des énoncés particuliers, mais des grammaires à propos d'énoncés particuliers.

3.1.2. Je rappelle brièvement ce qu'est la sous-catégorisation stricte. On part de la constatation suivante: une unité lexicale (verbe, nom, je laisse de côté les autres catégories majeures) se construit avec un (ou des) complément(s) d'un certain type catégoriel; elle ne forme pas une suite grammaticale quand elle est construite avec d'autres types catégoriels. Ainsi, et en reprenant les exemples classiques:

(13) a. L'idée que tu viennes m'ennuie.
a'. * Le plan que tu viennes m'ennuie.

b. Je choisis la liberté (* pour la liberté).
b'. *J'opte la liberté (OK pour la liberté).

c. J'ai obtenu un rapport de Paul (* Paul)
c'. J'ai soutiré un rapport à Paul (* de Paul)

On constate, donc, qu'un item lexical donné n'est compatible qu'avec un ensemble restreint de constituants, sans que le semantisme des items ne permette de prévoir cette compatibilité (cette dernière caracterisation devrait être nuancée). Deux descriptions, extremement différentes dans ce qu'elles présupposent de la langue, sont en lice.

La première (elle a donné son nom au phénomène) est directement issue de l'approche distributionnelle qui est au fondement des grammaires syntagmatiques. Un verbe, par exemple *choisir*, apparaît dans le contexte d'un GN, alors que tel autre verbe, par exemple *opter*, ne le peut pas. On peut exprimer le contexte dans deux dimensions: la chaîne ou le syntagme.

Dans la dimension de la chaîne, on trouve les traitements classiques dits « dépendants du contexte ». C'est celui de *Aspects*: chaque verbe est affecté d'un trait stipulant son contexte droit. La règle d'insertion lexicale est sensible non seulement à l'identité catégorielle (par exemple, être un verbe), mais aussi à la sous-classe de V à laquelle il appartient (etre un v qui peut apparaître dans tel contexte, par exemple pour *choisir*, dans la chaîne « ----n2 »). Le traitement dans la dimension du syntagme a été développé par la GSG. On dira que *choisir* a la propriété d'appa-

raître dans un GV de forme [v0 n2], alors que *opter* a la propriété d'entrer dans un GV de forme [v0 p2]. Chaque verbe est affecté d'un trait (précisément d'une valeur d'un trait SUBCAT) représentant l'identificateur de la règle de développement de syntagme où ce verbe peut apparaître en position de tête lexicale (Gazdar et al, 1985: 34).[16]

Ces deux traitements obéisssent au même schème: on distingue des sous-catégories de la catégorie V sur la base des possibilités d'occurrence des membres de V. Les verbes sont sous-catégorisés par leurs compléments.

La seconde approche, que j'appellerai projective, implique un renversement de perspective: un verbe (toute unité lexicale majeure) sélectionne ses compléments. Ce qui est exprimé dans l'approche distributionnelle par la propriété « être insérable dans tel contexte », est exprimé dans l'approche projective par la propriété « régir tel(s) complément(s) ». La composition catégorielle du GV (de la phrase si on refuse ce constituant) dépend crucialement d'une propriété lexicale prêtée à chaque item lexical, la propriété de rection.[17]

La problématique peut être étendue à ce qui n'est pas de l'ordre de la rection à proprement parler (les compléments), à tous les éléments dépendant d'une unité lexicale. Par exemple, les déterminants dans le groupe nominal. Il y a sens à dire que les noms en français sous-catégorisent la position de spécifieur (absence ou présence de cette position): par exemple, appellatifs et noms propres ne requièrent pas de déterminant *(Mademoiselle est venue, Marie est venue).*[18] On peut constater que seuls noms et verbes se projettent dans des groupes permettant des circonstanciels.

Progressivement, c'est la totalité de la structuration des syntagmes qui « passe sous la dépendance » de la tête lexicale. La place et l'extension d'un composant autonome de règles syntagmatiques en diminuent d'autant: la rection lexicale étant locale, il n'y a de sens à stipuler une règle de composition que pour un constituant mettant en relation deux localités: c'est la thèse de GB concernant le groupe S dans les langues comme le français. X-barre peut, alors, être conçu comme stipulant une forme abstraite d'organisation des syntagmes, quelque chose comme une bonne forme.

3.2. Eléments de synthèse.

Par delà le problème spécifique du traitement du fait à construire sur les contrastes de (13) ci-dessus, nous avons là deux hypothèses fondamentalement différentes sur la langue, la place du lexique dans la syntaxe, la définition de ce qu'est une structure syntaxique, les conditions de reconnaissance et d'interprétation des structures syntaxiques.
Dans l'approche à base distributionnelle, les règles définissent les structures syntagmatiques possibles pour une langue donnée. Ces structures définissent le cadre d'emploi et de fonctionnement des unités lexicales. Elles constituent également un principe d'organisation du lexique: le lexique étant la liste de tous les items, étant donnée la sous-liste des items appartenant à telle catégorie, on peut partitionner les listes catégorielles en sous-listes regroupant les items particuliers sur la base de leur occurrence dans les règles syntagmatiques (on reconnaît là l'approche de l'identité catégorielle de la GSG).

---

[16] On procède ainsi pour toutes les catégories majeures. De fait, dans GSG, pour tous les items lexicaux. Je renvoie à Gazdar (op. cit.) pour la critique du traitement « dépendant du contexte » de Aspects.

[17] Sur la reprise du concept de rection, voir, par exemple, Rouveret 1987 ou Milner 1985a.

[18] Voir sur ce point Milner 1978, Marandin (en prép).

Dans l'approche projective, les items lexicaux définissent des espaces structuraux et sélectionnent les entités linguistiques qui peuvent occuper les positions définies par ces espaces. Les espaces déterminés par les items lexicaux, sont soumis à des principes de bonne formation. Chaque item lexical étant a priori différent de tous les autres, il n'y a pas de principe d'organisation du lexique. S'il y a organisation du lexique, elle est décelable sur la base des items eux-mêmes et de leur propriétés. On peut constater ici une convergence avec le programme de recherche « lexique-grammaire » défini par M. Gross.

Dans le référentiel de l'approche projective, il n'y a guère de sens à définir un parseur comme un interpréteur de règles syntagmatiques: il apparaît davantage comme un « interpréteur » d'informations lexicales (les propriétés des items parsés) sur les structurations possibles d'une portion de chaîne: cette portion est définie comme le voisinage droit ou gauche qui fait l'objet de l'information portée par l'item. Il est donc fondamental, pour la théorie du parsage et, pour un parseur particulier, de se donner les moyens de choisir entre ces deux hypothèses.

L'hypothèse projective est adoptée par la théorie GB (Chomsky 1984, Rouveret 1987). Dans ce cadre, le principal argument donné pour l'abandon des règles syntagmatiques générant les constituants à tête lexicale, est la redondance des règles de réécriture par rapport à l'information associée aux items lexicaux. « [La règle V" -->N" S'] récapitule une information déjà présente dans le lexique: *convaincre* est un prédicat sous-catégorisé pour deux compléments, un complément N" et un complément S', auxquels il attribue des rôles thématiques » (Rouveret 1987: 54). Le raisonnement met en jeu les principes d'économie d'une théorie.[19] Mais il doit, aussi, être soumis à la validation empirique. Cette validation met crucialement en jeu l'ensemble des propriétés des items lexicaux. C'est un programme de recherche où le facteur quantitatif (le nombre d'items ou de famille d'items décrits) prend une valeur certaine.

Ce facteur quantitatif est, bien sûr, déterminant pour un parseur basé sur l'approche projective. En toute logique, tous les items lexicaux doivent être associés à l'information pertinente au parsage de la portion d'énoncé où ils apparaissent. Etant donné l'ampleur de la tâche, et l'état des problèmes afférents à la représentation et au stockage des informations lexicales, il faudra se résoudre -- d'un point de vue réaliste et pour quelques années encore -- à doter ce type de parseur de procédures heuristiques de parsage par défaut d'information.

4. CONCLUSION.

J'ai introduit trois problèmes distincts. Ils relèvent de la syntaxe du groupe nominal français, de la syntaxe de la chaîne parlée/écrite et de la relation entre lexique et syntaxe. Le groupe nominal doit être décrit dans toute sa complexité: est-ce une complexité phénoménologique ou une complexité structurale? A-t-on, pour cette unité, plusieurs principes formels d'organisation ou un seul? Le problème de la chaîne parlée/écrite est autrement plus redoutable: on constate qu'elle n'est pas exhaustivement structurée par la projection des items lexicaux en domaines et/ou une grammaire syntagmatique s'enracinant dans S. Le domaine S est immergé dans une organisation qui doit recevoir un statut théorique et formel (elle n'est pas concevable comme une strate enchâssante): c'est une des conditions de possibilité du traitement des processus syntaxiques et sémantiques de repérage, thématisation, prédication sous-jacente, reprise, modalisation énonciative. Si l'on admet que l'interprétation d'un énoncé est d'abord contextuelle, l'enjeu est considérable pour la formalisation de la sémantique associée au parseur.

---

[19] On notera que le traitement de la GSG évite cette redondance en recourant à la notation trait/valeur: trait de sous catégorisation/identificateur d'une règle syntagmatique. Et .... maintient un module de règles syntagmatiques. On trouvera dans Heny 1979 (mystérieusement peu cité) l'archéologie de cette discussion. Je n'introduis pas dans ce qui suit le Principe de Projection, bien évidemment solidaire de cet argument, afin de ne pas alourdir l'exposé.

Enfin, la place du lexique dans l'économie de la théorie syntaxique, par delà l'enjeu théorique portant sur la définition des propriétés de l'entité langue, détermine l'architecture des parseurs. Elle détermine très directement la forme et le contenu des représentations associées aux items lexicaux. Or, on ne souligne pas assez que le "lexique" (pris en lui-même) n'impose aucune forme; comme la confection d'un dictionnaire est une tâche considérable, toute décision portant sur la forme des entrées lexicales est lourde de conséquences pratiques.

L'accroissement de la couverture des parseurs semble donc étroitement dépendante des avancés théoriques et empiriques de la linguistique.

Références:

ALLEN, J. 1987, *Natural language Understanding*, The Benjamins/Cummings Publishing Company.

BLANCHE-BENVENISTE, C. 1975, *Recherches en vue d'une théorie de la grammaire française*, Paris: Librairie Honoré Champion.

CADIOT, P., FRADIN, B. éds., 1988, Le thème en perspective, *Langue Française* 78, Paris: Larousse.

CHOMSKY, N. 1984, *Lectures on Government and Binding*, Dordrecht: Foris.

------ 1987, *La nouvelle syntaxe*, Paris: Le Seuil.

DANLOS, L. éd., 1988, Les expressions figées, *Langages* 90, Paris: Larousse.

DIK, S. 1981, *Functional grammar*, Dordrecht: Foris Publications.

EMONDS, J. 1985, *A Unified Theory of Syntactic Categories*, Dordrecht: Foris Publications.

FRADIN, B. 1988, « Approche des constructions à détachement: la reprise interne » [Cadiot-Fradin]: 26-56.

------ (en prp), Les constructions à détachement: la génération perdue.

FRANCKEL, J-J. 1988, « Gérondif et repérage interpropositionnel», *Etudes sur l'ordre des mots*, : 97-128, Collection ERA 642, Paris-7.

GAZDAR, G. et al., 1985, *Generalized Phrase Structure Grammar*, Oxford: Basil Blackwell.

GRECIANO, G. 1983, *Signification et dénotation en allemand: la sémantique des expressions idiomatiques*, Paris: Klincksieck.

GROSS, G. 1988, « Degré de figement des noms composés », [Danlos]: 57-72.

------ 1975, *Grammaire transformationnelle du français; syntaxe du nom*, Paris: Larousse.

------ 1975, *Méthodes en syntaxe*, Paris; Hermann.

------ 1988, « Les limites de la phrase figée », [Danlos,d.], : 7-22.

HENY, F. 1979, "Review of The Logical Structure of Linguistic Theory", *Synthese* 40.

HUOT, H. 1981, *Constructions infinitives du français*, Genève-Paris: Droz.

KING, M. ed., 1983, *Parsing Natural Language*, Londres: Academic Press.

MARANDIN, J.-M. (en prp.) Il n'y a pas de déterminant zéro en français.

MARCUS, M. 1980, *A Theory of Syntactic Recognition for Natural Language*, MIT Pres.

MILNER, J.-C. 1978, *De la syntaxe à l'interprétation*, Paris: le Seuil.

------ 1982, « La redondance fonctionnelle », *Ordres et raisons de langue*, Paris: Le Seuil.

------ 1985a, « Réflexions sur le concept de catégorie vide», *Modèles linguistiques*, 7: 33-55.

------ 1985b, *De l'inutilité des arbres en linguistique*, Collection ERA 642, Paris-7.

MORIN, J.-Y. sd, *Théorie syntaxique et théorie du parsage*, Université de Montréal.

PIOT, M. 1988, « Conjonctions de subordination et figement », [Danlos d.], :39-56.

PLANTE, P. 1988, *Le langage FX: la programmation en faisceaux*, Centre d'ATO, UQAM.

PROUDHIAN, D. et POLLARD, C. 1985, "Parsing Head-driven Phrase-structure Grammar", *Proceedings 23 Annual Meeting of the ACM*, Université de Chicago: 167-171.

ROUVERET, A., 1987, « Présentation » et « Postscript », [Chomsky 1987].

ROUVERET, A. et VERGNAUD, J-R., 1980, "Specifying Reference to the Subject: French Causatives and Conditions on Representations", *Linguistic Inquiry*, 11: 97-202.

SALKOFF, M. 1973, *Une grammaire en chaîne du français*, Paris: Dunod.

------ 1979, *Analyse syntaxique du français: grammaire en chaînes*, Amsterdam: John Benjamins B.V.

SHIEBER, S. 1986, *An Introduction to Unification-based Approaches to Grammar*, Stanford: CSLI.

SMALL, S. 1983, "Parsing as Co-operative Distributional Inference. Understanding through Memory Interactions", [King]: 247-276.

«INDUSTRIES DE LA LANGUE» : UN CONCEPT À DÉFINIR

Marie-Claude L'Homme
Université Laval

Nous nous sommes rendu compte que le titre, choisi au préalable c'est-à-dire « Industries de la langue : un concept à définir », pouvait donner l'impression que nous apporterons des solutions aux problèmes soulevés par la définition et la description des industries de la langue. Or, il est assez difficile de synthétiser tout ce que comprend et, surtout, tout ce que sous-entend le terme *industries de la langue* à l'intérieur d'une seule définition.

Dans le cadre de cet exposé, nous nous limiterons à mettre en lumière deux aspects du concept d' « industries de la langue » en tentant de faire ressortir ce qui le caractérise dans un contexte francophone. En fait, notre exposé prendra la forme d'une introduction aux industries de la langue : nous nous attacherons à démontrer la difficulté de définir ce concept. Peut-être qu'un titre comme «Industries de la langue : un concept en voie de définition» lèverait l'ambiguïté posée par le premier.

Nous tenons à préciser que nous présenterons les industries de la langue en faisant référence au contexte francophone car, dans un autre contexte, les IDLL revêtent une valeur toute autre.

On a vu surgir, vers 1984, un terme nouveau, celui d'*industries de la langue* qui a donné naissance à des dérivés : *industrialisation* (de la langue), *s'industrialiser* (en parlant de la langue) et *industrialisé* (à valeur d'adjectif). (Ajoutons ici que nous avons déjà entendu *industriel de la langue* mais l'utilisation de ce terme semble limité à un cadre restreint).

Si le terme *industries de la langue* en tant que tel semble connu par plusieurs, le concept qu'il recouvre reste encore vague. Il fournit, cependant, des indices quant à son contenu. *Industrie* implique une forme de transformation, de fabrication ou d'adaptation d'un matériau, ici le matériau serait la langue. On pourrait y voir, en outre, l'indice d'une volonté de commercialiser des produits reliés d'une façon ou d'une autre à la langue. Les traits de définition que nous avons esquissés ont de quoi surprendre. On peut peut-être parler, à la rigueur, d'adaptation de la langue mais comment peut-on fabriquer, transformer ou commercialiser un matériau linguistique? Nous reviendrons sur cette question plus loin.

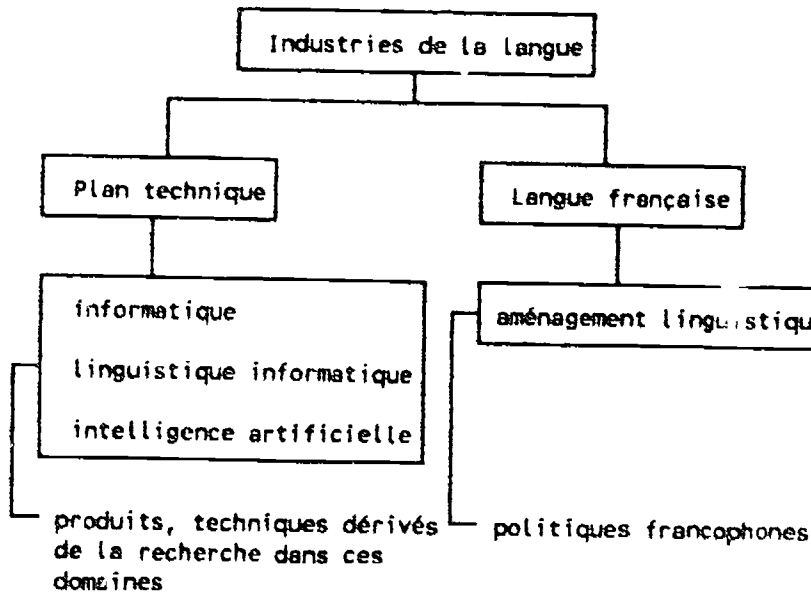
Le concept d' « industries de la langue » a été défini et redéfini par plusieurs personnes spécialistes ou non (plus souvent non spécialistes) mais il demeure encore difficile à circonscrire et tous les auteurs s'entendent là-dessus. Le fait qu'on utilise toujours le terme au pluriel (on parle des industries de la langue mais rarement d'une industrie de la langue) témoigne probablement du caractère encore flou du concept qu'il recouvre.

Lorsqu'il est question d'industries de la langue, il est souvent question d'informatique, de langue naturelle, de reconnaissance ou de synthèse vocale, d'intelligence artificielle ou de linguistique informatique, de grands termes qu'on insère dans des énumérations qu'on voudrait

explicites sinon impressionnantes. Mais comment trouver le lien entre toutes ces disciplines? Qu'est-ce qui a fait qu'elles se sont trouvées réunies et comment se fait-il qu'elles se sont trouvées étroitement liées à la question de la sauvegarde de la langue française comme langue véhiculaire de la science et de la technique?

Avant de répondre à cette question, on peut d'abord affirmer qu'il est possible de présenter les industries de la langue de deux façons, sous deux perspectives qui, à première vue, ne paraissent pas complémentaires (voir figure 1).

FIGURE 1:

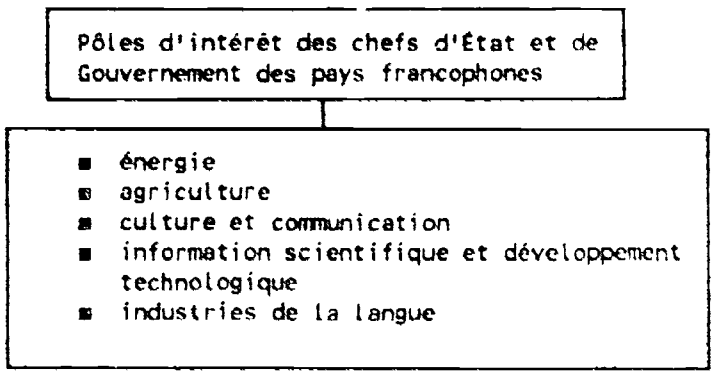


D'une part, les industries de la langue sont caractérisées par ce que nous appellerons un aspect technique: les IDLL regroupent l'ensemble des procédés et des techniques qui permettent un traitement automatique de la langue.

D'autre part, les industries de la langue présentent un second aspect qui est plus difficile à décrire que le premier. Elles apparaissent comme une solution au problème de sauvegarde de la langue française comme mode de diffusion dans les secteurs de pointe. Elles s'inscrivent en ce sens dans un vaste plan d'aménagement linguistique.

On peut expliquer l'interrelation de ces deux aspects en décrivant les circonstances qui ont motivé l'apparition du concept d'« industries de la langue ». Il est né dans un contexte politique, et plus particulièrement dans le cadre de la Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, titre officiel donné au Sommet francophone de Paris tenu en février 1986. (Il y avait bien eu quelques travaux auparavant mais c'est surtout à cette occasion que le concept a été diffusé.) (Voir fig. 2.) Ainsi les industries de la langue sont devenues un des cinq pôles d'intérêt principaux des pays francophones au même titre que l'énergie, l'agriculture, la culture et la communication et enfin, l'information scientifique et le développement technologique (voir Fig. 2).

FIGURE 2:



Les études menées lors du sommet ont conduit au constat suivant:

> « La langue française doit rapidement s'inscrire dans le mouvement actuel d'industrialisation des langues; autrement elle deviendra de moins en moins apte au développement de la recherche dans les secteurs de pointe et, à long terme, se marginalisera par rapport aux autres grandes langues de communication internationale dans ces champs d'activités essentiels à l'avenir de la francophonie.[1] »

Les chefs d'État ont vu dans l'industrialisation de la langue le moyen d'assurer (ou plutôt, il faut bien l'admettre, de redonner) à la langue française son statut de langue véhiculaire de la science et de la technique. Trois organismes ont été créés pour poursuivre les travaux amorcés dans le cadre du sommet et pour proposer des programmes visant à promouvoir les industries de la langue: il s'agit du Réseau des industries de la langue, organisme à vocation internationale; de la Mission industries de la langue, dont le centre d'activités est situé en France; et du Sous-comité québécois des industries de la langue dont le siège est, de toute évidence, au Québec.

Par ailleurs, à l'extérieur de la structure des sommets francophones, d'autres travaux ont été effectués. Un mois après la tenue du premier sommet francophone, un colloque portant sur les IDLL réunissait plusieurs spécialistes qui ont débattu la question. Le colloque intitulé *Industries de la langue. Enjeux pour l'Europe* s'est tenu à Tours en mars 1986. On y a souligné l'importance des enjeux représentés par les IDLL et identifié les secteurs d'intervention principaux. Ailleurs, de nombreuses universités et des groupes de recherche affiliés ou non ont élaboré des projets visant à faire avancer les recherches en ce domaine: pensons au Centre international de recherche sur le bilinguisme (maintenant appelé Centre international de recherche sur l'aménagement linguistique), ici même à l'Université Laval, qui a décidé d'accorder une grande place à la recherche dans les domaines des industries de la langue.

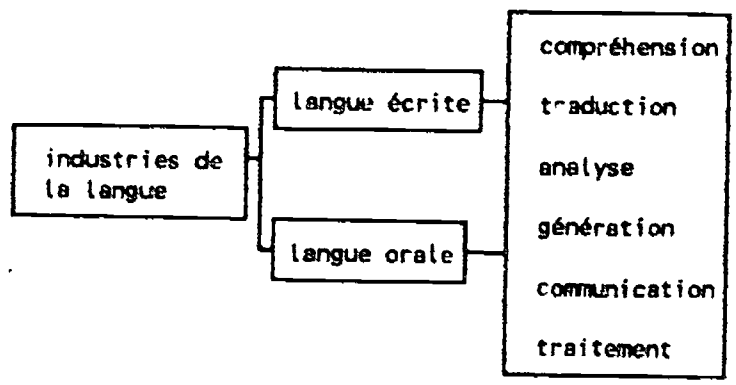
Mais laissons de côté ces considérations d'ordre historique pour nous attarder sur l'aspect technique des industries de la langue. Comme nous l'avons dit plus haut, les industries de la langue offrent des procédés et des techniques qui visent un traitement automatique du matériau linguistique. Donc, les IDLL font intervenir l'informatique ou, de façon plus générale, une forme d'automatisme. Cela peut paraître étonnant à première vue car il n'y a rien dans le terme

[1] Document de synthèse : Industries de la langue, Québec, s.d. [1987], p. 173.

industries de la langue qui nous laisse entendre qu'il s'agit d'informatique contrairement à d'autres disciplines qui arborent un *-tique* ou un *assisté par ordinateur* lorsqu'elles se trouvent associées à l'informatique.

Les produits offerts par les différentes sphères d'activité des IDLL sont des systèmes informatiques ou des automates qui traitent, manipulent, génèrent et comprennent le langage humain aussi bien sous sa forme écrite que sa forme parlée (voir fig. 3).

FIGURE 3:



Des systèmes qui :

- traitent le langage humain: les correcteurs automatiques, les lemmatiseurs;
- manipulent le langage humain: les lecteurs automatiques de texte;
- génèrent le langage humain: les synthétiseurs de parole ou les générateurs de texte;
- comprennent le langage humain: les systèmes de reconnaissance vocale ou les systèmes de dialogue personne-machine.
  (Ici *comprennent* est employé au sens informatique du terme.)

Les produits offerts par les différents secteurs d'activité sont souvent présentés de façon futuriste ou du moins optimiste. On décrit souvent des décors qui appartiendraient davantage à la science-fiction qu'à la recherche.

A titre d'exemple de vision futuriste, citons le bureau informatisé dans lequel un ordinateur pourrait saisir un texte automatiquement à la suite d'une lecture optique, corriger les erreurs qu'il contient, le traduire et l'expédier en Europe par voie télématique. Ce même ordinateur pourrait fournir une réponse contenue dans une base de données à la suite d'une requête formulée verbalement par l'utilisateur. Citons également le guichet automatique qui reconnaîtrait la voix de l'utilisateur et lui donnerait ce qu'il a demandé en réponse à un simple message formulé oralement. Enfin, qui n'a jamais rêvé de dicter à sa voiture les opérations à effectuer pour se rendre à la maison après une journée de travail?

Les titres d'articles portant sur certains domaines des IDLL trahissent souvent les vues futuristes des auteurs.

Par exemple:

« Demain l'ordinateur polyglotte »

« Les banques de terminologie de l'avenir »

« Des machines qui comprennent notre langage »

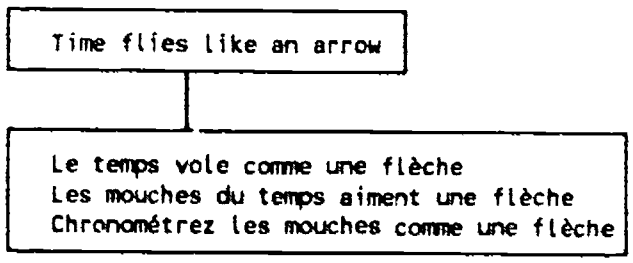
« Quand les machines à écrire auront des oreilles »

Pensons également à Jacques Pitrat qui a dédié son ouvrage intitulé, *Textes, ordinateurs et compréhension*, « au premier programme d'ordinateur qui sera capable de le comprendre »[2].

Nous n'avons donné là qu'un aperçu de ce qu'on peut rédiger sur le sujet et il ne s'agit pas là, précisons-le, de passages tirés de romans de science-fiction.

Même si nous sommes encore loin de ce monde idéal pour certains et cauchemardesque pour d'autres, il est difficile de nier les progrès considérables qui ont été réalisés dans plusieurs domaines des industries de la langue depuis la fin des années 1970. On a souvent cité la traduction automatique comme exemple pour démontrer qu'un système informatique ne peut traiter les structures linguistiques parce qu'il est incapable de comprendre. La phrase qui illustre le mieux l'incompréhension de la langue par la machine est bien la suivante (voir fig. 4). Cet exemple sert bien aux opposants du traitement automatique des langues naturelles.

FIGURE 4:



On voit que la machine confond *Time*, *flies* et *like* qu'elle considere comme étant tantôt des formes verbales, tantôt des formes nominales, tantôt des formes conjonctives. (On oublie de dire cependant qu'on retrouve des erreurs du genre dans les traductions humaines.)

Malgré de nombreux échecs et plusieurs années de réclusion dans les laboratoires, des produits innombrables inondent aujourd'hui le marché: par exemple, des systèmes de traitement de la parole et de traduction automatique pour micro-ordinateur, des interfaces aux bases de données en langage naturel, des lecteurs automatiques de textes, etc.

Ce qui explique ce revirement c'est qu'on a su adapter les objectifs visés à la capacité de la machine. En traduction automatique, par exemple, on travaille dans des domaines restreints;

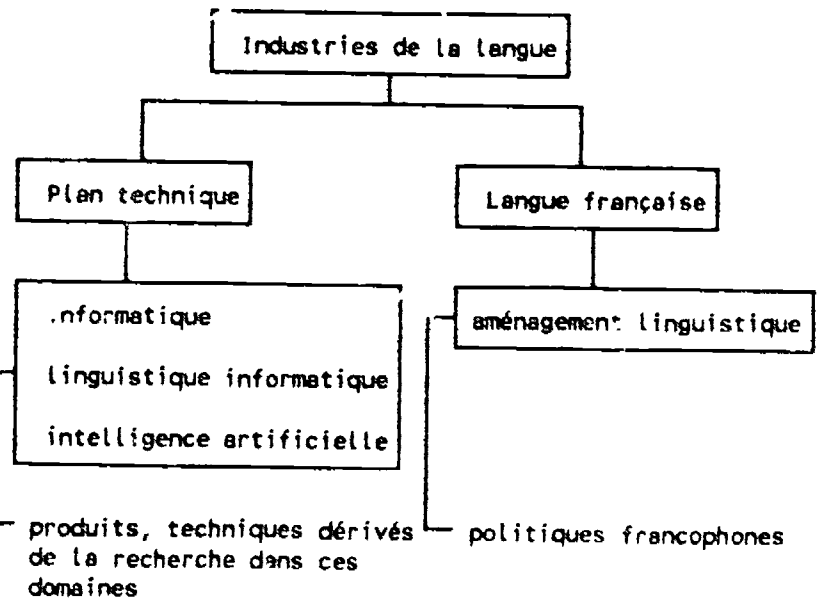
[2]Jacques Pitrat, Textes, ordinateurs et compréhension, Paris, Eyrolles, 1985.

on parle également de traduction assistée par ordinateur lorsqu'il est question de traiter des textes de nature variée. En reconnaissance vocale, on limite le nombre de mots à reconnaître, etc. Il faut également noter que les capacités de mémoire et de traitement des systèmes informatiques ont augmenté beaucoup depuis les premiers systèmes informatiques et continuent de s'accroître.

On a souvent tendance à confondre les industries de la langue et la linguistique informatique ce qui est tout à fait justifiable car elles présentent, entre autres, deux points communs: d'abord la langue et ensuite l'informatique. Ces deux activités, qui s'apparentent par le mode de traitement et par le matériau, poursuivent cependant des objectifs différents. Si la linguistique informatique se sert de l'ordinateur pour traiter la langue, les industries de la langue font de même car elles se servent des acquis de la linguistique informatique mais, de plus, se servent de la langue pour adapter l'informatique. Si la linguistique informatique est surtout axée sur la recherche, les industries de la langue sont orientées vers un marché identifiable, dans le but de fabriquer des produits commercialisables. De plus, les IDLL n'intéressent plus uniquement le linguiste mais également, le cogniticien, le didacticien et l'informaticien.

Attardons-nous sur le second aspect des industries de la langue (reprise de la figure 1) celui qui porte plutôt sur l'aménagement de la langue française.

FIGURE 1: (reprise)



Les industries de la langue, comme nous le disions plus haut, s'inscrivent dans un vaste plan d'aménagement linguistique. Il peut paraître étonnant que les IDLL ne semblent concerner que la langue française car d'après tout ce qui a été dit plus haut, elles devraient, logiquement, tou her toutes les langues existantes. Si la langue française peut se prêter à des traitements automatiques, toutes les langues le peuvent et certaines le font déjà: le japonais et l'anglais en sont de parfaits exemples. Le problème se situe à un autre niveau.

Ce n'est un secret pour personne que le français n'occupe plus la place privilégiée qu'il occupait auparavant dans la diffusion des sciences et des techniques et dans la communication

internationale. L'anglais a pris sa place comme il l'a fait pour plusieurs autres langues, dans de nombreux domaines et particulièrement dans les domaines techniques et scientifiques. Le domaine technique favorisé à l'heure actuelle est l'informatique car il touche toutes les sphères de l'activité humaine et la langue qui le diffuse est encore une fois l'anglais. Les développements récents de l'informatique et plus particulièrement de l'intelligence artificielle laissent présager des avenues très vastes aux produits « en langue naturelle » plaçant ainsi la langue traitée en position de force.

Notre propos ne vise pas à analyser la question de la perte de vitesse de la langue française par rapport aux autres langues ni à en examiner le bien-fondé ce qui a déjà été fait à plus d'une reprise. Nous voulons plutôt faire ressortir les raisons qui ont fait que les IDLL se sont vu accorder autant d'attention et d'intérêt depuis leur apparition.

Les chefs d'État ont souligné, lors du premier sommet francophone et plusieurs spécialistes partagent cet avis, que les langues qui ne pourront s'industrialiser se marginaliseront par rapport aux autres. Ils ont mis en évidence le fait qu'il faut mettre sur le marché des produits qui parlent et qui comprennent le français, ou, du moins, se préparer à le faire pour être en mesure d'affronter les impératifs commerciaux de demain. (Ici, nous tombons nous-mêmes dans le piège des vues futuristes puisque nous employons *demain*.)

Cette affirmation a de quoi faire sursauter le francophone a qui on a toujours dit que sa langue était une langue littéraire qui ne se prêtait pas à des représentations exactes. (Ce que nous venons de dire s'applique peut-être moins aujourd'hui mais encore ...) De plus, certains termes utilisés peuvent paraître assez étonnants lorsqu'on les applique à la langue française. *Industrialiser, transformer, adapter, traitement automatique de la langue* et *matériau linguistique* sonnent faux aux oreilles du néophyte.

Pour résumer ce deuxième aspect des industries de la langue, il suffit de retenir qu'elles permettent à la langue française de demeurer dans les rangs des grandes langues de diffusion internationale en lui offrant des outils de développement soit pour assister le travail linguistique, soit pour diffuser des produits issus des connaissances linguistiques, soit, enfin, pour développer la langue en vue de son traitement automatique.

C'est à ce niveau qu'intervient toute la question reliée à l'informatique en français, celle qui prône le développement d'outils conçus pour et par les locuteurs francophones. Cet aspect permet également d'expliquer les divers points de vue exprimés dans tous les articles portant sur les industries de la langue.

Par exemple :

> «Le français pour survivre doit être mis en puce».[3]
>
> « Et le risque est que, faute de produits « parlant français », la nécessité économique ne pèse de tout son poids en faveur de l'anglophone ».[4]

---

[3]Robert Gelly, **Les vrais trésors de la langue française**, dans Ça m'intéresse, novembre 1986.

[4]William Baranes, **Les industries de la langue**, dans Qui-vive international, no 4, p. 74.

ou encore :

> « Alors que l'on s'achemine vers un monde où le dialogue de la voix humaine avec la machine sera quotidien, il s'agit de savoir si les ordinateurs sauront aussi parler le français ».[5]

Il faudrait définir les industries de la langue de deux façons pour englober les deux aspects que nous avons décrits ci-dessus, c'est-à-dire l'aspect technique et le côté qui touche à l'aménagement linguistique.

Elles peuvent d'abord être définies sur le plan technique ce qui a déjà été fait. (La définition que nous donnons à un caractère quasi officiel et a été reprise par presque tous ceux qui s'intéressent de près ou de loin aux industries de la langue.)

> « Les industries de la langue fabriquent et commercialisent des automates qui manipulent, interprètent, génèrent le langage humain aussi bien sous sa forme écrite que sous sa forme parlée. »[6]

Il faudrait proposer une deuxième définition qui porterait sur un aménagement de la langue française dans les secteurs de pointe et qui correspondrait en gros, aux objectifs visés par les industries de la langue, à l'idéologie véhiculée par son concept.

Les industries de la langue regroupent l'ensemble des moyens entrepris pour traiter la langue française écrite ou parlée, afin de lui permettre de conserver son statut de langue véhiculaire de la science et de la technique et son statut de mode de communication internationale.

Nous voudrions, pour conclure, souligner le fait que les industries de la langue, dans un contexte francophone, ne sauraient se définir en tenant compte uniquement de l'aspect technique. Il faut également considérer le côté « aménagement linguistique » que nous avons décrit de façon sommaire ci-dessus.

Nous terminerons en espérant avoir contribué à une meilleure compréhension du concept d'industries de la langue et à une prise de conscience des enjeux que de telles industries représentent.

Nous espérons également que d'ici quelques années nous verrons apparaître une véritable industrie de la langue française.

---

[5] François Schlosser, **Les ordinateurs et la francophonie**, dans Le Nouvel observateur, no 1110, 14 février 1986 p. 38.

[6] Rapport de synthèse : Industries de la langue, Actes de la Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, Paris, 17-19 février 1986, La Documentation française, 1986, p. 86.

Auteure **Laurence Danlos**
L.A.D.L., Paris 7

Titre **Intéraction des décisions dans un système de Génération automatique de textes**

RÉSUMÉ

Il est généralement supposé qu'un système de génération automatique de textes peut être modularisé en une suite de composants, le premier prenant les décisions conceptuelles, les suivants prenant les décisions linguistiques (i.e. choix lexicaux et choix des constructions syntaxiques), l'avant-dernier effectuant les opérations syntaxiques, le dernier les opérations morphologiques. Cette modularisation d'un système de génération repose sur les hypothèses suivantes :

- *les décisions de "haut niveau" doivent être prises avant les décisions de "bas niveau";*
- *les décisions conceptuelles sont de "haut niveau", les décisions linguistiques de "bas niveau", les opérations syntaxiques de "très bas niveau", les opérations morphologiques d'un niveau encore plus bas que celui des opérations syntaxiques.*

Nos travaux précédents ont infirmé ces hypothèses. D'abord, nous avons montré (Danlos, 1987a) que les décisions conceptuelles et linguistiques sont dépendantes les unes des autres. De ce fait, nous avons conçu un modèle de génération modularisé de la façon suivante : un "composant stratégique", prend simultanément les décisions conceptuelles et linguistiques. Il fournit un "schéma de texte" qui est synthétisé en un texte par un "composant syntaxique". Ce composant syntaxique effectue des opérations telles que l'accord entre un verbe et son sujet, ou la réduction d'une complétive à une infinitive (Danlos 1987b). Il traite aussi des questions de pronominalisation, i.e., il détermine quand une forme pronominale doit être synthétisée. Nous avons montré dans (Danlos and Namer 1988) que les questions de pronominalisation mettent en jeu le niveau morphologique dans les systèmes de génération produisant des textes en langues romanes. Les décisions concernant la pronominalisation - pierre d'achoppement de tout système de traitement automatique de langage naturel - ne doivent certainement pas être prises en dernier ; de ce fait, les décisions morphologiques ne doivent non plus pas être prises en dernier.

Notre article commencera par une récapitulation de nos travaux précédents. Brièvement, il exposera, d'une part, pourquoi les décisions conceptuelles et linguistiques sont dépendantes les unes des autres, d'autre part, comment intervient le niveau morphologique dans les questions de pronominalisation. Ensuite, nous ferons un pas supplémentaire dans nos recherches en montrant l'intéraction des décisions en génération automatique. Nous montrerons que notre modularisation en deux composants - un composant stratégique et un composant syntaxique, ce dernier traitant les phénomènes de pronominalisation - est encore trop modulaire: certaines questions de pronominalisation demanderaient à être prises en compte au moment où les décisions conceptuelles et linguistiques sont prises. Nous n'avons pas encore conçu

d'algorithme de génération qui reflète l'intéraction totale entre les niveau conceptuels, linguistiques, syntaxiques et morphologiques; néanmoins, nous espérons au moins souligner la complexité linguistique d'un système de génération automatique, complexité qui est encore sous-estimée.

Bibliographie

DANLOS, L., 1985, *Génération automatique de textes en langues naturelles*, Masson, Paris.

------ 1987a, *The linguistic basis of text generation*, Cambridge University Press, Cambridge.

------ 1987b, A French and English Syntactic Component for Génération, *Natural Language Generation: New results in Artificial Intelligence, Psychology and Linguistics*, Kempen G. ed, Dortrecht/Boston, Martinus Nijhoff Publishers.

------ 1988, Morphology and cross dependencies in the synthesis of personal pronouns in Romance languages, *Proceedings of COLING-88*, Budapest.

INTERACTIONS DES REPRÉSENTATIONS ORTHOGRAPHIQUES ET PHONOLOGIQUES DURANT LA LECTURE

Martin Beaudoin
Université Laval

INTRODUCTION

Le langage est un phénomène fort complexe. Nous ne serions pas ici réunis pour en discuter s'il en était autrement. Le développement de modèles psycho-informatiques durant les dernières années a permis de décrire et d'expliquer certains processus jusqu'alors obscurs.

Je cite ici Martinet: « Ceci ne doit pas faire oublier que les signes du langage humain sont en priorité vocaux, que, pendant des centaines de milliers d'années, ces signes ont été exclusivement vocaux, et qu'aujourd'hui encore les êtres humains en majorité savent parler sans savoir écrire. On apprend à parler avant d'apprendre à lire: la lecture vient doubler la parole, jamais l'inverse. »[1]

En effet, le langage est essentiellement oral. Même que la compréhension de la lecture ne peut se produire sans représentation mentale des phonèmes. C'est ce que je démontrerai dans les prochaines minutes.

L'amélioration des connaissances des processus de lecture a des effets tant en éducation, en linguistique, en psychologie, en communication qu'en informatique. Le développement de modèles expliquant ces processus a permis entre autres de mettre en place un programme d'enseignement québécois du français au primaire qui soit parmi les meilleurs au monde. Cela permet aussi une plus grande accessibilité à la technologie informatique.

Je définirai d'abord certains termes, puis je décrirai sommairement le cadre théorique dans lequel s'inscrit cet exposé. Je résumerai ensuite quelques expériences ayant démontré que les représentations phonologiques et orthographiques interagissent dans l'identification des mots, étape préalable à la compréhension de la lecture. Je conclurai cet exposé en énonçant quelques-unes des implications théoriques et pratiques du modèle psycholinguistique qui découle de ces recherches.

La lecture a été définie par Geschwind[2] comme étant l'habileté d'extraire le sens de quelle que forme visuelle de représentation langagiere que ce soit. Cette définition est trop généralisante. Elle sous-tend que la compréhension de langages gestuels relève aussi de la lecture. Je crois qu'il faut etablir des distinctions entre langages parlés et langages gestuels. Cette distinction est nécessaire pour tenir compte des différences d'acquisition et de certaines différences de représentation. Selon Bellugi[3], l'acquisition du langage est plus rapide pour les enfants sourds

---

[1] 1970, p. 8.

[2] 1985, cité par Marshall, 1987, traduction personnelle

[3] 1987.

dont les parents sont sourds que pour les enfants entendants dont les parents sont entendants. De plus, Hanson et Fowler ont démontré en 1987 que les sourds ont accès à des représentations phonologiques pour lire, et ce, même lorsque l'âge de surdité est survenu avant trois ans.

Je propose donc de définir la lecture comme étant l'habileté à extraire le sens d'un graphème ou d'un ensemble de graphèmes.

Les représentations mentales sont des abstractions de l'environnement. Elles sont en quelque sorte des connaissances que nous avons tirées de notre environnement. Elles servent à modaliser des processus discrets et invisibles, donc, à élaborer une simplification de phénomènes complexes par leur décortiquement dans un but explicatif et prédictif. On décortique l'environnement en divers types de représentations dont trois sont spécifiques au langage: les représentations sémantiques, les représentations orthographiques et les représentations phonologiques. Chacun de ces ensembles de représentations a sa structure propre, ce qui n'est pas l'objet de cet exposé.

Ce qui m'intéresse davantage d'élaborer aujourd'hui, c'est le réseau des interactions que ces représentations entretiennent durant la lecture. Plus spécifiquement, mon but est de démontrer que les représentations sémantiques ne sont accessibles en lecture que lorsque les connaissances qui sont véhiculées dans le document écrit sont traitées orthographiquement et phonologiquement, obligatoirement.

CORPS

L'hypothèse selon laquelle cette interaction existe provient de Conrad[4]. Cette hypothèse a été fortement étudiée depuis[5]. Je présenterai donc quelques expériences, parmi les plus probantes, démontrant cette interaction. Mais d'abord, il m'apparait important de noter que l'acquisition de la lecture peut fournir de précieuses indications de cette interaction. S'il semblait à Goldstein en 1976[6] que la conscience de la phonologie d'une langue est le précurseur de la lecture, il n'y a selon Wagner et Torgesen[7] aucune évidence sur le sens de l'influence[8]. Ainsi, ces auteurs soutiennent que la lecture amène à la capacité de segmentation phonémique[9]. Cependant, dans l'ensemble, les chercheurs s'accordent maintenant pour dire que la représentation phonologique est à la source de la lecture[10]. Vellutino et Scalon[11] notent par ailleurs qu'une étude longitudinale qu'ils ont réalisé démontre avec évidence la valeur prédictive de l'habilité à

---

[4] 1962, cité par Banks, Oka et Shugarman, 1981.

[5] Waters et al., 1985; Perfetti et al., 1987; Davidson, 1986 entre autres.

[6] cité par Bryant et Goswami, 1987.

[7] 1987.

[8] Wagner et Torgesen, 1987; Bryant et Goswami, 1987.

[9] Morais et al., 1979, cité par Culter et al., 1987.

[10] Wagner et Torgegesen, 1987.

[11] 1987.

segmenter phonémiquement les mots. Ces résultats sont aussi confirmés par Patel et Soper[12]. Nous ne pouvons établir de relation causale entre les représentations orthographiques et phonologiques[13], mais ces résultats nous indiquent qu'un lien étroit les unit.

La première recherche à être résumée constitue pour le sujet qui nous intéresse le travail dont le protocole expérimental a été élaboré avec le plus de rigueur. Il s'agit d'une étude de Van Orden, Johnston et Hale, publiée il y a quelques mois[14], et portant sur les sujets unilingues. La recherche vise à démontrer que les représentations sémantiques ne sont pas directement accessibles par les représentations orthographiques durant l'identification de mots. L'information doit obligatoirement passer par les représentations phonologiques pour y accéder. L'induction des stimuli s'est bien sûr faite par écrit. La tâche demandée aux sujets de cette expérience est une catégorisation. La variable dépendante principale est le taux d'erreurs positives de catégorisation, c'est-à-dire, lorsqu'une personne catégorise un mot de façon erronée en répondant "oui" alors qu'elle aurait dû répondre "non". On présente ainsi une catégorie sémantique, puis un stimulus, soit homophone soit un item de contrôle. Vous avez des exemples de ceci au No 1 du document que l'on vous a remis à la porte. La moitié des stimuli sont des mots, et l'autre moitié, des non-mots ou logatomes. La ressemblance orthographique a été tenue constante pour tous les stimuli, ayant été mesurée par une méthode élaborée par Weber[15]. La ressemblance phonologique a été contrôlée pour les homophones seulement. Il y a eu vérification de la prononciation des non-mots auprès de 15 personnes indépendantes du bassin des sujets.

La non-lexicalité a aussi été vérifiée auprès de 20 personnes toujours indépendantes du bassin des sujets. Finalement, la catégorie des mots homophones a été mesurée avec l'échelle de Uyeda et Mandler[16] pour éviter les recoupements catégoriels.

La première expérience de cet article compte 30 sujets, tous élèves d'une même école secondaire. Nous n'en savons pas plus sur les sujets.

Ainsi, pour valider l'hypothèse, il faudra qu'il y ait moins d'erreurs positives produites à l'induction des items de contrôle qu'aux homophones. Les items de contrôle et les homophones étant comparativement semblables aux exemples de catégorie dont les homophones sont tirés, la différence de précision dans l'exécution de la tâche provient de l'influence des représentations phonologiques. L'influence potentielle des représentations sémantiques est mesurée par la présence de non-mots parmi les homophones. En effet, si les stimuli étaient traités sémantiquement avant d'être traités phonologiquement, il y aurait eu une différence entre les résultats des mots et des non-mots, ces derniers n'ayant pas de représentations sémantiques.

Les hypotheses ont été confirmées:il y a significativement plus d'erreurs positives pour les homophones que pour les items de contrôle, et parmi les homophones, il n'y a pas de différences significatives entre les mots et les non-mots. Certains temps de catégorisation ont été mesurés et calculés. La seule différence significative est que les homophones prennent plus de temps à être jugés incorrects que les items de contrôle. Cette différence, bien que significative est légère. Ceci peut être expliqué par le fait que l'homophone activeles représentations sémantiques de

---

[12] 1987.

[13] Wagner et Torgesen, 1987.

[14] 1988.

[15] 1970, cité par Van Orden et al., 1988

[16] 1980, cité par Van Orden et al., 1988.

l'homophone lui-même et du mot dont il provient, cela bien sûr après être passé par les représentations phonologiques. Les items de contrôle ne pouvant renvoyer qu'aux représentations sémantiques d'un seul élément lorsqu'il s'agit d'un mot, ils seront légèrement moins longs à traiter.

Une deuxième expérience dans le même article vérifiait si le temps de catégorisation et la précision à catégoriser un homophone sont différents pour catégoriser d'autres exemples des mêmes catégories dont étaient tirés les homophones. Vous trouverez des exemples de ceci au No 2 du document. Ce protocole permettrait de préciser l'apport de la sémantique dans la tâche demandée. Il n'y avait donc plus d'items de contrôle comme dans la première expérience. La phonologie a encore été mesurée. L'orthographe n'a cependant pas été mesurée puisque les différences entre homophones et items de contrôle avaient été démontrées significatives dans l'expérience 1, et que cela ajoutait trop de contraintes à l'élaboration des stimuli non-mots pour vérifier un effet déjà mesuré. Les nouveaux exemples de catégories ont été choisis en fonction d'une expérience pilote sur 15 personnes indépendantes du bassin des nouveaux sujets. Durant la présentation de la tâche aux sujets, les expérimentateurs ont mis l'accent sur la rapidité d'exécution pour augmenter la variation du taux d'erreurs.

Les résultats obtenus dans l'expérience 2 reflètent les premiers résultats:tant pour le taux d'erreurs positives que pour les temps de catégorisation, il n'y a toujours pas de différence significative entre les mots homophones et les non-mots homophones. Il n'y a pas non plus de différence significative entre les homophones et les nouveaux exemples de catégories.

Ceci démontre qu'il n'y a pas d'interaction directe entre les représentations sémantiques et les représentations orthographiques. Le lecteur doit donc avoir recours aux représentations phonologiques pour accéder au sens du document. Ces résultats corroborent les résultats de plusieurs recherches[17], tout en étant plus valides.

Certains auteurs[18] ont affirmés que la fréquence jouait un rôle primordial dans le type de représentation activées. Des corrélations ont donc été tirées des résultats pour éclaircir cette question. Une seule corrélation s'est avérée significative:la fréquence des mots dont sont dérivés les homophones sont corrélés avec le taux d'erreurs positives des homophones correspondants. Ceci porte à croire que la fréquence joue un rôle uniquement du point de vue sémantique.

Finalement, une dichotomie encore couramment utilisée par Colthart et ses collaborateurs[19] concernant le type de traitement phonologique est insoutenable avec les résultats de Van Orden et de ses collaborateurs. Colthart et ses collaborateurs distinguent le traitement automatique du traitement par recodage, ce dernier référant à la décortication de mots inconnus ou rares puisque ces mots seraient absents du "lexique phonologique". Cette dichotomie tient selon ces auteurs tant pour les adultes que pour les enfants, bons ou mauvais lecteurs quoique de moindre importance pour les adultes. Les résultats des expériences que nous venons de voir ne révèlent aucune différence dans les temps de catégorisation des mots et des non-mots telle que le prédit cette dichotomie. Nous pouvons dès lors en conclure qu'il n'y a probablement qu'un seul type de traitement phonologique, sans pour autant être en mesure d'en préciser la nature.

Résumons maintenant les connaissances acquises par cet article. Ces expériences démontrent avec évidence l'importance des représentations phonologiques dans l'identification de mots. Posons ceci sur graphique. Le graphique 1 du document présente le modèle découlant de ces résultats.

---

[17] Perfetti et al., McCutchen et Perfetti, 1982; Altenberg et Smith Cairns.

[18] Seidenberg 1985a et 1985b; McCaan et Besner, 1987; Hudson et Bergman, 1985, Balota et Chumbley, 1985.

[19] 1988.

L'article de Van Orden et de ses collaborateurs portait sur des sujets unilingues anglophones. Or quelques chercheurs[20] présentent l'hypothèses que l'interaction entre les divers types de représentations ne soient pas les mêmes pour toutes les langues. Ainsi, selon eux, des langues comme le serbo-croate où la correspondance grapho-phonémique est élevée supposent une interaction plus élevée des représentations orthographiques et phonologiques. De même, des langues comme l'hébreu où la correspondance est basse supposent une interaction faible, voir nulle. Cette variation étant posée en un continuum, l'anglais et le français se retrouvent aux environs du milieu. Le lecteur anglophone ou francophone aurait donc un accès facultatif aux représentations phonologiques durant la lecture. Les résultats des expériences que nous venons de voir invalident ces résultats. De plus, la validité interne de études dans ce courant d'idée est en général relativement moyenne. D'ailleurs, les résultats de Van Orden et de ses collaborateurs ne sont pas les seuls à invalider cette hypothèse. Seidenbergen 1985 a démontré que même le chinois qui s'écrit en graphèmes idéographiques a besoin des représentations phonologiques pour accéder au sens d'un document écrit.

Ceci nous amène à parler des sujets bilingues. Malgré qu'aucune étude n'ait vérifié directement l'hypothèse d'interaction entre les représentations orthographiques et phonologiques, l'étude de plusieurs recherches peut se révéler un bon indicateur de cette interaction, compte tenu de l'interaction chez les sujets unilingues.

Lukatela et ses collaborateurs ont démontré dès 1978 que des orthographes différentes sont codifiées dans des ensembles de représentations distincts. Leurs sujets n'étaient pas bilingues, mais ils utilisaient deux systèmes graphémiques pour une même langue. Si nous n'étudions que le côté orthographique de leur analyse, la situation s'avère comparable à une population bilingue dont les systèmes orthographiques sont différents. Malgré la rigueur qu'ils ont observé, ils sont les seuls à avoir mesuré cette distinction. Aucun autre groupe de chercheurs à ma connaissance n'a tenté de répliquer à ces résultats comme ce devrait habituellement être le cas. Il faut donc attendre une confirmation avant de considérer ces résultats comme un acquis. Nous ne pouvons cependant pas les ignorer. Nous y reviendrons.

Altenberg et Smith-Cairns[21] ont pour leur part établi que les bilingue sont deux ensembles de contraintes phonotactiques distinctes, mais qu'ils sont interactifs quand les conditions expérimentales demandent de la rapidité. En effet, les chercheurs laissaient tout le temps désiré pour juger de la lexicalité de mots, les sujets bilingues ont des résultats similaires à ceux des unilingues, alors que lorsque la tâche requiert un jugement rapide, les résultats des bilingues sont significativement plus lents que ceux des unilingues. Ceci permet de poser que les sujets bilingues ont deux ensembles de contraintes phonotactiques distinctes. Il ne faut cependant pas en conclure que les bilingues ont sans équivoque deux ensembles de représentations phonologiques distincts. Malgré que le protocole d'Altenberg et Smith-Cairns soit valide, malgré que les résultats de certaines études en psychologie tendent à opter en cette direction démonstration n'en a pas été faite. Les contraintes phonotactiques ne sont qu'une part de traitement phonologiques. Ces résultats sont cependant de très solides indicat des hypothèses plus poussées. Or, il est plus difficile d'établir un modèle d'identificatio en lecture pour les sujets bilingues avec les informations actuelles. La recherche en bilinguisme met en cause un nombre faramineux de variables, ce qui ralentit grandement les travaux. Étant donné le nombre relativement plus restreint de psycholinguistes dont le travail porte sur les processus cognitifs des bilingues, le retard qui existe par rapport à la recherche sur les processus cognitifs des unilingues est léger.

---

[20] Frost et al., 1987; katz et Feldman, 1988; Lukatela et al , 1980.

[21] 1983.

[22] Paivio, et al., 1988; Paivio et Desrochers, 1980; Paivio et Lambert, 1981.

CONCLUSION

Mais les processus de lecture ne doivent pas être totalement différents des unilingues aux bilingues. Si les personnes appris une langue seconde après avoir appris à lire dans leur langue maternelle, les processus et les types de représentations engagées doivent rester semblablement les mêmes. Il y a là matière à recherche.

Les modèles de perception et de compréhension de la lecture nous permettent d'abord d'améliorer les protocoles de recherche dans l'ensemble de la psycholinguistique cognitive. En effet, lorsqu'une personne s'intéresse à n'importe quelle formes de représentations du langage, il pourrait dorénavant induire ses stimuli par écrit. Sachant qu'il nous est beaucoup plus facile de contrôler les facteurs tels le rythme, l'attention du sujet et la régularité des stimuli, il sera plus simple et possiblement plus valide dans certains cas d'avoir recours à des stimuli visuels plutôt qu'auditifs. Ces procédés ne sauraient toutefois diminuer l'importance des recherches sur la perception et la compréhension du langage oral. L'induction par écrit permet cependant un plus grand contrôle de variables parasites. Cette méthodologie ne se substituera jamais à l'élaboration du corpus oral puisque premièrement une grande proportion des humains ne sait pas écrire, et deuxièmement, elle ne permet pas d'étudier la structure interne des représentations phonologiques.

Finalement, ces recherches permettent de mieux comprendre les phénomènes de traitement pré-sémantique de la lecture. Ainsi, ayant la capacité d'élaborer des modèles psycho-informatiques plus valides et plus efficaces, il sera plus aisé de poursuivre certains travaux en intelligence artificielle. De ces travaux, nous pourrons à notre tour tirer des hypothèses sans doute fructueuses. L'amélioration de l'interface personne-machine ne peut passer que par une meilleure connaissance de l'interface personne-personne.

Bibliographie

ALTENBERG, E.P. et H. SMITH-CAIRNS, 1983, The effects of Phonotactic Constraints on Lexical Processing in Bilingual and Monolongual Subjects, *Journal of Verbal Learning and Verbal Behavior*, 22, p. 174-188.

BALOTA, D.A. et J.I. CHUMBLEY, 1985, The Locus of Word Frequency Effects in the Pronunciation Task: Lexical Acces and/or Production?, *Journal of Memory and Language*, 24, p. 84-106.

BANKS, W.P., E. OKA et S. SHUCARMAN, 1981, Recoding of Printed Wordsto Internal Speech:Does Recoding Comes Before Lexical Acces?,dans Tzeng, O.J.L. et H. Singer, *Perception of Print*, Hillsdale (N.J.), Lawrence Erlbaum Ass., p. 137-170.

BELLUGI, U., 1987, The Acquisition of a Spatial Language, dans Kessel, F., *The Development of Language and Language Researchers: Essays in Honor of Roger Brown*, Hillsdale (N.J.), Lawrence Erlbaum Ass.

BRYANT, P. et U. GOSWAMI, 1987, Phonological Awarenell and Learning to Read, dans Beech, J. et A. Colley, *Cognitive Approches to Reading*, New York, John Wiley & sons, p. 213-243.

COLTHART, V., V. LAXON, M. RICHARD et C. ELTON, 1988, Phonological Recoding in Reading for Meaning by Adults and Children, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(3), p. 387-397.

COWAN, N., M.D.S. BRAINE et L.A. LEAVITT, 1985, The Phonological and Metaphonological Representation of Speech: Evidence from Fluent Backward Talkers, *Journal of Memory and Language*, 24, p. 679-698.

CUTLER, A., J. MEHLER, D. NORRISET, J. SEGUI, 1987, Phoneme Identification and the Lexicon, *Cognitive Psychology*, 19, p. 141-177.

DAVIDSON, B.J., 1986, Activation of Semanticand Phonological Codes during Reading, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12(2), p. 201-207.

FROST, R., L. KATZET S., BENTIN, 1987, Strategies for Visual Recognition and Orthographic Depth: A Multilingual Comparison, *Journal of Experimental Psychology: Human Perception and Performance*, 13, p. 104-115.

HANSON, V.L. et C.A. FOWLER, 1987, Phonological Coding in Word Reading: Evidence from Hearing and Deaf Readers, *Memory and Cognition*, 15(3), p. 199-207.

HUDSON, P.T. et BERGMAN, M.W., 1985, Lexical Knowledge in Word Recognition: Word Length and Word Frequency in Naming and Lexical Decision Tasks, *Journal of Memory and Language*, 24, p. 46-58.

KATZ, L. et L. FELDMAN, 1983, Relation Between Pronunciation and Recognition of Printed Words in Deep and Shallow Orthographies, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 9(1), p. 157-166.

LUKATELA, G., D. POPADIC, P. OGNJENOVIC et M.T. TURVEY, 1980, Lexical Decision in a Phonologically Shallow Orthography, *Memory and Cognition*, 8(2), p. 124-132.

LUKATELA, G., M.D. SAVIC, P. OGNJENOVIC et M.T. TURVEY, 1980, On the Relation Between Processing the Roman and the Cyrillic Alphabets: A Preliminary Analysis with Bi-Alphabetical Readers, *Languages and Speech*, 21(2), p. 113-141.

MARSHALL, J.C., 1987, The Cultural and Biological Context of Written Languages: Their Acquisition, Deployment and Breakdown, dans Beech, J.R. et A. Colley. *Cognitive Approches to Reading*, New York, John Wiley & sons, p. 15-30.

MARTINET, A., 1970, *Éléments de linguistique générale*, Paris, Armand Colin.

McCANN, R.S. et D. BESNER, 1987, Reading Pseudohomophones: Implications for Models of Pronunciation Assembly and the Locus of Word Frequency Effects in Naming, *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), p. 14-24.

McCUTCHEN, D. et C.A. PERFETTI, 1982, The visual Tonge-Twister Effect: Phonological Activation in Silent Reading, *Journal of Verbal Learning and Verbal Behavior*, 21, p. 672-687.

NORRIS, D. et G. BROWN, 1985, Race Models and Analogy Theories: a Dead Heat? Reply to Seidenberg, *Cognition*, 20, p. 155-168.

PATEL, P.G. et H.V. SOPER, 1987, Acquisition of Reading and Spelling in a Syllabo-Alphabetic Writing System, *Language and Speech*, 30(1), p. 69-81.

PERFETTI, C.A., L.C. BELLET, S.M. DELANEY, 1988, Automatic (Prelixical) Phonetic Activation in Silent Word Reading: Evidence from Backward Masking, *Journal of Memory and Language*, 27, p. 59-70.

SEIDENBERG, M.S., 1985a, The Time Course of Phonological Code Activation in Two Writing Systems, *Cognition*, 19, p. 1-30.

------ 1985b, Constraining Models of Word Recognition, *Cognition*, 20, p. 169-190.

VAN ORDEN, G.C., J.C. JOHNSTON et B.L. HALE, 1988, Word Identification on Reading Proceeds From Spelling to Sound to Meaning, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(3), p. 371-386.

VELLUTINO, F.R., D.M. SCANLON, 1987, Phonological Coding, Phonological Awareness, and Reading Ability: Evidence from a Longitudinal and Experimental Study, *Merrill-Palmer Quaterly*, 33(3), p. 321-363.

WAGNER, R.K. et J.K. TORGESEN, 1987, The Nature of Phonological Processing and its Causal Role in the Acquisition of Reading Skills, *Psychological Bulletin*, 101(2), p. 192-212.

WATERS, G.S., M.K. KOMODA et T.Y. ARBUCKLE, 1985, The Effects of Concurrent Tasks on Reading: Implications for Phonological Recoding, *Journal of Memory and Language*, 24, p. 27-45.

LA PHONÉTISATION AUTOMATIQUE DE TEXTES FRANÇAIS[1]

Éric Laporte,
CERIL

Phonétisation, conversion graphème-phonème, phonémisation, transcription orthographique-phonétique, transcription phonétique automatique ... : le terme n'est pas fixé. Quant à la notion, elle est simple: la phonétisation automatique de textes consiste à produire automatiquement une transcription phonétique d'un texte écrit. Depuis 1967 (B. Pratt et G. Sylva), il existe des programmes de phonétisation automatique de textes français. L'objectif de cet article est de faire le point sur les méthodes utilisées dans ce domaine si souvent abordé. En particulier, nous insisterons sur une distinction entre la phonétisation automatique proprement dite et les aides automatiques à la phonétisation, qu'on pourrait aussi qualifier de « phonétisation semi-automatique ». Sous le terme de phonétisation, nous entendons aussi bien celle qui se fait par consultation d'un dictionnaire électronique que celle qui se fait par l'intermédiaire d'un système de règles, ou «phonétiseur par règles».

1. LES BESOINS

Pourquoi phonétiser automatiquement des textes écrits? Cette opération répond en fait à plusieurs besoins industriels à plus ou moins long terme et débouche donc sur autant d'applications informatiques dans le domaine des industries de la langue. Il s'agit en fait de tous les cas d'utilisation de transcriptions phonétiques dans le traitement automatique des langues naturelles: en effet, chaque fois qu'on a à manipuler explicitement des transcriptions phonétiques, on doit les élaborer tôt ou tard , à partir des mots ou des textes orthographiés normalement.

1.1 Construction de dictionnaires électroniques

Parmi les besoins auxquels répond la phonétisation, la construction de dictionnaires électroniques semble être celui qui a le plus d'avenir: il recèle des potentialités technologiques et il a aussi un intérêt linguistique intrinsèque. Il s'agit d'introduire des transcriptions phonétiques dans un dictionnaire électronique, et ce automatiquement ou semi-automatiquement. Cette méthode n'est pas rentrée dans les moeurs, ne serait-ce que parce qu'il existe peu de dictionnaires électroniques qui donnent la prononciation. Elle commence cependant à être utilisée. Le dictionnaire phonétique obtenu est destiné à servir dans des projets de reconnaissance de la parole (F. Néel, M. Eskenazi et J. Mariani, 1986; M.R. Carapiperis, 1988). W. Daelemans (1988) mentionne l'utilisation d'un phonétiseur de mots néerlandais pour la construction d'un dictionnaire imprimé. Dans ces trois cas et dans d'autres, il s'agit en fait d'aides automatiques à la phonetisation: dans des cas d'ambiguïté tels que la prononciation ou la non-prononciation du *s* final dans les noms *tas* [ta] et *atlas* [atlas], de même que pour les emprunts étrangers tels que le nom *poster* [pɔstɛr], un phonétiseur donne des résultats approximatifs tels que [ta], [atla] ou [pɔste] qui sont ensuite vérifiés et corrigés à la main par un phonéticien. Le dictionnaire DELAP[2] (E. Laporte, 1988) est engendré d'une façon entièrement automatique à partir d'un autre dictionnaire, le DELASP, qui contient les formes orthographiques des mots

---

[1]Centre d'études et de recherches en informatique linguistique, 17, cours Blaise-Pascal, 91000 EVRY.

[2]Dictionnaire électronique du LADL pour la phonémique.

ainsi que des informations codées sur la façon dont leur orthographe doit être interprétée en cas d'ambiguïté, par exemple *sa(22.)s* pour *sas*. Ce dispositif permet la maintenance du DELAP, qui est assurée par la maintenance du DELASP et par une phonétisation automatique.

Il est prévisible que les phonétiseurs seront de plus en plus utilisés pour l'élaboration, l'extension et la maintenance de dictionnaires électroniques phonétiques. Quant à la fonction de ces derniers, ils permettent à leur tour d'effectuer une phonétisation automatique par consultation. De plus ils ont un intérêt linguistique car ils constitueront une référence dans l'attribution de transcriptions phonétiques aux mots. Notons une conséquence de cette fonction: un dictionnaire électronique étant un instrument de référence, la phonétisation qui permet de le construire doit être fiable.

1.2. Synthèse de parole par l'intermédiaire de textes phonétiques

La synthèse de messages oraux est l'application la plus évidente et celle qui a été mise en oeuvre depuis le plus longtemps. Il ne s'agit ici que des configurations où la synthèse de la parole se fait par l'intermédiaire d'un texte présenté sous forme d'une transcription phonétique et transmis à un synthétiseur de parole qui élabore alors un signal de parole artificiel. Nous ne parlons donc pas des configurations dans lesquelles les mots, les phrases ou les textes qui constituent les messages oraux sont préenregistrés et stockés séparément sous forme d'un signal de parole codé ou plus ou moins comprimé. Cette dernière situation ne relève d'ailleurs pas de la synthèse de parole proprement dite, mais de la compression de la parole. Notons que la synthèse de parole par l'intermédiaire de transcriptions phonétiques s'impose lorsque le nombre de messages s'accroît. Or, de nombreuses applications industrielles potentielles impliquent de pouvoir synthétiser une grande variété de messages: de l'ordre de 10^6, 10^{10} voire 10^{20} messages distincts. Au-delà d'un certain seuil, les messages ne pourront plus être enregistrés séparément, ni même réalisés par combinaison de groupes de mots, et on devra avoir recours à l'intermédiaire de transcriptions phonétiques.

Pour que les sorties vocales soient facilement intelligibles, les transcriptions doivent être précises et exactes. Ce résultat pourrait *a priori* être obtenu de deux façons: soit à l'aide d'un phonétiseur semi-automatique, les textes phonétiques produits devant être ensuite vérifiés et corrigés à la main par un phonéticien, soit à l'aide d'un phonétiseur automatique fiable. La première solution est tout à fait réaliste. La deuxième est plus difficile à mettre en oeuvre, surtout pour les langues dont l'orthographe est très ambiguë, comme le français et l'anglais.

1.3. Aide à la correction orthographique par phonétisation

Il s'agit ici d'une utilisation de transcriptions phonétiques dans des systèmes où la parole n'intervient pas. Des fautes détectées dans des textes écrits, par exemple *racompter* pour *raconter*, peuvent être corrigées semi-automatiquement. Une des aides envisageables consiste à extraire d'un dictionnaire électronique les formes correctes qui se prononcent comme la forme erronée, dans ce cas des formes du verbe *raconter*. Cette méthode a été utilisée pour le français (J. C. Marcovici, 1987; E. Laporte, 1988) et pour le néerlandais (B. Van Berkel et K. De Smedt, 1988). Les formes correctes sont retrouvées dans le dictionnaire par l'intermédiaire d'une transcription phonétique qui revêt un intérêt particulier si on considère l'importance des procédures de correction d'erreurs pour tout traitement automatique de textes écrits par des utilisateurs, car ces textes ne sont jamais exempts d'erreurs.

2. LES CARACTÉRISTIQUES D'UN PHONÉTISEUR

De nombreux phonétiseurs ont été conçus et réalisés pour différentes langues. Tous ne sont pas aussi fiables et n'ont pas les mêmes performances. Ils diffèrent par un certain nombre de

paramètres. Nous allons passer en revue les plus significatifs de ces paramètres, d'abord pour caractériser les phonétiseurs déjà existants, mais aussi pour les situer par rapport aux besoins industriels auxquels des phonétiseurs devront répondre. Nous examinerons notamment une importante contrainte à respecter: assurer la fiabilité de la phonétisation, du moins dans deux situations, la construction de dictionnaires électroniques et la production de textes phonétiques en vue des sorties vocales. Dans le cadre de la correction orthographique par phonétisation, la fiabilité est moins importante, car il ne s'agit alors que de produire des suggestions de correction qui seront de toutes façon soumises au choix d'un utilisateur. Nous évoquerons en priorité les caractéristiques liées à la fiabilité.

2.1. Fonctionnement automatique ou semi-automatique

Considérons les situations industrielles ou de recherche qui nécessitent une phonétisation fiable. Dans ce cadre, un phonétiseur ne peut être qualifié d'entièrement automatique que s'il donne par lui-même des résultats fiables. Dans le cas contraire, ces résultats doivent être revus et corrigés à la main, et l'ensemble de l'opération prend l'aspect d'un processus semi-automatique: une transcription approximative est élaborée automatiquement puis achevée à la main. La partie manuelle de l'opération demande alors un personnel spécialisé, mais si les résultats de la partie automatisée sont suffisamment bons, l'ensemble est plus rentable que de produire les transcriptions entièrement à la main.

En français, ce sont des méthodes soit entièrement manuelles, soit semi-automatiques qui ont été utilisées jusqu'ici pour obtenir des transcriptions exactes. En effet, des phonétiseurs par règles fiables faisaient entièrement défaut, et les dictionnaires électroniques ne font que commencer à se développer. La phonétisation semi-automatique peut encore rendre des services appréciables, mais l'apparition d'une phonétisation entièrement automatique, lorsqu'elle est réalisable, permettrait de réduire le temps de production des textes phonétiques et de s'acheminer vers une synthèse de parole « en direct », par exemple.

2.2. Maintenance

Un produit de traitement automatique des langues naturelles, qu'il s'agisse d'un logiciel qui comporte des données linguistiques ou d'un système de données linguistiques, nécessite une maintenance, c'est-à-dire un entretien qui consiste à corriger les erreurs, à élaborer des extensions qui correspondent à l'évolution des besoins des utilisateurs, et à répercuter l'évolution de la langue, qui est rapide dans les domaines techniques. L'entretien d'un phonétiseur consiste donc, hormis la correction d'erreurs, à prendre en compte de nouveaux mots, c'est-à-dire des mots auxquels on n'avait pas pensé et des néologismes. Le coût de cette activité de maintenance n'étant pas négligeable, la facilité d'entretien d'un phonétiseur est une donnée significative. Les cas les plus favorables de ce point de vue sont les suivants:

- si la phonétisation se fait par consultation d'un dictionnaire. La maintenance se ramène alors à l'entretien du dictionnaire, pour lequel il existe des méthodes;
- si le phonétiseur est couplé avec un dictionnaire électronique. Après toute modification, on peut en évaluer les conséquences par comparaison avec le dictionnaire;
- si le phonétiseur peut être construit automatiquement à partir d'un dictionnaire électronique phonétique. Des études sont en cours sur cette intéressante possibilité.

2.3. Taille du vocabulaire

Tout phonétiseur de mots peut être caractérisé par l'ensemble des mots qu'il transcrit correctement, ensemble que l'on peut appeler le vocabulaire du phonétiseur. La taille du vocabulaire et la facilité de la maintenance sont liées; plus le vocabulaire est étendu, plus la maintenance prévisible est réduite. Elle ne se réduit jamais à zéro, en raison de l'évolution de la langue, mais construire un phonétiseur sur un vocabulaire limité équivaut à reporter sur la maintenance le problème du vocabulaire éludé lors de la construction, et donc à rendre d'autant plus coûteux le développement du phonétiseur dans une utilisation industrielle éventuelle.

C'est ce qui s'est produit pendant longtemps. En l'absence de dictionnaires électroniques, il n'était pas possible de construire un phonétiseur sur un vocabulaire étendu, ni même d'évaluer la taille du vocabulaire du phonétiseur. L'utilisation de dictionnaires pour la phonétisation débute en 1979 pour le français (G. Tep), mais le dictionnaire est encore rudimentaire: il comporte 2000 formes. Plusieurs contributions plus récentes (G. Pérennou et M. de Calmès, 1986; E. Laporte, 1986; P. Trescases et M. Crocker, 1988), présentent une évaluation de la taille du vocabulaire et aussi montrent le souci d'étendre celui-ci pour rendre les résultats plus fiables et faciliter la maintenance. On note la même tendance pour d'autres langues (W. Daelemans, 1988).

2.4. Rapidité

Le temps mis par un système à phonétiser un texte nous semble moins significatif que les impératifs de performance et de fiabilité que nous venons de mentionner. En effet, lorsque les résultats sont approximatifs, ils doivent être revus à la main, ce qui est plus long et plus coûteux que les plus lents des phonétiseurs. L'efficacité d'un phonétiseur n'est donc une donnée essentielle que lorsque la fiabilité importe peu, par exemple pour la correction d'erreurs détectées. Même dans ce cas, elle évolue vite avec les progrès technologiques et les progrès dans la rapidité de consultation des bases de données; elle constitue en fait un problème indépendant des autres et qui relève de techniques distinctes, aucunement spécifiques de la phonétique.

2.5. Nombre de solutions en cas d'ambiguïté

Lorsque l'orthographe est ambiguë, par exemple dans le cas ds emprunts étrangers qui ne se prononcent pas comme ils s'écrivent, comme *poster* ou *charter*, il n'est pas toujours facile de déterminer automatiquement la transcription, mais il est toujours possible de produire un petit nombre de transcriptions parmi lesquelles se trouve la transcription c orrecte, par exemple: [pɔste], [pɔstɛr], [pɔstœr]. Cette possibilité, jusqu'à présent, n'a été mise à profit que dans quelques systèmes (F. Néel et al., 1986; E. Laporte, 1988). Elle est pourtant assez réaliste si deux conditions sont réunies: (1) que la fiabilité importe peu et (2) que les résultats du phonétiseur soient soumis à un opérateur humain qui choisisse parmi les solutions.

2.6. Précisions dans les transcriptions

Toutes les applications requièrent des transcriptions précises, sauf la correction orthographique, pour laquelle des transcriptions moyennement precises donnent de meilleurs résultats. Par exemple, si [œ̃] et [ɛ̃] sont confondus dans les transcriptions, la forme *emprunter* pourra être proposée pour corriger la forme *empreinter*. Certains phonétiseurs du français, conçus pour la correction orthographique, donnent ainsi des transcriptions relativement imprécises, par exemple dans lesquelles [s] et [z] sont confondus.

2.7. Variations phonétiques libres

Nous parlons de variations phonétiques libres lorsque deux prononciations sont interchangeables, par exemple pour le verbe *lier* qui peut se prononcer [lje] ou [lije]. Cette situation est courante pour le français: on peut citer, outre l'exemple précédent, l'effacement facultatif de certains *e* muets (*projeter*), certaines lettres orthographiques doubles prononcées soit doubles soit simples (*illégal*), certaines liaisons facultatives (*Luc est indemne*), etc. Ces problèmes sont connus mais la description de leur extension lexicale en est à ses balbutiements. En face d'une variation libre, trois solutions sont envisageables:

- choisir arbitrairement une des variantes et produire seulement la transcription correspondante;
- produire toutes les transcriptions équivalentes ou une sélection d'entre elles (F. Néer et al., 1986);
- produire non plus des transcriptions phonétiques mais une transcription phonémique, plus abstraite, dont on peut ensuite déduire automatiquement les variantes phonétiques libres (F. Dell et M. Plénat, 1985; E. Laporte, 1988).

La première solution est la plus simple: elle a l'avantage de pouvoir être appliquée en l'absence de toute connaissance sur les variations phonétiques libres. En raison du peu de données dont on dispose sur le sujet, c'est cette solution qui est souvent adoptée. Elle se justifie pleinement dans le cadre de la synthèse de sorties vocales: lorsque toutes les variantes sont interchangeables, il faut en choisir une. La deuxième solution, et la troisième qui met en jeu des formalismes plus élaborés pour retrouver les mêmes informations, se justifient pour les autres applications. En effet, un dictionnaire électronique étant un instrument de référence, il est naturel que les variations phonétiques libres y soient représentées, soit explicitement, soit par l'intermédiaire d'un formalisme abstrait. Cette spécification des variations libres sera notamment indispensable à la reconnaissance d'entrées vocales variées.

Les divers phonétiseurs diffèrent par la nature et la quantité des variations phonétiques prises en compte dans les transcriptions. Le manque de données systématiques sur les faits phonétiques est sensible ici.

Ces variations posent le problème du niveau d'abstraction des transcriptions. Ce problème n'a guère été abordé, jusqu'ici, dans le cadre du traitement automatique; pourtant, suivant leurs utilisations, les transcriptions doivent être à des niveaux différents. Ainsi, les transcriptions transmises à un synthétiseur de parole doivent être purement phonétiques, c'est-à-dire spécifier une prononciation bien précise. Au contraire, un dictionnaire électronique destiné à produire toutes les formes fléchies et leurs variantes phonétiques comportera plutôt des transcriptions phonémiques. Cette notion de niveau d'abstraction rejoint la différence entre phonétique et phonémique ou phonologie, qui est fondamentale en linguistique.

2.8. Phonétiseurs de mots, de termes ou de phrases

On peut parler d'un phonétiseur de mots lorsque les mots sont phonétisés indépendamment les uns des autres, par exemple pour la construction d'un dictionnaire ou pour la correction orthographique de mots. Cette méthode est insuffisante pour transcrire des phrases, pour plusieurs raisons sur lesquelles nous reviendrons dans les trois sections qui suivent. Or les sorties vocales à prévoir seront essentiellement des phrases, telles que *Le Cubain N... a remporté aujourd'hui la médaille de bronze de saut en hauteur*, ou des groupes nominaux, tels que *dépassement de capacité dans la cuve principale*

2.9. Variations phonétiques conditionnées

Les variations phonétiques conditionnées s'opposent aux variations libres en ce que les variantes ne sont pas interchangeables: elles ont des conditions d'emploi différentes. Or ces variations ne sont pas toujours indiquées explicitement dans l'orthographe. Ainsi, le pronom *les* reçoit une liaison obligatoire dans *Luc les a*, mais cette liaison est interdite dans *Luc les prend*: on peut donc dire que le pronom les admet deux variantes phonétiques conditionnées. Ces variantes sont plus préoccupantes que les variations libres, car on est obligé d'en tenir compte pour produire des transcriptions exactes des phrases. Or elles jouent un rôle non négligeable dans l'intelligibilité des phrases: une sortie vocale dans laquelle des liaisons obligatoires sont omises ou des liaisons interdites sont faites est difficile à comprendre. Ces variations sont conditionnées par divers paramètres tels que des conditions lexicales, grammaticales et syntaxiques, comme le montre la comparaison entre

Luc a six ans. et *Luc en a six à la main.*

D'une part, ces conditionnements sont mal connus; d'autre part, ils font appel à des informations qui ne sont pas représentées explicitement dans les textes.

2.10. Paramètres prosodiques

Dans les phrases destinées à servir de sorties vocales, l'intonation joue un rôle dans l'intelligibilité. L'intonation résulte des valeurs des paramètres prosodiques: la hauteur du son (ou fréquence du fondamental), l'intensité (ou volume) et la durée des éléments phonétiques (segments, syllabes, pauses), ainsi que des variations de ces paramètres en fonction du temps. C'est pourquoi des sorties vocales ne sont plus concevables sans une détermination des variations de ces paramètres, ce qui implique de reconnaître la structure prosodique des phrases. Cette reconnaissance suppose à son tour une analyse grammaticale et syntaxique des phrases, et donc la consultation de dictionnaires morphologiques, grammaticaux et syntaxiques.

2.11. Analyse grammaticale

Les deux problèmes que nous venons de mentionner: variations phonétiques conditionnées et paramètres prosodiques, rendent nécessaire la connaissance d'informations grammaticales et syntaxiques sur les phrases à phonétiser en vue de sorties vocales. A cette nécessité s'ajoute celle de désambiguïser les homographes non homophones tels que *couvent*. Ceci est connu depuis longtemps, mais il est difficile de préciser quelles informations grammaticales seraient suffisantes pour établir automatiquement des transcriptions fiables et exactes. Elles comprendraient au moins les éléments suivants:

- l'analyse des déterminants du groupe nominal pour effectuer certaines liaisons obligatoires ;
- l'analyse des particules préverbales pour en effectuer d'autres;
- la délimitation des groupes nominaux et des rapports syntaxiques entre eux pour déterminer la structure prosodique.

Etant donné l'étroitesse des relations entre les différentes sous-tâches de l'analyse syntaxique, ces trois nécessités en supposent d'autres, notamment la détermination des catégories grammaticales des mots après la consultation d'un dictionnaire grammatical. Finalement, les

opinions divergent en ce qui concerne la précision et la quantité des informations grammaticales à recueillir. Souvent, les auteurs qui en font une évaluation *a priori* minimisent cette précision et cette quantité, alors que ceux qui les évaluent *a posteriori* par des tests sur un système existant constatent généralement qu'une analyse syntaxique plus approfondie améliorerait la qualité des sorties vocales (J. Allen, M. Sh. Hunnicutt et D. Klatt, 1987, par exemple).

Comme la phonétisation automatique était conçue à l'origine pour la synthèse de la parole, de nombreux phonétiseurs de phrases comportent une phase d'analyse grammaticale destinée à recueillir ces informations. La nature et la quantité des informations obtenues dépendent d'abord du dictionnaire grammatical utilisé. En français, G. Tep (1979) est le premier à utiliser un dictionnaire grammatical. A cette époque, les dictionnaires électroniques grammaticaux actuels, beaucoup plus étendus, n'existaient pas encore, ce qui incitait certains concepteurs de phonétiseurs à tenter de réunir des informations grammaticales sur les mots sans avoir recours à un dictionnaire grammatical. Mais les catégories grammaticales et les traits flexionnels ne se déduisent pas de la forme des mots par des règles simples, ni même par des règles fiables, d'où les difficultés rencontrées pour mettre au point ces systèmes (B. Prouts, 1979; N. Catach, 1984; M. Divay, 1984). Des solutions approchées et des listes partielles n'ont permis d'obtenir que des données fragmentaires et peu fiables.

L'apparition de dictionnaires morphologiques et grammaticaux tels que le DELAP résout cette partie du problème, mais ne suffit pas à recueillir toutes les informations grammaticales et syntaxiques nécessaires, ni même à résoudre les ambiguïtés lexicales. Ainsi, l'analyse syntaxique nécessaire pour cela constitue toujours un obstacle. Le phonétiseur le plus avancé dans cette direction à l'heure actuelle semble être celui de J. Allen, M. Sh. Hunnicutt et D. Klatt (1987) pour l'anglais.

Rappelons que cette difficulté n'apparaît pas si les sorties vocales sont le résultat d'une génération automatique du texte ou d'une traduction automatique, car l'analyse syntaxique du texte produit est alors superflue.

3. RÉALISATIONS ET POSSIBILITÉS

Les nombreux phonétiseurs par règles réalisés depuis plus de vingt ans sont construits à partir de règles générales sur la correspondance entre l'orthographe et la prononciation. Ils s'opposent aux dictionnaires phonétiques, plus récents, dans lesquels chaque mot est traité séparément. Nous allons examiner les uns comme les autres.

3.1. Phonétiseurs par règles existants

A examiner les nombreux phonétiseurs par règles existants, on a l'impression que chacun représente une tentative de compromis entre des contraintes incompatibles.

D'une part, nous avons vu la nécessité d'assurer la fiabilité des transcriptions et la facilité de la maintenance, de prendre en compte certaines variations phonétiques, et d'exploiter des informations grammaticales. Ces contraintes sont d'autant plus pressantes que la majorité des phonétiseurs par règles ont été conçus en vue de la synthèse de sorties vocales.

D'autre part, un certain nombre de circonstances ont longtemps fait obstacle à la réalisation de ces objectifs. L'absence de dictionnaires électroniques empêchait à la fois de construire un phonétiseur sur un vocabulaire étendu, d'évaluer la taille du vocabulaire des systèmes, de recueillir des informations grammaticales sur les mots des textes, et même de réunir des informations

systématiques sur les variations phonétiques et leur extension lexicale. Ce manque se comble peu à peu. Un autre facteur qui incite à des compromis est le souci de limiter les temps de calcul. Chaque phonetiseur par règles respecte ainsi un « dosage » particulier: l'un a un vocabulaire plus étendu que les autres , mais ne tient pas compte des variations phonétiques; un autre produit certaines de ces variations, mais se passe d'informations grammaticales; etc. Chacun de ces « dosages » débouche sur un problème différent, résolu séparément dans le cadre d'une réalisation spécifique. C'est probablement pour cette raison que, depuis 1967, des réalisations si nombreuses ont vu le jour dans ce domaine, et qu'on en construit régulièrement de nouvelles.

Toutefois, cette stratégie de compromis est défavorable à des recherches systématiques sur les obstacles rencontrés pour répondre aux besoins industriels qui s'annoncent: lever ces obstacles signifierait en priorité, d'une part, combler un manque de données linguistiques formelles, et d'autre part améliorer la consultation rapide de dictionnaires électroniques étendus.

3.2. Recherches actuelles

Parmi les contraintes que nous avons énumérées dans la section 2, la plupart des recherches actuelles sur la phonétisation par règles portent en fait sur deux points.

Il s'agit d'abord des stratégies de transcription. Les phonétiseurs utilisent parfois des informations morphologiques sur les mots, ou des informations syllabiques obtenues par une analyse de l'orthographe des mots. Dans d'autres langues que le français, on utilise également la position de l'accent tonique. Ces stratégies se prêtent à de nombreuses variantes et combinaisons: les informations morphologiques, syllabiques et accentuelles ne sont pas toujours obligatoires, peuvent être explicites ou implicites, et peuvent être recueillies ou exploitées dans un ordre chronologique particulier ou par diverses méthodes. Pour la phonétisation de l'anglais, de l'allemand et du néerlandais, des informations morphologiques et accentuelles sont nécessaires, ce qui n'est le cas qu'exceptionnellement en français (par exemple dans *antiatomique*).

L'autre point souvent abordé est celui de la structure du logiciel, du formalisme dans lequel les règles sont exprimées, et des relations entre les règles et le reste du logiciel. Ces règles procèdent à une reconnaissance de formes et leur présence apparente l'ensemble à un système expert, qui peut être implanté de multiples façons, par exemple en ce qui concerne l'ordre d'application des règles; l'utilisation d'opérateurs booléens dans les règles; la représentation des lettres et des phonèmes soit par des symboles uniques, soit par des faisceaux de traits; la possibilité d'utiliser un même formalisme d'expression des règles pour plusieurs langues, ou pour plusieurs théories phonologiques; le degré de dépendance entre les règles et l'algorithme qui les applique; l'utilisation d'un réseau connexionniste... Toutes ces variantes de programmation permettent d'implanter en machine, sous des formes diverses, les règles de transcription qui sont toujours du même type mathématique: des applications locales, c'est-à-dire des applications dans lesquelles chaque élément de la chaîne de caractères de départ est traduit en fonction de sa valeur et d'un contexte limité.

L'enjeu de ces recherches se situe dans la rapidité d'exécution et dans la lisibilité et la maintenabilité des programmes. Toutefois, ces enjeux dépendent également d'autres facteurs qui jouent aussi un rôle déterminant. En effet, remarquons qu'il ne s'agit pas ici de la phonétisation par consultation de dictionnaire mais de la phonétisation par règles, dont l'utilisation se restreint de plus en plus aux seules procédures semi-automatiques de construction et de maintenance de dictionnaires et de correction par phonétisation. Or, comme la rapidité de l'ensemble du processus est de toute façon limitée par sa partie manuelle, la rapidité de la partie automatisée importe moins qu'il n'y paraît. Quant à la facilité de maintenance des phonétiseurs par règles, elle dépend plus de l'étendue du vocabulaire pris en compte que de la formulation des règles. Les critères informatiques de qualité, notamment la fiabilité, la rapidité et la maintenabilité, mettent en jeu,

dans le cas qui nous occupe, les données linguistiques. On ne peut donc pas considérer la qualité informatique des systèmes en faisant abstraction des données linguistiques qu'ils manipulent, et notamment de leur précision, de leur quantité ou de leurs possibilités d'extension.

En ce qui concerne les recherches menées sur le contenu des dictionnaires électroniques phonétiques, elles concernent

- le repérage, la description et la représentation formelle des variations phonétiques à l'intérieur du français standard,
- l'extension progressive du vocabulaire pris en compte, notamment à travers l'analyse automatique de vastes corpus,
- la mise en relation des informations phonétiques et des informations grammaticales et morphologiques, y compris les ambiguïtés grammaticales.

CONCLUSIONS

La production de textes phonétiques nous semble confrontée à trois défis qui commencent à être explorés mais qui impliquent des recherches descriptives, méthodologiques et théoriques:

- atteindre un objectif de fiabilité, qui passe notamment par la facilité de la maintenance,
- arriver à des connaissances systématiques sur les variations phonétiques, et les façons de les manipuler dans des systèmes informatiques,
- et, à plus long terme, exploiter les relations entre les faits phonétiques et les faits grammaticaux et syntaxiques.

Si l'ampleur de ce travail ne doit pas être sous-évaluée, les moyens à mettre en oeuvre sont clairs. La fiabilité des transcriptions et la facilité de la maintenance dépendent de la construction et de l'exploitation de dictionnaires électroniques. La description des variations phonétiques ne peut se faire que dans le cadre d'un dictionnaire, car chaque variation est limitée à un ensemble de mots, qui peut être considérable. Construire des méthodes pour représenter et manipuler ces variations consiste en fait à élaborer non plus des dictionnaires phonétiques, mais des dictionnaires phonémiques munis de logiciels qui permettent de retrouver les transcriptions exactes à partir des représentations phonémiques. Quant à l'exploitation des informations grammaticales, elle se conçoit dans le cadre plus général de l'analyse syntaxique de textes, qui a devant elle un avenir fructueux.

Bibliographie

J. ALLEN, M.Sh. HUNNICUTT et D. KLATT, 1987, *From text to speech. The MITalk system*, Cambridge: Cambridge University Press.

B. PRATT et G. SYLVA, 1967, *PHONTRS. Transcribing French Text*, Monash University, Australie.

M.R. CARAPIPERIS, 1988, *Rapport de stage*, IBM France, Paris.

N. CATACH, 1984, *La phonétisation automatique du français*, Paris: CNRS.

W. DAELEMANS, 1988, "Grafon: a Grapheme-to-Phoneme Conversion System for Dutch", *Proceedings of Coling 1988*, Budapest.

F. DELL et M. PLENAT, 1985, "Semi-voyelles et consonnes finales en français", *Rapport interne du GRECO "Communication parlée"*, Toulouse.

M. DIVAY, 1984, *De l'écrit vers l'oral, ou contribution à l'étude des traitements de textes écrits en vue de leur prononciation sur synthétiseur de parole*, Thèse d'état, Université de Rennes.

E. LAPORTE, 1986, "Applications de la morpho-phonologie à la production automatique de textes phonétiques", *Actes du séminaire "Lexiques et traitement automatique des langages"*, Université Paul-Sabatier, Toulouse.

E. LAPORTE, 1988, *Méthodes algorithmiques et lexicales de phonétisation de textes*, Thèse de doctorat, Université Paris 7.

J.C. MARCOVICI, 1987, *The Electronic Directory Service*, Rapport de la Direction générale des télécommunications.

F.NEEL, M. ESKENAZI, J. MARIANI, 1986, "Module de traduction phonétique avec variantes", *Actes du séminaire "Lexiques et traitement automatique des langages"*, Université Paul-Sabatier, Toulouse.

G. PERENNOU et M. de CALMES, 1986, "BDLEX: une base de données et de connaissances du français parlé", *Actes du séminaire "Lexiques et traitement automatique des langages"*, Université Paul-Sabatier, Toulouse.

B. PROUTS, 1979, "Traduction phonétique de textes écrits en français", *Actes des 10e journées d'étude sur la parole*, Grenoble.

G. TEP, 1979, "Système de génération des phrases phonétiques", *Actes des 10e journées d'étude sur la parole*, Grenoble.

P. TRESCASES et M. CROCKER, 1988, "Linguistic Contributions to Text-to-Speech Computer Programs for French", *Proceedings of Coling 1988*, Budapest.

B. VAN BERKEL et K. DE SMEDT, 1988, "Triphone Analysis: A Combined Method for the Correction of Orthographical and Typographical Errors", *Proceedings of the 2nd ACL Applied Conference*.

SYSTÈME D'ANALYSE DE CONTENU ASSISTÉE PAR ORDINATEUR (SACAO)

Jules Duchastel, Luc Dupuy, et François Daoust
UQUAM

1. LE PROJET

Le projet SACAO[1] (Système d'Analyse de Contenu Assistée par Ordinateur) vise l'intégration systématique de procédures existantes ou nouvelles de lecture assistée de données textuelles. Il s'agit d'offrir à des utilisateurs, dans un environnement logiciel relativement intégré, divers modules de description, d'exploration et d'analyse de données textuelles, tout en leur laissant le soin de paramétrer ces procédures en fonction de leurs propres hypothèses de lecture. Ces procédures ne comportent qu'un minimum de préconstruction théorique et facilitent un maximum d'itérativité entre leur application et l'analyse du texte. L'intégration est assurée par l'établissement de liens informatiques entre fichiers comportant des structures de données communes. Cet environnement convivial répond ainsi aux besoins différents de diverses catégories d'usagers confrontés aux problèmes d'analyse de données textuelles.

1.1. Le problème:

L'évolution récente de l'informatique et le développement d'un domaine aux contours encore imprécis, le Traitement Automatique des Langues (TAL), n'interpellent pas seulement la communauté des chercheurs de diverses disciplines, mais aussi celle, beaucoup plus large, des usagers de la langue écrite (documentalistes, gestionnaires, décideurs, etc.). La micro-informatique a pénétré aussi bien les lieux de savoirs que les organisations, favorisant de nouvelles habitudes de travail et générant de facto une quantité croissante d'information textuelle sur support magnétique. Celle-ci se retrouve dans des banques de données ou des répertoires de textes qui demeurent pour l'instant sous-exploités.

Cette situation a créé des attentes de la part des usagers quant à l'amélioration des diverses procédures d'aide à l'écriture ou à la lecture. Du côté de la production de textes et de leur gestion, ces attentes vont bien au-delà des traitements de texte. Déjà des systèmes, opérationnels ou à l'état de prototypes, proposent une aide à la rédaction (support lexical: dictionnaires, conjugueurs, terminologie, synonymie,...), à la révision (correcteurs orthographiques, stylistiques,...) ou encore à l'annotation (résumés automatiques, indexation, construction de thésaurus, ...).[2] D'un autre côté, les problèmes d'accès et de valorisation des banques de données textuelles suscitent également des espoirs envers les systèmes d'aide à la lecture. En gros, ces systèmes s'intéressent aux descriptions morphologique, syntaxique, sémantique, logique ou pragmatique des textes, à leur exploration pour en extraire l'information pertinente ou pour y faire surgir un sens quelconque et, enfin, à l'analyse des données ainsi extraites.

D'un côté, on trouve des usages du traitement informatique de la langue et une quantité croissante de données textuelles déjà disponibles, de l'autre, des procédures diversifiées d'écriture

---

[1] La conception du projet remonte à 1986. Sa mise en opération effective date de janvier 1988

[2] Voir Pierre Plante, Jules Duchastel, Lorne H. Bouchard, Potentiel d'applications de Déredec dans le contexte de la bureautique, Ministère des Communications du Québec, avril, 1986

et de lecture assistées. Par contre, il existe peu de méthodologie pour l'usage intégré de ces procédures selon des protocoles définis. Ces procédures sont partielles, peu standardisées et souvent difficilement accessibles. Leur utilisation, quand elle a lieu, est peu stratégique faute de modèles d'utilisation susceptibles de guider les usagers.

1.2. L'état de la question

Depuis leur origine, les recherches[3] reliées à la modélisation informatique des langues naturelles se profilent suivant deux axes: l'adaptation des modèles linguistiques et logiques à des contextes informatiques et la mise au point des techniques d'"ingénierie du langage". Coulon et Kayser[4] définissent deux optiques possibles correspondant à ces axes: le modèle philosophique dont le but est d'accroître la connaissance de la langue et le modèle ergonomique qui est orienté vers la production et l'utilisation d'outils. Dans un cas, il s'agit du projet de programmer une machine pour la compréhension automatique des phénomènes langagiers, dans l'autre, il s'agit plutôt de proposer des outils pour faciliter, par étape, cette compréhension.

L'histoire de ce domaine de recherche est traversée, de part en part, par ces deux optiques, mais elle est également caractérisée par une succession d'approches théoriques différentes qui ont dominé le champ durant des périodes données. En effet, chaque période est définie par la prévalence de l'une ou l'autre de ces approches, bien que chacune d'entre elles se soit superposée aux autres et continue, encore aujourd'hui, de se développer simultanément. Une première période (1945-1955), relativement étanche, a été caractérisée par l'approche statisticomorphologique. Elle fut suivie d'une dominance de la syntaxe de 1955 à 1970. Mais dès 1963, la recherche s'affairait à la programmation de modèles logico-sémantiques. Enfin, depuis 1974, le souci majeur est la représentation et l'organisation de la connaissance en faisant appel à des modèles cognitifs. Ces étapes renvoient, comme on peut le constater, aux divers niveaux classiques de la compréhension des phénomènes de langage. On trouve, aussi bien du côté philosophique que du côté ergonomique, de très nombreux exemples de ces travaux. Dans le premier cas, on donnera en exemples le développement important des approches lexicologiques, des techniques de parsage appliquées à des langages restreints (grammaires LL(n) et LR(n)) auxquelles s'ajoutent des syntaxes formelles comme les grammaires en chaînes, transformationnelles ou encore sémantiques (grammaires de cas et grammaires lexicales-fonctionnelles, etc.). Dans le second cas, l'ingénierie logicielle a, entre autres, contribué au développement de traitements morphologiques, de la gestion des lexiques, des analyseurs syntaxico-sémantiques (ATN), des analyseurs déterministes, des grammaires de métamorphoses et des Definite Clause Grammars (DCG) et, enfin, des modules d'inférence. Il ne s'agit pas là d'un inventaire, mais d'une indication de l'abondance des recherches fondamentales ou appliquées à tous ces niveaux.

Ces recherches ont permis des avancées notables, mais elles ont mis en évidence un très grand nombre de problèmes. La prévalence épisodique de l'une ou l'autre approche souligne, à loisir, les espoirs maintes fois déçus d'avoir trouvé l'angle d'attaque privilégié pour atteindre la compréhension automatique des langues. Les développements disciplinaires ou d'écoles ont favorisé des avances significatives, mais les contradictions entre diverses approches théoriques ainsi que l'opacité de certains modèles ont peu favorisé l'intégration des connaissances ainsi produites. La relative courte durée des projets indique l'existence fréquente d'impasses théoriques. La projection très problématique des avances théoriques dans les applications pratiques a mis en évidence

---

[3] Voir les analyses détaillées de Daniel Coulon et Daniel Kayser, "Informatique et langage naturel: présentation générale des méthodes d'interprétation des textes écrits", Technique et science informatique, vol. 5, no 2, 1986, ainsi que de B.J. Gross et al. Readings in Natural Language Processing, California, Morgan Kaufmann Publishers, inc., 1986, 664 pages.

[4] Op.cit.

l'incomplétude des systèmes. A travers ce cheminement complexe, pourtant, les limites de couverture linguistique, conceptuelle ou inter-disciplinaire qui se sont révélées au grand jour, ont permis de réévaluer les difficultés liées à la compréhension des phénomènes de langue et de discours et certains problèmes sont ainsi apparus comme prioritaires. On pense à la contextualisation nécessaire des phénomènes de discours, à la représentation des connaissances, à la nécessité d'incorporer une quantité considérable de données extra-linguistiques dans les modèles de TAL, à la prise en compte de la logique dite naturelle.

2. L'APPROCHE PRIVILÉGIÉE

Précisons d'abord que nous avons réduit le domaine de notre recherche, en choisissant la langue écrite (y compris les retranscriptions de l'oral) par opposition à la langue parlée et les aides à la lecture par opposition aux aides à l'écriture. Celà dit, l'approche privilégiée par SACAO se définit selon deux axes: premièrement, plutôt qu'une approche de compréhension en profondeur des phénomènes langagiers, elle propose une orientation pragmatique de valorisation des données textuelles; deuxièmement, face à une approche trop stictement syntaxique ou sémantique, elle favorise une analyse des morphologies du discours.

En ce qui concerne le premier axe, SACAO vise, avant tout, l'application de modules fonctionnels à de grands ensembles textuels. En somme, nous choisissons une approche pragmatique plutôt que fondamentale ou, dans les termes de Coulon et Kayser, une optique ergonomique plutôt qu'une optique philosophique. La logique de la démarche fondamentale favorise d'abord l'approfondissement des connaissances et ne recherche que secondairement des applications robustes et généralisables aux données du "monde réel". Une démarche pragmatique s'intéresse, au contraire, au développement d'outils ou d'applications qui nous permettent d'ores et déjà d'accroître notre capacité de lecture de plusieurs manières: accès rapide et systématique au contenu de grands ensembles textuels, rigueur et régularité de la lecture, production d'informations nouvelles par rapport aux formes traditionnelles de la lecture, introduction de la mesure et de procédures de validation, etc. Ils ont donc valeur pratique pour qui s'intéresse à la connaissance des textes.

Bien que les recherches fondamentale ou appliquée nous semblent indissociables, il est certain que notre objectif d'accroître le potentiel d'analyse du contenu des textes plaide inévitablement en faveur d'une approche pragmatique. Celà dit, il ne peut y avoir d'application qui ne soit fondée sur certains choix théoriques, mettant en jeu non seulement la langue, mais aussi le discours et la connaissance. Inévitablement, les choix pratiques qui sont effectués dans SACAO ne peuvent obvier à cette réalité. Il nous faut donc nous questionner minimalement sur les conséquences épistémologiques de notre option avant d'en revenir aux orientations théoriques qui guident notre entreprise.

Il serait abusif aujourd'hui d'associer trop strictement, d'un côté, démarche fondamentale et "systèmes automatiques" appliqués à des micro-mondes et, d'un autre côté, démarche pragmatique et "systèmes assistés" appliqués à des macro-mondes. Certaines recherches en intelligence artificielle ont pourtant privilégié le caractère automatique des procédures et visé la complétude des systèmes, du fait même qu'elles recherchaient la simulation plus ou moins isomorphique de phénomènes réels. SACAO a renoncé, méthodologiquement, aux prémisses épistémologiques propres à cette orientation. L'automatisation n'est recherchée que sur une base pragmatique et ne constitue pas une condition première. Nous mettons de l'avant une approche hybride, alliant procédures automatiques et assistées, et une substitution de l'idée d'intégration maximale des outils à l'objectif de complétude des systèmes. Ce point de vue n'est pas uniquement pratique, en ce qu'il serait motivé uniquement par l'impératif d'une couverture large du monde réel. Il répond à une conception extensive du problème de la compréhension des phénomènes de langue et de discours. Il est fondé également sur la conviction du caractère

créatif qui revient à l'usager dans le processus d'analyse. Les systèmes automatiques, aussi puissants soient-ils, proposent avant tout une boîte noire aux utilisateurs. SACAO propose une méthode interactive où le chercheur investit ses hypothèses et construit progressivement son analyse à l'aide d'outils performants.

Le projet SACAO s'est donc défini une posture épistémologique de nature empirico-constructiviste. De manière succincte, cette approche conçoit la connaissance des phénomènes langagiers comme le produit d'un processus non-univoque de construction des objets. Cela implique d'abord la coexistence de plusieurs procès de construction complémentaires (par exemple, multiplication des niveaux d'analyse) et potentiellement contradictoires[5] (par exemple, la coexistence d'approches non exclusivemement compatibles), ensuite la nécessité d'une démarche d'aller-retour entre la constitution des modèles et leur validation empirique. Cette démarche favorise la méthode inductive et le caractère interactif du système. Par exemple, nous évitons la projection du modèle aux données, et de manière plus ou moins déterministe, de modèles théoriques préconstruits sur le réel. Nous favorisons, au contraire, l'ajout de descriptions successives du texte en alternance avec l'exploration de résultats provisoires.

Revenons-en aux orientations théoriques de SACAO. Deux arguments nous incitent à expliciter nos prémisses théoriques. D'une part, la production ou la sélection d'outils doivent nécessairement trouver leur cohérence dans des cadres théoriques de référence. D'autre part, du point de vue des intérêts immédiats des chercheurs impliqués dans le projet SACAO, une orientation plus théorique doit guider et faire converger les développements qui seront favorisés ultérieurement. Le deuxième axe de notre approche renvoie à un présupposé théorique favorable à une analyse des morphologies du discours.

Un premier choix théorique place donc SACAO résolument du côté de l'analyse de contenu par opposition à la description linguistique. Bien que ces deux options ne soient nullement antagonistes, cette priorisation donnée à la saisie du sens délimite l'espace de travail qui sera le nôtre, en fonction d'objectifs de connaissance des textes. L'étagement des niveaux (morpho-lexical, syntaxique, sémantique, logique et pragmatique) caractérisant les phénomènes socio-linguistiques ne fait pas seulement énumérer les diverses dimensions de la langue et du discours, mais semble proposer un ordre souhaitable dans les étapes de la recherche. Par choix de méthode, la linguistique générale et la linguistique informatique ont souvent mis de l'avant le caractère prioritaire du fonctionnement proprement linguistique des phénomènes de langage et de discours. SACAO considère les divers niveaux de description comme la résultante d'un découpage et d'une construction différentielles de cet objet, et non comme les étapes ordonnées d'un parcours obligé qui mènerait de la description lexico-syntaxique à la compréhension globale de la langue naturelle.

Aussi, lorsque nous préconisons une analyse des morphologies du discours,[6] nous nous déplacons d'un intérêt pour la langue vers un intérêt pour le discours. Les descriptions linguistiques du texte serviront de support à l'analyse d'un système sémiotique, par ailleurs, beaucoup plus complexe. Nous faisons l'hypothèse que le texte est un espace diversement structuré, qui se déploie selon un processus de séquentialisations multiples (par ex., le point de vue de la narration, le point de vue de l'argumentation,...) et dans lequel des objets se schématisent pour former des noyaux de sens. Il nous intéresse donc de repérer les modes de segmentation qui caractérisent l'organisation d'un texte et les condensations de sens qui se produisent en certains

[5] On trouve dans les réflexions épistémologiques sur la physique des quanta l'idée de l'éclectisme et du complémentarisme des approches. Voir Fritjof Capra, The Tao of Physics, Shambala, Boulder, 1976 et Le temps du changement, Science, société, nouvelle culture, éd. du Rocher, 1983; Heinz Pagel, L'Univers quantique, Paris inter-éditions, 1985.

[6] Nous tenons à souligner la contribution importante de nos collègues Alain Lecomte (GRAD, Grenoble) et Jean-Marie Marandin (L.I.S.H., INaLF, Paris) au domaine de l'analyse du discours et spécialement au développement des hypothèses discutées dans ces lignes.

lieux privilégiés. Nous nous appuyons, pour ce faire, sur la connaissance lexicale du texte, élargie aux expressions terminologiques, et sur une description morpho-syntaxique non-exhaustive de ses unités. Nous privilégions deux axes principaux: l'axe nominal et l'axe verbal. Le premier renvoie à l'organisation sémantique du texte. L'analyse des proximités ou des relations de dépendance contextuelles (détermination, thème-propos,...) permettent de reconstruire des réseaux de signification. L'axe verbal renvoie davantage à la structure d'action du texte. L'analyse des caractéristiques et de l'environnement des verbes permet de reconstruire l'articulation des textes ainsi que le fil de l'argument.

3. LA MÉTHODOLOGIE

Les quelques remarques qui précèdent auront plutôt indiqué une direction de recherche ou un espace de travail que défini un cadre conceptuel précis. SACAO vise le minimum de préconstruction théorique justement parce qu'il propose, non pas un modèle d'analyse, mais un environnement offrant une panoplie de moyens de lecture diversifiés et minimalement contraints. C'est en ce sens que l'on parle d'une méthodologie pour l'usage intégré et stratégique d'outils d'analyse de données textuelles. Le caractère intégré de l'usage est autorisé par l'architecture du système qui offre la possibilité de retenir une ou plusieurs procédures de description, d'exploration ou d'analyse des données textuelles et de les faire interagir dans un plan d'ensemble. Son aspect stratégique consiste précisément à laisser le choix des modules, à offrir la possibilité de les modifier en fonction d'hypothèses particulières et à favoriser la structuration globale de la démarche de recherche.

Le système, adoptant une approche utilitaire, ne vise pas une compréhension strictement automatique du texte, mais propose des aides à la lecture et à l'analyse de textes. Il met à la disposition de l'utilisateur des outils éprouvés dans l'état actuel de leur développement. Il ne s'agit donc pas de proposer une méthode indépendante du contexte de recherche de l'utilisateur et qui garantirait des résultats générés par l'application aveugle de procédures. SACAO offre plutôt des outils de manipulation des données dont les a priori théoriques sont identifiés. Ces outils seront sciemment employés dans des stratégies de recherche définies.

Le système favorise, en effet, le maximum d'interactivité entre les besoins de l'usager et les dispositifs de lecture et d'analyse qui lui sont fournis. L'utilisateur doit pouvoir tester la valeur des résultats générés par toute procédure afin de décider de la retenir ou pas. Il doit pouvoir également ordonner, dans sa propre démarche, le recours aux divers moyens qui sont mis à sa disposition. Dans la mesure où c'est possible, il doit également choisir les paramètres qui seront activés dans chaque procédure. Cela signifie que la conception des procédures laisse place à une redéfinition des paramètres.

C'est donc en fonction des caractéristiques énoncées ci-haut que nous procédons à la mise en place du système. Nous présenterons maintenant les principaux éléments de cette mise en place. D'abord, la faisabilité du projet n'est possible que grâce à la disponibilité de modules informatiques spécialisés d'analyse de textes et de l'expertise que nous réunissons dans le domaine. Mentionnons les logiciels SATO (Système de base de données textuelles destiné à l'analyse de contenu), Déredec (Environnement général à base d'automates pour l'analyse et la construction de systèmes cognitifs), FX (progiciel de programmation de faisceaux), D_expert (Environnement pour la génération de systèmes experts) et les progiciels de description linguistique (Catégorisation de base syntaxique du français, Lemmatisation et caractérisation morphologique du français,

Grammaire de surface du français, Analyseur lexico-syntaxique du français). Tous ces systèmes ont été développés au Centre d'ATO, par les membres du Centre ou en collaboration avec des chercheurs du Centre.[7]

Nos travaux ou bien s'appuient sur des applications déjà développées ou en voie de développement (voir progiciels), ou bien donnent lieu à de nouveaux développements. Dans le premier cas, les modules sont soumis à une évaluation dans des situations de production sur de larges corpus et donnent lieu à l'optimisation des procédures ou, encore, à l'identification de sous-modules opérationnels dont l'utilité pour l'analyse de textes est prioritaire, par exemple, la catégorisation, la description thématique ou argumentative. Dans le second cas, nous introduisons des développements originaux qui s'avèrent nécessaires dans l'économie générale du système. Les modules "locutions" et "foncteurs sémantiques" sont des exemples de ces développements en cours.

SACAO met de l'avant une philosophie d'intégration des divers modules fondée sur la création de liens informatiques dans un même environnement machine et sur la portabilité des modules d'une machine à l'autre. Chaque adaptation des modules existants ainsi que les nouveaux développements devraient être intégrés et implémentés dans ces environnements. Mais, de façon réaliste, l'objectif prioritaire est de réaliser l'intégration de l'ensemble des modules sur le VAX, alors que plusieurs modules particuliers seront disponibles sur micro-ordinateurs.

Nous expérimentons sur une base systématique les divers modules de SACAO sur de grands corpus. Nous possédons une banque de données textuelles très importante contituée des corpus provenant de différents projets de recherche. Pour l'essentiel, l'expérimentation se fait à partir de données textuelles provenant de la sphère publique. Sans restreindre son utilisation à d'autres types d'application, cela implique que les utilitaires (par ex., dictionnaire de locutions terminologiques, dictionnaires sémantiques de domaines,...) sont d'abord enrichis à même des données relevant du domaine public. Il s'en trouve alors que l'environnement semblera plus familier à l'analyste du discours qu'au critique littéraire.

Il faut mentionner, en terminant, que cette expérimentation donne lieu à l'écriture systématique de fiches techniques qui permettent de documenter en profondeur les diverses procédures et qui serviront de base à la rédaction d'un manuel d'utilisation de SACAO.

4. L'ARCHITECTURE DU SYSTÈME

4.1. Les objectifs

Le projet SACAO poursuit, sur le plan informatique, les objectifs suivants :

1) Favoriser l'accroissement de la robustesse du système, en assurant une plus grande intégration des modules entre eux. Assurer la portabilité d'une machine à l'autre (PC, Macintosh et VAX), afin de permettre à l'usager d'accomplir certaines tâches dans des environnements familiers, tout en lui donnant accès à une capacité augmentée de traitement sur VAX.

[7] Délédec et FX sont des progiciels mis au point par Pierre Plante du Centre d'ATO. Il est à noter que les concepteurs de SATO et D_expert, respectivement François Daoust et Louis-Claude Paquin, sont membres actifs du projet SACAO.

2) Évaluer systématiquement les modules existants afin, soit de les enrichir, soit d'en extraire des procédures particulières comportant une utilité plus immédiate. Enrichir également le système de procédures de description, d'extraction et d'analyse comportant une complexité et une couverture plus grande.

3) Encourager l'accessibilité au système, en fournissant une documentation détaillée et exhaustive de toutes les procédures, appuyée sur leur expérimentation systématique sur des corpus témoins.

Nous décrivons ci-apres la dimension fonctionnelle de l'architecture de SACAO. Il faut préciser d'entrée de jeu que le terme architecture suppose plusieurs dimensions. La dimension fonctionnelle, privilégiée ici, décrit les caractéristiques des différents modules regroupant des unités de traitement. Nous n'aborderons pas les dimensions organique et algorithmique.

4.2. L'interface personne-machine

A l'heure actuelle, l'environnement informatique le mieux intégré est celui du VAX. On y retrouve les langages utilisés pour développer l'ensemble des applications (Pascal, C et Le_Lisp); on y trouve également les applications utilisées dans le contexte du projet, telles que mentionnées à la section méthodologie: SATO (Système d'Analyse de Textes par Ordinateur), Déredec et FX (langage de programmation des faisceaux), D_expert (progiciel pour la génération de systèmes experts) ainsi que divers utilitaires (programme de conversion des formats ASCII, courrier électronique, etc.). Du côté de l'environnement IBM et compatibles nous retrouvons SATO, une version réduite de Déredec et FX ainsi que des utilitaires pour la conversion des formats ASCII. Dans le cas de l'environnement Macintosh, nous y retrouvons principalement les applications réalisées en LISP soit Déredec, FX et le D_expert.

Une telle variété d'environnements de travail pourrait entrainer des difficultés importantes du point de vue de l'utilisation des ressources SACAO. Afin de prévenir les inconvénients liés à cette situation nous avons choisi deux options ergonomiques qui pourront pallier à ces difficultés: la transparence et la portabilité.

La transparence doit être assurée de manière à offrir à l'utilisateur une interface qui soit relativement indépendante de l'environnement matériel utilisé. En général, l'ensemble des décisions s'effectue de manière interactive à partir de choix offerts dans des menus hiérarchisés. Cette gestion "par menus" favorise le dialogue utilisateur-unité de traitement qui doit être sensible au contexte.

Au principe de transparence s'ajoute le principe de portabilité. Ce principe stipule que les options de développement doivent faciliter le transfert du savoir-faire contenu dans les modules de gestion et les unités de traitement. La portabilité d'une implantation matérielle à l'autre (PC vers VAX, VAX vers Macintosh, etc.) assure la possibilité du traitement coopératif (par ex., développer une maquette d'analyse sur PC et poursuivre le traitement des données sur VAX), les transferts des données entre les différentes unités de traitement, etc.

4.3. La gestion des données textuelles

Dans la perspective de rendre accessibles, au plus grand nombre d'utilisateurs, les outils et les données textuelles rassemblés dans SACAO, nous nous sommes intéressés dès le départ au problème de la gestion des données. Notre objectif était de structurer des programmathèques

ayant un caractère public. Celles-ci contiennent la panoplie des modules utilisés dans le cadre du traitement des données textuelles et les procédures pour les traitements en lot (batch processing). Elles intègrent également les corpus que différents chercheurs ont choisi de rendre publics. L'ensemble de ces dispositifs assure le caractère cumulatif de la production d'outils pour l'analyse des données textuelles.

Aux utilitaires d'archivage s'ajoute un utilitaire pour la conversion des formats ASCII propres aux trois implantations matérielles. Grâce à cet utilitaire, les usagers francophones sont assurés de pouvoir maintenir l'intégrité des textes sources et de procéder à l'analyse et au traitement des données de la même manière dans les différentes implantations matérielles.

4.4. La description des données textuelles

Tout mode d'investigation suppose une intervention technique sur les données à analyser. En effet, la notion de "donnée" implique nécessairement un processus de construction des unités de l'analyse et, par là même, une intervention de re-structuration qui transforme les unités d'information en unités d'analyse. Le module de description des données textuelles est le moment où s'accomplit la structuration initiale des données. Dans le cadre du projet SACAO, trois niveaux de description sont prévus : les niveaux lexical, morphologique et syntagmatique. Ces niveaux sont relativement autonomes les uns par rapport aux autres, mais ils peuvent être conjugués de manière différente eu égard aux besoins spécifiques d'une problématique de recherche ou d'analyse.

Au niveau lexical, la description des données vise à mettre en forme les différents aspects du vocabulaire (lexique) d'un texte. On pense ici plus particulièrement à la structuration du vocabulaire à partir de dictionnaires de locutions ou encore de thésaurus spécialisés. Dans un cas comme dans l'autre, il s'agit de procédures pour dresser l'inventaire des éléments d'un corpus de données textuelles. Au vocabulaire de base du français, s'ajoutent des expressions qui marquent les traits idiomatiques d'une communauté linguistique donnée. Les formes lexicales se réalisent souvent comme des groupes de mots qui fonctionnent de la même façon que les mots uniques. Afin de faciliter l'inventaire de ces unités, le module de description des données textuelles offre la possibilité de procéder au regroupement des différentes formes synaptiques (locutions). Il est ainsi possible d'indexer, dans le lexique des textes d'un corpus, les locutions canoniques (prépositionnelles, adverbiales, etc.), les locutions usuelles propres à un locuteur ou une famille de locuteurs, les locutions techniques, les termes institutionnels, les locutions onomastiques (noms propres), etc.

Au niveau morphologique, il faut faire en sorte que les dimensions grammaticales (morphèmes lexicaux et grammaticaux) puissent être bien identifiées. Nous disposons à l'heure actuelle d'une unité de traitement pour la caractérisation morpho-syntaxique du français contemporain.[8] Cette unité permet d'effectuer l'indexation des éléments d'un vocabulaire ou d'un lexique, en adjoignant aux formes lexicales des étiquettes syntaxiques (étiquettes pour la classification des noms, des verbes, des adjectifs, etc.). Une seconde unité de traitement rend possible le marquage de traits relatifs à la dimension lexicale des mots (morphème lexical ou radical).[9]

---

[8] CBSF (Catégorisation de base syntaxique du français), progiciel conçu par Lucie Dumas du Centre d'ATO, permet de reconnaître la catégorie syntaxique des formes lexicales de la langue française. Le caractère automatique de la procédure se réalise dans 80% des occurrences, dans le cas du français écrit contemporain.

[9] LCMF (Lemmatisation et caractérisation morphologique du français), également développé par Lucie Dumas, permet de regrouper automatiquement autour d'une unité minimale de représentation toutes les formes flexionnelles qui y sont associées.

Finalement, nous disposons d'unites de traitement pour décrire les dimensions syntagmatiques des données textuelles. A un premier niveau, nous pouvons faire appel à deux analyseurs du français, aptes à produire, de manière automatique ou semi-automatique, une description syntaxique des phrases ("expressions bien formées") du français écrit contemporain. Le premier (GDSF),[10] de nature avant tout heuristique, parvient à dépister pour toute proposition, le thème et le propos, des indications sur des compléments verbaux et plusieurs types de détermination nominale. Le second (ALSF),[11] actuellement en développement, a une portée linguistique plus grande. Conçu comme un environnement global de traitement des énoncés en français, il prévoit des modules d'information syntaxique, d'analyse syntaxique et d'interprétation des structures syntaxiques. Dans l'état actuel, certaines unites sont déjà accessibles (par exemple, la description du groupe nominal).

A un second niveau, il existe quelques exemples d'analyseurs textuels qui prennent appui, soit sur une première description morpho-syntaxique des phrases du texte, soit sur l'organisation sémantique des textes. Un exemple du premier cas se retrouve dans SAADI[12] qui, fonctionnant sur la base du groupe nominal et de la structure des propositions (concessives, restrictives, conclusives,...) permet de décrire la structure argumentative du texte. Il existe, par ailleurs, des grammaires de représentation sémantique de divers objets textuels, développées par différents chercheurs. Donc, dans le cas où ce qui nous intéresse relève des niveaux de structuration du texte autres que morpho-syntaxiques (par exemple, les analyses thématiques, la classification d'expressions ou d'énoncés, etc.), nous disposons d'unités de traitement permettant de programmer sur mesure des algorithmes de description. Deux langages (Déredec et FX) permettent la programmation de grammaires (du genre des "Augmented Transition Networks") automatiques ou assistées.

4.5. L'exploration des données textuelles

Le module d'exploration permet un travail complémentaire à celui effectué par les unités de traitement du module de description. Une fois les données constituées, il faut pouvoir disposer de mécanismes (regroupement d'opérations spécifiques) pour la sélection, le regroupement et la classification des données. Dans le module d'extraction, on retrouve des unités de traitement pour la constitution d'inventaires ou pour le regroupement catégoriel des informations.

Pour les unités qui sont structurées de manière linéaire (séquences lexicales), il est possible d'obtenir: des lexiques fréquentiels; des concordances (ou KWIC : Key Word In Context) basés sur la recherche de mots-clés ou sur des étiquettes symboliques ou numériques associées à ces mots-clés; des co-occurrences (mot-clé et lexique des mots étroitement associés au mot-clé); etc. Pour le dépistage de ces expressions, nous disposons d'opérations permettant de déterminer la forme et le nombre des chaînes de caractères qui seront employées comme paramètres des procédures d'extraction.

---

[10] GDSF (Grammaire de surface du français), conçue par Pierre Plante du Centre d'ATO, est un ensemble de procédures, programmées en Déredec, dont l'objectif est l'obtention des structures de surface du français écrit.

[11] ALSF (Analyseur lexico-syntaxique du français), produit en collaboration et sous la responsabilité de Jean Marie Marandin de l'INaLF, construit les structures syntagmatiques projetées par les catégories majeures du français: les noms, les verbes, les adjectifs et les prépositions. Il construit également les relations qu'entretiennent entre elles ces catégories dans des unités séquentielles.

[12] SAADI (Système d'analyse assistée des interviews), mis au point par Alain Lecomte et Catherine Péquegnat de l'Université de Grenoble, considère les enchaînements questions réponses et dépiste les réponses directes dans le processus d'entrevue.

Dans le cas des unites structurées à partir de contraintes morphologiques bien définies (configurations syntaxiques, données structurées de manière arborescente) ou floues (unités thématiques, énoncés axiologiques, etc), le module d'extraction permet le dépistage des données à partir de patrons définis par le chercheur ou l'analyste.

En plus des inventaires et des classifications, le module d'exploration permet la définition et la circonscription de partitions du corpus analysé. Ainsi, une personne analysant un corpus quelconque pourra à volonté appliquer à des sous-ensembles arbitrairement définis, les opérations de fouille mentionnées au paragraphe précédent. Autrement dit, il est possible de générer à partir du corpus une diversité de sous-textes. Il faut préciser que la génération de ces textes peut s'effectuer de manière à répondre aux exigences des traitements statistiques (techniques d'échantillonnage) ou de façon à permettre la vérification d'hypothèses sur un sous-ensemble relativement restreint (principe de la maquette) avant de poursuivre les opérations sur l'ensemble du texte.

4.6. L'analyse des données textuelles

Le module d'analyse de données textuelles offre actuellement les traitements suivants :

A) Un module de statistiques lexicales qui permet d'obtenir pour un lexique donné les statistiques suivantes: moyenne, écart-type, variance, fréquences minimum et maximum, score z et distribution procentuelle des classes de fréquences et d'occurrences.

B) Des mesures de distance inter-textuelle. La distance permet de comparer deux à deux des textes ou des parties de textes de manière à faire apparaître quels éléments lexicaux sont "responsables" des écarts de surface entre deux textes ou parties de texte. L'analyse de la distance peut être basée sur différentes distributions de fréquences correspondant à diverses segmentations du lexique et être pondérée par un lexique de référence identifié par le chercheur.

C) Indices de lisibilité. Les indices de lisibilité[13] sont des mesures empiriques permettant d'apprecier la difficulté ou la facilité de lecture, de comprehension et de mémorisation d'un texte ou des parties d'un texte. Ces mesures sont calculées à partir de paramètres comme la longueur des mots, la longueur des phrases, etc.

5. LE FONCTIONNEMENT DU PROJET S A C A O

Revenons rapidement sur les principales conclusions qui ressortent de l'exposé précédent, avant d'en montrer les conséquences sur la définition de l'équipe SACAO et sur l'organisation de ses activités. Nous avons établi, dès le départ, le besoin avéré d'une aide à la lecture de données textuelles. Ce besoin se manifeste aussi bien dans les nombreuses disciplines universitaires dont une des sources de connaissance est le materiau textuel, que dans les multiples usages du texte au sein des organisations. Nous avons opte pour une approche ergonomique de la question, préconisant l'usage intégré d'outils diversifiés dans une perspective de support à l'analyse. Donnant priorité à l'analyse de contenu par rapport à la connaissance purement formelle de la langue, nous avons privilégié une approche interdisciplinaire. Notre point de vue pragmatique

[13] Ces indices sont discutés en détail dans le texte de François Richaudeau, Le langage efficace, Paris, C.E.P.L., 1973, 300 p.

encourage donc une attitude heuristique dans le processus de la recherche et met de l'avant la plus grande autonomie des chercheurs en regard des moyens mis à leur disposition. La philosophie hybride, faisant appel autant à des procédures automatiques qu'assistees, favorise la participation active de l'analyste de texte.

Les moyens que nous nous donnons sont donc orientés en fonction de ces besoins et de cette approche. La mise sur pied d'une methodologie pour l'usage intégré de procedures d'aide à la lecture se traduit dans un environnement qui permet la gestion stratégique de ces moyens. L'usager doit pouvoir choisir librement les procedures qu'il retiendra, choisir également les paramètres qui seront actives dans ces dernieres. Il doit pouvoir articuler diversement, en fonction de ses propres besoins, les multiples procédures les unes par rapport aux autres et, ainsi, structurer globalement sa démarche de recherche. Les spécifications du système, pour répondre à cela, favorisent l'interactivité entre les chercheurs et les outils, demeurent ouvertes à la possibilité de varier les paramètres et comprennent le plus grand support documentaire.

L'architecture de SACAO a ainsi été conçue pour favoriser cette orientation. Elle définit diverses strates qui correspondent, en quelque sorte, à la démarche concrète de l'utilisateur. Fournissant à l'utilisateur des méthodes standardisées de fonctionnement et des facilités de gestion, elle définit les trois principaux champs d'activité autour de la description des données textuelles, de leur exploration et de leur analyse.

Le projet SACAO a été pensé et développé dans un contexte qui reflète bien les préoccupations résumées ici. D'abord inscrit de manière diffuse dans le cadre des activités de recherche du Centre d'ATO, le projet s'est progressivement spécifié dans un processus de différenciation par rapport à d'autres domaines de recherche en compréhension des langues naturelles. A côté du développement nécessaire de modules de description linguistiques ou cognitives, le besoin spécifique d'outils pour l'analyse de texte s'est fait urgemment sentir. L'équipe SACAO regroupe ainsi des chercheurs dont la formation disciplinaire et les domaines de spécialisation sont différents, mais qui ont pour objectif ultime l'analyse de textes. Cette équipe comporte également la caractéristique de correspondre à des demandes hétérogènes en termes de développement. Certaines de nos activités s'inscrivent dans la structure de la recherche universitaire, alors que d'autres sont immédiatement associées aux demandes de développement de systèmes destinés aux organisations.

Cette equipe dont chaque membre poursuit, par ailleurs, une activité relativement indépendante dans son champ de spécialisation, a dû concevoir un projet commun qui reflète l'aspect polymorphique des besoins, de l'approche et des moyens préconisés. Elle a donc défini quatre domaines d'activités et mis en place des mécanismes pour leur réalisation. Ces activités sont: le développement informatique, l'adaptation et le développement d'unités de traitement, l'expérimentation et la documentation et, enfin, les activités de réflexion et de formation. Les mécanismes de réalisation consistent en un séminaire hebdomadaire d'échange et de planification et en un partage des tâches selon les diverses compétences. Nous illustrerons très rapidement le type d'activités qui relèvent de chacun de ces domaines.

Le développement informatique renvoie à l'aspect informatique lié à la mise au point et à la gestion des procédures d'aide à la lecture. Il peut s'agir de l'entretien des environnements logiciels dans les diverses implantations, de la mise au point d'interfaces et de la portabilité. Ce sont également les divers développements informatiques liés aux développements des procédures: nouvelles structures de représentation, nouveaux automatismes, etc.. C'est encore le développement des procedures de gestion des fichiers.

L'adaptation d'unités de traitement peut s'illustrer par l'exemple d'un travail d'évaluation que nous avons effectué des descriptions GDSI de la structure thématique des textes d'un corpus de discours politiques. Sur la base de cette validation, certains sous-ensembles de procédures, enrichis de nouveaux développements, sont utilisés pour établir une description arborescente des

propositions du point de vue de leur hiérarchie thématique dans la tradition de la grammaire fonctionnelle. Le développement de nouvelles unités de traitement peut s'illustrer par les nouvelles procédures de repérage, de blocage et de thésaurisation des locutions. Ce système utilise les propriétés de nos logiciels et progiciels dans le but de fournir un instrument nouveau aux utilisateurs

L'expérimentation renvoie au travail systématique de validation des procédures sur des corpus de référence. Ce travail permet de varier les contextes d'application et de tester la robustesse des systèmes face à la redéfinition des paramètres. En plus de la validation, cette expérimentation permet de produire des fiches techniques destinées à documenter le système et des fiches d'utilisation réservées aux usagers.

Enfin, les activités d'échange et de formation nous sont apparues comme étant primordiales. L'interdisciplinarité à la base du projet et la multiplicité des voies qui y sont explorées nous obligent à faire le point sur des questions théoriques et méthodologiques fort variées. Nous abordons ainsi des questions comme: les problèmes de la catégorisation sémantique, les diverses stratégies d'analyse du discours, les diverses approches de l'analyse thématique, la théorie du parsage, etc.. La formation s'effectue quant à elle à travers la mise sur pied de cours spécialisés en ATO.

En somme, SACAO n'est pas un projet fermé, mais plutôt un programme de travail ouvert. Il correspond à l'identification de besoins précis et ouvre un espace de travail interdisciplinaire qui doit être investi pour lui-même. Même s'il bénéficie abondamment de la recherche fondamentale en linguistique informatique et en sciences cognitives, il ne doit jamais perdre de vue que ce qui l'intéresse, c'est l'analyse de textes assistée par ordinateur.

Bibliographie

ACTES DU COLLOQUE: *Représentation du réel et informatisation*; 26 et 27 mai 1988; Saint Etienne (France)

ALLEN, Sture, (1982) *Text processing: text analysis and generation: text typology and attribution*, Stockholm, Almqvist & Wiksell International, 1982, 653 pages.

ARRIVÉ, Michel, GADET, Françoise, GALMICHE, Michel, (1986) *La grammaire d'aujourd'hui. Guide alphabétique de la langue française*, Paris, Librairie Flammarion, 720 pages.

BERWICK, Robert C., (1985) *The acquisition of syntactic knowledge*, Cambridge, Mass., MIT Press, 368 pages.

BONNET, Alain; HATON, Jean-Paul; TRUONG-NGOC, Jean-Michel. *Systèmes-experts: vers la maîtrise technique.* Paris: InterEditions; 1986.

BOREL, Marie-Jeanne, GRIZE, Jean-Blaise, MIÉVILLE, Denis, (1983) *Essai de logique naturelle*, Berne, éditions Peter Lang SA, 1983, Sciences pour la communication, N° 4, 241 pages.

COLLOQUE INTERNATIONAL CNRS, (1986) *Méthodes quantitatives et informatiques dans l'étude des textes*, Genève - Paris, Slatkine - Champion, 947 pages.

COULON, Daniel, KAYSER, Daniel (1986) "Informatique et langage naturel: Présentation générale des méthodes d'interprétation des textes écrits", *Technique et Science Informatiques*, Février, 1986, p. 103-126.

CRUSE, D. A., (1986) *Lexical Semantics*, Great Britain, Cambridge University Press, 1986, Cambridge textbooks in linguistics, 310 pages.

DAVIES, R.; LENAT, D. *Knowledge-based systems in artificial intelligence*: McGraw-Hill; 1982.

DUBOIS, D.; PRADE, H. *Théorie des ssibilités*. Paris: Masson; 1985.

DUCROT, Oswald, (1972) *Dire et ne pas dire. Principes de sémantique linguistique*, Paris, Hermann, 1980. Collection Savoir, 311 pages.

DANLOS, Laurence, (1987) *The linguistic basis of text generation - Laurence Danlos translated by Dominique Debize and Colin Henderson - Génération automatique de textes en langues naturelles*, Angleterre, Cambridge University Press, 222 pages.

DAOUST, François, (1987) *SATO: Système d'Analyse de Textes par Ordinateur (version 3.4). Manuel de référence pour les micro-ordinateurs PC et PC compatibles*, Université du Québec à Montréal, Centre d'Analyse de Textes par Ordinateur, 1987, 81 pages.

FARRENY, H., (1987) *Les systèmes experts. Principes et exemples*, Cepadues-Editions.

GROSS, Maurice, (1975) *Méthodes en syntaxe. Régime des constructions complétives*, Paris, Hermann, 1975, 414 pages.

GROSZ, Barbara J.; JONES, Karen Sparck; WEBBER, Bonnie Lynn, (1986) *Readings in Natural Language Processing*, California, Morgan Kaufmann Publishers, Inc., 1986, 664 pages.

GUIRAUD, Pierre, (1961) *Les locutions françaises*, Paris, Presses Universitaires de France, 126 pages.

HALLIDAY, M.A.K., (1985) *An introduction to functional grammar*, London, E. Arnold, 1985, 387 pages.

KRIPPENDORFF, Klaus, (1980) *Content Analysis. An Introduction to its Methodology*. Sage Publications, 189 pages.

HAYES-ROTH, F. ; WATERMAN, D. A.; LENAT, D. *Building Expert Systems*. Reading, Mass.: Addison Wesley; 1983.

Numéro spécial "Knowiedge Acquisition for Knowledge-based Systems" *International Journal of Man Machine Studies*; 1987; (26)

LECOMTE, A., (1988) "Le marmot et la mamelle, critique des représentations du raisonnement", Centre de Coordination pour la Recherche et l'Enseignement en Informatique et Société (CREIS), *Représentation du réel et informatisation*, Saint-Etienne, I.U.T. de Saint-Etienne, 1988, 21 pages.

LECOMTE, A., MARANDIN, J. -M, "Analyse de discours et morphologie discursive", Montréal, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 1984, 6, pages.

MARANDIN, J.M., (1988) "A propos de la notion de thème de discours. Eléments d'analyse dans le récit", *Langue Française*, (à paraître), 1988.

MELCHUK, Igor Aleksandrovich, ARBATCHEWSKY-JUMARIE, Nadia, (1984) *Recherches lexico-sémantiques*, Montréal, Presses de l'Université de Montréal, 1984, 172 pages.

PAQUIN, Louis-Claude, *Déredec-EXPERT (Version 2.0)*, Université du Québec à Montréal, Centre d'Analyse de Textes par Ordinateur, 1987, 119 pages.

PÊCHEUX, Michel, (1969) *Analyse Automatique du Discours*, Paris, Dunod.

PLANTE, P., *Manuel de programmation Déredec*, Centre d'ATO.

------ (1975) *Proposition d'algorithme pour le dépistage de relations de dépendance contextuelle dans un texte*, Montréal, Université du Québec à Montréal, 111 pages.

RASTIER François et al., (1987) *Sémantique et intelligence artificielle*, Paris, Librairie Larousse, 1987, Langage #87, Septembre, 128 pages.

SOWA, J. F., (1984) *Conceptual Structures. Information Processing in Mind and Machine*, Addison-Wesley Publishing Company, Inc., 481 p.

WATERMAN, D. A. *A Guide to Expert System*. Reading, Mass.: Addison-Wesley; 1985.

LA DESCRIPTION DES LANGUES NATURELLES EN VUE D'APPLICATIONS INFORMATIQUES
SATO, un outil au service de l'Administration publique

Maurice Gingras
Secrétariat du Conseil du trésor, Gouvernement du Québec.

Le Secrétariat du Conseil du trésor a la responsabilité de la mise à jour et de la diffusion de la politique administrative générale du gouvernement, laquelle trouve son point de chute dans le *Répertoire des politiques administratives* et dans le *Recueil des politiques de gestion*. On y retrouve plusieurs centaines de pièces (lois, règlements, directives, décisions du Conseil du trésor, etc.) qui constituent l'encadrement normatif de la gestion gouvernementale. Au gré des décisions du Conseil des ministres, du Conseil du trésor ou des divers ministères responsables d'un volet de la gestion, ces pièces sont édictées, modifiées ou mises à jour au fil des ans. Il s'agit d'une somme imposante d'information textuelle qui a pour but d'informer les employés sur les politiques de l'organisation, sur ses lignes de conduite, ses standards et les nombreuses procédures à suivre, qu'il s'agisse de biens et de services, de gestion de personnel ou d'administration financière.

Dans le cadre de sa participation au projet lexical et sémantiques des domaines, l'auteur a entrepris de réaliser l'analyse de cette masse documentaire à l'aide du logiciel SATO. Nous présentons aux participants du colloque le récit de cette expérience dont les diverses applications sont susceptibles d'en intéresser plus d'un.

LES DIVERSES ÉTAPES DE L'ANALYSE DE TEXTES

■ La formation et l'équipement

Le logiciel SATO a été développé par François Daoust du Centre d'ATO de l'UQAM, et peut fonctionner sur divers types d'ordinateurs (Vax, Mac et Pc). La version utilisée donne des résultats plus que satisfaisants sur un PC-AT (compatible). L'usager déjà familier avec le micro-ordinateur peut réaliser de façon à peu près autonome, après quelques heures de formation et quelques semaines d'expérimentation, l'analyse d'un premier corpus de textes de son domaine.

■ Les données textuelles

La disponibilité sur support informatique des données textuelles à analyser représente un atout important. Dans le cas qui nous occupe, les documents de la politique administrative sont déjà sur traitement de textes dédié (Micom). Il s'agit donc, dans un premier temps, de les transférer en format PC-WordPerfect et de les déposer sur disque rigide pour le traitement ultérieur (il existe sur le marché des logiciels de transfert). L'opération est, en soi, assez fastidieuse mais il faut penser que ce changement de support donne en même temps accès à tous les avantages du traitement de textes sur micro-ordinateur: la souplesse, la rapidité d'accès, de mise à jour, de partage des informations et la sécurité des données.

■ La codification des documents

Par la suite, chacune des pièces est codifiée en vue de son traitement par SATO. Cette codification, réduite à sa plus simple expression, consiste à déterminer un ordre alphabétique (qui

commandera, entre autres, la disposition du lexique des formes rencontrées dans le texte), et à donner un titre et un numéro de document au texte. Il s'agit là de commandes préalables qui vont servir de guides de lecture pour le programme. On mémorise enfin le document en format ASCII.

Une fois le texte codifié, on appelle le programme SATOGEN qui "lit" le document et génère un certain nombre de fichiers qui seront utilisés par la suite dans le programme d'interrogation. Ce premier traitement s'opère au rythme d'environ 50 mots à la seconde et produit un lexique de toutes les formes (mots) rencontrées avec la fréquence d'apparition ainsi que la trace de ces mots dans le texte; certaines marques d'édition sont en même temps ajoutees telles la pagination exacte de chacun des mots du texte, les majuscules de ponctuation, les en-têtes de paragraphes, la longueur de chaque mot, etc.

- **Le programme d'interrogation**

On appelle ensuite le programme d'interrogation (SATOINT) lequel comprend un ensemble de commandes qui sont autant d'outils permettant d'interroger et d'annoter le texte. En combinant ces outils, on peut établir un protocole d'analyse de textes mémorisé et reproductible, lequel peut être appelé automatiquement par une commande d'exécution.

LE TRAITEMENT DE LA POLITIQUE ADMINISTRATIVE

- **La production d'un index**

L'une des premières retombées du projet a consisté à produire un index de la politique administrative dans le but de faciliter l'identification des pièces traitant d'un sujet donné, à partir de mots clés. Pour ce faire, on a procédé au traitement par SATO de la table des matières des divers recueils de cette politique. Une fois extrait le lexique de tous les mots contenus dans les titres des pièces et éliminés les mots vides (le, la les, dans, etc.), on a demandé aux spécialistes de valider les mots ou locutions qui leur semblaient caractéristiques du domaine ou utiles pour la consultation. Le résultat de cette opération est conservé dans un fichier "dictionnaire" et peut être utilisé par la suite pour automatiser l'indexation d'un nouveau texte.

Partant de cette liste des mots clés, on demande au programme d'en rechercher toutes les concordances dans la table des matières. Le programme génère alors un fichier imprimable dans lequel on retrouve, par ordre alphabétique, chacun des 400 ou 450 mots retenus et, pour chacun, le titre des pièces où se retrouve le mot avec sa référence aux volumes. Il suffit par la suite de rappeler ce texte dans un programme d'édition (tel WordPerfect) et de lui donner le format souhaité pour l'impression.

- **La constitution d'un corpus de la politique administrative**

Une autre étape consiste à rassembler toutes les pièces de la politique administrative en un seul corpus, chacune étant identifiée par un numéro de document, et à traiter l'ensemble par le module SATOGEN. Cette façon de procéder permet d'interroger le contenu intégral des textes de la politique (plus de 180 documents dans notre cas) sur un mot ou une expression pour en retrouver les concordances. Elle permet également de constituer un vocabulaire exhaustif du domaine visé.

Au fil de l'expérience la liste des applications concrètes et des demandes d'information s'allonge. C'est ainsi qu'ayant remarqué la présence, dans ce type de documents normés, de

nombreuses définitions ayant en commun une certaine régularité d'écriture (alinéa et guillemets, par exemple), on peut demander au programme d'extraire ces définitions à partir d'un patron de recherche et d'en produire un fichier imprimable.

Pour certaines fins, on peut limiter le domaine à certains documents (ceux, par exemple couvrant la gestion des biens et des services) et rechercher les passages où l'on traite de seuils d'autorisation, etc. On comprendra toute l'utilité de disposer d'un tel instrument pour extraire, par exemple, la "connaissance" relative à un domaine de gestion donné.

Ces quelques exemples n'épuisent pas les possibilités du logiciel; ils permettent seulement d'en démontrer l'intérêt pour un "travailleur du texte", particulièrement lorsque l'information textuelle est volumineuse et que le souci d'exhaustivité est impérieux.

CONCLUSION

- **Le développement de systèmes à base de connaissance**

L'expérimentation du logiciel SATO est une piste prometteuse pour qui s'intéresse au développement de systèmes experts. Ce logiciel initie l'usager à la mise en forme et au traitement des données textuelles de son organisation en vue de l'extraction des connaissances. Il est possible de générer, à partir de SATO, une information directement traitable par le système DÉREDEC, développé par Pierre Plante de l'UQAM. Il s'agit d'un logiciel de traitement linguistique, d'analyse de contenu des textes; il trouve son prolongement dans un progiciel générateur de système expert, le D-EXPERT (Louis-Claude Paquin, de l'UQAM). On peut utiliser l'information numérique générée par SATO comme entrée pour la mise au point de systèmes experts en langue naturelle, ou pour des logiciels d'analyse de données et de traitement statistique.

L'auteur vient de compléter sur D-EXPERT un prototype de système expert sur un volet de la politique administrative; il s'agit d'un système d'aide à l'attribution des contrats de services qui prend appui sur les travaux réalisés à l'aide du logiciel SATO.

REMERCIEMENTS:

La contribution majeure des chercheurs du Centre d'ATO de l'UQAM a été pour les membres de notre organisation chargés d'élaborer des instruments de recherche documentaire, d'outils d'indexation ou de systèmes experts, de leur faire prendre conscience de l'existence et de l'utilité d'outils informatiques efficaces, simples d'utilisation et en mesure d'augmenter l'efficience du processus d'extraction de la connaissance d'un domaine.

Cette contribution a pour effet de stimuler la créativité des spécialistes d'un domaine donné, par une approche où ce ne sont pas les automatismes ni la confiance aveugle dans l'outil qui priment, mais les manipulations répétées, libres et variées de l'usager.

LOGICIEL D'AIDE À LA CONCEPTION DE BASES DE CONNAISSANCES DÉONTIQUES À PARTIR DE L'ANALYSE DE TEXTES DE RÈGLEMENT

Marie-Michèle Boulet, Bernard Moulin, Daniel Rousseau, Gérard Simian
Département d'informatique
et Régine Pierre
Département de psycho-pédagogie
Université Laval

RÉSUMÉ

Dans le cadre de cette recherche, nous explorons la possibilité de constituer des bases de connaissances à partir d'informations contenues dans des textes utilisés dans des organisations. Les documents que nous étudions correspondent à des "textes prescriptifs" que l'on peut trouver dans les entreprises: manuels de normes, règlements, manuels d'utilisation d'appareils ou de logiciels, etc. Notre étude actuelle se concentre plus particulièrement sur les textes de règlement émis par le Gouvernement du Québec.

Dans ce projet, nous visons à développer un logiciel d'acquisition des connaissances qui permette aux experts de transformer un texte prescriptif sous la forme d'une base de connaissances manipulable par un moteur d'inférence.

Nous énonçons dans cet article les lignes directrices de notre recherche sur la mise au point d'un logiciel d'acquisition des connaissances à partir de textes prescriptifs. Nous présentons une approche de la structuration des textes qui nous fait distinguer la macrostructure de la microstructure et de la composante domaniale du texte. Nous livrons une première analyse de la microstructure d'un texte règlementaire. Cela nous conduit à nous interroger sur la représentation des connaissances déontiques dans une base de connaissances. Nous suggérons un premier noyau de spécification d'un métalangage réglementaire ainsi que sa transformation en énoncés logiques. Nous proposons aussi les schémas de principe d'un système d'acquisition et d'un système de manipulation de connaissances déontiques. Finalement, nous discutons brièvement des divers éléments d'une phrase à mettre en évidence lors du traitement de la microstructure d'un texte prescriptif.

1. INTRODUCTION

L'expérience montre que les experts ont plus de facilité à expliciter les connaissances et le raisonnement qu'ils mettent en oeuvre pour résoudre leurs problèmes si on les libère de l'utilisation de formalismes compliqués de représentation et si on leur permet d'exprimer leurs connaissances sous une forme textuelle structurée et lisible (un sous-ensemble compréhensible du français).

Nous appelons "base d'acquisition de connaissances" ('BAC') une forme textuelle structurée, élaborée et mise à jour par les experts au cours des sessions d'acquisition des connaissances. La BAC apparaît comme une forme éditable de base de connaissances en français, compréhensible et manipulable par les experts du domaine.

Dans le cadre du projet A.C.A.T. ("Acquisition de connaissances et analyse de textes"), nous visons à développer un logiciel qui permette aux experts d'éditer et de manipuler une BAC et d'aider les informaticiens à convertir la BAC sous une forme qui puisse être utilisable par des moteurs d'inférence commerciaux. La compilation de la BAC permettra de générer automatiquement un ensemble structuré de bases de connaissances (ou forme compilée de la BAC). Ces bases de connaissances ('BC') seront ensuite converties sous une forme interne compatible avec l'utilisation de moteurs d'inférence commerciaux (ex. GURU). Par ailleurs, une grande quantité de connaissances est enregistrée sous la forme d'écrits de natures diverses: textes informatifs, textes éducatifs, normes, procédures de travail, etc. Dans le cadre de cette recherche, nous explorons la possibilité de constituer des bases de connaissances à partir d'informations contenues dans des textes utilisés dans des organisations. Les documents que nous étudions correspondent à des "textes prescriptifs", que l'on peut trouver dans les entreprises: manuels de normes, règlements, manuels d'utilisation d'appareils ou de logiciels, etc. Cette étude exploratoire se concentre plus particulièrement sur les textes de règlement émis par le Gouvernement du Québec.

Les objectifs poursuivis dans ce projet sont les suivants:

- La mise au point d'un langage de spécification du contenu d'une base de connaissances correspondant à un sous-ensemble du français utilisé pour décrire des connaissances prescriptives ou déontiques.

- La mise au point d'un logiciel d'acquisition des connaissances implantant la notion de "base d'acquisition des connaissances" (BAC), et d'un système permettant de générer à partir de la BAC des bases de connaissances sous forme de règles de production (système d'acquisition des connaissances).

- La mise au point d'un moteur d'inférence et d'un environnement de conception de systèmes à base de connaissances (système de consultation).

- La mise au point d'une méthode de conception de bases de connaissances, qui tienne compte de la possibilité de spécifier les connaissances à partir d'un texte prescriptif, et qui mette en oeuvre le concept de BAC.

- L'application de ces méthodes et outils à l'analyse de textes prescriptifs, et plus particulièrement aux textes de règlement.

Dans la suite de cet article, nous distinguons diverses composantes dans un texte prescriptif: la macrostructure, la microstructure et la composante domaniale. Puis, nous livrons une première analyse de la microstructure d'un texte réglementaire en nous interrogeant sur la représentation des connaissances déontiques dans une base de connaissances. Nous proposons un premier noyau de spécification d'un métalangage réglementaire ainsi que sa transformation en énoncés logiques. Nous suggérons les schémas de principe d'un système d'acquisition et d'un système de manipulation de connaissances déontiques. Finalement, nous discutons brièvement des divers éléments de la microstructure à traiter par le système d'acquisition des connaissances afin d'élaborer la BAC correspondant à un texte prescriptif.

2. LES COMPOSANTES D'UN TEXTE PRESCRIPTIF

Les textes de règlement constituent de bons exemples de documents de synthèse rassemblant pour un domaine précis des connaissances formulées en langue naturelle sous une forme prescriptive. Pour la formulation des articles de règlement, les auteurs emploient habituellement

un "style juridique" qui répond à certaines règles générales de présentation et d'expression (voir par exemple [T-J-L 84]). Le texte décrivant un règlement peut être considéré comme un ensemble de connaissances décrivant en théorie d'une façon exhaustive les caractéristiques d'un domaine pratique d'application de la loi. Lorsqu'on étudie un texte réglementaire, on peut distinguer des éléments qui peuvent correspondre à l'un des trois types de composantes: la macrostructure du texte, la microstructure, la composante domaniale [MOU 88].2.1

2.1 La macrostructure d'un texte

Nous définissons la macrostructure d'un texte comme l'ensemble des informations qui servent à organiser le contenu du texte par l'appoint d'une "superstructure" enrichissant la présentation des énoncés et facilitant la consultation: titres, en-têtes, paragraphes, table des matières, index, références, notes, etc.

La macrostructure du texte est en général établie en fonction de quelques règles de présentation communément adoptées (titres, paragraphes, tables des matières, index). Par contre, le découpage du texte et le contenu sémantique de la macrostructure sont élaborés par l'auteur, souvent de façon intuitive , en fonction de ce qu'il considère être la meilleure façon de présenter son texte au lecteur.

Bien que cela soit rarement fait en pratique pour un même document, remarquons que la présentation d'un texte pourrait varier pour s'adapter à des types différents de lecteurs: novices, spécialistes d'un domaine, décideurs, etc. Cette adaptation de la présentation et du contenu du texte demanderait que l'on s'intéresse aux objectifs du lecteur et à un modèle de ses connaissances préalables: "le texte intelligent" adapterait son contenu un peu comme le pédagogue s'adapte au niveau de connaissances de son élève.

2.2 La microstructure d'un texte

Nous définissons la microstructure du texte comme l'ensemble des "mots réservés", des locutions et des symboles qui servent à structurer le contenu du texte pour en faire ressortir la structure logique. Voici des exemples de mots ou de locutions utilisés pour supporter la microstructure: "si", "alors", "sinon", la virgule, "lorsque", "il est interdit de", "il est possible de", etc. Ces éléments servent à structurer l'exposition ou l'argumentation du texte. L'auteur peut utiliser une formulation plus ou moins systématique de la microstructure suivant la nature du texte. Les textes de loi et les règlements obéissent à des règles assez systématiques d'exposition et présentent une microstructure apparente. Il en est souvent de même pour des textes de normes ou de procédures.

L'étude de la microstructure d'un texte permet de mettre en évidence la cohérence logique de l'argumentation et d'y repérer éventuellement certaines inconsistances. On peut ainsi considérer les énoncés supportant la microstructure du texte comme constituant un métalangage utilisé pour décrire l'enchaînement logique des propositions en fonction de certains objectifs d'argumentation que l'auteur veut atteindre. Ce métalangage peut être étudié du point de vue de la logique modale, et plus particulièrement de la logique déontique pour les textes de règlement [KAL 72].

2.3 La composante domaniale d'un texte

Nous définissons la composante domaniale du texte comme l'ensemble des informations caractéristiques du sujet traité et n'appartenant ni à la microstructure, ni à la macrostructure.

Notons que nous ne nous intéressons pas ici aux composantes graphiques et tabulaires que peuvent contenir certains textes: images, graphiques, tableaux de données, etc.

La composante domaniale du texte comporte un ensemble de propositions qui peuvent être soit traitées globalement comme des énoncés logiques, soit analysées de façon plus détaillée pour faire ressortir les composantes sémantiques du texte. La nature du traitement appliqué à la composante domaniale dépend du degré de compréhension du contenu sémantique du texte que doit atteindre le système envisagé.

Ainsi, pour réaliser un système expert, on peut se contenter de manipuler les propositions d'un texte seulement à un niveau logique. La base de connaissances ainsi développée permet à l'usager d'accéder aux enchaînements logiques supportés par le texte, l'interprétation sémantique des propositions étant effectuée par l'usager lui-même: un tel système travaille seulement avec des connaissances de surface. Par contre, il faut tenir compte du contenu sémantique des propositions pour réaliser un système de compréhension qui permette de répondre aux questions de l'usager concernant les connaissances profondes du texte.

2.4 La manipulation de formes textuelles structurées

Nous postulons que plusieurs catégories de textes habituellement utilisés dans les organisations constituent "des bases de connaissances textuelles" pour lesquelles les auteurs ont déterminé par expérience et souvent par intuition des formes structurées, afin d'en faciliter la consultation. Dans certains cas, des guides de rédaction ont été proposés pour uniformiser la structuration de certaines catégories de textes, notamment dans le secteur législatif (voir par exemple [T-J-L 84]).

Certaines catégories de textes sont naturellement structurées: textes prescriptifs (normes, directives, modes d'emploi, etc.), textes déontiques (règlements, règles de jeu, lois, etc.).

Nous faisons l'hypothèse qu'en faisant subir certaines transformations à des textes prescriptifs afin de rendre leur formulation systématique (et éventuellement plus formelle), il est possible d'élaborer des bases de connaissances exploitables par des moteurs d'inférence. La forme textuelle structurée issue de la transformation du texte original pourra ainsi servir pour la spécification du contenu de la base de connaissances, et la rendre naturellement compréhensible par des usagers non intéressés à apprendre des langages formels de spécification. Cette forme textuelle structurée sera appelée "base d'acquisition des connaissances" (B.A.C.).

3. CONNAISSANCES DÉONTIQUES DANS UN TEXTE DE RÈGLEMENT

3.1 Introduction

Nous présentons dans cette section quelques réflexions concernant un noyau de langage qui permettrait d'établir une spécification formelle du contenu d'un texte déontique, à partir de laquelle on pourrait générer une base de connaissances exploitable par un moteur d'inférence.

Un texte de règlement peut être considéré comme un ensemble de prescriptions qui doivent être honorées par les personnes (physiques ou morales) dans des situations d'application précisées par le règlement.

Ainsi le règlement sur le traitement des déchets solides (voir l'annexe 1) fixe des règles qui doivent être suivies par les promoteurs qui désirent mettre en opération certains types d'installations permettant le traitement des déchets solides: dépôt en tranchée, incinérateur, usine de pyrolyse, etc.

Une première analyse de la microstructure du texte (voir section 4.2) permet de relever des régularités de formulation qui soulignent la présence d'un métalangage qui supporte l'argumentation "réglementaire". A notre connaissance ce métalangage n'a pas été formalisé, mais il est utilisé par les juristes conformément à un usage accepté par consensus.

Nous ne suggérons pas ici que les juristes utilisent un langage formel et rigide, mais que par la nature même des concepts qu'ils doivent exprimer dans leurs textes prescriptifs, ils ont mis au point par expérience un "langage de style juridique" que nous allons étudier comme s'il constituait un métalangage d'expression des prescriptions réglementaires.

De nombreux auteurs se sont intéressés au contenu des textes juridiques. En particulier, au début des années 50 apparaissent les premiers systèmes traitant formellement "d'une logique des normes" ou "logique déontique".

En fait le premier système de Von Wright (1951) se base sur deux idées fondamentales: l'idée de l'analogie entre l'obligation, la prohibition et la permission d'un côté, et respectivement la nécessité, l'impossibilité et la possibilité; l'idée de la transposition sur le terrain des énoncés déontiques (normatifs) des termes de la logique des prédicats. De nombreux autres logiciens ont proposé des modèles de logique déontique. Nous renvoyons l'auteur intéressé à l'ouvrage de Kalinowski [KAL 72], qui fait une excellente présentation des divers travaux qui ont été menés au sujet de la "logique des normes".

3.2 Formes logiques dérivées des expressions du métalangage

Dans cette analyse exploratoire du "métalangage réglementaire", nous avons choisi d'étudier le texte du reglement sur le traitement des déchets solides (voir un extrait en annexe 1), pour essayer d'en faire ressortir toutes les expressions qui supportent les énoncés prescritifs. L'annexe 2 présente les principales catégories d'expressions que nous avons relevées.

Ce métalangage est basé sur l'utilisation d'un certain nombre d'expressions ou de tournures de phrases qui supportent la formulation logique des termes du règlement. Nous proposons une approche de réécriture en termes formels des propositions du règlement sous une forme logique qui s'inspire des travaux cités précédemment.

Remarquons tout d'abord que la plupart des prescriptions du règlement correspondent à des énoncés prescriptifs du type :

"Tout X doit proposition P",

ou

"Pour tout X, il est interdit de proposition P".

Ce type d'énoncés correspond à la prescription d'une caractéristique ou d'une propriété P (ou de Non (P) noté $\top$P) que doit vérifier obligatoirement tout élément x appartenant à la catégorie X.

Cela suppose qu'il existe diverses situations (ou cas) que l'on peut observer dans un "monde d'observation", dans lequel on peut mesurer des caractéristiques P(x) d'occurrences x d'objets décrits par des catégories X. On veut comparer ces caractéristiques observées aux caractéristiques prévues par le règlement pour juger de leur conformité.

Ainsi les prescriptions du règlement décrivent les caractéristiques obligatoires (ou interdites) que doivent présenter certains objets faisant partie du "monde étalon" auquel se rapporte le règlement. La confrontation des situations du "monde d'observation" avec les caractéristiques du "monde étalon" permettent de détecter les violations du règlement ou les possibilités non exploitées.

Nous proposons une formulation logique pour décrire ces différents types de prescriptions. Nous employons les notations suivantes. Les catégories d'objets seront représentées par les symboles X, Y, Z, W. Les propositions seront dénotées par les symboles P, Q, R, S, T, étant exprimées lorsque cela est nécessaire, par des prédicats indiquant en arguments les catégories d'objets sur lesquelles porte la proposition: P(X,Y) par exemple. Nous emploierons le symbole ! pour exprimer l'obligation ("il est obligatoire de ...") et le symbole £ pour exprimer la permission ("il est permis de ..."). Par exemple, P! exprime "il est obligatoire que P", P£ exprime "il est permis que P", ⊤P! exprime "il est interdit que P".

Nous utiliserons aussi le prédicat "élem" exprimant l'appartenance d'un elément à une catégorie d'objets: élem(x, X) exprime que x appartient à la catégorie X. Lorsqu'une partie de formulation est optionnelle, nous la mettrons entre crochets []. Lorsque plusieurs alternatives de formulation sont envisageables, nous les mettrons entre parenthèses {}.

A l'annexe 2 nous donnons les interprétations pour les expressions les plus courantes du métalangage réglementaire tel que dérivé à partir de l'étude du document de l'annexe 1.

On peut distinguer différentes catégories de prescriptions: les prescriptions exprimant l'obligation, celles qui expriment la possibilité, celles qui expriment les interdictions, les règles assertives, les règles de définition de catégories, les métarègles.

- **les prescriptions exprimant l'obligation**

Les catégories 1, 2, 3, 4, 5, 9, 10 et 12 expriment l'obligation. Par exemple, la catégorie 1 correspond à l'interprétation d'expressions du type:

{ [(sous réserve de, sauf si, en dehors de, à moins que) Q,] tout X doit P },

ou

{ X ne peut que P, à l'exception de Q }.

Nous exprimons sous une forme logique le contenu sémantique de ces expressions par la formule:

(Pour tout x) élem(x,X) [&⊤Q(x)] -> P(x) !

où élem(x,X) exprime l'appartenance d'une occurrence observée x à la catégorie X, et où P(x) ! exprime l'obligation de vérifier la proposition P(x) sur le cas observé.

- **les prescriptions exprimant la possibilité**

Les catégories 2', 9', 11, 11', 11" et 12' expriment la possibilité. Par exemple, la catégorie 11' correspond à l'interprétation d'expressions du type

(Les X peuvent P)

exprimées par

(pour tout x) elem(x,X) -> P(x) £;

La catégorie 11" correspond à l'interprétation d'expressions du type

(Dans le cas où P, Q [,à condition que R])

exprimées par

P [R] -> Q £.

- **les prescriptions exprimant les interdictions**

La catégorie 6 exprime l'interdiction:

([Toutefois] {il est interdit de, il n'est pas permis de, nul ne peut} P,
{en vue de, dans le but de, uniquement pour} Q, [si R])

Nous exprimons sous une forme logique le contenu sémantique de ces expressions par la formule:

P & Q [& R] -> violation du règlement (VIR)

où nous avons choisi d'énoncer la forme positive des propositions P, Q et R, et d'indiquer en conclusion de la règle qu'il y a violation du règlement (VIR).

- **les règles assertives**

La catégorie 7 décrit des expressions du type

({si, lorsque, alors que, pendant} P
{et {si, que}, ainsi que, de même que} Q {alors, il faut que, ','} R).

Celles-ci expriment des règles qui permettent de déduire nouvelles assertions; nous les exprimons sous la forme logique suivante:

P & Q -> R.

- **les règles de définition de catégories**

La catégorie 8 donne un exemple de type de règles qui permettent de définir des catégories par rapport à d'autres. Les expressions du type

(Tout X qui P, (est, est réputé) Y)

permettent de décrire les objets de la catégorie Y comme des spécialisations de la catégorie X (on définit une sous-classe au sens de la théorie des ensembles).

Nous exprimons ces règles sous la forme logique suivante :

(pour tout x) élem(x,X) & P(x) -> élem(x,Y)

- **les métarègles**

Les catégories 14, 15, 16 et 17 correspondent à des règles qui portent non pas sur les objets du monde observé, mais sur les clauses mêmes (articles, sections ou alinéas) du règlement. Nous les nommons métarègles. Elles ont une incidence certaine sur l'interprétation du règlement, mais ne peuvent pas se formuler simplement sous une forme logique analogue à celles que nous avons proposées pour les catégories précédentes. En effet, leur application dépend du contexte sur lequel porte l'article correspondant du règlement et peut se traiter de diverses manières: modifications de règles, duplication et ajustement de certaines règles précédemment énoncées, etc.

3.3 Générer des bases de connaissances déontiques

La reformulation du règlement telle que nous la proposons permet d'établir à partir d'expressions logiques du type

"conjonction de prémisses -> conclusions"

des regles d'inférence sous la forme classique (règles de production)

"SI conjonction de prémisses ALORS conclusions".

Par contre, le fait que les règles ainsi obtenues énoncent des prescriptions ne nous permet pas de les utiliser directement par un moteur d'inférence (à base de règles de production par exemple), pour faire vérifier si les cas présentés par les promoteurs sont conformes au règlement.

Par exemple, supposons que notre base de connaissance contienne une seule règle: "Tout X doit P" qui est reformulée en

"(pour tout x) élem(x,X) -> P(x) !".

Supposons que le promoteur présente un cas caractérisé par le fait (x1).

Un moteur d'inférence classique lancé en chaînage avant conclurait P(x1) et rajouterait ce fait comme élément descriptif du cas du promoteur.

Or dans la réalité, le cas peut présenter la propriété ⊤P(x1) et l'inférence serait erronée. Cela tient simplement au fait que les énoncés issus du règlement prescrivent des contraintes (du "monde étalon") par rapport auxquelles on doit valider la conformité des cas présentés (issus du "monde d'observation").

Nous nous intéresserons donc à un système à base de connaissances qui interrogera l'utilisateur afin de lui permettre de décrire le cas observé. Le moteur d'inférence utilisera les règles dérivées du règlement pour vérifier la validité des faits décrits par l'utilisateur par rapport aux prescriptions du règlement.

Pour pouvoir utiliser les moteurs d'inférence traditionnels qui mettent en oeuvre essentiellement une stratégie de raisonnement basée sur la règle du modus ponens de la logique classique, nous proposons les transformations suivantes concernant l'obligation, l'interdiction et la permission:

- *L'obligation P(x) -> Q(x) !* transformée en

  SI P(x) ⊤Q(x) ALORS VIR(art i) CAU(Q) interprétée par

  "Si P(x) et non Q(x) alors on a violation de l'article i du règlement, la cause étant le fait non Q(x)".

- *La permission P(x) -> Q(x) £* transformée en

  SI P(x) ⊤Q(x) ALORS POS(art i) CAU(Q) interprétée par

  "Si P(x) et non Q(x) alors on a la possibilité de la clause Q, d'après l'article i du règlement".

- *L'interdiction* a déjà été formulée dans ces termes:

  SI P & Q [& R] ALORS INT(art i) & CAU(P) interprétée par

  "Il est interdit de P, en vue de Q [à condition que R]" ou "Si P et Q [et éventuellement R], alors on a violation de l'article i du règlement, la cause étant l'interdiction P".

Ces transformations sont basées sur certaines caractéristiques que nous avons retenues pour le type de bases de connaissances déontiques que nous voulons générer et du mode de consultation que nous envisageons. Ce sont sur ces bases de la logique déontique que le système de manipulation de BAC s'appuiera pour transformer une BAC obtenue à partir d'un texte prescriptif en base de connaissances compatible avec l'utilisation de moteurs d'inférence commerciaux.

4. CARACTÉRISTIQUES DU SYSTÈME DE MANIPULATION DE BAC

Le système de manipulation de BAC que nous sommes en train de développer comporte deux composantes principales: un système d'acquisition des connaissances et un système de consultation.

4.1 Le système de consultation de bases de connaissances déontiques

La figure 1 présente le plan du sous-système de consultation de bases de connaissances déontiques, décrivant les principaux modes d'interaction qui seront offerts à l'utilisateur (Les rectangles représentent les processus du système; les rectangles à coins arrondis les accumulations d'information; les doubles carrés l'environnement).

La fonction P1 permet au système d'interroger l'utilisateur (E1) sur les caractéristiques du cas observé. Les réponses de l'utilisateur sont mémorisées dans le cas traité (A3). La conduite du dialogue avec l'utilisateur est faite en fonction du contenu de la base d'acquisition des connaissances relative au règlement concerné: BAC 'règlement' (A2).

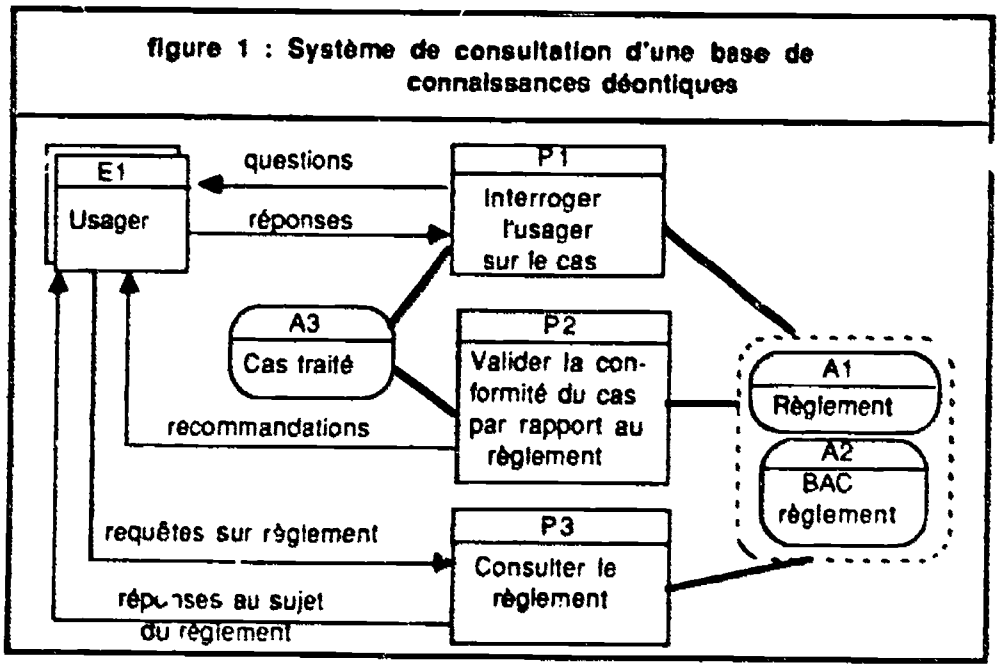
La fonction P2 permet de valider la conformité du cas traité (A3) par rapport aux prescriptions "réglementaires" contenues dans la BAC (A2). Des recommandations sont alors faites à l'utilisateur (E1).

La fonction P3 permet à l'utilisateur (E1) de consulter le règlement original (A1).

Ces différentes fonctions sont mises en oeuvre par le moteur d'inférence du système à base de connaissances en fonction des besoins de la consultation. Nous discutons dans le rapport d'état d'avancement des travaux (octobre 1988) des caractéristiques du moteur d'inférence et du module d'acquisition des connaissances.

Le type de consultation permis par nos bases de connaissances déontiques correspond à la validation des caractéristiques des cas traités par rapport aux prescriptions formulées sous la forme de règles dans la BAC. Les inférences qu'elles permettent de faire conduisent à des conclusions relatives à des violations du règlement par obligation non respectée, ou par interdiction transgressée, ou à des permissions non exploitées.

figure 1 : Système de consultation d'une base de connaissances déontiques

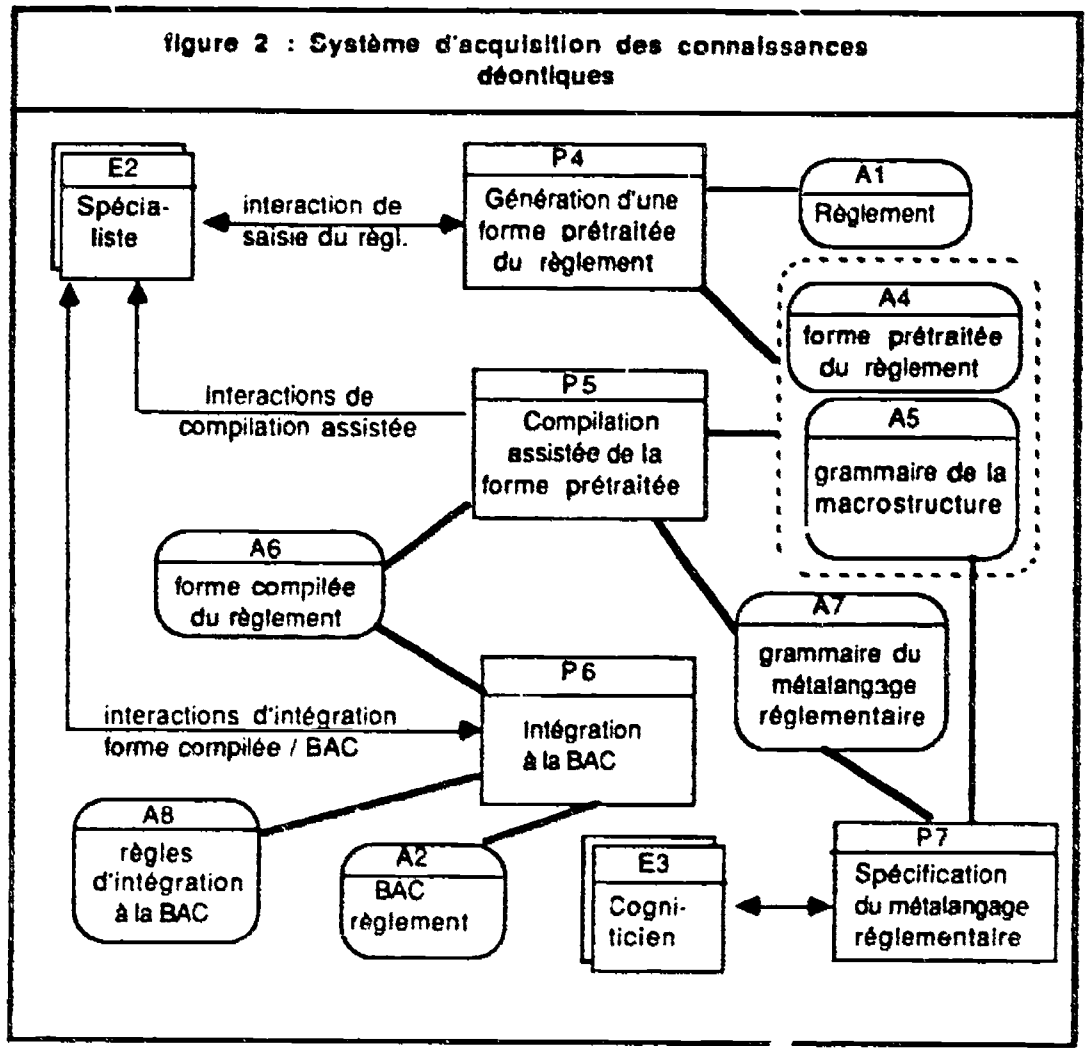


4.2 Le système d'acquisition des connaissances déontiques

Dans cette section nous exposons les principales caractéristiques d'un logiciel d'acquisition des connaissances à partir de l'analyse de textes de règlement. L'approche d'acquisition des

connaissances que nous proposons consiste en un prétraitement du texte du règlement, une compilation de la forme prétraitée, et l'intégration de la f-rme compilée au contenu de la BAC (figure 2).

figure 2 : Système d'acquisition des connaissances déontiques



Le premier processus consiste en la génération d'une forme prétraitée du texte (P4). Plusieurs traitements préliminaires peuvent être effectués sur le texte du règlement (A1) lui-même: mise en évidence des éléments de la macrostructure, détection des expressions du métalangage supportant la microstructure, identification des propositions élémentaires.

Pour générer la forme prétraitée du règlement (A4), on utilise des connaissances relatives à la présentation de la macrostructure du texte réglementaire: mots-clés, séparateurs d'énoncés, descripteurs spécifiés par la grammaire de la macrostructure.

Le processus P5 consiste en une "compilation assistée" de la forme prétraitée du règlement (A4) en fonction de la grammaire du métalangage réglementaire (A7) (grammaire de la microstructure), et permet d'obtenir une forme compilée du règlement (A6). Cette compilation tient compte dans certains cas d'éléments présents dans plusieurs phrases du règlement. Au cas où une phrase de la forme prétraitée du règlement (A4) ne pourrait pas être reconnue, le système demande de l'assistance au spécialiste (E2). Plusieurs scénarios pourront alors se présenter: une ambiguïté apparaît dans le texte, ou la grammaire du métalangage n'est pas suffisamment complète pour reconnaître une expression du texte, etc. Dans chaque cas, on demande au spécialiste de fournir une solution au problème rencontré et une explication de sa démarche de résolution que

le cogniticien pourra analyser plus tard. Le cogniticien pourra utiliser la fonction de spécification du métalangage réglementaire (P7) pour apporter les modifications pertinentes à la grammaire de la macrostructure (A5) ou à la grammaire du métalangage réglementaire (A7). Cette approche permet d'augmenter la richesse du métalangage réglementaire avec l'expérience de traitement de nouveaux textes réglementaires.

La forme compilée du règlement (A6) pourra avoir diverses caractéristiques en fonction des propriétés attendues pour la base d'acquisition des connaissances (A2). La forme compilée du règlement dépendra des caractéristiques du moteur d'inférence qui permettra d'exploiter la BAC. Ainsi pour un moteur d'inférence qui permettra de supporter un système de consultation de bases de connaissances déontiques tel que présenté à la figure 1, on pourra consulter [MOU 88].

Le processus d'intégration (P6) de la forme compilée du règlement (A6) au contenu de la BAC règlement (A2) doit permettre d'enrichir le contenu de la base de connaissances du système de consultation de bases de connaissances déontiques. Cette intégration se fera en fonction de règles d'intégration (A8) à préciser en fonction des caractéristiques de la BAC considérée.5.

5. TRAITEMENT DE LA MICROSTRUCTURE

La principale fonction du système d'acquisition des connaissances consiste à compiler le texte du règlement (prétraité) en fonction des caractéristiques du métalangage réglementaire spécifiées par la grammaire de la microstructure. Le résultat de la compilation permet de transformer les phrases du règlement sous la forme de structures logiques exploitables par le moteur d'inférence du système de consultation conformément aux règles énoncées aux paragraphes 3.2 et 3.3.

La grammaire de la microstructure permet au système d'acquisition des connaissances de procéder à un prédécoupage du texte traité. Ce prédécoupage met en évidence les principales composantes de chaque phrase: opérateur modal, connecteurs et portées.

La plupart des phrases dans un texte de règlement contiennent un opérateur modal. On appelle **opérateur modal** toute expression qui signale la modalité d'une phrase. Les trois modalités les plus courantes sont:

- l'obligation (Exemple: **Il faut qu'**un talus soit recouvert de végétation);
- l'interdiction (Exemple: La présence d'un tel déchet dans le sol d'un lieu d'enfouissement sanitaire **est prohibée** en vertu de la Loi);
- la possibilité (Exemple: L'exploitant d'un dépôt en tranchée peut y recevoir des déchets solides).

On distingue trois grandes catégories d'opérateurs modaux: les opérateurs monadiques, les opérateurs diadiques et les opérateurs spéciaux. Tout opérateur modal a une ou deux portées.

Une **portée** est une partie de la phrase courante sur laquelle s'applique l'opérateur modal. Une portée précédant un opérateur modal est appelée **portée-avant**, alors que **portée-arrière** désigne une portée qui suit l'opérateur. Par convention, on indique dans une phrase la portée en l'entourant d'accolades.

Un opérateur modal monadique a une portée-arrière mais aucune portée-avant. La portée-arrière est une proposition contenant au moins un verbe.

> Exemple: Il **est** qu'(un talus soit recouvert de végétation).

Un opérateur modal diadique a une portée-avant et une portée-arrière. La portée-avant est un terme ne contenant aucun verbe, alors que la portée-arrière est une proposition verbale.

> Exemple: (L'exploitant d'un dépôt en tranchée) **peut** (y recevoir des déchets solides).

Un opérateur modal spécial est un opérateur qui ne correspond pas à la description d'un opérateur monadique ou diadique. Il a toujours une portée-avant, mais est privé très souvent de portée-arrière. S'il en a une, celle-ci est un terme non-verbal.

> Exemple: (La présence d'un tel déchet dans le sol d'un lieu d'enfouissement sanitaire) **est prohibée** (en vertu de la Loi).

Une phrase peut contenir également, mis à part l'opérateur modal et ses portée, un ou plusieurs connecteurs. On distingue trois grandes catégories de connecteurs: les connecteurs inter-énoncés, les connecteurs de condition et les connecteurs d'exception.

Un **connecteur inter-énoncés** est un connecteur, placé généralement en début de phrase, qui relie la phrase courante à une autre phrase. "Toutefois" et "Cependant" sont des exemples de connecteurs inter-énoncés. Un tel type de connecteur n'a pas de portée.

> Exemple: Le recouvrement final d'un lieu d'enfouissement sanitaire doit être constitué d'au moins 60 centimètres de terre. **Cependant,** lorsque l'épaisseur des couches de déchets solides superposées atteint ou dépasse 6 mètres, le recouvrement final doit être constitué d'au moins 120 centimètres de terre.

On appelle **connecteur de condition** une expression qui introduit une condition. Comme connecteurs de condition, on retrouve, entre autres, les expressions "si", "lorsque" et "dans le cas où". Un connecteur de condition peut être placé n'importe où dans une phrase. Il a une portée-arrière composée de la condition qu'il introduit.

> Exemple: Cependant, **lorsque** (l'épaisseur des couches de déchets solides superposées atteint ou dépasse 6 mètres), le recouvrement final doit être constitué d'au moins 120 centimètres de terre.

Tout connecteur qui introduit une exception est appelé **connecteur d'exception**. "Sauf si" et "à l'exception de" sont des exemples de connecteurs d'exception. Tout comme un connecteur de condition, le connecteur d'exception a une portée-arrière et peut être situé n'importe où dans la phrase.

> Exemple: Le recouvrement final doit être constitué d'au moins 120 centimètres de terre sauf si (l'épaisseur des couches de déchets solides n'atteint pas 6 mètres).

6. CONCLUSION

Nous avons présenté les lignes directrices de notre recherche sur la mise au point d'un logiciel d'aide à la conception de bases de connaissances déontiques à partir de l'analyse de textes prescriptifs. Notre approche s'intéresse au métalangage réglementaire, considérant les propositions du règlement comme étant les composantes unitaires de manipulation.Une direction d'exploration complémentaire à cette recherche concerne la manipulation du contenu sémantique de la BAC. L'objectif envisagé serait de développer un système qui a une compréhension de la structure profonde des connaissances contenues dans le règlement, et non pas des seules connaissances de surface, à la manière de la génération actuelle de systèmes experts. Pour cela, on pourrait adopter une méthode de représentation des connaissances qui permette de spécifier le contenu sémantique des propositions du règlement, suivant une approche semblable à celle qui a été proposée par J. Sowa avec sa théorie des graphes conceptuels [SOW 84]. Le contenu sémantique de la BAC pourrait alors être traité avec un système de manipulation de graphes conceptuels [K-M 87a] [K-M 87b].

REMERCIEMENTS

Cette recherche est supportée par le CEFRIO, le Centre francophone de recherche en informatisation des organisations, subvention 1987-88.

Bibliographie

[C-L 88] J. C. CORITON, F. M. LFSAFFRE, *SYMA: système expert en milieu administratif*, in (RAU 88).

[D-Q 87] J. DESCHAMPS, M. QUENILLET, *Systèmes experts juridiques: une réalité*, in proceedings of the seventh Workshop Expert Systems and their Applications, Avignon May 1987.

[G-Q 87] GOUVERNEMENT DU QUÉBEC, *Règlement sur les déchets solides*, révisé mai 1987.

[H-K 85] P. HARMON, D. KING, *Expert Systems*, John Wiley and sons pub., New York 1985.

[H-W-L 83] F. HAYES-ROTH, D.A. WATERMAN, D. B. LENAT editors, *Building Expert Systems*, Addison Wesley 1983.

[KAL 72] G. KALINOWSKI, *La logique des normes*, Presses Universitaires de France, 1972.

[K-M 87a] A. KABBAJ, B. MOULIN, *Structures de connaissances dans un système de manipulation de graphes conceptuels*, rapport de recherche DIUL-RR-87-14, Université Laval, Département d'informatique, mai 1987.

[K-M 87b] A. KABBAJ, B. MOULIN, *Un système de manipulation de graphes conceptuels au coeur d'un dictionnaire de graphes conceptuels*, rapport de recherche DIUL-RR-87-15, Université Laval Département d'informatique, mai 1987.

[LIE 87] A. Von der Lieth GARDNER, *An artificial Intelligence Approach to Legal Reasoning*, The MIT Press, Cambridge, Mass, 1987.

[MAT 78] R. MATTESICH , *Instrumental reasoning and systems methodology*, An epistemology of applied and social sciences, D. Reidel Pub. comp., 1978.

[McC 77] J. McCARTHY, L. THORNE, *Reflections on TAXMAN : An experiment in artificial intelligence and legal reasoning*, Harvard Law Review, 90, 1977.

[MOU 85a] B. MOULIN, *La méthode è.P.A.S. pour la modélisation et la conception de systèmes*, rapports de recherche DIUL-RR-85-07, 08 et 09, Université Laval,septembre 1985.

[MOU 85b] B. MOULIN, *L'ingéniérie du savoir*, revue L'Ingénieur, n° 366, 71e année, Mars-Avril 1985,p 19-24.

[MOU 86a] B. MOULIN, *Les systèmes à base de connaissances dans les organisations*, revue L'Ingénieur, n° 375, 72e année, Septembre-Octobre 1986, pp 21,27.

[MOU 86b] B. MOULIN, *L'informatique cognitive des organisations*, revue Systèmes, printemps 1986, pp 3-9.

[MOU 87a] B. MOULIN, *Systèmes experts: une technologie à la portée des entreprises*, revue L'Ingénieur, n° 379, 73e année, mai juin 1987, pp 4,7.

[MOU 87b] B. MOULIN, *Une démarche pour la conception de systèmes experts*, revue L'Ingénieur, n° 379, 73e année, mai juin 1987, pp 17,21.

[MOU 88] B. MOULIN, *Réflexions sur la conception et l'utilisation de bases de connaissances déontiques*, Rapport de recherche CEFRIO, soumis en mars 88, accepté en juillet 88.

[M-R 89] B. MOULIN, D. ROUSSEAU, *Analyse de textes prescriptifs pour la génération de bases de connaissances déontiques*, Communication soumise et acceptée au colloque ICO 89, Québec, juin 1989.

[M-S 80] J. McCARTHY, N. S. SRIDHARAN, *The representation of an evolving system to legal concepts : I. Logical Templates*, in proceedings ot the Third Biennal Conference of the Canadian Society for Computational Studies of Intelligence, Victoria, B. C. 1980.

[RAU 88] J. C. RAULT editor, *proceedings of the eigth Workshop èxpert Systems and their Applications*, Avignon May 1988.

[SCH 88] U. J. SCHILD, *JURIX : a legal expert system*, in (RAU 88).

[SOW 84] J. F. SOWA, *Conceptual Structures: Information Processing in Mind and Machine*, Addison Wesley, 1984.

[S-S-K 86] M. J. SERGOT, F. SADRI, R. A. KOWALSKI, F. KRIWACZEK, P. HAMMOND, H.T. CORY, *The British Nationality Act as a Logic Program*, Communications of the ACM 29, 1986.

[TID 88] T.H. TIDRICK, *Expert consultant for law enforcement : the casen of alleged revenue officer embezzlement*, in (RAU 88).

[TUR 84] R. TURNER, *Logics for artificial intelligence*, Ellis Horwood Lim. pub. comp., John Wiley & Sons 1984.

[T-J-L 84] R. TREMBLAY, R. JOURNEAULT, J. LAGACÉ, *Guide de rédaction législative, publication du Gouvernement du Québec*, Société québécoise d'information juridique, 1984.

[WAT 85] D.A. WATERMAN, *A Guide to Expert Systems*, Addison Wesley 1985.

ANNEXE 1

Extrait du règlement sur les déchets solides

SECTION 4

ENFOUISSEMENT SANITAIRE

23) Zonage et plaines de débordement:

1- Il est interdit d'établir un lieu d'enfouissement sanitaire dans une plaine de débordement ou dans tout territoire zoné par l'autorité municipale pour fins résidentielles, commerciales ou mixtes (résidentielles-commerciales) et à moins de 150 mètres d'un tel territoire.

24) Aéroport:

1- Il est interdit d'établir un lieu d'enfouissement sanitaire à moins de 3 kilomètres d'un aéroport.

25) Voie publique:

1- Aucun lieu d'enfouissement sanitaire ne peut être établi à moins de 152,40 mètres de tout chemin entretenu par le ministre des Transports et à moins de 50 mètres de toute autre voie publique.

26) Distance de certains lieux:

1- L'aire d'exploitation d'un lieu d'enfouissement sanitaire doit être située à plus de 150 mètres de tout parc municipal, terrain de golf, piste de ski alpin, base de plein air, plage publique, réserve écologique créée en vertu de la Loi sur les réserves écologiques (LRQ, c R-26), parc au sens de la Loi sur les parcs (LRQ, c P-9), parc au sens de la Loi sur les parcs nationaux (SRC, 1970, c N-13), mer, fleuve, rivière, ruisseau, étang, marécage ou batture.

27) Distance de certains immeubles:

1 L'aire d'exploitation d'un lieu d'enfouissement sanitaire doit être située à plus de 200 mètres de toute habitation, institution d'enseignement, temple religieux, établissement de transformation de produits alimentaires, terrain de camping, restaurant ou établissement hôtelier détenteur d'un permis délivré en vertu de la Loi sur l'hôtellerie (LRQ, c H-3), colonie de vacances et établissement au sens de la Loi sur les services de santé et les services sociaux (LRQ, c S-5).28)

28) Lacs:

1- L'aire d'exploitation d'un lieu d'enfouissement sanitaire doit être située à plus de 300 mètres de tout lac.

29) Conditions hydrogéologiques:

1- L'enfouissement sanitaire des déchets solides doit s'effectuer sur un terrain où les conditions hydrogéologiques sont telles que les eaux de lixiviation s'infiltrent dans le sol et que le temps de migration des eaux y est supérieur à 5 ans avant de parcourir 300 mètres ou avant d'atteindre tout puits ou source servant à l'alimentation en eau potable et situé à une distance inférieure à 300 mètres, à moins que ces eaux n'aient fait résurgence auparavant. Dans ce dernier cas, elles doivent avoir circulé dans le sol pendant plus de 2 ans à une vitesse moyenne inférieure à 150 mètres par an.

2- Dans le cas où on ne retrouve pas les conditions hydrogéologiques décrites au premier alinéa, l'enfouissement sanitaire peut s'effectuer à condition que l'on procède à des aménagements afin d'empêcher l'infiltration dans le sol de toute eau de lixiviation. Toutefois, il n'est pas permis d'aménager un terrain en vue d'y effectuer de l'enfouissement sanitaire si une infiltration accidentelle risquait de compromettre la qualité des eaux d'une nappe exploitée à des fins d'alimentation en eau potable.

3- Dans le cas où les conditions hydrogéologiques d'un terrain d'enfouissement sanitaire sont telles que les eaux provenant de ce terrain s'écoulent en surface ou font résurgence avant 2 ans, on doit y installer un système permettant le captage complet de ces eaux et le traitement de celles-ci conformément aux exigences de l'article 30.

ANNEXE 2 :

Quelques réflexions relatives au métalangage utilisé pour la spécification d'un règlement

Dans cette annexe nous présentons une première esquisse d'analyse du métalangage qui est utilisé par les juristes pour décrire un texte de règlement. Ces réflexions ne présentent pas une étude exhaustive de ce métalangage, mais offrent un noyau de départ qui permettra au concepteur d'effectuer l'analyse du contenu d'un texte de règlement du point de vue de la logique déontique. Cette étude a porté sur la microstructure du texte de règlement fourni à l'annexe 1. Les conventions utilisées sont celles exposées au 3.2.

Nous pouvons donner les interprétations suivantes pour les expressions les plus courantes du métalangage réglementaire.

1. {[(sous réserve de, sauf si, en dehors de, à moins que) Q,] tout X doit P},

   {X ne peut que P, à l'exception de Q}.

   (pour tout x) élem(x,X) [& ⊤Q(x)]-> P(x) !

2. {[Dans tous les cas,] tout X, en vue de Q, doit P},
{Tout X doit P,{de sorte que, de manière à, afin de, pour [que], de façon à} Q}

(pour tout x) élem(x,X) & Q(x) -> P(x) !

2'. Les X peuvent P, seulement {au cas où, dans le cas où, dans le cadre de} Q)

(pour tout x) élem(x,X) & Q(x) -> P(x) £

3. Dans le cas de X, P.

(pour tout x) élem(x,X) -> P(x, !

4. {Tout X doit P. {Dans ce [dernier] cas, à cette fin, pour cela} {il faut [que], il est nécessaire {que, de}, il est obligatoire {que, de}, on doit} Q},
{X doit P, {avant, avant {que, de}} Q}.

(pour tout x)) élem(x,X) -> P(x) ! & Q(x) !

5. P est {obligatoire,nécessaire} pour Q.

Q -> P !

6. [{Toutefois,Cependant}] {il est interdit de, il n'est pas permis de, nul ne peut, il est défendu de} P, {en vue de, dans le but de, uniquement pour} Q, [si R].

P & Q [& R] -> violation du règlement (VIR)

7. {si, lorsque, alors que, pendant} P {et {si, que}, ainsi que, de même que} Q {alors, il faut que, ','} R.

P & Q -> R

8. Tout X qui P, {est, est réputé} Y.

(pour tout x) élem(x,X) & P(x) -> élem(x,Y)

9. {Dès, Aussitôt} que P, X doit Q. {X doit Q, après P}.

(pour tout x) élem(x,X) & P(x) -> Q(x) !

9'. {Dès, Aussitôt} que P, X peut Q. {X peut Q, après P}.

(pour tout x) élem(x,X) & P(x) -> Q(x) £

10. X doit P. Cependant, {lorsque, si} Q, X doit R.
{X doit P, sauf [si] Q. Dans ce[s] cas, R}

(pour tout x) élem(x,X) &T Q(x) -> P(x) !
(pour tout x) élem(x,X) & Q(x) -> R(x) !

11. (il est possible de, il est permis de, il est admis de) P

P £

11'. Les X peuvent P.

(pour tout x) élem(x,X) -> P(x) £

11". Dans le cas o P, Q [,à condition que R]

P [& R] -> Q £

12. X doit P. Il en est de même (de, pour) Y.

(pour tout x) élem(x,X) -> P(x) !
(pour tout x) élem(x,Y) -> P(x) !

12'. X peut P. Il en est de même (de, pour) Y.

(pour tout x) élem(x,X) -> P(x) £
(pour tout x) élem(x,Y) -> P(x) £

13. Toute combinaison de 'et', 'ou', 'ne pas', permet d'obtenir de nouvelles propositions à partir de la composition de propositions élémentaires.

Certaines expressions du métalangage réglementaire peuvent permettre de faire référence à d'autres articles.

14. Les {exigences, articles, clauses) C s'appliquent mutatis mutandis {à, aux) X.

Cette métarègle spécifie la validité des règles édictées par C pour élem(x,X). On peut par exemple :

- réécrire les règles de C en substituant dans les prémisses élem(x,X);
- augmenter les règles de C en ajoutant la clause élem(x,X) en conjonction dans les prémisses.

15. X doit P, conformément à l'article A.

Vérifier si les règles dérivées de l'article A sont conformes à cet énoncé, sinon les modifier.

16. {Nonobstant, malgré} les autres dispositions du règlement, X doit P.

Cette métarègle confère une priorité exclusive à cet énoncé.

17. (la présente section, le présent article) ne s'applique qu'aux X [tels que] énum.

Cette métarègle définit les valeurs acceptables pour X dans l'énumération énum.

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■

ANNEXE 3

Grammaire de la microstructure

La grammaire de la microstructure utilisée par le système d'acquisition des connaissances est définie comme suit:

```
<phrase> ->              [<partie1>] <partie-mod> [<partie2>] /

<partie1> ->             <partie11> "," /

<partie11> ->            [<conn-inter-énoncés> ","] <condition> [<exception>] !
                         [<conn-inter-énoncés> ","] <exception> !
                         <conn-inter-énoncés> /

<conn-inter-énoncés> ->  "Toutefois" ! "Cependant" ! "Dans" "tous" "les" "cas" !
                         "Dans" "ce" ["dernier"] "cas" /

<condition> ->           {<condition1>} /

<condition1> ->          <conn-cond> <corps-cond-prop> !
                         <conn-cond-terme> <corps-cond-terme> /

<conn-cond> ->           <dans> "le" "cas" "où" ! <lorsque> ! <si> ! <alors> <que> !
                         <au> "condition" <que> ! "et" "si" ! "et" <que> ! <pendant> <que> !
                         <dès> <que> ! <après> <que> ! <avant> <que-ou-de> /

<conn-cond-terme> ->     <dans> "le" "cas" <de> /

<exception> ->           <conn-except-prop> <corps-except-prop> !
                         <conn-except-terme> <corps-except-terme> /

<conn-except-prop> ->    <au> "moins" <que> ! <sous> "réserve" <de> !
                         <sauf> ["si"] /

<conn-except-terme> ->   <en> "dehors" <de> ! <au> "l" "'" "exception" <de> /

<partie2> ->             <condition> [<exception>] ! <exception> /

<partie-mod> ->          <op-mod-mon> <corps-portée-arrière> !
                         <corps-portée-avant> <partie-mod-suite> /

<partie-mod-suite> ->    <op-mod-diad> <corps-portée-arrière> !
                         <op-mod-spéc> [<corps-portée-arrière-spéc>] /

<op-mod-mon> ->          <il> "faut" [<que>] ! <il> ["n" "'"] "est" ["pas"]
                         <complément> <que-ou-de> /

<complément> ->          "obligatoire" ! "nécessaire" ! "interdit" !
                         "défendu" ! "permis" ! "possible" ! "admis /
```

```
<op-mod-diad> ->        ["ne"] <verbe-mod> ["pas"] /

<verbe-mod> ->          "doit" ! "doivent" ! "peut" ! "peuvent" /

<op-mod-spéc> ->        "est" <compl-op-spéc> !
                        ["ne"] <applique> [<pas-ou-que>] [<mut-mut>]/

<compl-op-spéc> ->      "permis" ! "permise" ! "interdit" ! "interdite" !
                        "défendu" ! "défendue" ! "prohibé" ! "prohibée" !
                        "réputé" ! "réputée" ! "obligatoire"! "nécessaire" /

<applique> ->           "s" "'" "applique" ! "s" "'" "appliquent" /

<mut-mut> ->            "mutatis" "mutandis" /

<q :e-ou-de> ->         <que> ! <de> /

<que> ->                "que" ! "qu" "'" /

<de> ->                 "de" ! "d" "'" /

<pas-ou-que> ->         "pas" ! <que> /

<il> ->                 "Il" ! "il" /

<dans> ->               "Dans" ! "dans" /

<lorsque> ->            "Lorsque" ! "lorsque" ! "Lorsqu" "'" ! "lorsqu" "'"/

<alors> ->              "Alors" ! "alors" /

<si> ->                 "Si" ! "si" /

<aa> ->                 "A" ! "a" /

<pendant> ->            "Pendant" ! "pendant" /

<dès> ->                "Dès" ! "dès" /

<après> ->              "Après" ! "après" /

<avant> ->              "Avant" ! "avant" /

<sous> ->               "Sous" ! "sous" /

<sauf> ->               "Sauf" ! "sauf" /

<en> ->                 "En" ! "en" /;
```

Les règles d'écriture de cette grammaire sont les suivantes:

- La grammaire se compose de plusieurs règles.
- Une règle est formée d'un membre gauche non-terminal, d'une flèche et d'un membre droit composé de terminaux et de non-terminaux.
- Une chaîne de caractères, placée entre guillemets, représente un terminal que l'on peut retrouver dans un texte prescriptif (exemple. "il").
- Chaque mot que l'on peut retrouver dans u texte prescriptif, chaque signe de ponctuation et l'apostrophe sont considérés comme étant chacun des terminaux qui doivent être placés entre guillemets (Exemple: "Il" "faut").

- Un non-terminal est un mot placé entre crochets pointus. Il est défini en termes de terminaux et / ou de non-terminaux (Exemple: <il>).

- Le choix entre plusieurs membres droits est indiqué par un point d'exclamation (Exemple: "interdit" ! "interdite")

- (...) signale la répétition de 0 à n fois des symboles placés entre accolades (Exemple: {<condition1>}).

- [...] équivaut à au plus une apparition des symboles placés entre crochets (Exemple: "sauf" ["si"]).

- Chaque règle est suivie du symbole "/".

- Le point-virgule indique la fin de la grammaire.

Certains non-terminaux ne sont pas définis dans la grammaire présentée dans cet article, car ils devraient l'être par un nombre trop imposant de terminaux possibles. Par exemple, **<corps-portée-avant>** englobe toute chaîne de caractères qui précède un opérateur modal. Nous distinguons ces terminaux plus généraux des autres terminaux en les écrivant en caractères gras.

LOGITEXTE
UN LOGICIEL DE CONCEPTION TEXTUELLE ASSISTÉE PAR ORDINATEUR

Jean-Yves Fréchette et Raymond Hamel
Cegep F.X. Garneau

UN LOGICIEL QUI MIME L'ACTE D'ÉCRITURE?

L'abondance des « outils pédagogiques » informatisés mis au point pour l'enseignement de la langue maternelle illustre que la didactique des langues naît d'abord d'une *philosophie* de la langue, d'une conception *a priori* qu'on se fait du fonctionnement du système linguistique et que le logiciel se charge de reproduire dans une espèce de mimétisme fonctionnel.

Ainsi, chaque nouveau logiciel qui prétend *traiter* la langue propose plus qu'un scénario informatique; il affiche carrément un certain *esprit* d'intervention dans la langue: d'où ces logiciels nés qui d'une approche structurale, qui d'une approche stylistique, qui d'une approche psychologique...

Il serait faux de prétendre qu'il n'y a pas à la base de *LogiTexte*[1] une pareille conception de la langue, une conception qui dicte non seulement les grands paramètres de l'architecture du logiciel comme tel, mais qui détermine également les comportements de l'utilisateur pour une utilisation optimale.

En effet, nous avons voulu que, dans sa mécanique même, *LogiTexte* puisse **mimer l'acte de parole** de la façon la plus simple possible. Nous avons voulu que les professeurs puissent utiliser un outil informatique qui serve:

1) leurs propres besoins d'outils didactiques pour des démonstrations et des explications lors des activités d'acquisition de connaissances;

2) les besoins d'exploration concrète des usagers (les élèves) qui veulent toucher, palper et manipuler les différents mécanismes de la structure linguistique.

Dès le départ, nous espérions que la pratique quotidienne de LogiTexte en vienne à suggérer l'idée que la langue, dans sa matérialité, est un ensemble qui fonctionne à partir de règles précises. Mais nous voulions également suggérer l'idée que la langue, comme système d'expression, n'existe qu'au moment même de son utilisation et que la fabrication du texte crée des conditions d'exploitation du matériau linguistique qui valent tant par la reprise d'éléments stables de la langue que par leur transgression.

Nous avons voulu que les utilisateurs de *LogiTexte* décrivent la langue comme une **structure** et soient capables de l'objectiver en ces termes; nous avons voulu qu'ils soient capables de décrire leur pratique d'écriture comme un ensemble mobile et rigoureux. Chaque fois que nous

---

[1] LogiTexte est un logiciel de Conception Textuelle Assistée par Ordinateur mis au point par les auteurs de cet article et développé selon une approche de développement en contexte par prototypage tout au long d'un travail de programmation et de validation avoir des élèves de 13-14. Cette expérience a eu cours à l'École secondaire Dollard-des-Ormeaux de Valcartier dans la classe de monsieur Bruno Laliberté.

nous servons du code de la langue pour nous exprimer, nous créons autant que nous empruntons. Nous empruntons les règles, l'espace commun entre les usagers, la procédure minimale acceptée par tous ceux et celles qui parlent la même langue; nous empruntons les *mots* qui désignent les différents contenus que nous voulons voir se répandre d'une oreille à l'autre et nous empruntons également les *règles de combinaison* qui se manifestent à chaque moment de la construction du discours.

Les mots sont nombreux, on le sait. Ils représentent un vaste répertoire que nous ne cessons jamais d'acquérir, même à l'âge adulte. D'où cette nécessité d'apprendre les multiples composantes lexicales de la langue; d'où la nécessité de commencer cet apprentissage au plus tôt et avec, si possible, cette rigueur qui ferait en sorte que l'apprenant puisse étendre ses prises sur le réel en se constituant un lexique personnel organisé recoupant différentes thématiques. La première manoeuvre d'utilisation de LogiTexte consiste donc à bâtir des fichiers lexicaux.

REPRODUIRE LE FONCTIONNEMENT DE LA LANGUE

Avec un logiciel qui permettrait à chaque moment du processus d'écriture 1) de **choisir** les mots du texte et 2) de **combiner** ces unités lexicales entre elles pour construire la structure de la phrase nous avons cru qu'il serait possible alors d'illustrer l'un des aspects les plus dynamiques de la langue: celui de la sélection paradigmatique et celui de la combinaison syntagmatique[2].

Jakobson qui simplifiait habilement le processus du discours, disait que chaque acte de parole se construit de la façon suivante: le locuteur procède d'abord à une **sélection** d'éléments de signification - les mots - (ce qu'il appelait *l'axe paradigmatique*: sorte de réservoir dans lequel le sujet parlant puise les éléments de son vocabulaire) puis à une **combinaison** de ces mêmes éléments (c'est ce que Jakobson appelait *l'axe syntagmatique* ou plus simplement la séquence des mots de la phrase). Jakobson insistait pour dire que ces deux activités, bien que complémentaires, n'en étaient pas moins opposées: l'une faisant appel à la mémoire des mots, donc au passé, l'autre au présent de l'acte de parole.

Si on voulait créer un logiciel qui permette à l'usager de reproduire le plus fidèlement possible l'acte de parole dans les deux dimensions essentielles évoquées par Jakobson, il fallait donc recourir à des structures informatiques qui permettraient de les illustrer symboliquement par l'une des fonctions du logiciel. Il nous fallait donc imaginer une structure informatisée où fonctionneraient simultanément une fonction qui représenterait **l'axe paradigmatique** (la mémoire des mots) et une autre fonction qui représenterait elle **l'axe syntagmatique** (la combinaison des mots entre eux, l'assemblage de la phrase ou même du texte) tout en tenant compte des contraintes propres à l'acte de combinaison lui-même, c'est-à-dire au respect des règles de grammaire.

UNE BASE DE DONNES LEXICALES + UN ÉDITEUR DE TEXTE

L'idée de combiner une base de données lexicales et un éditeur de texte nous apparaissait alors plus que justifiée. La banque de mots correspondait assez bien à l'idée de ce que Jakobson se faisait du *paradigme* et l'éditeur de texte quant à lui permettrait d'accomplir toutes les fonctions de combinaisons de *l'axe syntagmatique*.

---

[2]Notre visée théorique reprend simplement le modèle si simple fourni par Roman JAKOBSON, Essais de linguistique générale, Paris, Éditions de Minuit, 1963. Jakobson y fait remarquer que l'acte de parole naît de deux actions simultanées: 1) choix/**sélection** des éléments lexicaux (mots) dans le réservoir personnel de chacun et 2) enchaînement/**combinaison** de ces différents mots dans la chaîne parlée de la phrase.

L'utilisateur pourrait à n'importe quel moment du processus d'écriture avoir recours à une banque de données où il aurait emmagasiné tout un répertoire lexical organisé; il pourrait, à son gré, commander des suggestions de mots si précises qu'il verrait se *placer* à la position souhaitée un mot (ou des mots) correspondant à la catégorie lexicale désirée et parfaitement accordé(s) en genre et en nombre avec les éléments du contexte immédiat. Mais plus que tout, l'usager de ce logiciel pourrait jouir de toute la latitude d'un système « ouvert » qu'il pourrait construire à sa guise et adapter à ses besoins d'expression.

PRODUIRE UN TEXTE CODÉ

Avant d'accomplir toutes ces tâches cependant, l'usager devra d'abord consentir à se plier au jeu d'une discipline minimale: il devra produire un **texte chiffré**. Un code d'appel simple - un nombre de trois chiffres - permettra à l'usager d'indiquer avec précision quelles seront les caractéristiques du mot souhaité.

- Ainsi la colonne des centaines permettra d'indiquer la **catégorie lexicale** (nom = catégorie 100; adjectif = catégorie 200; verbe = catégorie 300).
- La colonne des dizaines signale le **genre** (masculin = catégorie 10; féminin = catégorie 20).
- La colonne des unités indiquera le **nombre** (singulier = le chiffre 1; pluriel = le chiffre 2).

Il est ainsi possible d'obtenir une série de dix variables morphologiques dont l'usager pourra commander l'apparition dans son texte au moment voulu.

Caractéristiques morphologiques	Équivalences codées
a) un nom masculin singulier	1 1 1
b) un nom féminin singulier	1 2 1
c) un nom masculin pluriel	1 1 2
d) un nom féminin pluriel	1 2 2
e) un adjectif masculin singulier	2 1 1
f) un adjectif féminin singulier	2 2 1
g) un adjectif masculin pluriel	2 1 2
h) un adjectif féminin pluriel	2 2 2
i) un verbe singulier	3 0 1
j) un verbe pluriel	3 0 2

Liste des variables morphologiques

UN TEXTE CHIFFRÉ

Lorsque les usagers se seront familiarisés avec ce code[3], c'est à ce moment précis qu'on pourra leur présenter le concept d'une nouvelle écriture paradoxalement basée sur la composition de structures *insignifiantes* (sans signification), de structures d'où se seraient absentées momentanément trois catégories lexicales capitales. Le jeu consiste alors à créer des structures de phrase où le principal élément qui doit apparaître n'est pas le mot mais son équivalent numérique. On verra alors surgir de curieuses graphies. Des suites de caractères où les lettres et les chiffres cohabitent dans la même séquence et où on croirait reconnaître une certaine ressemblance avec la phrase française type: **des 112 212 302 parmi les 122 222**. S'il active la commande Substitution Complète, l'usager sera surpris de voir apparaître à l'écran quelque chose comme **des coquillages huileux voyagent parmi les îles brumeuses.**

PROGRAMMER L'ARMATURE SYNTAXIQUE

Transposé en contexte pédagogique, ce protocole d'utilisation permettrait à l'usager de *dessiner* a priori l'armature d'une structure syntaxique vide et voire même de tout un texte. Il pourrait *programmer* la structure syntaxique de sa phrase - ou de son texte - de manière à y faire apparaître, en temps voulu, les mots qui viendraient donner un sens à l'ensemble de ce squelette. Il pourrait *meubler* en quelque sorte une structure vide de sens pour la garnir des mots de son choix.

Ce processus, on le comprendra facilement, aura l'immense avantage de permettre à l'usager d'isoler la structure de la phrase comme un objet autonome; dissociant ainsi l'objet syntaxique du reste des composantes linguistiques de la phrase, l'usager pourra le concevoir comme objet d'apprentissage facilement *manipulable* et procéder à des acquisitions de connaissances d'ordre strictement syntaxique.

CONSTRUIRE SON UNIVERS LINGUISTIQUE

Avec un tel logiciel, les mots - bien que suggérés par un processus de pige au hasard - ne font jamais que représenter l'univers linguistique de l'usager puisque c'est lui qui, dans un premier temps, a garni chacun des fichiers lexicaux. De sorte que, s'identifiant instantanément aux mots suggérés par LogiTexte, l'usager peut alors se reconnaître - retrouver en quelque sorte la trace de son propre travail antérieur - à chacun des moments d'écriture. Dans ces conditions, le logiciel de CTAO agit comme un outil qui favorise le réinvestissement de ses propres matériaux linguistiques.

POLYVALENCE DES INTERVENTIONS D'ÉCRITURE

Structure vide que l'usager « meuble » au fur et à mesure de sa progression scolaire, LogiTexte agit comme outil d'apprentissage et d'exercices dans la manipulation du matériau linguistique écrit.

---

[3] Ceux qui aimeraient travailler avec un code alphabétique pourront le faire: un 111 deviendra ainsi un #nms; un 222, un #afp; etc.

LogiTexte vise également à faire aimer l'écriture. Il va de soi que toute manipulation concrète du matériau verbal qui s'inscrit dans la poursuite des objectifs de production, ne peut faire autrement que développer un nombre important de pratiques connexes comme:

a) accroître son réseau lexical personnel
b) perfectionner ses modèles syntaxiques du code écrit de la langue
c) affiner ses stratégies personnelles de lecture et/ou de décodage
d) multiplier les activités « brouillonnage »
e) manipuler le jeu des interactions textuelles (perspectives stylistiques)

Il serait possible de concevoir des scénarios pédagogiques qui intégreraient LogiTexte depuis les tout premiers instants du primaire. Il conviendrait alors de voir comment ce logiciel-outil peut favoriser l'apprentissage des notions de base du vocabulaire des usagers tout en favorisant l'acquisition des habiletés fondamentales de la langue comme la maîtrise du schéma de la phrase simple par exemple.

LOGITEXTE: UN LOGICIEL D'APPLICATION OUVERT

Comme tout logiciel éducatif ouvert, LogiTexte ne sera jamais riche que des efforts de chacun des usagers pour nourrir ses propres fichiers. A priori, il n'y aura pas dans LogiTexte d'exercices « tout fait d'avance » programmés par le maître ou le concepteur du logiciel. Il n'y aura pas non plus de banques de lexiques que l'usager pourra piller et utiliser comme bon lui semble. **Tout doit être bâti par l'usager**: les lexiques, les structures et les textes, bien sûr!

Tout dans LogiTexte est construit de telle sorte que l'usager puisse **intervenir le plus souvent possible dans le processus d'écriture**. LogiTexte ne permettra jamais autre chose que la reproduction exacte des compétences de l'usager. En fait, LogiTexte ne *travaille* qu'à partir des matières premières qu'on lui apporte; voilà pourquoi il ne peut être autre chose que le reflet ponctuel de la compétence linguistique de chacun, qu'il pourra, avec la complicité du maître, parfaire et bonifier.

C'est pourquoi, en ouvrant LogiTexte pour la première fois, l'usager sera devant une structure vide, une **nouvelle structure** ou un **nouveau lexique**. Pour parvenir à l'écriture, il devra faire ce que fait toute personne lorsqu'elle apprend sa langue:

1) apprendre des mots
2) les répertorier dans des fichiers lexicaux personnels
3) formuler des schémas de phrases conformes à la syntaxe française
4) insérer au bon moment (dans les bonnes positions) les bons mots.

OUVRIR LA CTAO AUX BESOINS DES INDUSTRIES DE LA LANGUE

Nous croyons que le principe de la CTAO pourrait être appliqué avec profit pour des segments précis de la clientèle des industries de la langue. Nous comptons bien poursuivre nos recherches et appliquer ces principes aux besoins rédactionnels de certain type d'entreprises (conception publicitaire, rédaction de rapports, conception textuelle au sens large). Des logiciels comme LogiTexte pourraient dépasser le cadre de formation scolaire et s'ajuster aux besoins de formation permanente et de recyclage des employés et des cadres d'entreprise. La CTAO pourrait

alors permettre aux industries de la langue de franchir un pas dans la tâche qu'elles se sont fixées de répondre aux besoins de production et de manipulation au sens large des ensembles textuels des sociétés industrialisées.

Nous travaillons dès à présent sur un prototype de logiciel (LogiTexte II) qui ajoutera aux fonctions déjà décrites d'autres fonctions capables notamment de repérer dans un texte le sens précis d'un mot donné. Il s'agit là, croyons-nous, d'un outil qui sera particulièrement apprécié des traducteurs et de ceux qui s'intéressent aux problèmes d'apprentissage d'une langue seconde.

Auteur **Jean-François Montreuil**
Université Laval

Titre **La notion de sémantique en intelligence artificielle**

RÉSUMÉ

Le terme «sémantique» est largement utilisé par tous les informaticiens s'intéressant au traitement automatique des langues naturelles. Par ailleurs, le terme est aussi connu et employé par des chercheurs provenant d'autres disciplines, telles la philosophie du langage et la linguistique. S'agit-il toujours de la même sémantique? Afin de répondre à cette question, nous passerons rapidement en revue ce que l'on entend par «sémantique» en informatique linguistique. Ceci nous amènera à nous interroger sur la place de la linguistique en intelligence artificielle.

Auteur **Jean-François Montreuil**
Université Laval

Titre **La notion de sémantique en intelligence artificielle**

RÉSUMÉ

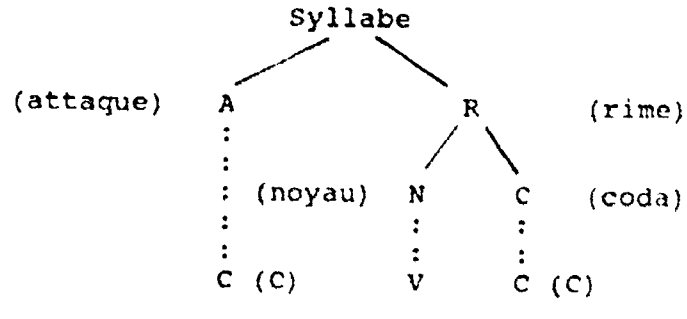
Le terme «sémantique» est largement utilisé par tous les informaticiens s'intéressant au traitement automatique des langues naturelles. Par ailleurs, le terme est aussi connu et employé par des chercheurs provenant d'autres disciplines, telles la philosophie du langage et la linguistique. S'agit-il toujours de la même sémantique? Afin de répondre à cette question, nous passerons rapidement en revue ce que l'on entend par «sémantique» en informatique linguistique. Ceci nous amènera à nous interroger sur la place de la linguistique en intelligence artificielle.

L'ORGANISATION DES DURÉES SEGMENTALES AU SEIN DE LA RIME SYLLABIQUE

Marise Ouellet
Université de Montréal

La production d'une parole de synthèse de bonne qualité exige une connaissance approfondie des structures et des éléments qui composent la chaîne parlée (Santerre: 1988). L'étude que nous avons entreprise s'inscrit dans la foulée des recherches fondamentales en linguistique. Elle a pour objet les variations des durées vocaliques en français et vise l'élaboration d'un modèle pouvant être utilisé pour la synthèse par règles.

Compte tenu du nombre considérable de facteurs pouvant modifier les durées segmentales (Ouellet: 1989), nous avons limité notre recherche au cadre défini par la rime syllabique (Selkirk: 1982):



L'exercice auquel nous nous sommes livrée consiste à décrire, regrouper et classer certaines voyelles du français, en nous basant sur leurs durées respectives ainsi que sur les modifications que subissent ces durées selon la nature des consonnes formant la coda. L'analyse des données recueillies nous permet donc d'opérer certains classements pour les voyelles sur la base des variations des durées intrinsèques et co-intrinsèques (Klatt: 1976).

Les variations des durées intrinsèques concernent les voyelles prises isolément. Du point de vue phonétique, les durées vocaliques varient selon les degrés d'aperture; les voyelles fermées étant systématiquement plus brèves que les voyelles ouvertes. On qualifie aussi de variations intrinsèques les différences de durée entre voyelles contribuant à des distinctions phonologiques. Le trait de durée peut supporter à lui seul les oppositions phonologiques ou être assorti de variations de timbre, comme cela est le cas en franco-québécois (Santerre: 1974).

Les variations de durée de type co-intrinsèque sont celles associées à la concaténation des unités dans la chaîne. La juxtaposition des segments vocaliques et consonantiques a pour conséquence de modifier les durées segmentales. Une voyelle, par exemple, sera plus brève ou plus longue selon que la consonne qui lui succède comporte le trait de sourdité ou de sonorité.

Il existe toutefois, entre les segments juxtaposés, des liens plus ou moins étroits. Ainsi, les voyelles subissent davantage l'influence des consonnes qui leur succèdent que celle des consonnes qui leur sont antéposées. Ce phénomène observé en phonétique trouve une correspondance dans les théories phonologiques fondées sur la structure des syllabes. Selon certaines approches de la phonologie métrique, le noyau et la coda d'une syllabe sont plus intimement liés que ne le sont l'attaque et le noyau (Selkirk: 1980)

Nous avons donc conçu un corpus nous permettant d'observer le noyau et la coda dans leurs rapports temporels au sein de la rime. Un ensemble de phrases à lire dont les structures syntaxiques et accentuelles demeurent invariantes, ont été soumises à des locuteurs. Ces phrases du type "Le mot _____ sonne bien" ou "Le mot _____ détonne" renfermaient des mots ou des logatomes de structure #CVC#. Ces séquences de trois segments débutaient par la consonne /p/.

Les segments vocaliques que nous avons retenus sont répartis sur trois degrés d'aperture. Il s'agit des voyelles fermées /i, u/ ainsi que de voyelles ouvertes et mi-ouvertes, orales et nasales. Pour les voyelles ouvertes et mi-ouvertes orales, nous avons vérifié l'existence, chez nos locuteurs, d'oppositions de durée en induisant par le biais de la graphie (p.e. patte - pâte) des voyelles longues et des voyelles brèves par nature (Santerre: 1974). Le système vocalique maximal qu'il nous était possible d'obtenir de cette façon est le suivant:

i	u	
E/ɜ		Ẽ
a/ɑ		ã ★

Les segments formant la coda dans les séquences #CVC# nous permettent d'évaluer l'influence que peuvent exercer sur les voyelles les traits consonantiques d'occlusion, de constriction, de sourdité et de sonorité. Les consonnes retenues sont /p t k b d g/ pour les occlusives et /f s ʃ v z ʒ / pour les constrictives. Ces contextes consonantiques peuvent être regroupés selon leur tendance à allonger ou à abréger les voyelles (Klatt: 1976). Le trait de sourdité associé au trait d'occlusion forme une combinaison abrégeante alors que celui de sonorité jumelé au trait de constriction produit un contexte allongeant. L'influence que peuvent exercer les occlusives sonores et les constrictives sourdes est alors considérée comme moins importante.

Nous avons procédé à l'étude des productions orales de deux locuteurs, l'un d'origine française, l'autre d'origine québécoise. L'introduction de cette nouvelle variable déterminée par la provenance de nos informateurs est motivée par une tendance bien connue consistant à délaisser les oppositions complexes du timbre et de durée dans le français hexagonal (Martinet: 1969). Par contre, Delattre et Monnot (1981) ont démontré que les voyelles nasales comportent des durées plus importantes que leurs contreparties orales dans cette variété du français.

Pour le franco-québécois, en plus de confirmer le maintien des oppositions entre deux E et deux A, les études menées par Jacques (1974) et Santerre (1974) ont permis de démontrer que les voyelles longues et brèves par nature réagissaient différemment aux contextes consonantiques dans la rime. Les brèves telles que /a/ et /E/ sont plus malléables et sont beaucoup plus influencées par la consonne qui leur succède que ne peuvent l'être les longues /ɑ/ et /ɜ/ ou encore les nasales /ã/ et /Ẽ/. A la différence de ces deux études, nous avons tenu à contrôler toutes les variables dont nous ne désirons pas mesurer l'influence (variables rythmiques, accentuelles, syntaxiques, etc.).

Les résultats présentés ici sont issus d'un corpus préliminaire où nous avons volontairement limité notre champ d'investigation aux éléments formant la rime syllabique. Les voyelles sont regroupées sous les phonèmes que nous avons tenté d'induire dans les tableaux. Ainsi, nous avons mis sous le phonème /ɑ/, les réalisations obtenues dans des séquences #CVC# telles que "pâte". Ces regroupements provisoires sont réévalués en raison des critères que nous avons retenus. Nous commenterons ces tableaux en tentant de catégoriser les voyelles sur la base de ces mêmes critères.

---

*Pour les voyelles orales: /a/ de "pattes", /ɑ/ de "pâtes", /E/ de "faites" et /ɜ/ de "fêtes".

TABLEAU I:

Durées moyennes des voyelles

fr. France		fr. Québec	
/u/	114.1 ms.	/i/	108.9 ms.
/i/	129.7	/u/	109.3
/E/	141.6	/E/	117.7
/a/	146.5	/a/	136.8
/ɑ/	146.6	/ɑ/	204.7
/ɔ/	160.9	/ɔ/	206.3
/Ẽ/	210.1	/Ẽ/	211.4
/ã/	223.6	/ã/	212.8

Chez nos deux informateurs, les durées moyennes des noyaux vocaliques tendent à confirmer le principe selon lequel la durée augmente avec le degré d'aperture. Cette tendance est toutefois inversée au niveau des voyelles longues attendues soit /ɑ/ et /ɔ/. Ce sont, par ailleurs, les voyelles nasales qui prédominent au plan des durées.

Les principales différences observées entre nos deux informateurs se retrouvent chez les voyelles longues orales que nous avons tenté d'induire. Chez notre locuteur français, le /a/ n'est pas distinct du /ɑ/ alors que le /ɔ/ long a une durée moyenne supérieure à celle du /E/ bref. Dans l'idiolecte de notre informateur québécois, on voit s'opérer une coupure assez nette entre les voyelles /i, u, a, E/ qui font partie, selon Santerre, des voyelles brèves et /ɑ, ɔ, ã, Ẽ/ qui sont, pour les deux premières, des longues par nature.

L'observation des durées moyennes des noyaux vocaliques ne nous permet pas de poser l'existence d'un groupe de longues orales dans l'idiolecte de notre informateur français. Il serait plutôt hasardeux, en effet, de croire que les différences de durées entre /ɔ/ et /E/ pourraient être significatives alors que, par ailleurs, /a/ et /ɑ/ ne se distinguent pas l'une de l'autre sur cette base.

Dans l'échantillon recueilli chez notre informateur franco-québécois, les durées moyennes nous portent à distinguer deux groupes de voyelles, des brèves dont la durée moyenne n'excède pas 150 ms et des plus longues dont les durées dépassent 200 ms.

TABLEAU II:

Écart à la moyenne en % selon la nature de la coda

fr. France						fr. Québec					
	$\bar{x}$	ptk	bdg	fsʃ	vzʒ		$\bar{x}$	ptk	bdg	fsʃ	vzʒ
/i/	129.7	-42%	-12%	-13%	+65%	/i/	108.9	-36%	-23%	- 8%	+67%
/u/	119.1	-29%	-11%	-10%	+47%	/u/	109.3	-43%	+ 7%	-27%	+82%
/a/	146.5	-27%	- 6%	-19%	+52%	/a/	136.8	-33%	+ 7%	-12%	+35%
/E/	141.1	-35%	-13%	-18%	+65%	/E/	117.7	-23%	-15%	-12%	+60%
/ɑ/	146.6	-17%	- 2%	-15%	+34%	/ɑ/	204.7	+ 5%	-15%	- 3%	+12%
/ɔ/	160.9	-39%	- 7%	-20%	+67%	/ɔ/	206.3	-14%	-12%	0%	+25%
/ã/	223.6	-12%	-11%	- 9%	+10%	/ã/	212.8	0%	0%	- 6%	0%
/Ẽ/	210.1	-13%	- 1%	- 5%	+18%	/Ẽ/	211.4	0%	- 9%	- 6%	0%

Dans cette analyse, nous avons introduit les variables constituées par différents traits consonantiques. Nous pouvons alors évaluer l'influence exercée sur les noyaux vocaliques par les caractères d'occlusion, de constriction, de sourdité et de sonorité. Ce tableau indique, en pourcentage, ce qu'il faut ajouter ou retrancher à la durée des voyelles pour obtenir une durée moyenne pour chacune d'entre elles dans des contextes consonantiques précis.

Nous constatons immédiatement que les voyelles varient davantage lorsqu'elles précèdent les occlusives sourdes /p t k/, où elles s'abrègent, et les constrictives sonores /v z ʒ/ devant lesquelles elles s'allongent. On peut distinguer, à cet égard, deux groupes de voyelles chez nos informateurs: celles qui varient de façon marquée et celles qui subissent des modifications moins importantes dans leur durée. Ce dernier type de voyelles serait, en quelque sorte, moins perméable à l'influence exercée par les consonnes abrégeantes et allongeantes formant la coda. Elles correspondraient, par le fait même, aux longues phonologiques telles que décrites par Santerre (1974). Il s'agit des voyelles /ɑ, ɔ, ɑ̃, Ɛ̃/ dans l'idiolecte de notre sujet québécois et des voyelles nasales /ɑ̃, Ɛ̃/ chez notre locuteur français.

Malgré que les variations de durée soient moins marquées dans les réalisations du phonème /ɑ/ attendu chez notre informateur français, nous considérons que le comportement de cette voyelle s'apparente davantage à celui des voyelles /i, u, a, E/ qu'à celui des voyelles nasales, beaucoup plus stables. Il nous est donc impossible, ici encore, de poser une distinction sans équivoque entre les voyelles /a, E/ et /ɑ, ɔ/ dans l'idiolecte de notre locuteur français.

TABLEAU III:

Partie de la rime occupée par la voyelle

fr. France

	ptk	bdg	fsʃ	vzʒ		ptk	bdg	fsʃ	vzʒ
/i/	29%	56%	33%	70%	/i/	44%	44%	48%	62%
/u/	35%	49%	32%	61%	/u/	37%	58%	37%	65%
/a/	38%	58%	31%	67%	/a/	52%	57%	46%	66%
/E/	33%	51%	32%	65%	/E/	42%	50%	36%	63%
/ɑ/	44%	58%	36%	65%	/ɑ/	73%	71%	57%	76%
/ɔ/	32%	61%	34%	75%	/ɔ/	66%	68%	57%	76%
/ɑ̃/	60%	85%	49%	77%	/ɑ̃/	78%	75%	66%	74%
/Ɛ̃/	56%	66%	43%	72%	/Ɛ̃/	79%	77%	65%	75%

Comme en fait foi ce tableau, nous avons considéré la rime syllabique comme une entité au sein de laquelle le noyau et la coda occupent une portion déterminée de la durée totale. La pertinence d'étudier les segments en ayant recours au concept théorique de la rime syllabique a été pressenti par certains chercheurs dans le domaine de la synthèse de la parole (Klatt: 1987). Toutefois, il n'y a eu aucun effort dirigé en ce sens jusqu'à présent. L'expression des rapports

entre voyelles et consonnes nous permet, comme nous le verrons plus loin (cf. tableau IV) de caracteriser les rimes de telle sorte qu'on puisse, sur la base des durées, en déterminer la composition segmentale.

Les résultats présentés dans le tableau III nous permettent de constater que les voyelles occupent systématiquement plus de la moitié de la durée totale de la rime lorsque la coda est une constrictive sonore. Chez notre informateur québécois, la voyelle domine toujours dans la rime si elle est soit une nasale soit encore une longue induite. La portion temporelle qu'occupent ces voyelles tend cependant à diminuer si elles sont entravées par /f s ʃ /. Lorsque ces mêmes consonnes suivent les voyelles /i, u, E, a/, ces dernières représentent moins de la moitié de la durée totale de la rime.

Pour ce qui est de notre locuteur français, les voyelles nasales tendent à dominer dans la rime. Ce sont, en outre, les seules voyelles à occuper plus de 50% de la durée totale devant les occlusives sourdes /p t k/. Lorsqu'elles précèdent les constrictives sourdes /f s ʃ / elles sont alors dominées par la coda. Ce phénomène correspond bien à ce qui a été observé chez notre sujet québécois dans ce contexte consonantique. Au niveau des voyelles orales, il n'y a pas de différence systématique dans le comportement des brèves et des longues induites.

Suite aux observations que nous venons d'effectuer dans les tableaux I, II et III, nous formulerons les conclusions suivantes: il existe bel et bien deux groupes de voyelles dans les deux idiolectes analysés; un groupe de voyelles longues et un groupe de voyelles brèves. Chez notre informateur français, le groupe des longues est formé strictement par les voyelles nasales alors que dans l'idiolecte de notre sujet québécois, ce groupe comporte, en plus des nasales, les voyelles /ɑ, ɜ/. Ces deux voyelles orales se distinguent phonologiquement des brèves /a E/ et se comportent, de façon générale, comme des voyelles nasales.

Les caractéristiques des voyelles longues produites par nos deux informateurs sont:

1) des durées moyennes plus importantes;

2) une certaine "imperméabilité" face à l'influence que peut exercer la coda;

3) une tendance nette à dominer dans la rime syllabique.

L'ensemble des données que nous avons recueillies nous a permis d'élaborer un modèle de la rime (plus ou moins précis à ce stade de notre recherche) pour chacun de nos locuteurs. Ce modèle tient compte de la durée totale de la rime, de la composition des codas et des portions occupées par les voyelles et les consonnes dans la rime.

Bien que notre but soit de produire des modèles qui soient exploitables pour la synthèse par règle, nous pouvons aussi envisager des applications possibles en reconnaissance de la parole. Les critères auxquels nous avons eu recours pour déterminer ces "profils" pour les rimes pourraient servir d'indices quant à la nature des noyaux syllabiques ainsi que sur la composition consonantique de la coda.

TABLEAU IV*:

Locuteur français			Locuteur québécois		
Coda	Durée de R	% de V	Coda	Durée de R	% de V
/fsʃ/	359 ms	33% (voy. orale)	/vzʒ/	291 ms	64% (voy. brève)
	416 ms	46% (voy. nasale)		318 ms	75% (voy. longues et nasales)
/vzʒ/	323 ms	67% (voy. orale)	/fsʃ/	272 ms	41% (voy. brève)
	334 ms	75% (voy. nasale)		307 ms	61% (voy. longue et nasale)
/ptk/	283 ms	35% (voy. orale)	/bdg/	216 ms	52% (voy. brève)
	328 ms	58% (voy. nasale)		261 ms	76% (voy. longues et nasales)
/bdg/	232 ms	56% (voy. orale)	/ptk/	186 ms	44% (voy. brève)
	303 ms	75% (voy. nasale)		280 ms	74% (voy. longue et nasale)

Ces modèles, bien qu'ils soient dans leur état actuel très approximatifs, confirment la pertinence d'une étude phonétique basée sur les constituants de la syllabe. Ils révèlent, en outre, l'importance de poser des divisions entre voyelles longues et brèves par nature si l'on veut être en mesure d'élaborer des algorithmes précis pour la synthèse du français.

*Les rimes sont présentées dans un ordre de durée décroissant.

Bibliographie

DELATTRE, P. et M. MONNOT (1981), "The Role of Duration in French Nasal Vowels", *Studies in Comparative Phonetics*, B. Malmberg Ed., Heidelberg, Philadelphia, Julius Groos Verlag, pp. 17-33.

JACQUES, B. (1974), "Variations de durée des voyelles et des consonnes fricatives post-vocaliques finales de syllabe en position accentuée et inaccentuée", *Cahier de Linguistique*, 4, pp. 89-115.

KLATT, D.H. (1987), *Review of Text-to-Speech Conversion for English*, J.A.S.A. 82, pp. 737-793.

------ (1976), "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", *J.A.S.A.*, 59, pp. 1208-1221.

MARTINET, A. (1969), *Le français sans fard*, Paris, P.U.F.

OUELLET, M. (1989), "Les variations des durées segmentales: un état de la question", *Actes des journées de linguistique 1988*, C.I.R.B., No B-170, pp. 121-135.

SANTERRE, L. (1988), "La synthèse haute-fidélité de la parole", *Interface*, mai-juin 1988, pp. 15-20.

------ (1974), "Deux A et deux E phonologiques en français québécois". *Cahier de Linguistique*, No 4, pp. 117-145.

SELKIRK, E. (1982), "The Syllable", *The Structure of Phonological Representations*, Dodrecht, Holland, Cinnaminson, U.S.A., Foris, Vol. II, pp. 337-384.

Auteures **Anna Firenze**, LADL-ERLI-Paris
Béatrice Pelletier, LADL-CORA-Paris

Titre **Recherche d'une description syntaxique contrastive des noms composés N de N du français et N di N, N da N de l'italien**

RÉSUMÉ

Cette communication a pour but de présenter différents problèmes rattachés à la traduction des noms composés N de N du français et N di N/N da N de l'italien.

Notre objectif étant de réaliser un dictionnaire électronique bilingue des noms composés, nous avons rencontré un certain nombre de difficultés, tant théoriques que pratiques, que nous avons traitées de la façon suivante:

Des traits sémantiques ont permis de distinguer les grandes classes dans chaque langue :

- *le classement :*
  - *-les concrets non-animés*
    - *-une dent de lait- un dente di latte*
  - *-les animés*
    - *-un homme d'affaires-un uomo di affari*
  - *-les abstraits*
    - *-une peine de coeur-una pena di cuore*

- *la traduction :*
  Nous avons envisagé différents niveaux de traduction selon qu'il y a ou non correspondance structurale entre les deux langues.

- *l'ergonomie du dictionnaire :*
  Comment donner à l'utilisateur la possibilité d'obtenir immédiatement et facilement la bonne traduction du nom composé demandé

Le dictionnaire qui résultera de cette étude comparative du français et de l'italien comprendra environ 20 000 mots, formes fléchies comprises.

Ce dictionnaire, outre le fait qu'il sera un outil indispensable d'aide à la traduction, servira aussi à l'utilisateur de correcteur orthographique.

Auteure **Elisabete Ranchod**
Universidade de Lisboa

Titre **Relations entre verbes supports - Prédicats nominaux supportés par ESAR et TER en Portugais**

RÉSUMÉ

The analysis of 2000 predicative nouns supported by the support verb **estar** *in Portuguese confirmed Z.S. Harris' hypothesis that there are nouns which form the nucleus of sentence : simple sentences with support verbs. In a support structure, the supported* N *behaves as the main element and it selects the other constituents in the same way as ordinary verbs do.*

A significant amount of these 2000 nouns accept both elementary support verbs **estar** *and* **ter** *(to be, to have). Relations between the two* **Vsup** *constructions can be :*

(i) simple **Vsup commutation** *:*

A Ana esta com medo de fazer isso
A Ana tem medo de fazer isso

(Ana is with fear of doing that)
(Ana has fear of doing that)

(ii) more complex ones :

A situaccao esta sob o controlo da Ana
A Ana tem o controlo da situacao

(The situation is under the control of Ana)
(Ana has the control of the situation)

The presentation will develop points (i) and (ii).

SCÉNARIO DE DÉVELOPPEMENT DES INDUSTRIES DE LA LANGUE

Richard Parent
Ministère des Communications du Québec

RÉSUMÉ

Le potentiel d'application de l'informatique linguistique dans le travail de bureau est énorme. Les outils d'aide au travail sur le texte, l'exploitation de bases de données textuelles, et les applications hors-texte constituent les principales catégories d'applications en informatique linguistique. Ce type nouveau d'informatique suppose que soient développées les connaissances linguistiques qui viendront se rattacher aux deux composantes principales : l'analyseur syntaxique et la base de données lexicales. Il y a une tension entre l'anglais et les diverses langues nationales. Des éléments de stratégie pour la langue française sont proposés : intérêt d'un atelier de génie linguistique et cognitif, positionner la langue française par un effort francophone de mise en commun et de coopération. La langue doit être découverte sous un aspect nouveau de technologie de l'information capable d'améliorer la productivité d'au moins 40 % de la main-d'oeuvre totale dans une économie avancée.

1. CONTEXTE

Dans sa brève mais fulgurante évolution, l'informatique a constamment envahi de nouveaux secteurs d'activités humaines : cette diversification comprend les bases de données, la robotique, le traitement de texte, etc.

Un nouveau secteur, les industries de la langue, est en voie d'émergence. On peut les définir comme les industries qui produisent des biens et services fortement dépendants de bases de données et algorithmes sur l'un ou plusieurs des niveaux de description d'une langue naturelle.

Ces descriptions font l'objet de la linguistique, soit un ensemble de disciplines scientifiques et de champs spécialisés d'analyse qui ont pour objet la langue en elle-même.

1.1 Besoin

Pourquoi des "industries de la langue" émergeraient-elles? Fondamentalement, pour tirer profit d'une nouvelle forme d'accumulation dans le système d'information des organisations, non plus de données chiffrées, mais de texte. Du "texte", c'est quelque chose dont on peut dire que c'est en français ou une autre langue. Des mots, phrases, paragraphes, documents, banques d'information avec "fiches" composent un volume textuel important, actuellement sous-exploité. Telle est la cible des "industries de la langue" qui requièrent le développement de l'informatique linguistique.

1.2 Informatique linguistique

L'informatique linguistique devrait avoir un impact important dans le travail de bureau en fournissant des outils d'aide devant permettre aux gens:

- de mieux écrire en français;
- de mieux consulter ou analyser des textes emmagasinés sur support électronique;
- de mieux indexer/repérer, avec des mots et des règles de la langue usuelle, le contenu d'objets textuels.

Il s'agit de faire en informatique linguistique ce qui a été fait en informatique numérique. Des outils puissants d'exploitation de matrices contenant des chiffres ont été mis au point par des mathématiciens, statisticiens, ingénieurs, spécialistes en comptabilité, en physique, etc. L'état de l'art en informatique linguistique est comparativement peu avancé. Les connaissances sont en grande partie disponibles sur papier, mais on ne sait pas encore comment les organiser, comment les modéliser, de façon à les rendre utilisables en contexte. On peut penser à des outils "pas bêtes" qui sont constitués, par exemple, de bases de données lexicales et de moteurs d'analyse syntaxique ou même d'analyse conceptuelle.

1.3 Deux hypothèses de développement

Comment va se développer cette partie des industries de la langue qui dépend des avancées scientifiques et pratiques en informatique linguistique?

Une étude du Stanford Research Institute prévoit les segments de marché suivants pour 1992 :

- traduction
- interrogation de bases de données
- interfaces de dialogue
- analyse de contenu des textes
- aide à la rédaction
- reconnaissance/synthèse de parole

Cette subdivision est intéressante, bien que relativement conservatrice. L'interprétation qui l'accompagne ne perçoit pas de liens entre ces marchés par le partage de ressources d'informatique linguistique.

Une autre étude, souvent citée, est celle de Tim Johnson (Ovum Inc., Londres) qui prend une perspective organisationnelle et met l'emphase sur le vocabulaire propre à une organisation. Ce vocabulaire, avec ses variations dans l'usage, serait rattaché à une structure conceptuelle inscrite dans les outils et ressources informationnelles d'usage commun. Son analyse se singularise par l'importance qu'il accorde à un appareil qui transcrit une dictée vocale en texte. Les autres applications principales que Johnson prévoit sont :

- les interfaces aux bases de données et logiciels
- l'analyse de contenu des textes
- la correction grammaticale
- l'aide à la traduction

Ces applications sont assez concordantes avec celles de la première étude. Comme la précédente aussi, le partage de ressources linguistiques est sous-estimé, mais pas autant puisque Johnson a bien vu l'importance des vocabulaires en usage relativement à un ensemble de concepts arrangés en modèles.

1.4 Le risque d'anglicisation

L'anglophonie domine largement encore l'industrie informatique. Même le projet franco-européen Émeraude d'atelier de génie logiciel comporte un langage de programmation dont les mots sont anglais! L'anglais profite également d'être la langue de communication scientifique la plus partagée et cela se répercute en informatique.

Mais au-delà d'un certain seuil de complexité, un usager devrait pouvoir conceptualiser et s'exprimer dans sa langue d'usage. Les langues autres que l'anglais ne seront pas bien desservies par les forces du marché existantes. Pour le personnel informaticien existant, ces nouvelles exigences et opportunités devront faire l'ojet d'un complément de formation.

2. SCÉNARIO

Pour corriger le marché, il faut organiser des actions de développement linguistique sur la langue française en vue de l'exploitation mieux soutenue par des logiciels de textes rédigés en français. C'est d'ailleurs ce que soulignait l'étude britannique déjà citée (ma traduction) :

> "plusieurs nations [non anglophones] vont considérer leur propre langue trop importante pour la laisser exposée aux forces du marché, et peuvent intervenir pour s'assurer que des développements clés en traitement de langue naturelle se fassent avec l'appui des gouvernements nationaux."

2.1 Catégories d'applications

"Industries de la langue" est une désignation commode d'une série de contextes d'usage de l'informatique : c'est en décrivant l'extension de ce qui est couvert par le terme qu'on peut en circonscrire le sens. Quels en sont les types d'usagers et de produits prévisibles?

2.1.1 *Types d'usagers*

a) "langagiers" ou personnel spécialisé en travail lié à la langue : traducteurs, terminologues, chercheurs, rédacteurs techniques, réviseurs, etc. avec des besoins en commun ainsi que des besoins spécifiques liés à des caractéristiques distinctes de tâche. Ce groupe adopte tôt les innovations et peut payer cher le bon produit en proportion du gain de productivité escompté. Source importante d'évaluation pragmatique des produits avant qu'ils parviennent au type d'usagers suivant.

b) personnel effectuant du travail semi-spécialisé avec les textes, au niveau des tâches de secrétariat par exemple. Ce groupe n'adopte que les outils robustes et pas trop complexes, directement utiles, et moins coûteux.

c) le public en général, recourant aux mêmes outils que le groupe précédent, ou à des outils plus simples encore. Le grand public recouvre aussi des besoins d'apprentissage de la langue, ainsi que nombre de besoins pratiques et ludiques.

2.1.2 *Types de produits*

Il est difficile de classifier les produits prévisibles en raison d'une interrelation complexe entre critères en cause, notamment :

- s'agit-il d'un outil de développement ou d'une application?
- s'agit-il de contenu interrogeable ou d'une fonction de calcul?
- des mécanismes sophistiqués d'inférence, d'appariement et d'analyse sont-ils utilisés ou disponibles?
- l'application concerne-t-elle une base de données textuelles ou une base de données numériques?

Notre synthèse est de regrouper en trois classes les produits existants et prévisibles :

- traitement de texte
- exploitation de texte
- applications ailleurs que sur texte.

a) *Traitement de texte*

Principale application de l'informatique dans les bureaux si l'on considère le nombre d'usagers. Ça fait peu d'années que l'usager francophone est à peu près bien desservi pour le français et ses fameux accents! La personne qui utilise un logiciel de traitement de texte peut s'en tenir à une simple saisie ou transcription, auquel cas la correction orthographique et des fautes d'accord conviendrait. Si la personne qui utilise le traitement de texte compose un texte, si elle en est l'auteur, une correction plus poussée, grammaticale et stylistique, serait utile. À la frontière du traitement de texte, l'aide à la rédaction s'appuie sur des systèmes experts pour la rédaction d'une lettre d'affaire, d'un curriculum vitae, d'un rapport de tel type, etc.

b) *Exploitation de texte*

Dans les bureaux, quelques années d'usage du traitement de texte ont permis l'accumulation de grands volumes de textes. Les outils pour maîtriser leur volume sont difficiles à trouver. Les systèmes courants en informatique pour la gestion des bases de données sont pensés en fonction de structures de données matricielles. Si cela convenait encore pour les systèmes de gestion documentaire à base de thésaurus, la percée a été le recours à un "fichier inverse" des formes lexicales avec adresses dans le texte. Cette technique tire parti du fait que le nombre de vocables dans la langue est relativement restreint par rapport au nombre de mots dans les textes emmagasinés.

Pour la langue française, il existe différents logiciels ayant cette capacité de fichier inverse, (exemples pris en gestion documentaire : Edibase, Seconde). Pour l'anglais (ou un français tronqué dans ses normes), il y a par exemple Stairs (d'IBM), Basis (d'Information Dimensions Inc.), ou Office Indexer (de Wang).

Quelles sont les fonctions spécifiques que devrait comprendre un système de gestion de base de données textuelles?

- emmagasiner de forts volumes de texte;
- favoriser l'indexation assistée par de nombreux modèles de description du contenu, dont le plus simple est l'analyse lexicométrique;
- favoriser le repérage assisté par de nombreux modèles d'exploration du contenu: soit pour une requête assez simple, soit pour mener une analyse de contenu de façon assistée;
- faciliter l'accès local ou par réseau à de nombreux contenus textuels, qu'il s'agisse de dictionnaires généraux ou spécialisés de langue française, de banques d'information scientifique et technique, ou de l'ensemble des documents administratifs dans un bureau.

c) *Applications ailleurs que sur texte*

Si l'on restreint le sens de "texte" à une certaine structure de données (mots, en phrases, en paragraphes, en sections, en document), il est possible d'identifier d'autres structures de données que du texte qui peuvent tirer partie des logiciels d'informatique linguistique :

- les interfaces avec menus et "mots à fournir" pour l'interrogation de bases de données et l'utilisation de logiciels dans les termes du langage familier à l'usager. Ce langage comporte un vocabulaire de l'organisation à plusieurs strates de spécification, par exemple au gouvernement: organismes centraux (ex.: Conseil du Trésor), ministère ou organisme particulier, directions centrales, chaque direction, chaque division... Il peut aussi comporter le langage des opérations comptables, du calcul statistique, de calculs d'ingénieurs, de modèles économiques, au besoin de l'usager.

- le besoin d'une composante linguistique dans les futurs environnements de programmation dans la ligne des développements d'atelier intégré de génie logiciel. La modélisation conceptuelle des données et des connaissances pourrait profiter d'une banque de concepts définie formellement en s'attachant à la sémantique lexicale de la langue naturelle de l'usager et au vocabulaire propre à l'organisation (relié à ses activités et sa "culture organisationnelle").

2.2 Catégories de connaissances linguistiques

La linguistique s'est constituée en un ensemble important d'approches comme la lexicographie, la syntaxe, la sémantique, etc., correspondant à ces niveaux de description de la langue et leurs phénomènes caractéristiques (phonèmes, morphèm vocables, termes, phrases...).

Les domaines de la linguistique ne sont pas tous égaux. Il y a des niveaux de description de la langue qui sont plus critiques que d'autres en fonction des applications. En effet, deux composantes majeures vont devoir se retrouver dans une majorité des applications d'informatique linguistique : un analyseur syntaxique et une base de données lexicales. Du moins peut-on rassembler autour de ces deux noyaux fonctionnels la plupart des connaissances utiles. C'est ainsi que la morphologie, la sémantique et la modélisation conceptuelle seront mises en relation avec ces deux noyaux à titre de compléments nécessaires et utiles en fonction du contexte d'application.

2.2.1 *Syntaxe*

Un analyseur syntaxique construit une représentation de la structuration des phases dans une langue donnée, et fournit une description linguistique élaborée en fonction des besoins d'indices qui seront pris en compte dans une application. L'analyseur syntaxique est le moteur de l'analyse de texte, le partenaire obligé de la plupart des autres niveaux de description.

Par exemple, l'assistance en traitement de texte requiert un analyseur syntaxique pour atteindre une qualité supérieure de correction orthographique et pour permettre la correction des erreurs d'accord ainsi que le relevé des faiblesses de style.

2.2.2 *Lexique*

Un complément essentiel d'un analyseur syntaxique est une base de données lexicales, contenant des renseignements linguistiques et encyclopédiques couvrant à la fois ce qu'on retrouve dans les dictionnaires généraux de langue et dans les ouvrages ou banques de données terminologiques à propos de divers domaines spécialisés des sciences et des techniques, ou d'institutions.

Les informations contenues dans le lexique sont diverses : morphologie, catégories grammaticales, etc.; afin que les définitions soient utilisables par le moteur d'analyse, une forme systématique est requise, comme des réseaux sémantiques, des graphes conceptuels (Sowa, 1984) ou d'autres variantes de modèles pourvu qu'ils soient traitables par les automates qui font partie de l'analyseur.

En 1988, les moteurs d'analyse sont beaucoup plus avancés que le travail de description sémantique dans les bases de données lexicales. C'est le plus difficile, mais aussi le plus critique pour des programmes d'applications vraiment intéressantes parce que capables, en contexte, d'un traitement conceptuel.

Ce défi de la description sémantique, particulièrement au niveau du lexique général de la langue, consiste à rendre explicite les grands modèles cognitifs et sémantiques qui structurent le vocabulaire d'une langue naturelle. Pour Ballmer (1984) par exemple, les structures sémantiques du vocabulaire sont reliées à des structures cognitives permettant de générer des modèles d'événements, d'entités, de relations, de contextes, etc. Divers auteurs pensent que certains domaines de connaissances comme le temps, l'espace, la causalité se prêtent bien à une systématisation. On voit de mieux en mieux le besoin de vastes bases de données lexicales comprenant notamment une sémantique ouverte mais contrôlée. On parlera de plus en plus de logique naturelle et de ses correspondances linguistiques.

2.3 Prévisions sur l'évolution du marché

Il est vrai que les produits en langue française accusent un certain retard sur l'anglais: une cause importante provient de normes inadéquates dans les matériels et systèmes d'exploitation informatiques, actuellement en bonne voie de résolution sous l'action d'organismes internationaux de normalisation. Un autre problème, mais moins près d'une solution, est celui de la quasi-inexistence d'un marché francophone international informatique. Ou s'il y a un tel marché, le Québec y est peu présent. Divers facteurs historiques, liés au matériel notamment, expliquent cette situation, mais à l'heure des normes et de la micro-informatique, il y a beaucoup de raisons de prévoir une intensification des courants d'échanges entre francophones.

Quand on dit "industries de la langue", on peut penser qu'en effet des descriptions propres à chaque langue naturelle constitueront l'épine dorsale dans ce marché. C'est vrai, mais il y a deux forces économiques qui lui seront adjacentes: parfois complémentaires parfois en concurrence:

- une catégorie de logiciels spécialisés dans diverses tâches qui ont été modélisées clairement et où l'interface est semi-indépendante et peut être adaptée à plusieurs langues selon la demande; cela sera surtout pratique pour des tâches aisément standardisées et où le vocabulaire n'est pas trop immense ni trop finement nuancé; l'importance relative d'universaux sémantiques sera déterminante pour ce potentiel; ces logiciels seront évidemment diffusés à grande échelle;

- une autre catégorie de besoins ne peut pas faire l'objet d'une commercialisation étendue : les besoins propres à une équipe de travail, à une organisation, à un groupe culturel ou à un groupe professionnel qui est de décrire, uniformiser, utiliser couramment un vocabulaire particulier qui concerne les individus qui composent l'organisation, le groupe, etc.

Globalement, on peut penser que la concurrence sera vive entre des produits comprenant des capacités nombreuses dans des assemblages variés et difficiles à comparer. L'évaluation de la qualité des produits risque d'être un sujet controversé en raison de la complexité des fonctions en cause.

3. ÉLÉMENTS DE STRATÉGIE

3.1 Vers un atelier de génie linguistique et cognitif

Le recours aux connaissances linguistiques intégrées aux logiciels de bureau devrait connaître une croissance phénoménale au cours des années 1990. Le potentiel d'applications, tout ce qui s'écrit, textes ou formulaires, tout écrit qui doit être analysé, est en effet très vaste. On peut s'attendre à ce que d'ici cinq ans les descriptions sémantiques de la langue française dans les formats et sur les objets requis en général aient atteint la masse critique pour rendre possibles de nouveaux paliers d'utilisation.

L'intérêt d'un atelier de génie linguistique et cognitif tire son origine de la forte synergie qui peut être escomptée de la combinaison contrôlée des logiciels d'analyse linguistique et de modélisation conceptuelle (avec inférence). En particulier, le bénéfice d'une sémantique lexicale pour la représentation de l'expertise est d'en faciliter la structuration conceptuelle et sa communication pour ceux qui développent et avec ceux qui utilisent une expertise. Les connaissances peuvent ainsi être structurées de façon familière, le système étant alors capable de reprendre à son compte une partie suffisante des acquis de la langue naturelle (lexique, syntaxe, sémantique générale) et des langages spécialisés. Les domaines administratifs et scientifiques apparaissent particulièrement prometteurs en raison du fait que ces langages ne sont que semi-ouverts et donc plus facilement formalisables.

Le but est de faciliter la conception, le développement et l'entretien de systèmes opérationnels grâce à une gamme étendue d'outils logiciels comprenant notamment des lexiques, thésaurus, dictionnaires conceptuels généraux et spécialisés, et jouant un rôle important dans l'aide à la conception et au maintien de la cohérence dans les connaissances véhiculées au sein de l'organisation.

Dans la mesure où ces conceptualisations peuvent être exprimées dans la langue naturelle de l'usager, le temps de formation sur l'outil logiciel est réduit, l'aisance de la communication est facilitée, et surtout, il y a décuplement de la complexité des représentations traitables couramment par ordinateur dans le quotidien des organisations. Il y a des mises en commun importantes qui requièrent une normalisation des vocabulaires utilisés afin de réutiliser des sections de bases de connaissances déjà développées. La recherche de ces combinaisons gagnantes d'investissement dans les systèmes experts pour une grande organisation engagera à des décisions progressivement plus importantes au cours des années quatre-vingt-dix.

Une méthodologie de développement de système expert aurait donc intérêt à expliciter le besoin d'intervention sur le plan terminologique et à chercher des éléments de méthode et des techniques transférables de l'analyse lexicométrique et de la terminologie aux activités de modélisation des connaissances. La maîtrise du réel présuppose la connaissance et la maîtrise d'un ensemble de termes pour l'appréhender et le manipuler (Rey).

Aux activités actuelles en terminologie, il est important d'ajouter des travaux de développement en sémantique pour des parties du lexique général de la langue française. Ces travaux pourraient d'abord chercher la complémentarité avec des développements de bases de données et de connaissances, tout en généralisant les besoins dans la description d'aspects pour des sous-ensembles de mots.

3.2 Positionner la langue française

La langue française est aussi une "technologie de l'information". Du moins faut-il la considérer dans cette perspective quant on parle de son industrialisation.

La description de la langue française est un objet naturel de coopération en recherche et développement pour la collectivité francophone mondiale. Cette coopération est amorcée surtout en terminologie actuellement, et elle y est structurée de façon modèle tant au niveau de la formation que de la méthodologie , ainsi que par l'échange des résultats via des publications et une base de données internationale. Le deuxième Sommet francophone a été l'occasion du démarrage d'une action encore modeste mais prometteuse dans le domaine plus étendu des industries de la langue.

Mais il faut hâter le pas car la langue anglaise s'insinue dans de multiples langages de spécialité via les logiciels. L'évolution de certains de ces logiciels promet d'être d'une rapidité foudroyante: on risque d'avoir à se contenter trop souvent de traductions qui trahissent le génie de la langue française (ses modes privilégiés de représentation étant données sa structure grammaticale et sa structure lexico-sémantique). Le domaine de la géomatique serait un cas intéressant à étudier sous cet angle dans les années à venir au Québec.

C'est pourquoi nous nous associons aux propositions visant à favoriser une mise en commun francophone des fonds linguistiques déjà constitués, que l'on pense aux dictionnaires de langue existants, au Trésor de la langue française, aux lexiques grammaticaux du LADL, à l'analyseur lexico-syntaxique du français développé conjointement à Paris (SLID-LISH) et Montréal (Centre d'ATO), à la Banque de terminologie du Québec.

Tel que déjà noté, le besoin le plus évident à l'heure actuelle est une activité de développement sémantique des bases de données lexicales, tout particulièrement les quelques dizaines de milliers de mots les plus courants. Il s'agit de donner une plate-forme aux producteurs francophones qui leur permet de mieux réaliser des logiciels et programmes d'application où la langue française n'est pas désavantagée par rapport à la langue anglaise.

La coopération francophone devrait permettre d'initier des actions sur le niveau général de la langue française en priorité. C'est urgent dans la course économique inter-culturelle de favoriser une mise en commun au niveau de la recherche ou au niveau de la commercialisation. L'enjeu est de fournir une base linguistique et conceptuelle riche pour la modélisation de l'information et des connaissances en français.

De façon plus générale, il faut organiser, outiller et financer l'analyse des besoins et le développement d'outils au moyen de projets pilotes répondant aux besoins de la situation. Avant, pendant, et après ces projets pilotes, il faut veiller à la formation des personnes et les habiliter à utiliser les nouveaux outils linguistiques et cognitifs.

CONCLUSION

Les "travailleurs du texte" sont loin de se limiter aux traducteurs et terminologues, mais constituent probablement plus des deux cinquièmes de la main-d'oeuvre totale au Québec par exemple. La tâche de description et modélisation des connaissances linguistiques est gigantesque selon une perception commune. Une mise en commun et une coopération active au niveau francophone est encouragée pour constituer une plate-forme "pré-compétitive" élevée qui donne des atouts dans la course de vitesse avec les produits en langue anglaise. Cette plate-forme fournirait une base linguistique et conceptuelle riche en un sens propice à une augmentation de productivité d'une grande importance économique... et culturelle!

Références

Susan HANDEL LOHRER et al. *Commercial prospects for artificial intelligence technologies.* Report no 757, Fall 1987. Standford Research Institute International, Business Intelligence Program.

Tim JOHNSON. *Natural language processing: commercial applications.* Ovum, Londres, 1986.

John F. SOWA. *Conceptual structures. Information processing in mind and machine.* Addison-Wesley, Reading, Mass, 1984.

Th. T. BALLMER. The position of argumentation in the framework of a text linguistics, speech act theory and lexicology. *Journal of Pragmatics*, 8, (1984), 9-29. Elsevier, Horth-Holland.

Alain REY. *La terminologie. noms et notions.* Que sais-je? no 1780, Paris, P.U.F., 19..

LES STRUCTURES ET LES MESURES DE LA PROSODIE DU FRANÇAIS*
(en vue de la synthèse par règles)

Laurent Santerre
Université de Montréal.

Les recherches sur la prosodie du français s'orientent lentement vers les applications aux industries de la langue, synthèse et reconnaissance. Voilà un vaste domaine où l'on peut accumuler des monceaux de données numériques sur les fréquences du Fo dans l'intonation, sur les durées segmentales et suprasegmentales et sur l'intensité répartie dans la phrase.

Mais il y a loin des données aux règles. L'intonation est liée à la syntaxe et à la sémantique et est ancrée aux points d'accentuation; les accents eux-mêmes sont de diverses natures et prennent des formes acoustiques qui varient selon les différentes fonctions linguistiques. La systématique des durées est à son tour conditionnée par la nature des constituants syllabiques, par les contraintes liées à la phonologie et à la morphologie, au débit, à l'intensité, à la place de la syllabe dans l'énoncé, au poids sémantique des mots, etc.

Des données acoustiques qui ne seraient pas motivées ou expliquées par l'influence des diverses structures prosodiques qui les sous-tendent ne nous apprendraient à peu près rien sur la nature de la parole et ne sauraient être systématisées dans les règles utilisables dans la synthèse.

On peut étudier séparément chacun des systèmes intonatif, accentuel et rythmique ou temporel, un peu abstraitement comme en laboratoire, au moyen de phrases expérimentales; on est bien obligé de le faire; on ne peut cependant pas perdre de vue que, dans la parole réelle, toute mesure de fréquence, de durée relative ou d'intensité est le résultat pondéré de toutes les commandes à la fois. D'où la nécessité de modèles séparés pour chacun des systèmes qui composent la prosodie de la langue; nécessité aussi de l'analyse de l'influence de chacun d'eux sur les valeurs acoustiques mesurées à la surface phonétique de l'énoncé.

Les recherches sur la prosodie ont longtemps été le fait des laboratoires de phonétique qui disposaient d'instruments de mesure, en un temps où la linguistique et la phonétique n'avaient pas des approches concertantes. Il n'empêche que beaucoup de travaux de cette époque dans le domaine français restent encore très instructifs. **Coustenoble (1934), Faure (1963, 1967, 1968), Delattre (1938, 1968), Fonagy (1978, 1979, 1983), Rigault (1961, 1970), Léon (1969, 1979), Martin, (1977), Boudreault (1968), Beauchemin (1970),** Rossi (1972 a, b, 1981), Emerard, (1977), Vaissière(1983), Carton (1976), Di Christo (1975, 1981, 1982 a, b) Lavoie (1965), Malmberg (1961, 1966), Mettas (1964), Holder (1968), Warren (1981) Santerre (1981). Le renouveau dans les recherches prosodiques nous est surtout venu des langues germaniques, de l'américain en particulier: Liberman (1977, 1979, 1983), Selkirk (1972, 1980), Pierrehumbert (1982), Gärding (1984), Thorsen (1981), Carlson, (1973), Cutler (1982), Ladd, (1980), Verluyten (1982) est influencé par Shane (1968), Dell l'est davantage par Liberman et Prince (1977). Rossi (1985 et 1987) est plus personnel et inventif.

---

*Cette recherche est subventionnée par le C R S H C

Les recherches sur la prosodie en vue de leur exploitation dans les industries de la langue ne se feront pas sans une étroite collaboration entre les spécialistes de la linguistique générale (syntaxe, morphologie, phonologie), et ceux de la phonétique expérimentale et de l'informatique. D'une part, la prosodie ne peut être étudiée sans l'analyse acoustique et les tests psychoacoustiques qui exigent des laboratoires de phonétique pourvus des meilleurs outils d'analyse et de synthèse; d'autre part, les descriptions les plus fines pourraient rester lettres et chiffres morts en dehors d'un cadre théorique capable de les élever au rang de données systématiques en termes de savoir scientifique.

Les données purement matérielles qui pourraient faire parler un ordinateur "comme du monde" ne sont pas le but ultime de ces recherches; la connaissance de l'homme à travers ces merveilles les plus étonnantes que sont les langues qu'il fabrique comme une araignée fabrique du fil d'araignée, c'est-à-dire au besoin, passe bien avant les prouesses les plus avancées de ces techniques.

Il se trouve aujourd'hui que ce sont les sociétés elles-mêmes qui poussent les scientifiques à développer les industries de la langue. Et comme ces derniers commencent à se rendre à l'évidence que les langues sont des objets de connaissance infiniment plus complexes qu'on peut encore l'imaginer, il faudra bien que les recherches fondamentales soient faites par ceux qui peuvent les faire et que les lois qui s'en dégagent soient formalisées en termes numériques exploitables dans la synthèse et la reconnaissance de la parole.

Cette communication comporte deux parties. Dans un premier point, je voudrais montrer que les mesures ne peuvent être séparées des structures qui les sous-tendent, de même que les systèmes linguistiques eux-mêmes ne sont pas indépendants les uns des autres et que les modèles abstraits de fonctionnement doivent être mis à l'épreuve des expérimentations. Dans un second temps, je présente un début de recherche sur les durées relatives dans les rimes pour illustrer la systématique qui semble s'y exercer à travers le système phonologique, les contraintes morphologiques et accentuelles.

1. STRUCTURES ET MESURES.

Dans les industries de la langue, ce ne sont pas les moyens informatiques qui manquent, mais les connaissances linguistiques et phonétiques. Ce que nous savons des langues, du langage et de la parole reste trop théorique, d'une part, et d'autre part, les précisions acoustiques que nous accumulons sur la parole et la variation linguistique ne sont pas toujours intégrées dans des lois ou des principes, ou des tendances générales qui pourraient les expliquer et permettraient de les prédire.

C'est en particulier le cas des études en prosodie. D'un côté, on développe des théories sur l'accentuation et l'intonation, sans souci de les mettre à l'épreuve de la confrontation avec la parole réelle; et de l'autre côté, on accumule les mesures sur la parole réelle, on fait des moyennes sur des ensembles souvent disparates, de sorte qu'on se trouve en présence de généralités qui renseignent plus sur le comportement général des locuteurs que sur la production ou la perception de la parole elle-même.

Ceux qui excellent dans les mesures doivent savoir que les données expérimentales ne sont jamais théoriquement neutres, c'est-à-dire qu'elles supposent nécessairement un cadre de référence qui devrait les rendre instructives. D'un autre coté, ceux qui préfèrent construire des modèles prosodiques devraient reconnaître l'importance des tests expérimentaux pour éprouver ces modèles. Car, que faire en sciences de modèles et de théories qu'on ne peut infirmer, valider, ou améliorer? Et que faire des données expérimentales qu'on ne peut pas interpréter dans un ensemble théorique qui leur donne une valeur de connaissances systématiques?

C'est ce lien entre la théorie et la pratique que propose Liberman (1983) aux organismes de recherches, qui commencent à comprendre que leurs puissants algorithmes de traitement du signal ne nous révèleront pas à eux seuls la nature et le fonctionnement des paramètres acoustiques du langage. L'intelligence artificielle nous sera d'un piètre recours pour aller plus loin, tant que l'intelligence naturelle elle-même sera prise au dépourvu. Ce sont les connaissances fondamentales sur la langue et le langage qui nous manquent avant tout dans les industries de la langue. Il n'est pas possible que nous puissions court-circuiter cet immense programme rattaché aux sciences humaines. Même quand il ne s'agit que de faire dire du texte correctement écrit selon la syntaxe et la sémantique, nous ne savons pas fournir à l'ordinateur les instructions prosodiques que l'oreille juge élémentaires, tant leur variété et leur complexité sont grandes.

C'est pas que nous n'ayons pas appris beaucoup de choses importantes jusqu'à maintenant sur les paramètres acoustiques qui font les formes de surface de la prosodie; c'est plutôt que nous manquons de modèles sous-jacents qui nous permettraient de formaliser des commandes séparées pour l'intonation, l'accentuation, l'intensité, l'organisation temporelle et rythmique, modèles qui tiendraient compte des principaux facteurs susceptibles de faire varier chaque paramètre.

L'évolution du Fo est relativement facile à suivre aux instruments, mais elle est loin d'être à elle seule le "pitch" qui, lui, fonctionne linguistiquement dans la production et la perception et varie selon la syntaxe, la sémantique, le rythme dans l'assertion, l'ordre, l'interrogation, l'exclamation, la focalisation (Rossi 1985), la stylistique (Fonagy 1983) et le poids sémantique des mots.

L'autre structure prosodique, peut-être encore plus importante que l'intonation, même si elle est moins immédiatement perceptible, réside dans le domaine complexe des commandes accentuelles. Nombreuses sont les espèces d'accents, leurs fonctions, et les formes acoustiques qu'elles peuvent prendre sous divers conditionnements.

Quant à l'intensité, plus étroitement liée à la mécanique de la production, elle paraît jouer un rôle plutôt stylistique; il ne semble pas, en effet, qu'elle puisse jouer un rôle morphologique comme l'intonation et la durée liées à l'accentuation (Santerre 1981). Elle pourrait pour cela être analysée seulement dans la mesure de ses incidences sur l'intonation et l'accentuation.

Les deux paramètres prosodiques d'intonation et d'accentuation sont étroitement liés à l'organisation syntaxique de l'énoncé et aux relations sémantiques de ses constituants majeurs. Mais les règles accentuelles peuvent en prendre très large dans la parole réelle avec certains noeuds terminaux de l'arbre syntaxique. Les règles d'accentuation phonologique (fin de mot et de syntagme en français) peuvent facilement céder le pas aux exigences rythmiques liées au débit et au nombre de syllabes dans les constituants. Cette organisation rythmique semble répondre à des lois propres, même au détriment de la correspondance qu'on attendrait avec les regroupements plutôt rigides qu'impose l'arbre syntaxique, surtout quand il n'y a pas risque d'ambiguïté et que le contexte est riche en redondance sémantique. Dell (1984) en fournit de nombreux exemples, mais il n'est pas écrit dans le ciel que le québécois se comporte comme le parisien.

Une fois les règles d'accentuation appliquées, il reste à faire passer les règles d'association tonale; bien sûr, les différents schèmes intonatifs phonologiques (ex./ B H M ou M H B etc/) sont ancrés sur les syllabes accentuées, mais plusieurs syllabes de suite peuvent être sous l'influence d'un même ton haut ou bas, ou bien une même syllabe peut être attachée à deux tons voisins; ces

règles d'association tonale, de même que les schèmes intonatifs eux-mêmes, varient d'un dialecte à l'autre; jusqu'à quel point, on ne sait pas. Il apparaît assez clairement que le schème accentuel fait le lien entre la structure syntaxique et le profil mélodique, mais il est lui-même très variable.

On peut penser que l'accentuation en québécois présente des différences importantes avec le français de France qui n'exploite plus comme nous le faisons les oppositions phonologiques de durée. L'organisation temporelle en québécois est assujettie à un système syllabique qui respecte les durées de huit voyelles longues par nature: (quatre orales / ɔ ø o a / (/ ɔ/ de fête) et quatre nasales); sept voyelles brèves sont allongeables et abrégeables par coarticulation consonantique; deux autres enfin /ə et e/ ne se trouvent pas en syllabe entravée. Il faut donc s'attendre à ce que l'organisation rythmique de l'énoncé, qui pèse si lourd sur la répartition des accents dans les tronçons intonatifs, imprime sa marque spécifique à la prosodie du québécois.

Quel corpus analyser?

L'examen des paramètres prosodiques de surface de la parole réelle dans son contexte social nous donne à observer, pour chacun des paramètres à l'étude, par exemple l'intonation, le résultat de toutes les commandes sous-jacentes qui l'ont faite ce qu'elle est à tous les moments de l'énoncé: ce sont le patron intonatif commandé par le sens, (ex. le profil phonologique / B H M / (bas - haut - moyen), l'arrangement de ce profil sur les syllabes qui le portent, la nature de l'accent impliqué dans ces syllabes (accent phonologique dont la forme est conditionnée par la frontière inter ou intra-syntagmatique); ce sont encore l'effet de la déclinaison du Fo en raison de la place de la syllabe par rapport à la dépense pulmonaire dans le groupe de souffle, la part de variation du Fo accordée à l'accentuation indépendamment du profil mélodique qui s'y superpose en vertu des exigences sémantiques et stylistiques (insistance de diverses natures); c'est enfin la part de variation mécaniquement ajoutée par une poussée facultative d'intensité, etc. Comment démêler la part respective de toutes les causes phonologiques et mécaniques, systémiques ou aléatoires, qui ont fait de ce profil intonatif ce qu'il est dans sa représentation acoustique? Un large corpus fournira-t-il deux exemples presque semblables qui nous permettraient par comparaison de voir varier un seul paramètre à la fois pour en observer l'impact dans la production et la perception?

A mon avis, on ne peut se dispenser d'étudier à la fois un corpus choisi de parole spontanée et un corpus construit de phrases simples où l'on peut examiner l'impact d'un seul facteur qui varie à la fois. Certains facteurs sont contrôlables au moment de la production, comme les choix de mots, la syntaxe, l'organisation sémantique (thème, rhème), la place des ictus mélodiques, le débit, la coupe syllabique dans les morphèmes à voyelles longues étymologiques entravées, etc. D'autres facteurs ne peuvent être contrôlés méthodiquement qu'au moyen de la synthèse aussi fidèle que possible; ce sont, par exemple, les degrés de variations du Fo, de l'intensité, des durées relatives dans les syllabes, le jeu des paramètres acoustiques dans les accents, etc.

Les manipulations de laboratoire les plus fines et les tests qui les complètent ne sauraient nous dispenser de retourner continuellement au corpus naturel pour reconnaître les systématiques qu'on pense avoir découvertes en laboratoire. L'observation des grands corpus peut permettre des généralités qui sont presque des universaux de la production de la parole et qui s'appliquent à toutes les langues.(Vaissière 1983). Il faut connaître ces généralités, mais ce n'est qu'un début; si l'on s'en tenait à ces grandes lignes pour faire de la synthèse ou de la reconnaissance, il n'y

aurait pas de différence entre les langues, les dialectes, les individus, ou les sentiments; la parole de synthèse resterait robotique comme elle l'est largement aujourd'hui. La variation au sein des systématiques nous étonne autant que les invariants un peu trop simplifiés qu'on pense découvrir dans le langage. L'invariance n'est pas où l'on pensait; on ne pourra la définir que lorsqu'on connaîtra la nature et le fonctionnement du langage; on n'en n'est peut-être qu'aux balbutiements.

2. EXEMPLE DE SYSTÉMATIQUE DANS LES DURÉES RELATIVES AU SEIN DE LA RIME.

L'organisation temporelle se retrouve à plusieurs niveaux de la production de parole: durée accentuelle, durée phonématique, durée liée au débit, à la rythmique, durée segmentale propre, durée conditionnée par l'intensité et par la composition syllabique. Rien n'est laissé au hasard. Il importe de découvrir les systématiques d'ordre phonologique ou mécanique ou articulatoire pour l'avancement des recherches fondamentales et la synthèse et la reconnaissance de la parole naturelle.

J'ai cherché à découvrir l'organisation systématique des durées relatives dans les rimes de quatre locuteurs, deux Parisiens, un homme et une femme, et deux Québécois, de même un homme et une femme. Le corpus est construit avec toutes les voyelles entravées par des consonnes obstruantes. Je ne présente ici comme illustration brève que les conclusions tirées de la production du locuteur québécois. A propos des trois autres locuteurs, je peux seulement dire pour l'instant que la systématique existe, mais qu'elle est toujours un peu différente d'un locuteur à l'autre, tout en respectant globalement les grandes commandes du système. La systématique est différente entre le parisien et le québécois, très certainement à cause des deux systèmes phonologiques; en parisien, il n'y a plus que 12 voyelles phonologiques, tandis que le québécois a conservé intactes les 17 voyelles héritées des fondateurs, avec les oppositions de durée et de timbre. (Santerre 1974, 1979, 1981).

Les phrases du corpus ont été prononcées deux ou trois fois par le même locuteur et c'est la moyenne que je présente ici. Je pense qu'il importe de découvrir la systématique de plusieurs locuteurs séparés et de ne pas faire de moyennes entre ces locuteurs, car dans l'optique de ces recherches, il vaut mieux préserver la systématique de chaque locuteur, étant bien entendu que l'ordinateur ne devra jamais parler comme une moyenne. Je trouve plus intéressant de comparer la production des locuteurs que de la ramener à une moyenne et à un écart-type. Les systématiques au niveau supérieur de la langue sont trop abstraites. Chaque locuteur exploite à sa façon la systématique de sa langue à travers son dialecte et son idiolecte, et il n'y a vraiment que cela qui explique sa production sonore. Le corpus comprend quelque 250 phrases.

J'ai groupé les voyelles québécoises en quatre catégories: les voyelles hautes / i, y, u /, les quatre voyelles brèves par nature / ɜ, a, ɔ, œ /, les quatre voyelles longues par nature, opposées par le timbre et la durée aux quatre brèves précédentes, comme dans *faite* et *fête*. *patte* et *pâte*. *sotte* et *saute*. *jeune* et *jeûne*; enfin les quatre nasales. Les deux voyelles restantes /e et œ/ ne se rencontrent pas en syllabe entravée. Les consonnes obstruantes font aussi quatre groupes homogènes; p t k abrégeantes; b d g neutres; f s ʃ neutres; v z ʒ allongeantes. On trouve donc plusieurs types de rimes (noyau + coda), puisque les voyelles brèves peuvent être abrégées ou allongées, ou laissées intactes; de même, les voyelles longues, orales et nasales, sont peu abrégées ou peu allongées, ou laissées intactes. Je n'ai pas examiné l'influence de la première consonne dans la syllabe (CVC); Di Christo dit qu'elle n'est pas très grande.

TABLEAU 1:

Moyennes et écarts - types de la durée des *voyelles* dans la rime et pourcentage de réduction selon l'accent

	Accent terminal		**Accent intérieur**		
	Ex. (Aimes-tu les pâtes) (Coupe-lui les pattes)		(Le mot pâle me plaît) (Le mot patte me plaît)		
	Moyenne (M)	écart - type (s)	M	s	réduction (R)
Brèves	11.3	2.0	7.9	1.8	.70
Longues	21.5	2.1	15.0	1.37	.70
Nasales	24.25	3.4	17.5	1.5	.72

Remarques: On observe une nette distinction de durée entre, d'un côté, les voyelles brèves abrégées ou non-allongées par coarticulation, et de l'autre, les longues par nature et les brèves allongées par coarticulation; les nasales, abrégées ou allongées, sont plus longues que les orales. Les durées moyennes des orales phonétiquement brèves ou longues varient presque du simple au double. La réduction de durée des voyelles est la même sous l'effet du déplacement de l'accent, soit 71% environ.

TABLEAU 2:

Durées moyennes des *consonnes* en position de coda après les différentes voyelles

	ptk	bdg	fsʃ	vzʒ
i y u	18.56	11.95	23.12	9.3
ɛ a ɔ œ	17.0	11.84	21.35	10.6
ə ɑ o ø	16.83	9.87	18.70	10.8
ɑ̃ ɛ̃ ɔ̃ œ̃	13.12	8.95	17.62	9.8

Remarques: quelles que soient les voyelles qui les précèdent, les constrictives sourdes sont les consonnes les plus longues, et les constrictives sonores sont les plus courtes. Les occlusions sourdes sont aussi toujours plus longues que les sonores. A l'intérieur de la rime, plus la coda est longue, plus le noyau est court; cela se vérifie avec ptk, bdg et fsʃ ; les consonnes allongeantes échappent à cette règle parce qu'elles allongent toutes les voyelles qu'elles entravent.

TABLEAU 3:

Durées moyennes des *consonnes* en coda après les voyelles brèves ou longues et leur réduction selon le type d'accent.

	Voyelles brèves			**Voyelles longues**		
	Acc. terminal	Acc. intérieur	R	Acc. term.	Acc. int.	R
ptk	17.78	13.64	.76	15.0	11.55	.77
bdg	11.89	7.0	.59	9.41	7.75	.82
fsʃ	22.23	19.79	.89	18.16	12.58	.69
vzʒ	10.0	8.1	.81	10.34	8.3	.80

Remarques: La réduction des consonnes selon l'accent n'est pas régulière comme celle des voyelles (tableau 1). Les constrictives sourdes résistent beaucoup à la réduction après les voyelles brèves (.89), mais s'abrègent davantage après des voyelles longues par nature qu'elles ne peuvent abréger. Les occlusives sourdes se réduisent à environ .77, indépendamment de la voyelle qu'elles ont le pouvoir d'abréger. La comparaison des réductions ne vaut pas avec les consonnes sonores, parce qu'elles sont brèves et ne sauraient s'abréger beaucoup; / bdg/ perdent 4 à 5 cs avec les voyelles brèves et seulement 1.5 ou 2cs avec les voyelles longues, peut-être parce qu'elles sont alors à la limite de leur réduction sous l'accent; quant aux allongeantes, elles sont toujours brèves parce qu'elles allongent les voyelles dans la rime. Je ne suis pas prêt à tenter des explications profondes de ces mécanismes de compensation de durée; sans doute, les limites des durées syllabiques, longues ou brèves, sont-elles à prendre en considération.

On le voit, d'un accent à l'autre, l'organisation des durées relatives du noyau et de la coda dans la rime semble se faire à l'intérieur de l'espace VC et respecte, à travers la dynamique des coarticulations, les durées phonologiques des voyelles et la nature phonétique des consonnes; si la voyelle est brève, la consonne sourde s'abrège, s'il s'agit d'une occlusive; si la voyelle est longue, phonologiquement ou par coarticulation, les consonnes longues cèdent du terrain. Ainsi, aucune rime n'est vraiment très courte, et les plus longues ne peuvent l'être trop. On peut distinguer plusieurs classes de rimes faciles à reconnaître automatiquement au moyen des durées relatives de V et de C, de la durée de la rime VC, et au moyen des occlusives ou des constrictives, longues ou brèves. Une telle analyse acoustique n'a pas besoin d'être très fine et ne repose pas sur le trait fragile de la sonorité. Pour la synthèse par règles, on peut donner à l'ordinateur un grand nombre de configurations de rimes, et même tenir compte de la différence dans une classe de consonnes, comme P ou K par exemple, ou dans une classe de voyelle, comme /i/ et /u/.

Dans certaines rimes, le noyau vocalique est nettement prédominant en durée sur la coda; c'est le cas des voyelles longues ou allongées entravées par une consonne allongeante, (Voir colonne 2, tableau 4.) Ex. *nage* = 25 +14, soit 64% de noyau contre 36% de consonne; si on donne une marque positive pour la vocalité, cette rime a +26 de proéminence positive. A l'autre extrême, *vache* obtient -35. Au centre, *relâche* est à +2.5.

Le tableau 4 a pour but d'illustrer le comportement des durées relatives en fonction de différents accents et de la coupe syllabique dans les morphèmes. On peut y observer que dans les deux premières colonnes où la rime n'est pas divisée par la coupe syllabique, la durée des consonnes varie en sens inverse de celle de la voyelle. Dans les colonnes 3 et 4, les mots à l'étude voient leur rime abrégée parce qu'elle se trouve sous l'accent intérieur, mais elle ne change pas de signe pour la prépondérance vocalique ou consonantique; la consonne nasale qui suit la coda semble l'abréger, mais je n'ai rien fait pour examiner cette influence. Il est possible, d'autre part, que la métrique dans les groupes prononcés ait influencé le poids de certaines syllabes, même si ces groupes ne se prêtaient pas à l'alternance comme dans les mots à trois ou quatre syllabes.

TABLEAU 4:

Exemples d'organisation des durée dans les rimes à noyau bref ou long selon la position du morphème par rapport à l'accent

Exemples d'accent et de coupes syllabiques.	1 Pâte	2 Aimes-tu les pâtes	3 Le mot pâte me plait	4 Des pâtes maison	5 De la pâte à tarte	6 Un empâté	7 Empâtement
Accent:	1	1	2	2	2/0	0 - 1	0 - 0
patte	13 • 2	9 • 17	8 • 14	8 • 12	8 • 8	9 - 13	7 • 13
pâte	25 • 15	18 • 13	13 • 11	11 • 10	12 • 7	12 - 12	10 • 8
chante	24 • 11	20 • 14	14 • 10	11 • 11	12 • 6	12 - 12	13 • 12
faite	12 • 20	7 • 17	7 • 14	7 • 15	7 • 7	8 - 12	
fête	23 • 13	17 • 12	13 • 12	11 • 11	13 • 8	11 - 13	
tinte	26 • 11	24 • 9	16 • 13	15 • 10	16 • 6	15 - 11	17 • 11
laide	12 • 12	9 • 11	8 • 8	8 • 6	11 • 6	7 - 9	8 • 6
l'aide	21 • 9	19 • 9	15 • 9	15 • 8	12 • 6	12 - 9	
vache	15 • 23	12 • 25	13 • 13	11 • 1·	1^ • 10	10 - 15	10 • 14
relâche	22 • 18	20 • 19	18 • 14	12 • 10	13 • 12	13 - 15	13 • 12
étanche	26 • 17	27 • 18	18 • 13	15 • 11	15 • 11	17 - 16	15 • 13
nage	24 • 14	25 • 14	14 • 8	13 • 9	11 • 7	14 - 10	16 • 9
âge	27 • 14	28 • 10	15 • 7	14 • 10	14 • 7	16 - 7	
mélange	28 • 11	26 • 11	15 • 9	15 • 9	14 • 6	17 - 7	17 • 10

Dans la colonne 5, la rime du morphème est divisée par la coupe syllabique [dla pA ta tArt]; mais, du moins quand le noyau est une longue par nature, la cohésion morphologique semble empêcher cette coupe syllabique dans la prononciation, de sorte que l'entrave peut rester intacte en québécois.

Par contre, dans la colonne 6, le morphème est vraiment divisé par la coupe syllabique, de sorte que la voyelle se trouve en position pénultième et la consonne, sous l'accent final. Dans ce cas, les voyelles longues pourraient être abrégées par rapport aux cas où elles sont sous l'accent 2, mais ce n'est pas ce qu'on observe; la pénultième, même si elle ne porte pas toutes les caractéristiques de l'accentuation, garde une durée qui aide à faire sentir le morphème. Dans la colonne 7, le morphème tout entier tombe en position pénultième. C'est le peu de différence entre les colonnes 5, 6 et 7 qui me porte à croire que les morphèmes à noyau long par nature ne se laissent pas vraiment diviser par la coupe syllabique et préservent une durée vocalique qui les démarque nettement des noyaux brefs, même en dehors de l'accent. Je ne trouve pas une telle tendance chez les locuteurs français

Cette étude ne constitue vraiment qu'une première approche. Tout reste à chercher dans ce domaine, mais on peut déjà entrevoir que la phonologie du français québécois devra être respectée dans toutes ses implications et qu'elle va peser lourd sur la rythmique, l'accentuation et l'intonation, donc sur toutes les composantes de la prosodie qu'il nous reste à définir.

En conclusion, je ne puis que souhaiter la formation d'équipes de concepteurs de modèles théoriques et de phonéticiens secondés par des informaticiens pour travailler utilement ensemble dans les industries de la langue. Cette collaboration ne rendra notre démarche que plus scientifique et plus prometteuse. Il faudrait peut-être aussi avertir la société et les pouvoirs publics que l'entreprise durera des décennies, et que personne ne fera à notre place l'industrialisation de notre langue. Tout progrès dans les connaissances fondamentales de la langue et dans son utilisation dans la société est de l'ordre des exposants en mathématique; la recherche scientifique dans ce domaine pourrait être subventionnée comme la défense du territoire.

Bibliographie

BEAUCHEMIN, N. (1970). *Recherches sur l'accent d'après des poèmes d'Alain Grandbois*. Les presses de l'Univ. Laval, Québec et Klincksieck, Paris.

BAUDREAULT, M. (1968). *Rythme et mélodie de la phrase parlée en France et au Québec*. Les presses de l'Univ. Laval, Québec et Klincksieck, Paris.

NEILSON, R., B. GRANSTRÖM (1973). "Word Accent, Emphatic Stress and Syntax in a Synthesis by Rules Scheme for Swedish Speech Transmission Lab". Stockholm, D.R.S.R., 2/3, 31-35.

CARTON, F., D. HIRST, A. MARCHAL et A. SÉGUINOT (1976). *L'accent d'insistance*. . Studia Phonetica 12. Didier.

CARTON, F. et F. LONGCHAMP (1977). "Validation d'indices perceptifs de l'intonation par analyse multidimensionelle". Actes des 8^{e} journées d'Etudes sur la parole. Aix-en Provence, 133-138.

COUSTENOBLE, H. M. and L. E. ARMSTRONG (1934). *Studies in French Intonation*. Heffer, Cambridge.

CUTLER, A. et D. R. LADD (1982). *Prosody: Models and Measurements*. Language and Communication 14. Springer-Verlag.

DELATTRE, P. (1938). "L'accent final en français : accent d'intensité, accent de hauteur, accent de durée". French Review, 12 (2) 141-145.

------ (1966). "Les dix intonations de base du français". French Review, 40 (1) 1-14.

DELL, F. (1984). "L'accentuation dans les phrases en français". In Dell et al. *Forme sonore du langage*. Herman, 65-119.

DELL, F., D. HIRST et J.-R. VERGNAUD (1984). *Forme sonore du langage*. Herman, Paris.

DELL, F. (1982). "On Delimiting International Stretches in French". Manuscrit.

DI CHRISTO, A. (1975). *Soixante et dix ans de recherche en prosodie*. Presses de l'Univ. de Provence.

------ (1982). *Prolégomènes à l'étude de l'intonation*. Editions du CNRS. Paris.

------ (1981). *De la microprosodie à l'intonosyntaxe*. Th. de Doct. d'Etat, Univ.

DI CHRISTO, A., J.-P. HATON, M. ROSSI, J. VAISSIÈRE (1982). *Prosodie et reconnaissance automatique de la parole*. Actes du séminaire du GALF, Aix-en-Provence. GALF et GRECO Communication parlée. CNRS. Paris.

EMERARD, F. (1977). *Synthèse par diphones et traitement de la prosodie*. Thèse de 3^{e} cycle, Université de Grenoble.

FAURE, G. (1962). "Aspects et fonctions linguistiques des variations mélodiques dans la chaîne parlée". Proc. 9th Int. Congr. Ling. Cambridge, 72-78.

------ (1967). "La description phonologique des systèmes prosodiques". Proc. 6th Int. Congr. Phon. Sciences, Pragues.

------ (1968). "Accent, rythme et intonation". *Le français dans le monde*, (57) 15-19.

FONAGY, I. et J. SAP (1978). "À la recherche de traits prosodiques du français parisien". Phonetica 36, 1-20.

------ (1983). *La vive voix*. Payot.

FONAGY, I. et P.R. LÉON (1979). *L'accent en français contemporain*. Studia Phonetica 15. Didier.

GARDE, P. (1968). *L'accent*. P.U.F., Paris.

GÄRDING, E. (1984). "Comparing Intonation". Lund Working Papers 27. 75-101.

HOLDER, M. (1968). "Étude sur l'intonation comparée de la phrase énonciative en français canadien et français standard". In Recherches sur la structure phonique du français canadien. Studia Phonetica 1, Léon ed., Didier. 175-191.

LADD, D.R. (1980). "The Structure of Intonational Meaning: Evidence from English". Indiana University Press, Bloomington.

LAVOIE, G. (1965). *Quelques aspects du rythme et de la mélodie chez Mgr Félix-Antoine Savard, écrivain canadien*. Th. de Doct. Manuscrit, Univ. de Strasbourg.

LÉON, P.R. et Ph. MARTIN (1969). *Prolégomènes à l'étude des structures intonatives*. Studia Phonetica 2, Didier.

LÉON, P.R. et M. ROSSI (1979). *Problèmes de prosodie*. Vol. I: approches théoriques. Studia Phonetica 17. Didier.

------ (1979). *Problèmes de prosodie*. Vol. II: Expérimentations, modèles et fonctions. Studia Phonetica 18. Didier.

LIBERMAN, M. and A. PRINCE (1977). "On Stress and Linguistic Rhythm". *Linguistic Inquiry* 8, 249-336.

LIBERMAN, M.Y. (1979). "The Intonational System of English". Ph.D. dissertation, M.I.T. Garland Press, New York.

------ (1983). *In Favor of some Uncommon Approaches to the Study of Speech*. In Mac Neilage (ed.) *The Production of Speech*. Springer-Verlag. 265-274.

MALMBERG, B. (1961). "Analyse instrumentale et structurale des faits d'accent". Proc. 4th Int. Congr. Phon. Sciences. Helsinki, 456-475.

------ (1966). "Analyse des faits prosodiques - problèmes et méthodes". Cah. Ling. Théor. Appl. Bucarest, 99-107.

MARTIN, Ph. (1977). "Résumé d'une théorie de l'intonation". Bulletin Inst. Phon. Grenoble, vol. VI, 57-87.

METTAS, O. (1964). "Étude sur l'intonation en français". Tr. ling. litt. Strasbourg, 2(1), 99-105.

PIERREHUMBERT, J. and M. LIBERMAN (1982). "Modeling the fundamental frequency of the voice". *Contemporary Psychology*, 27(9), 690-692.

RIGAULT, A. (1961). "Rôle de la fréquence, de l'intensité et de la durée vocalique dans la perception de l'accent en français". Proc. 4th Int. Congr. Phon. Sciences. Helsinki, 735-748.

------ (1970). "L'accent dans deux langues à l'accent fixe". Studia Phonetica 3, 1-12.

ROSSI, M. et M. CHAFCOULOFF (1972a). "Recherche sur le seuil différentiel de fréquence fondamentale dans la parole". Tr. de l'Inst. de Phon. d'Aix, 1:179-185.

------ (1972b). "Les niveaux intonatifs". Tr. de l'inst. de Phon. d'Aix 1:167-176.

ROSSI, M., A. DI CHRISTO, D. HIRST, Ph. MARTIN, Y. NISHINUMA (1981). *L'intonation: de l'acoustique à la sémantique.*

ROSSI, M. (1985). "L'intonation et l'organisation de l'énoncé". *Phonetica* 42: 135-153.

------ (1987). "Peut-on définir l'organisation prosodique du langage spontané?". Études de linguistique appliquée. Didier Érudition. Avril-Juin, 20-48.

SANTERRE, L. (1974). "Deux /E/ et deux /A/ phonologiques en français québécois". In *Le Français de la région de Montréal.* Presses de l'Université du Québec. 117-145.

------ (1979). "Comparaison des /E/ et des /A/ en québécois et en français". In *25 ans de linguistique au Canada.* Centre éducatif et culturel, Montréal. 325-361.

------ (1979). "La fusion des voyelles en frontières inter et intra-syntagmatiques". Amsterdam Studies in the Theory and History of Linguistics Science, vol. 9. 1131-1138. John Benjamins B.V. Amsterdam.

------ (1981). "Stabilité et variation des opposition du /ɛ/ bref et du /ɛ/ long, du /a/ antérieur et du /ɑ/ postérieur en français montréalais". *Logos Semantikos*, vol. V. 375-384. Gregos, Walter de Gruyter.

SANTERRE, L. et J.-L. CHANDON (1981). "Fonctions morphologiques des paramètres suprasegmentaux en français québécois". *12e Journées d'Études sur la parole.* Presses de l'Université de Montréal. 54-65.

------ (1981). "Duration Distinguishes tenses in Montreal French". Actes du symposium Prosodie. Martin, Ph. éditeur, University of Toronto. 28-41.

SANTERRE, L. et D. VILLA (1981). *Les paramètres acoustiques en frontières de mots.* Studia Phonetica 18. 3-10.

SANTERRE, L. (1987). "Durées systématiques dans les rimes *CVC* en fonction des segments et de l'accent". Actes des 16e Journées d'Études sur la parole. Société française d'acoustique, LIMSI, Orsay, Paris.

SANTERRE, L. (1987). "Systématique des durées segmentales dans les rimes syllabiques à voyelles longues et brèves par nature". Actes du Congrès International des Sciences Phonétiques, vol. 5. 120-129. Tallium, URSS.

SELKIRK, E.D. "The Phrase Phonology of English and French". Ph.D. dissertation, M.I.T.

SELKIRK, E.O. (1980). "The Role of Prosodic Categories in English Word Stress". Linguistic Inquiry 11-3, 563-605.

SHANE, S. (1968). *French Phonology and Morphology*. M.I.T. Press

THORSEN, N. (1981). "Intonation Contours and Stress Group Patterns in Declarative Sentenses of Varying Length in ASC Danish". ARIPUC 15, 13-47.

VAISSIÈRE, J. (1983). "Language - Independent Prosodic Features". In Cutler and Ladd, Prosody: *Models and Measurements*. Springer - Verlag. 53-66.

VERLUYTEN, P. (1982). *Recherches sur la prosodie et la métrique du français*, Universitaire Instelling Antwerpen. University Microfilms Internaticnal, Ann Arbor, Michigan.

WARREN, R. et L. SANTERRE (1981). "Les paramètres acoustiques de l'accent en français montréalais". Studia Phonetica 15, Didier. 53-63.

ZWANENBURG, W. (1964). *Recherches sur la prosodie de la phrase française*. Universitaire Pers. Leiden Hollande.

LE TRAITEMENT INTERACTIF DES DOCUMENTS

Michael Mepham
Université Laval

L'informatisation du traitement de la langue est fondée sur des hypothèses simplificatrices quant à la nature de l'activité linguistique humaine. Nous soutenons que l'innovation et l'improvisation inhérentes à cette activité sont incompatibles avec l'atteinte à court et à moyen terme de résultats intéressants par la voie de l'automatisation pure.

Les problèmes de l'automatisation seront illustrés à l'aide d'exemples tirés des travaux de l'équipe à Laval. Les avantages d'une approche interactive seront mis en exergue, et certains travaux et projets dans ce sens seront esquissés.

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■

INTRODUCTION

Les travaux dont il est question ici font partie d'un ensemble connu sous le nom de projet RELAI. Ils se situent dans un créneau étroit à l'intérieur du domaine du traitement automatique des langues naturelles.

Nos travaux visent le développement d'outils et de méthodes pour la création de bases de données textuelles; c'est-à-dire, des bases de données fondées sur des textes, dans lesquelles le contenu pertinent est accessible par automatisme. La stratégie mise de l'avant est celle du prétraitement des éléments du texte, notamment les mots, afin de préciser leur nature et leurs relations en vue de la détermination du sens.

Nous commençons par justifier l'approche adoptée, en examinant les difficultés fondamentales qu'opposent les langues aux approches plus générales et ambitieuses en vogue. Ensuite, nous discutons des réalisations obtenues jusqu'ici: programmes, bases de connaissances, et bases de données textuelles.

BASES DE DONNÉES TEXTUELLES

Pour qu'un texte soit considéré comme une base de données, il faut que l'on puisse en accéder automatiquement au contenu. Autrement dit, un texte est une base de données dans la mesure où on peut en extraire des connaissances. Les systèmes les plus rudimentaires qui répondent à cette définition se servent des seules formes qui sont directement reconnaissables dans le texte: les mots-formes. Ainsi, les systèmes de documentation automatique les plus simples fonctionnent avec les formes des mots et leurs combinaisons booléennes pour sélectionner les passages en réponse à une requête. Le principe de fonctionnement est semblable à celui des correcteurs orthographiques dans les logiciels de traitement de texte, et à celui des fonctions d'aide dans certains logiciels. De tels systèmes connaissent des limites qui découlent du fait de considérer le sens comme une propriété des mots.

LES PROBLÈMES DE TYPE LEXICAL

Les mots ayant un rôle clé dans le discours, il n'est pas étonnant que les difficultés de traitement se manifestent déjà à leur niveau. En outre, il y a des difficultés liées à d'autres niveaux, mais qui peuvent être ramenées à celui des mots quant à leur traitement. Parmi ces problèmes de nature lexicale, nous nous penchons sur certains cas typiques, choisis pour éclairer la discussion subséquente de nos logiciels.

Le problème fondamental est celui de l'ambiguïté: chaque mot-forme peut correspondre à plusieurs sens. Cet état de choses n'a rien de surprenant lorsque l'on considère que le vocabulaire d'une personne peut être de l'ordre d'une dizaine de milliers de mots, et qu'il doit servir à spécifier des millions d'entités différentes. Ainsi, un terme générique comme *unité* fait double emploi pour représenter *disquette* et *dispositif de lecture* (exemples 1 et 2). De même, un bureau peut être un meuble (exemple 4), une pièce (exemple 5), ou une institution (exemple 6). Dans le cas de *copie*, nous avons une ambiguïté entre le nom et le verbe (exemples 1 et 3).

(1) Faire une copie du fichier sur une unité simple face.
(2) Mettre la disquette dans l'unité B.
(3) Lorsque l'on copie un fichier, il faut...
(4) Pierre était assis derrière son bureau.
(5) Il a quitté son bureau à cinq heures.
(6) La compagnie a deux bureaux à Québec.

Les exemples précédents illustrent des cas où les sens sont multiples, mais discrets et énumérables. Encore plus difficilement traitables sont les cas où le sens dérive sans borne prévisible. Doit-on prendre le mot *Dallas* (exemple 7) comme synonyme de *1963*, ou d'*assassinat*? En plus des métaphores bien connues, comme *carte* (exemple 9) dans le sens d'*atout*, il y a des improvisations qui constituent une source intarissable d'innovations, lesquelles sont compréhensibles en dépit de leur originalité. C'est le cas pour le sens positif d'*écoeurant* dans l'exemple 8.

(7) Depuis Dallas, la politique américaine a bien changé.
(8) J'aime Sting; c'est écoeurant comme il est bon!
(9) Le libre-échange constitue sa meilleure carte électorale.

Les ambiguïtés structurales font multiplier les effets des ambiguïtés de mot sur les interprétations d'une phrase. Lorsque l'ambiguïté structurale peut être associée à un mot dans la phrase, le phénomène est semblable à l'ambiguïté des sens de mot. Le rattachement des syntagmes prépositionnels nous sert d'exemple. Après le complément d'objet, le syntagme prépositionnel peut être en relation avec le nom qui précède (exemple 10), avec le verbe (exemple 11), ou avec la proposition entière (exemple 12).

(10) Jean a acheté la serrure avec la clé.
(11) Jean a acheté la serrure avec l'argent de sa mère.
(12) Jean a acheté la serrure avec un sourire aux lèvres.

Au-delà de la syntaxe de la phrase, il y a des relations de discours qui constituent une source importante d'ambiguïté dans l'interprétation des textes. Le phénomène d'anaphore est typique: le sens précis d'un pronom doit être déduit par un processus complexe à partir du contexte

linguistique et situationnel. Ainsi, dans l'exemple 13, *il* fait référence à *Paul* et non à *professeur*. Le phénomène de déixis, par lequel un mot comme *celui-là* renvoie à un objet réel en dehors du discours, est analogue.

(13) Paul a demandé au professeur à voir sa copie d'examen. Il voulait vérifier sa note.

Finalement, le problème primordial pour le traitement des textes en tant que bases de données est celui de la reconnaissance même des mots comme éléments. Ce que nous appelons mots simples, délimités par des blancs et des signes de ponctuation, n'est pas en cause. Mais s'il n'y avait que les mots simples pour désigner les choses, l'ambiguïté des sens prendrait des proportions astronomiques. La combinaison des mots simples en mots complexes constitue le palliatif consacré à cette lacune virtuelle. L'ennui pour le traitement automatique, c'est que les mots complexes ne sont pas reconnaissables aussi facilement que les mots simples.

LES MOTS COMPLEXES

Le traitement des mots complexes rencontre plusieurs embûches. D'abord, en français, les mots complexes ne sont pas formellement délimités, sauf exception, comme pour *compte-gouttes* dans l'exemple 14. De plus, il n'existe pas de relevé exhaustif: les dictionnaires généraux n'en présentent qu'une petite fraction, si on admet qu'il peut y en avoir quatre fois plus de mots complexes que de mots simples différents dans le lexique de certains domaines. Bien entendu, un tel estimé dépend de la définition de la notion de mot complexe, souvent considéré comme une unité dont le sens n'est pas déterminé par la simple composition des sens de ses mots constituants. En pratique, une telle définition laisse beaucoup de flou, et ne satisfait pas les terminologues, qui s'intéressent davantage à la correspondance entre mot et objet de référence. Les exemples 15 et 16 montrent des cas où la segmentation du textes en mots dépend de l'interprétation du texte par un lecteur humain. Il s'agit en l'occurrence de l'expression *nouvelle disquette*, utilisé tantôt (exemple 15) comme synonyme de *disquette vierge*, et tantôt (exemple 16) comme déictique pour la disquette la plus récemment utilisée. La segmentation est parfois compliquée par la présence d'ambiguïtés structurales: faut-il isoler *premier ministre* ou *ministre de l'Éducation* dans l'exemple 17? A la lumière des problèmes de ce type, nous maintenons que l'utilisation de textes comme bases des données exige la reconnaissance des mots complexes.

(14) Il a payé sa dette au compte-gouttes.
(15) Pour utiliser DISKCOPY, mettez la disquette à dans le lecteur A et la nouvelle disquette dans le lecteur B.
(16) Faites une copie de la disquette originale et la nouvelle disquette.
(17) M. Paul Gérin-Lajoie a été le premier ministre de l'Educatio. du Québec.

L'APPORT DES BASES DE CONNAISSANCES

Les types de problème que nous avons évoqués plus haut constituent des entraves au traitement automatique des textes en tant que bases de connaissances. Les systèmes de compréhension automatique du langage naturel actuellement en développement proposent des éléments de solution qu'il convient de mentionner. Bien que les solutions peuvent prendre la forme de programmes ou de bases de connaissances, la distinction n'est pas pertinente pour les fins de cet exposé. Il importe de considérer les connaissances qu'il faut formaliser en vue des solutions, plutôt que les stratégies propres aux différents logiciels. Par conséquence, nous n'examinerons que

les apports possibles des bases de connaissances. D'ailleurs, notre intérêt étant situé autour du traitement des mots, il s'agit surtout de voir comment les types de bases de connaissances peuvent contribuer au traitement des mots.

Nous envisageons les bases de connaissances selon la hiérarchie suivante: lexical, syntaxique, conceptuel, et réel. La hiérarchie va du plus simple, au niveau des mots, au plus inclusif quant à la contribution au sens. Puisque les lacunes de niveau inférieur exigent un recours au niveau supérieur, il est implicite qu'un système identifié à un niveau quelconque peut comprendre les niveaux inférieurs.

Les connaissances lexicales sont typiquement consignées dans des dictionnaires automatiques, et servent à l'identification des mots et leurs relations. Ainsi, un inventaire des mots complexes pour un domaine permet évidemment la reconnaissance de ces mêmes éléments, hormis les cas d'ambiguïté. Cette reconnaissance élimine une bonne partie de l'ambiguïté propre aux mots simples constitutifs des mots complexes. Les contraintes de voisinage contextuel permettent d'éliminer encore d'autres ambiguïtés: jusqu'à 50 pour cent dans une expérience sur des textes traitant de géographie (Dahlgren 1988). D'autre part, l'utilisation de thésaurus permet de formaliser des relations entre mots, par exemple entre synonymes, et ainsi de dépasser les limites imposées par l'orthographie conventionnelle.

Est-il possible de reconnaître automatiquement les mots complexes sans recours à une base de connaissances? On pourrait espérer qu'avec un corpus suffisamment grand, le relevé des collocations fréquentes coinciderait avec les expressions fixes. Une expérience (Choueka 1988) montre qu'environ 10 % des collocations fréquentes n'étaient pas des mots complexes. Il faut donc recourir à l'intervention humaine dans la capture des mots complexes pour inclusion dans des bases de connaissances.

Le recours à l'analyse syntaxique semble promettre la résolution des ambiguïtés de mot, tout en contribuant à la représentation du sens de la phrase. Typiquement, un ensemble de règles de type syntagmatique est appliqué aux suites de mots (Katz and Fodor 1963). Mais l'analyse syntaxique de phrases dont l'identification des mots comporte de multiples ambiguïtés résulte en une explosion combinatoire des structures possibles. L'exercice permet d'éliminer une portion des ambiguïtés de mot, mais pas toutes. Deux voies de solution existent: l'interactivité avec une personne pendant l'analyse, ou le pré-traitement des phrases.

Maurice Gross a démontré l'inaptitude des ensembles de règles syntagmatiques à saisir la complexité des relations syntaxiques, et a proposé le développement d'un lexique syntaxique très systématique (Gross 1976). Chaque sens de mot distinct est associé à un ensemble de restrictions syntaxiques. Celles-ci peuvent servir à la sélection du sens approprié dans un texte, dans la mesure où les occurrences textuelles manifestent des restrictions distinctives. La base de connaissances permettra de traiter celles des ambiguïtés qui admettent un traitement sans recours à la représentation du sens. Il s'agit d'une approche qui, lorsqu'elle sera opérationnelle, exigera un matériel et des logiciels importants. La résolution automatique des ambiguïtés récalcitrantes exigera toujours un recours à une base de connaissances conceptuelles.

Dans la notion de base de connaissances conceptuelles, il faut comprendre tous les systèmes qui formalisent les relations sémantiques liées à la langue plutôt qu'à la réalité extra-linguistique. Ainsi, il y a la sémantique préférentielle (Wilks 1975), les dépendances conceptuelles de Schank, la sémantique naïve (Dahlgren 1988), les structures conceptuelles (Sowa 1984), et la sémantique décompositionnelle de Jackendoff (Jackendoff 1983). Chacun des systèmes intègre d'une façon ou d'une autre le composant syntaxique discuté plus haut. Ils vont au-delà de ce que permet un analyseur syntaxique, en faisant intervenir ce qui a pu être formalisé au niveau conceptuel.

Aucun de ces systèmes n'est suffisamment général et développé pour servir de façon pratique dans l'état actuel des choses. Et aucun ne peut promettre d'arriver à une représentation

complète du sens d'un texte sans inclure une base de données comportant des connaissances universelles. Entre le niveau conceptuel et celui des connaissances du monde, il est difficile de tracer une ligne, et pour les besoins de notre propos, il n'est pas nécessaire. En fin de compte, pour obtenir de façon automatique une représentation adéquate d'un texte, il faut pouvoir disposer des mêmes connaissances auxquelles les humains font appel lorsqu'ils communiquent par la parole. Ainsi, pour résoudre l'ambiguïté du mot *bureau* dans l'exemple 4, il faut faire intervenir des connaissances conceptuelles et réelles semblables à celles de la liste suivante:

bureau, sens de meuble
endroit pour travailler
comporte une surface plane
utilisé par une personne en position assise
ordre de grandeur d'un être humain
bureau, sens de pièce
endroit pour travailler
partie d'un édifice
peut contenir une ou plusieurs personnes
Pierre, sens du nom propre
nom d'une personne
objet animé
pierre, sens d'objet
objet inanimé
pierre, sens de matériel
matière inanimée

Divers formalismes ont été mis de l'avant: les "frames" (Minsky 1975), "scripts" (Schank and Abelson 1977), "scenes", etc. Aucune des approches envisagées ne permet d'espérer une solution générale à la représentation de telles connaissances.

L'ALTERNATIF DU TRAITEMENT INTERACTIF

Nous avons vu que l'engagement dans la voie de l'automatisation pure nous entraîne nécessairement au-delà du possible. Faut-il renoncer à l'utilisation de l'ordinateur pour la manipulation de la langue sauf dans les cas où on peut tolérer les erreurs? Quelles sont les alternatives? En ce qui concerne la reconnaissance des mots, il y a deux autres voies qui n'ont pas été mentionnées, mais qui ne suffisent pas non plus. D'une part, il est possible de réduire le degré du problème en travaillant à l'intérieur de domaines restreints. D'autre part, et de façon analogue, il est possible de faire intervenir la notion de sujet de paragraphe dans le choix du sens. Par contre, il n'est pas évident de définir le domaine, ou le sujet, selon le cas.

Le traitement interactif constitue également une possibilité. L'utilisation des aides à la rédaction et à la traduction en est un exemple concret. L'analyseur syntaxique de Tomita en est un autre.

L'approche que nous préconisons est celle du traitement interactif, mais comme étape de pré-traitement aux autres traitements. L'idée est d'éliminer les obstacles au traitement automatique des textes par l'entremise d'un enrichissement sélectif du texte au niveau des mots.

Pour que cette alternative soit valable, elle doit constituer une solution concrète et pratique. L'intervention humaine doit être minimisée à l'aide d'outils informatiques simples à utiliser. L'assistance humaine ne doit pas exiger d'expertise autre que la connaissance de la langue

et une certaine connaissance du domaine, comme c'est le cas pour des secrétaires. Le logiciel doit être formulé pour être souple et général, et doit pouvoir tourner sur des équipements de micro-informatique. Les résultats doivent être facilement transportables dans d'autres systèmes sous une forme qui est intelligible à l'utilisateur, ou directement présentable selon ses besoins.

Dans cet esprit, nous avons déjà développé plusieurs logiciels, et nous continuons à travailler au développement de l'ensemble. Dans ce qui suit, nous faisons un survol très sommaire de deux logiciels, LEMMATISEUR et SYREX, afin de montrer en quoi ils répondent aux objectifs mis de l'avant, et comment ils sont utiles dans la conduite de travaux concrets.

L'approche générale des deux logiciels est semblable. Chacun applique automatiquement aux mots les connaissances contenues dans la base de connaissances, laissant l'utilisateur reprendre en mode interactif les cas résiduels. En mode interactif, chaque occurrence à traiter est affichée dans son contexte, et accompagnée des renseignements sur son identification effective et possible. Les décisions prises par l'utilisateur au sujet d'un mot sont ajoutées à une base de connaissances appartenant au corpus de texte ou au domaine dont celui-ci relève. Ainsi, à mesure que l'utilisateur avance dans le traitement de sous-corpus successifs, la base de connaissances s'étoffe et la part d'intervention humaine diminue.

LE LOGICIEL L E M M A T I S E U R

Le logiciel LEMMATISEUR est un outil pour la création de bases de données textuelles. Il possède un grand nombre de fonctions qui permettent d'étiqueter chaque occurrence du texte avec une forme standard (lemme), une catégorie, et des précisions diverses selon les besoins de l'utilisateur. La base textuelle peut être exportée sous la forme de base enrichie, sous la forme de texte ordinaire, ou encore sous la forme d'index ou de concordance des mots analysés.

Le fonctionnement détaillé du logiciel est bien documenté (Mepham et Bérubé 1987). Considérons plutôt les types d'application auquel il peut se prêter, en suivant l'énumération des problèmes de traitement évoqués plus haut.

L'ambiguïté d'un mot peut porter sur sa catégorisation, sur le choix de lemme (forme standard) à l'intérieur de la catégorie, ou sur des précisions de sens pour un même lemme. Dans tous les cas, les ambiguïtés se traitent principalement en mode interactif. On parcourt la base textuelle en s'attardant sur chaque occurrence de la forme proposée, afin de sélectionner la bonne identification parmi celles qui existent déjà, ou d'en assigner une nouvelle. Ou bien, on parcourt la base en s'attardant sur chaque occurrence déjà marquée comme ambiguë lors d'une phase de traitement automatique. Il faut bien noter que le système ne fait pas la capture automatique des ambiguïtés; il ne fait que marquer dans la base textuelle celles qui sont déjà consignées dans la base de connaissances. Le travail humain de désambiguïsation devrait être réduit éventuellement par l'ajout d'un module de règles contextuelles.

Les cas de métaphore et de dérive de dénotation ne diffèrent pas de l'ambiguïté quant à leur traitement. Il suffit que l'utilisateur sache comment distinguer les sens, les lemmes, et les catégories qu'il veut assigner. Ainsi, le logiciel permet de traiter des corpus littéraires, historiques philosophiques ou autres selon les termes d'analyse propres à son domaine.

Certains cas d'ambiguïté structurale se prêtent au même traitement. Nous avons choisi l'exemple de l'attachement des syntagmes prépositionnels. Ou bien les prépositions que l'on veut analyser sont marquées dans la base textuelle pendant une phase de traitement automatique, ou bien on choisit nommément chaque préposition que l'on veut traiter comme ambiguë. Ensuite, en mode interactif, on assigne à chaque occurrence un code distinctif de son régime d'incidence, selon qu'il est attaché au nom, au prédicat ou à la proposition dans le contexte affiché.

Les cas d'anaphore et de déixis ne diffèrent pas des prépositions quant à leur traitement. Ainsi, les pronoms se voient attribuer une précision quant à leur référent. Par contre, les relations de discours ne sont pas toujours marquées dans le texte sous la forme d'un mot. C'est le cas pour la relation de causalité entre la deuxième et la première proposition de l'exemple 8. Il est possible d'insérer un marqueur sous la forme d'un mot bidon dans la suite textuelle, pendant le traitement interactif, ou d'assigner un code approprié à un mot existant (le verbe, par exemple).

Dans sa version actuelle, LEMMATISEUR ne permet pas le traitement de mots dont la longueur dépasse 30 lettres. Cette situation sera corrigée dans la prochaine version, présentement en préparation, par la prévision de zones d'information dont la longueur est definissable par l'utilisateur. Il sera alors possible de traiter des séquences de mots au même titre que les mots simples. La reconnaissance de ces séquences est assurée par un logiciel autonome, nommé SYREX pour *sy*stème de *r*econnaissance des *ex*pressions.

LEMMATISEUR a été conçu en réponse aux besoins du projet d'automatisation de l'enquête du français dans la région de Québec (Deshaies 1981). Il s'agissait de constituer en base de données les 120 sous-corpus de transcriptions issues des enregistrements de langue parlée faites dans le cadre d'une enquête sociolinguistique. L'objectif était de rendre le corpus utile pour des études portant sur les éléments linguistiques. A l'heure actuelle, la base comprend déjà 66 sous-corpus, pour au-delà d'un demi-million de mots courants. Les mots ont été traités afin d'assigner des formes lemmatiques aux variantes morphologiques des noms, verbes et adjectifs, ainsi que des catégories grammaticales. Les ambiguïtés restent à être traitées par chaque utilisateur de la base en fonction de ses besoins. Un code de référence identifie le sous-corpus d'appartenance de chaque occurrence textuelle, de sorte que les données de type social peuvent être associées à celles extraites de la banque, et des analyses par progiciel statistique appliquées à la nouvelle base ainsi créée. Par exemple, une étude pilote a été menée sur la distribution de près de 100 000 occurrences de pronoms personnels en fonction des variables d'âge, sexe, quartier d'habitation, statut socio-économique et autre des locuteurs (Deshaies 1986).

Le développement de LEMMATISEUR a été influencé par un autre utilisateur important: le projet Nag Hammadi sous la direction de Paul-Hubert Poirier de l'Université Laval. Il s'agit de traiter des textes en alphabet non-romain, et en langue copte, qui possède une morphologie de langue sémitique. De plus, l'analyse ne peut pas se satisfaire de la limite de dix zones d'étiquetage possible pour chaque forme textuelle, zones qui étaient prévues pour les lemmes. La solution adoptée est celle de l'utilisation de zones de lemmatisation complémentaire de niveau lexical, dont le nombre peut atteindre 20. Cette application démontre que le logiciel est suffisamment général pour servir dans une grande variété de contextes de travail.

Cette conclusion est confirmée par le grand nombre de travaux d'étudiants et étudiantes de deuxième et de troisième cycles qui ont utilisé LEMMATISEUR. Certaines des applications impliquent l'analyse détaillée de chaque occurrence d'une même forme à travers le corpus. Dans ce cas, le concept de lemme pour la forme, limité par le logiciel au nombre de 10, ne convient pas. Par contre, le zone de lemmatisation complémentaire de type contextuel permet d'assigner 20 caractères d'information à chaque occurrence, sans limite quant à la variété des valeurs.

Mentionnons une autre application en guise d'exemple de ce à quoi peut servir l'association d'un dictionnaire et d'un texte à l'intérieur d'une même base de données. Nous avons obtenu des indices de différents types de difficulté que présentent les mots des manuels DOS. Ces indices font partie d'un fichier du logiciel DBASE III, comme les fichiers de LEMMATISEUR. Nous pouvons alors faire un fichier dictionnaire avec les indices de difficulté et l'appliquer à d'autres textes semblables aux manuels DOS. De cette façon, avec un minimum de travail pour les formes ambiguës et les mots qui ne sont pas communs aux deux textes, on peut faire le profil de difficulté du nouveau texte.

LE LOGICIEL S Y R E X

Le logiciel SYREX (*sy*stème de *r*econnaissance des *ex*pressions) a été développé expressément pour traiter les mots complexes. Comme LEMMATISEUR, il fonctionne en mode automatique et en mode interactif. En mode automatique le logiciel effectue la segmentation d'un texte en mots complexes d'après le contenu du dictionnaire qui fait partie du système. En mode interactif, l'utilisateur prend des décisions en parcourant les occurrences textuelles, décisions qui sont reflétées par l'ajout de mots complexes dans la base textuelle.

Le fonctionnement en mode interactif est extrêmement simple pour la personne qui utilise le logiciel. Elle choisit parmi les modalités de parcours offertes celle qui convient: tous les mots, les mots simples, les mots complexes, les mots ambigus, les mots non vérifiés, les mots contenant telle mot-forme, etc. Ensuite, elle visionne à l'écran chaque mot dans sa forme courante et entouré des mots de contexte également dans leur forme courante. Par des touches simples, elle regroupe les mots à sa guise pour obtenir de nouvelles formes courantes. Elle peut reculer ou avancer dans le contexte afin de refaire des séquences avoisinantes. S'il y a une coquille dans le texte, elle peut le corriger en supprimant et en rajoutant des occurrences textuelles. Tous les changements sont intégrés à la base textuelle, et tous les mots complexes nouvellement créés sont ajoutés au dictionnaire.

Avec SYREX, nous visons le développement d'un outil polyvalent. La préoccupation primordiale était celle de faciliter le pré traitement des mots complexes lors de la création de bases de données textuelles.

Une des applications les plus évidentes du logiciel se situe en lexicographie. Le logiciel constitue un outil pour la capture des éléments lexicaux dans les textes. Nous savons que les dictionnaires généraux ne sont pas complets, et pour combler cette lacune, il faut relever des exemples d'emploi réels. Le traitement systématique de bases textuelles devra constituer une source précieuse de données sur les mots complexes pour les lexicographes.

En terminologie, le logiciel présente un intérêt analogue. Dans des domaines de pointe en technologie, la production terminologique distance toujours les relevés terminologiques. En traitant les textes à mesure qu'ils sont produits, il sera possible de suivre de plus près les innovations.

Enfin, le logiciel est prometteur comme outil pédagogique. Nous prévoyons l'adapter pour utilisation dans le cadre des travaux des étudiants en lexicologie et en terminologie.

LES BASES TEXTUELLES COMME BASES DE CONNAISSANCES

Les logiciels SYREX et LEMMATISEUR constituent des modules dans une constellation de logiciels potentiels pour le traitement des bases textuelles au niveau des mots. Nous pouvons combiner le pré traitement des mots complexes avec ceux que nous obtenons à l'aide de LEMMATISEUR. Ensuite, il incombe à ceux qui utilisent des bases textuelles de mettre à l'épreuve des bases sélectivement enrichies comme point de départ de leurs systèmes. Par exemple, les analyseurs syntaxiques prenant comme entrée des textes exempts de problèmes lexicaux pourraient utiliser des algorithmes plus simples, plus rapides et moins exigeants en espace de travail. Les systèmes de traduction assistée bénéficieraient également d'une réduction des alternatives en tablant sur la reconnaissance préalable des termes syntagmatiques (mots complexes).

En documentation automatique, il est clair qu'une base enrichie en vue d'un type d'interrogation aurait un effet favorable sur le rendement, tant du point de vue du bruit que du silence. En général le pré traitement, par voie de classificateurs ou descripteurs, cause une perte

de généralité du système, en introduisant des éléments qui ne peuvent pas anticiper toutes les interrogations possibles. Nous croyons que si le pré traitement sert à marquer non pas les notions en tant que telles, mais plutôt les relations linguistiques par lesquelles les notions sont médiatisées, nous évitons le problème du pré traitement arbitraire. La communication auteur/lecteur se fonde sur le partage d'un système linguistique et cognitif, et en autant que nous nous limitons à rendre explicites les relations nécessaires à cette communication, nous ne pouvons pas nuire. Nous retouchons le code écrit, et non le message. Il reste aux recherches en documentation automatique à déterminer le rôle optimal du pré traitement et à conclure quant à l'économie globale de son emploi.

CONCLUSIONS

Nos travaux se situent dans un créneau prometteur. Le développement du matériel micro-informatique rend accessible des traitements réservés autrefois à des installations centrales. De même, le développement de progiciels généraux et de langages informatiques de haut niveau facilite le traitement du langage humain. Par exemple, LEMMATISEUR a été développé comme application de dBASE III.

Nous participons au développement du créneau sur trois plans. D'abord, nous contribuons à l'élaboration de programmes pour le traitement des bases de données textuelles. Deuxièmement, nous alimentons des bases de connaissances lexicales par la capture d'information sur les mots, information qui est complémentaire aux programmes dans les logiciels. Et finalement, nous produisons des bases de données textuelles enrichies qui deviennent le point de départ d'autres travaux.

En guise de conclusion, les textes étant encore la forme dominante de représentation des connaissances humaines, toute approche qui facilite le traitement des textes par ordinateur commande l'intérêt.

Bibliographie

ARENS, Y. 1981. Using Language and Context in the Analysis of Text. *Proc. IJCAI.*

BOBROW, D.G. and T. WINOGRAD. 1977. An Overview of KRL, a Knowledge Representation Language. *Cognitive Science* 1:3-46.

CHOUEKA, Yaacov. 1988. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. Conférence RIAO 88, MIT, mars 22, 1988. Cambridge, Mass.

DAHLGREN, K. 1988. *Naive Semantics for Natural Language Understanding.* Boston: Kluwer Academic Publishers.

DESHAIES, D. 1981. *Le français parle dans la ville de Québec: une étude sociolinguistique.* Québec: CIRB, Université Laval.

DESHAIES, D. 1986. L'homogène et l'hététogène dans le langage: analyse d'un corpus recueilli auprès d'adolescents et d'adultes francophones de la ville de Québec. In C. Bureau, dir., *Cinq études sur la langue orale d'enfants, d'adolescents et d'adultes francophones de la région de Québec.* Hamburg: Helmut Buske Verlag.

DIK, Simon C. 1986. Linguistically Motivated Knowledge Representation: Working Papers on Functional Grammar 9. Amsterdam.

FILLMORE, C. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica* 6.2:222-254.

GROSS, M. 1976. *Méthodes en syntaxe*, Paris: Hermann.

HAYES, P.J. 1977. On Semantic Nets, Frames and Associations. *Proc. IJCAI.* 99-107.

HIRST, G. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge, England: Cambridge University Press.

JACKENDOFF, R. 1983. *Semantics and Cognition.* Cambridge, Mass.: MIT Press.

KATZ, J.J. 1972. *Semantic Theory.* New York: Harper and Row.

MATURANA, H.R. and F.J. VARELA. 1980. *Autopoiesis and Cognition: The Realization of the Living.* Dordrecht, Holland: D.Reidel.

MEPHAM, M. and R. BÉRUBÉ. 1987. *Le logiciel LEMMATISEUR: guide d'utilisation.* Québec: CIRB, Université Laval.

MINSKY, M. 1975. A Framework for Representing Knowledge. In P. Winston, ed., *The Psychology of Computer Vision*, New York: McGraw-Hill.

REICHMAN, R. 1985. *Getting Computers to Talk Like You and Me.* Cambridge, Mass.: MIT Press.

SANFORD, J.A. and S. GARROD. 1981. *Understanding Written Language: Explorations of Comprehension.* New York: Wiley Press.

SCHANK, R.C. and R.P. ABELSON. 1977. *Scripts. Plans. Goals and Understanding.* Hillsdale, N.J.: Erlbaum.

SCHANK, R. and C. RIESBECK. 1981. *Inside Computer Understanding.* Hillsdale, N.J.: Erlbaum.

SIMMONS, R.F. 1973. Semantic Networks: Their Computation and Use for Understanding English Sentences. In R.C. Schank and K.M. Colby, eds., *Computer Models of Thought and Language.* San Francisco: Freeman.

SOWA, J.F. 1984. *Conceptual Structures: Information Processing in Mind and Machine.* Reading, Mass.: Addison-Wesley.

WILKS, Y. 1975. Preference Semantics. In E. Keenan, ed., *Formal Semantics of Natural Language,* Cambridge, England: Cambridge University Press.

WINOGRAD, T. and F. FLORES. 1986. *Understanding Computers and Cognition: A New Foundation for Design.* Reading, Mass.: Addison-Wesley.

LA TRANSCRIPTION DE CORPUS ORAUX DANS UNE PERSPECTIVE COMPARATIVE LA DÉMARCHE DU PROJET P L U R A L*

Michel Francard
Université Laval

Louise Péronnet
Université de Moncton

0. Le développement des recherches sociolinguistiques au départ de corpus oraux a eu comme effet bénéfique de susciter une réflexion approfondie sur les problèmes méthodologiques liés à la constitution de ces corpus, notamment en ce qui concerne leur transcription. Diverses contributions récentes (BLANCHE-BENVENISTE & JEANJEAN 1986; WELKE 1986; THIBAULT & VINCENT 1988) soulignent que cette opération est une *transformation* nécessitant diverses démarches d'analyse et d'interprétation.

Ce codage passant par une notation forcément sélective - parmi la masse des données observées, seule une partie d'entre elles sera retenue par le(s) transcripteur(s), en fonction des objectifs de recherche poursuivis - se pose le problème du choix des conventions de transcription. Longtemps appréhendées comme des problèmes techniques non pertinents pour le linguiste (WELKE 1986:195; BLANCHE-BENVENISTE & JEANJEAN 1986:93), les difficultés de transcription constituent actuellement un thème majeur de réflexion et de discussion.

A tel point que toute présentation d'un corpus oral s'accompagne aujourd'hui d'un exposé détaillé non seulement des conventions de transcription, mais surtout des principes généraux qui ont présidé à cette opération. La constitution d'un corpus oral étant subordonnée aux objectifs des chercheurs, on ne s'étonnera pas que les opérations de transcription, réputées à juste titre pour être les plus longues et les plus ardues, soient elles-mêmes gouvernées par ces mêmes objectifs.

On peut donc s'attendre à de nombreuses convergences dans les protocoles de transcription régissant des corpus destinés à des exploitations similaires. Certains corpus, de constitution récente, le prouvent en effet. Mais une comparaison précise fait apparaître quelques différences essentiellement dues à ce que les responsables des corpus ont bâti une logique interne valant pour leurs trancriptions, sans se soucier explicitement d'une perspective comparative (entre corpus).

A l'heure où l'ensemble du domaine francophone se préoccupe des variétés effectivement attestées et que des voix plaident même pour une confrontation panromane des méthodes et des résultats des enquêtes orales,[1] il nous a paru intéressant de communiquer les bases d'un protocole de transcription conçu pour un projet impliquant deux régions de la francophonie géographiquement éloignées : l'Acadie et la Wallonie. Il s'agit du projet PLURAL (Plurilinguisme et Attitudes linguistiques), visant à étudier les attitudes de locuteurs se trouvant confrontés à des situations

---

*Ce texte a bénéficié des commentaires et des critiques de Nathalie DUBOIS, Christine FONTAINE et Marc Van CAMPENHOUDT. Nous les en remercions vivement.

[1] Citons celle de Cl. BLANCHE-BENVENISTE: « Les enquêtes orales sur les langues romanes ont beaucoup à échanger: pour les techniques d'enquêtes, les objectifs envisagés, les types d'analyse adoptés. Mais surtout, elles ont à échanger sur le contenu linguistique: il paraît déraisonnable d'étudier certains phénomènes de morpho-syntaxe ou de syntaxe dans une de ces langues isolément; certains faits de grammaire ou de pratique discursive ne peuvent se comprendre que dans un ensemble roman. » (BLANCHE-BENVENISTE 1985:292).

de langues en contact, plus particulièrement les formes variées d'insécurité linguistique que ces locuteurs manifestent. Ce projet associe actuellement le Centre de Recherche en Linguistique Appliquée (C.R.L.A.) de l'Université de Moncton et le Groupe de Recherche VALIBEL (Variétés linguistiques du français de Belgique) de l'Université de Louvain-la-Neuve.

Dans les pages qui suivent seront exposés les principes qui ont guidé la mise au point du code de transcription, ainsi que les conventions les plus importantes.

1. PRINCIPES DE BASE

1.1 Une perspective comparative

La perspective d'une comparaison au départ des variétés du français de Wallonie (ci-après fr.w.) et celles du français d'Acadie (ci-après fr.a.) a bien évidemment déterminé prioritairement nos conventions de transcription. Et cela dans deux directions parfois contradictoires. D'une part le souci d'une grande fidélité à la variété régionale, par la notation d'un maximum de traits différentiels, d'autre part une nécessaire standardisation qui permette, au minimum, un code unique de transcription.

Il n'y aura donc pas de standardisation des variantes qui ferait disparaître la majorité des marques régionales et/ou socioculturelles. Ce choix posera évidemment divers problèmes de consultation. Il conviendra en effet que les concordances soient exploitées en tenant compte de ces variations, notamment pour regrouper les formes proches et pour distinguer les homonymes (voir plus loin). Il nous paraît que ces inconvénients peuvent être palliés par l'application de certaines règles d'équivalence (phonétiques ou autres) et qu'ils sont mineurs par rapport à celui qu'entraînerait une standardisation maximale où seraient gommées bon nombre de variations.

Nous sommes conscients de ce que notre souci de « réalisme » dans la transcription des variantes régionales nous oblige à recourir à des «trucages orthographiques» dont certains effets pervers doivent être dénoncés.[2] Mais notre optique est bien d'enregistrer des variétés attestées de français et non de renforcer, par des formes reconstruites, le mythe d'un français standard.[3]

Ce choix nous permet en outre de limiter les pré-analyses, qu'elles soient phonétiques, morphophonologiques ou syntaxiques. Lorsque celles-ci s'avéraient néanmoins indispensables, nous avons opéré une évaluation en terme de «marqué-non marqué». Généralement, la variante « standard » est considérée comme non marquée par rapport à la variante non standard, à la

---

[2] Cf. BLANCHE-BENVENISTE & C. JEANJEAN (1986:130) rappellent que, dès 1977, le GARS prônait d'éviter des « bâtards phonético-orthographiques » (selon une expression de Raingeard & Lorscheider). Constater que certains trucages orthographiques en vigueur dans des transcriptions de corpus sont précisément ceux utilisés par certains auteurs pour rendre le français parlé (en visant une exactitude de prononciation ou un effet littéraire) n'implique nullement, à notre avis, que l'effet de dévalorisation constaté lorsque ces illustrations orales sont intégrées dans un texte littéraire affectera également un corpus oral retranscrit, dont les conditions de production et -surtout- de réception sont très différentes. On signalera en outre que l'introduction d'illustrations empruntées à la parlure quotidienne n'entraîne pas automatiquement leur dévalorisation. C'est, à l'inverse, un souci de valorisation du parler franco-acadien qui anime A. Maillet dans certains de ses romans (voir la revue Présence francophone 31 (1987) et 32 (1988)) ou encore certains auteurs wallons comme A. MASSON, A. REMY ou L. WARNANT (même si l'on peut s'interroger sur l'ambiguïté d'une certaine folklorisation).

[3] Comme le soulignent THIBAULT & VINCENT (1988:24), «standardiser dans le contexte de la transcription de l'oral, ne signifie donc en aucun cas, corriger la langue des locuteurs pour la rendre conforme à une norme». Si les intentions des linguistes-transcripteurs ne sont pas à mettre en doute, on peut néanmoins s'interroger sur l'effet produit lors de la réception du texte oral, lorsqu'on a affaire à des «lecteurs non avertis». Pour ce qui concerne le projet PLURAL, nous comptons diffuser les corpus retranscrits en dehors du cercle des linguistes.

condition expresse qu'elle soit attestée dans le corpus. Dans le cas de deux variantes non standard, l'évaluation en termes de marque tiendra compte de la situation observée dans d'autres aires de la francophonie.

1.2. La lisibilité du texte

Dans une perspective maximaliste, la logique de ce qui est exposé en 1.1. aurait pu nous mener à adopter une transcription fine, c'est-à-dire une transcription phonétique. Nous ne reviendrons pas ici en détail sur les divers arguements qui plaident en faveur d'une transcription proche de l'orthographe standard (voir BLANCHE-BENVENISTE & JEANJEAN 1986:115 sv.). Pour ce qui nous concerne, la nécessité d'une exploitation informatisée du corpus rend actuellement impossible le recours à l'API; et notre volonté d'une large diffusion des textes exige qu'une transcription phonétique soit accompagnée d'une «traduction» en orthographe conventionnelle.

Nous avons donc respecté, autant que faire se peut, les principes de l'orthographe conventionnelle, que ce soit pour les mots isolés ou pour les séquences de mots. Pas question de jouer aux émules de Zazie avec des *doukipudonktan*! Nous avons également eu recours à des signes graphiques conventionnels déjà utilisés par d'autres transcripteurs (l'apostrophe; les majuscules; le trait d'union; l'espace; la barre oblique; la barre verticale; les parenthèses, rondes ou angulaires).

Nous avons suivi la tendance générale dans les transcriptions antérieures, qui est d'éviter de multiplier les signes, par souci de lisibilité du texte (et d'économie). Par contre, nous avons, dans le même souci de lisibilité, évité d'attribuer plusieurs fonctions au même signe graphique.[4]

Au passage, signalons que les mots d'emprunt seront orthographiés selon la norme en usage dans la langue dont ils proviennent (avec une indication sur leur prononciation effective dans l'idiolecte de l'informateur). Cela vaut notamment pour les emprunts à l'anglais en fr.a. ou les emprunts au wallon (transcrits dans ce cas en orthographe Feller) dans le fr.w.[5]

1.3. L'analyse de l'interaction

Nous avons rappelé plus haut que la transcription d'un corpus est fonction des utilisations prévues. Les phénomènes interactionnels, particulièrement importants dans le cadre d'une recherche portant sur les attitudes, retiennent tout naturellement notre attention: tours de parole, chevauchements, «back channel», etc. Notre transcription tentera d'en préciser les manifestations, non seulement au plan linguistique, mais également dans le langage non verbal (gestuel, proxémique).

---

[4] Sur ce point, nous nous écartons de la position de THIBAULT & VINCENT (1988:26), lesquelles attribuent plusieurs fonctions au même signe graphique.

[5] Par contre, les conventions orthographiques adoptées par certains dictionnaires des parlers régionaux ou par les ouvrages de littérature régionale nous semblent être des pièges à éviter, les objectifs des linguistes n'étant pas nécessairement compatibles avec ceux des lexicographes amateurs ou des auteurs littéraires. Ainsi, des graphies comme genses chez S. Poplack, coudon, criss chez THIBAULT & VINCENT, ou coudons chez BEAUCHEMIN, même si elles s'appuient sur une notation attestée dans un dictionnaire ou un ouvrage littéraire, ne sont pas satisfaisantes parce qu'elles viennent brouiller un système de transcription par ailleurs beaucoup plus systématique et fonctionnel.

Dans la même perspective, toute une série de traits caractéristiques de l'oral, généralement appréhendés comme des actes « manqués » -- répétitions, interruptions, ruptures syntaxiques, hésitations, etc. -- seront considérés comme des «signes conversationnels»[6] (voir WELKE 1986:208) et seront transcrits (p. ex. on notera *euh* plutôt qu'une didascalie telle que 'hésitation'; ou encore *ouille* plutôt que 'cri de douleur').

Chaque texte sera en outre accompagné d'une fiche d'identification précisant le profil socio-culturel des informateurs, celui de l'enquêteur, ainsi que le contexte général de l'entrevue.

1.4. Un corpus informatisé

Les dimensions du corpus réuni dans le projet PLURAL nécessiteront le recours au traitement informatique des données pour diverses exploitations: concordance, analyse de contenu, etc.

Cette caractéristique, partagée par l'ensemble des grands corpus récents (dans les universités de Montréal, Laval, Sherbrooke, Québec, Ottawa, St Mary's à Halifax, ainsi que le corpus du groupe de recherche ontarien CREFO) est évidemment une source importante de consensus. Tous ces corpus ont été traités avec un même programme de concordance, l'OCP (Oxford Concordance Program), lequel impose certaines conventions (utilisation des parenthèses pour les commentaires non pris en compte dans la concordance; utilisation du trait d'union pour les lexies, etc.). Sans vouloir assujettir notre démarche de transcription au fonctionnement d'un logiciel particulier, nous n'adopterons pas de convention qui entraverait l'utilisation de tel logiciel particulièrement répandu auprès des chercheurs.

Le développement de corpus oraux montre que les corpus de grande taille, pour des raisons évidentes (de temps et d'efficacité dans le traitement informatisé notamment), ne peuvent faire l'objet d'une transcription fine et, en conséquence, font l'impasse sur un grand nombre de phénomènes phonétiques, particulièrement dans le domaine suprasegmental. Nous adoptons ce point de vue (voir 1.2.) tout en permettant l'accès aux enregistrements sonores pour qui voudrait travailler ces phénomènes sur base de micro-corpus.

1.5. L'oralité du corpus

Les (socio)linguistes confrontés aux pratiques de transcription des corpus oraux insistent sur la nécessité de distinguer ce travail de celui qu'accomplissent certains folkloristes ou même certains écrivains,[7] dont les pratiques aboutissent quelquefois à un texte assez éloigné du support oral de départ.

---

[6] La forme graphique des interjections et des onomatopées sera celle consignée dans les dictionnaires standard. En introduisant ces phénomènes dans la transcription (et non entre parenthèses, où ils échapperaient à la concordance), nous refusons -à la différence de THIBAULT & VINCENT 1988:28- d'opérer une distinction a priori entre le « back channel » et le tour de parole « véritable ».

[7] Cf. la distinction « texte de littérature orale-document linguistique » opérée par BLANCHE-BENVENISTE & JEANJEAN (1986:162 sv.); voir aussi WRENN (1986:22), à propos de la transposition à l'écrit du dialecte oral franco-acadien dans le livre La sagouine d'A. Maillet: « Il s'agit non d'une transcription -la représentation fidèle de toute manifestation de tout trait- mais d'une représentation sélectionnée de façon à recréer l'effet général de la réalité que l'auteur cherche à traduire (...) La manipulation du transcodage confère au langage une fonction non plus référentielle, mais poétique. »

Les conventions de transcription peuvent donc préserver ou au contraire gommer l'oralité du corpus de départ. Certains des principes déjà énoncés ci-dessus s'inscrivent dans le souci d'une mise en valeur de l'oralité des corpus retranscrits: la non-standardisation des variantes et l'attention aux formes linguistiques de l'interaction.

Il en va de même pour notre choix d'une « ponctuation » substituant aux signes traditionnels (le point, la virgule) l'usage des barres obliques. Nous n'utilisons pas la ponctuation de l'écrit, non seulement parce que celle-ci est inadéquate pour l'oral, mais parce qu'elle peut entretenir, bien plus que l'orthographe standard p. ex., l'illusion que l'écrit et l'oral «fonctionnent» de la même manière et renforcer ainsi le stéréotype d'un oral qui ne serait que l'écrit de seconde zone.

On nous objectera que ce choix entraîne une difficulté de lecture pour le consultant non familiarisé avec cette convention. Pour nous, il convient que le « lecteur », face à un corpus oral retranscrit, adopte un autre décodage que celui qui prévaut pour le texte écrit et dont la « ponctuation » est, de manière évidente, un signal. En d'autres termes, la valorisation d'une linguistique attentive aux productions orales passe, selon nous, par la prise de conscience d'une distance tangible entre le texte écrit et l'oral transcrit.

2. QUELQUES CONVENTIONS DE TRANSCRIPTION

2.1. L'apostrophe

Nous employons l'apostrophe dans la logique de son utilisation conventionnelle où elle sert à marquer une élision. Toutefois nous étendons cette convention à l'ensemble des segments, vocaliques et consonantiques.

Exemples: *t' aurais dû venir; v's auriez pu attendre.*

L'apostrophe permettra, dans plusieurs cas, d'éviter les homonymes.

Exemple 1: la forme *a* (verbe avoir) sera distinguée du *a'* (pr. personnel fém. en fr.a.) où l'apostrophe fait référence au *L* élidé devant consonne.

Exemple 2: en fr.a., le pronom LUI inaccentue sera transcrit *'i*, avec une apostrophe qui fait référence au *LUI* initial qui, s'il n'est pas attesté en fr.a., l'est en fr.w. (et à la forme accentuée en fr.a.). Cela permettra en outre d'eviter l'homonymie avec le *i* (IL sujet devant consonne).[8]

L'apostrophe peut également contribuer à l'identification de l'unité réduite (en signalant explicitement qu'un segment a été élidé).

[8] Ainsi qu'il apparaît dans cette illustration, nous considérons que la forme i (pour IL, devant consonne) est la forme de base (et non une réduction de il). Cette position est justifiée par les occurrences largement répandues de cette forme en francophonie, qui nous la font considérer comme forme non marquée, à la différence de 'i (LUI)

Exemples: *p't-être* (pour *peut-être*); *'steûre* (pour *asteûre*).

On ne notera pas dans la transcription, par l'usage de l'apostrophe, la chute d'un (ou de plusieurs) segment(s) qui est prédictible suite à des règles générales dont il faudra établir l'inventaire. On songe ici particulièrement aux règles de phonétique combinatoire, telle la réduction des groupes consonantiques en finale de mot.

Exemples: en fr.w., *fEnêtre* est à « lire » [foenɛt]; *prendre* [prât]; *couple* [kup], etc.

2.2. Les majuscules

Les majuscules sont utilisées pour marquer l'actualisation d'un segment latent (à l'intérieur du mot ou dans sa finale), propre à certains informateurs, mais non partagée par la majorité des locuteurs au sein de la même variété.

Exemples: le *h* initial aspire en certaines régions de Wallonie, plus particulièrement par des témoins âgés. On opposera donc *Haie* et *haie*; *quanD même* [kât] et *quand même*; *tandiS-que* et *tandis-que*, etc.

La même convention servira à marquer des variations entre les variétés impliquées.

Exemple: on distinguera *alphabeT* (fr.a.) et *alphabet* (fr.w.); *touT* (indéfini masculin en fr.a.) et *tout* (fr.w.).[9]

Dans certains cas, il peut y avoir hésitation entre l'utilisation de l'apostrophe et celle de la majuscule. Le mot MARS (nom de mois), prononcé [mars] en fr.w. mais [ma:r] en fr.a. pourrait être transcrit *marS* (fr.w.) vs *mars* (fr.a.) ou *mars* (fr.w.) vs *mar'* (fr.a.). La forme marquée étant dans ce cas celle attestée en fr.a., nous choisirons la seconde solution (qui désigne le cas effectivement marqué). Parallèlement, pour le mot NERF, au lieu d'opposer *ner'* (fr.w.) à *narf* (fr.a.), on distinguera *nerf* [nɛr] (fr.w.) et *narF* [narf] (fr.a.).

L'usage de la majuscule permettra également une notation non ambiguë des *e* caducs. Ceux-ci seront toujours transcrits (pour éviter une multiplication des apostrophes), mais leur réalisation phonétique effective sera indiquée au moyen de la majuscule.

Exemple: *pElouse*; comp.: *jE me dEmande* [zoe m doemãd] et *je mE demande* [zmoe dmâd].[10]

---

[9] Cette convention nous paraît préférable à l'adjonction d'un -e final (cf. la forme toute chez Poplack) qui introduit une forme féminine arbitraire et non fonctionnelle. Ou encore préférable à l'introduction d'une apostrophe (cf. tout' chez Thibault & Vincent) qui, dans son utilisation standard, désigne un son manquant (plutôt que l'actualisation d'un son latent).

[10] Ce système de notation ne postule donc pas, pour être correctement interprété, la connaissance préalable des règles de fonctionnement du e caduc dans une variété donnée, ou l'existence de « normes de diction » communes aux locuteurs et connues par eux (telles que les postule Cl. LEROY 1985:13).

On aura également recours à la majuscule là où un segment se maintient dans certaines productions en dépit de règles générales prévoyant son amuïssement (voir 2.1 in fine).

Exemple: *terribLE* [tɛri:bloe] vs *terrible* [tɛri:p].

2.3. Le trait d'union et l'espace

L'utilisation de ces deux symboles est largement tributaire des conventions standard, lorsque celles-ci n'entrent pas en conflit avec les contraintes de certains logiciels d'exploitation (tel l'OCP). Ainsi les lexies que le linguiste veut analyser comme une seule unité seront réunies par le trait d'union.

Exemples: *pourquoi-que tu dis ça? est-ce-que tu viens? tout-à-fait; à travers, il-y-a, parce-que*, etc.

A l'inverse, les unités que le linguiste veut analyser séparément (p. ex. pour isoler le pr. sujet du verbe) seront séparées par un espace.

Exemple: *j' arrive; penses - tu? n' y songe pas*

L'espace servira aussi à mettre en évidence certaines épenthèses consonantiques (tels les cas de liaison non prédictibles au vu de la consonne finale prévocalique).

Exemples: *un gros t arore* (fr.a.) (vs *les gros arbres*); *cent z ans*;

On étendra cette convention aux cas de « liaison différée » (p. ex.: *il est suffisamment / t aisé*) et aux épenthèses consonantiques après finale vocalique (p. ex.: *qui z ont*; *devra t être*).

2.4. La barre oblique

Nous avons choisi[11] de marquer les pauses qui interrompent le continuum sonore au moyen de la barre oblique (avec une simplification qui oppose la pause brève -une seule barre oblique- et la pause longue -deux barres obliques-). La barre oblique, dans cette utilisation, est immédiatement précédée et suivie d'un espace typographique.

[11] Le GARS a fait un choix différent en n'indiquant aucun signe de ponctuation. La ponctuation étant « un système de représentation des articulations du discours qu'on ne peut utiliser correctement qu'après avoir étudié les articulations spécifiques dans chaque langue » (HAZAEL-MASSIAUX 1985:273), ce choix se comprend aisément: « mieux vaut ne pas trancher trop tôt, en suggérant une analyse avant de l'avoir faite » (BLANCHE-BENVENISTE & JEANJEAN 1986:142). Pour nous qui utilisons la barre oblique pour marqueur d'une pause dans le continuum sonore, ce système de « ponctuation » n'est pas le reflet d'une pré-analyse syntaxique.

Cette barre oblique n'est pas un simple substitut de la ponctuation traditionnelle (une barre oblique pour la virgule; deux barres obliques pour un point). On va donc la rencontrer dans des contextes où un signe de ponctuation traditionnel ne serait pas attendu.

Exemple: *le parlement de Bruxelles s'est / réuni*

Elle assume pour l'oral des fonctions qui n'ont pas de pertinence à l'écrit. On l'emploie notamment dans des cas d'actes manqués, tels l'interruption-reprise: *la Haute-Assem/ Assemblée.* Dans ce type d'énoncés, le mot tronqué sera *immédiatement* suivi de la barre oblique (sans espace), afin de pouvoir isoler ce type d'occurrence dans la concordance.

A la fin de l'intervention d'un locuteur, nous n'utiliserons pas le symbole // (pause longue). Il ferait double emploi avec le retour à la ligne, qui indique qu'une nouvelle prise de parole a lieu (voir 2.5).

Le point d'interrogation sera conservé pour marquer les énoncés interrogatifs.

2.5. Les parenthèses

Les utilisateurs de l'OCP réserve l'usage des parenthèses (arrondies) aux notations en tous genres non destinées à être reprises dans la concordance. Nous conservons cet usage, pour faire figurer entre parenthèses

- des précisions linguistiques (remarque sur telle prononciation idiosyncrasique);
- la mention d'un passage incompréhensible, transcrit (x) (une syllabe) ou (xxx) (un groupe de syllabes);
- la délimitation - par (a) -- (z) - de portions du texte, codifiée par le transcripteur en fonction d'une recherche particulière (analyse du contenu thématique; recherche sur les alternances de codes, etc.);
- une transcription plausible, non retenue par la majorité des transcripteurs auquel le corpus a été soumis;[12]
- des renseignements sur le contexte situationnel (explication de tel bruit à tel moment de l'entretien; notes sur la gestuelle du locuteur, l'occupation de l'espace, etc.) et, plus généralement, toute précision nécessaire à la compréhension de la séquence enregistrée par quelqu'un qui n'a pas assisté personnellement à l'entretien.

A la différence de ceux qui utilisent les parenthèses angulaires pour identifier les tours de parole des divers locuteurs, nous indiquerons chaque nouvelle prise de parole par un retour a la ligne, avec désignation du locuteur concerné.

Exemple: 00223 L 3 *et t' as appris ce qui vient dE se passer?*
(= intervention du locuteur no 3 à la ligne 223).

[12] Nous n'adoptons pas le système de multi-transcription proposé notamment par le GARS (voir BLANCHE-BENVENISTE 1986:143 sv.). Cette solution s'accommode mal du traitement informatisé et nous pensons en outre, comme THIBAULT & VINCENT (1988:25), que le contexte permet dans une majorité des cas d'élucider l'ambiguïté.

Les parenthèses angulaires seront utilisées pour désigner les interventions des locuteurs impliqués dans un processus de chevauchement. Voir 2.6.

2.6. La transcription des chevauchements

Nous avons signalé plus haut notre volonté d'être attentifs aux marques de l'interaction. Parmi celles-ci, le chevauchement est une des plus intéressantes à observer. Pour en rendre compte de façon satisfaisante, nous tenterons de préciser, autant que possible, les frontières de ce chevauchement, lesquelles seront indiquées par une barre verticale.

Deux cas peuvent se présenter. Tantôt le chevauchement d'une intervention de L2 sur celle de L1 n'interrompt pas ce dernier (qui garde la parole). Dans ce cas, les deux interventions sont présentées linéairement, sans retour à la ligne, deux barres verticales indiquant la partie de l'intervention de L1 et l'intervention L2 qui sont concernées par le chevauchement, des parenthèses angulaires annonçant l'intervention de L2.

Exemple: 00042 L1 i s' en va | à la fin de l' année <L2> non je crois pas
| pour
00043 passer l' hiver en Californie

Ce qui pourrait se représenter visuellement de la manière suivante:

L1 i s' en va | à la fin de l' année | pour passer l' hiver...
<L2> | non je crois pas |

Tantôt le chevauchement est lui-même le début d'un nouveau tour de parole, qui contraint le locuteur précédent à céder la parole.[13] Dans ce cas, nous indiquons le changement de locuteur par un retour à la ligne, les barres verticales indiquant cette fois encore les frontières du chevauchement.

Exemple: 00345 L1 peut-être quE tu | pourrais venir
00346 L2 i n' en est pas question | n' y songe pas

Ce qui pourrait se visualiser de la manière suivante:

00345 L1 peut-être quE tu | pourrais venir |
00346 L2 | i n' en est pas question | n' y
songe pas

[13] Nous tentons ici de rendre compte de la distinction proposée par B.-N. GRUNIG (1986) entre auto-STOP (où le STOP est « consenti » par le locuteur qui a la parole) et hétéro-STOP (le STOP est provoqué par la prise de parole d'un interlocuteur).

2.7. Graphie des variantes régionales

Nous avons souligné plus haut notre souci de transcrire les variantes régionales au plus près de leur réalisation phonétique, tout en observant les conventions graphiques traditionnelles.

Les mots pourront donc apparaître dans une graphie proche de la prononciation. Par exemple, en fr.a.: *poumme* 'pomme'; *tarrible* 'terrible'; *histouère* 'histoire' [istwɛr]; *miroué* 'miroir' [mirwe], etc.

Les mots identifiés comme emprunts (à l'anglais, au wallon) seront écrits selon les standards orthographiques de la langue-source (voir plus haut) et seront accompagnés d'une transcription phonétique.

Exemple: (fr.a.) *c' est moi qui drive* [draiv]
(fr.w.) *elle a pris un petit pèleû* [pɛlø].

La même convention régira les mots (ou expressions) qui peuvent appartenir, sur base de leur seule forme graphique, à l'une ou l'autre langue.

Exemple: (fr.a.) *tout ça c' était top secret* [tɔp si:krit].

Il en ira de même lorsqu'il y aura intégration de l'emprunt aux schèmes de prononciation de la langue emprunteuse.

Exemple: (fr.w.) *je vais toujours voir le football* [fɔtbal] *avec le voisin.*

Les noms propres dont la prononciation est remarquable seront eux aussi accompagnés de leur transcription phonétique.

Exemple: (fr.w.) *j' habite près de Villers* [vile].

En morphologie, on notera, pour le fr.a., *i* (IL), *a'* (+ C) et *alle* (+ V) (ELLE); *i chantont* 'ils chantent', etc.[14] De même, *sèye* '(il) soit', *èye* (il) ait', etc. Pour le fr.w., on distinguera, à propos du pr. ILS devant voyelle (p.ex. ILS ONT), les transcriptions *i z ont* et *ils ont*, suivant la prononciation du locuteur.

2.8. Conventions additionnelles

Bien des possibilités graphiques restent inexploitées dans le système proposé ci-dessus et sont donc disponibles pour une transcription plus fine. Nous proposons les conventions additionnelles suivantes, compatibles avec celles qui précèdent.

[14] Nous n'opérons donc pas de « restitution » des formes standard, comme la pratiquent notamment S. Poplack, P. Thibault & D. Vincent et l'équipe du GARS. De même, en ce qui concerne l'identification des morphèmes absents (du type:« faut je parte » pour « IL faut QUE je parte »), nous ne codons pas cette absence dans la transcription (à la différence notamment de S. Poplack, P. Thibault & D. Vincent).

2.8.1. Les deux points (:) sont utilisés pour marquer un allongement de la voyelle qui précède.

Exemple: *c' est ma petite ami:e // elle est terri:bLE*

2.8.2. Le soulignement d'une syllabe (ou d'un groupe de syllabes) par une série discontinue de + marque un renforcement de l'intensité à cet endroit. L'utilisation d'une série discontinue de - indique une intensité plus faible.

Exemples: quelle horreur ce type
+ + + + + + +

i marchait sans un bruit
- - - - - - -

2.8.3. Pour marquer les variations de débit, on utilisera le soulignement continu avec des flèches de direction différente:

(a) accélération du débit: -->--->--->--
(b) ralentissement du débit: --<---<---<--

2.8.4. L'intonation sera transcrite au-dessus de la ligne concernée, en utilisant les symboles suivants:

↗ intonation ascendante
↘ intonation descendante
--- intonation plane

2.9. Divers problèmes se poseront au fur et à mesure de la transcription du corpus. Un inventaire de ceux-ci et des solutions proposées pour les résoudre sera établi par les transcripteurs et mis à la disposition des consultants de la banque de données.

3. POUR NE PAS CONCLURE

A l'heure actuelle, les industries de la langue s'appuient presque exclusivement sur le français standard. Si l'on veut aboutir à ce que les variétés linguistiques du français soient prises en compte, il faut s'atteler d'urgence à la constitution de bases de données où ces variétés seront authentifiées et illustrées.

Pour ce qui concerne plus spécifiquement l'oral, la nécessité de réunir une documentation à l'échelle de la francophonie est plus impérieuse encore. Mais cet effort, s'il ne se fonde pas sur un minimum de convergences entre les chercheurs, particulièrement au plan des conventions de transcription, risque de limiter considérablement l'exploitation de certaines bases de données. La démarche du projet PLURAL nous paraît illustrer certaines de ces convergences dans une perspective comparative. Nous espérons qu'elle suscitera de nouvelles collaborations, en vue d'aboutir à une description des usages linguistiques oraux de la francophonie.

Références bibliographiques

BLANCHE-BENVENISTE, Cl. 1985. « Etat des enquêtes sur les langues romanes parlées », dans *Contacts de langues. Discours oral.* Actes du XVII[e] congrès international de linguistique et philologie romanes (Aix-en-Provence, 29 août - 3 septembre 1983), vol. 7, Aix-en-Provence, p. 201-214.

BLANCHE-BENVENISTE, Cl. & JEANJEAN C. 1986. *Le français parlé. Transcription et édition.* Paris, Inalf - Didier érudition.

GRUNIG, B.-N. 1986. « Inachèvements ». *DRLAV* 34-35, p. 1-48.

HAZAEL-MASSIAUX, M -Ch. 1985. « Peut-on appliquer directement les règles de ponctuation des langues romanes à l'écriture des langues néo-romanes ? Problèmes de la notation des créoles et français régionaux en relation avec le français standard », dans *Contacts de langue. Discours oral.* Actes du XVII[e] congrès international de linguistique et philologie romanes (Aix-en-Provence, 29 août - 3 septembre 1983), vol.7, Aix-en-Provence, p.269-281.

LEROY, Ch. 1985. « La notation de l'oral ». *Langue française* 65, p. 6-17.

POPLACK, S. 1984. « The care and handling of a mega-corpus: the Ottawa-Hull French Project », dans R. FASOLD & D. SCHIFFRIN (dir.), *Proceedings of NWAVE-XI*, Washington DC, Georgetown University Press.

THIBAULT P. & VINCENT D. 1988. « La transcription ou la standardisation des productions orales ». *LINX* 18 (1), p. 19-32b.

WELKE, D. 1986. « La semi-interprétativité dans les transcriptions en "analyse conversationnelle" et pragmatique linguistique: travaux américains et allemands ». *DRLAV* 34, p. 195-213.

WRENN, Ph. 1985. « Le transcodage d'une parlure en texte: *La sagouine* et le mythe du dialecte ». *Francophonia* 8, p. 3-22.

TRANSLEGS
UNE STATION DE TRAVAIL LINGUISTIQUE

Yvette Mathieu
LADL

TRANSLEGS est une station de travail linguistique qui permet:

- l'étude du lexique-grammaire d'une langue
- la comparaison de 2 ou plusieurs lexiques-grammaires de langues différentes.

La réalisation présentée ici porte sur le français et l'italien. Elle repose sur les travaux linguistiques des chercheurs du LADL (Université Paris 7) et de l'Institut de Linguistique de Salerne (Université de Salerne)

LES DONNÉES LINGUISTIQUES

Les lexiques-grammaires

Leur conception repose sur la constatation qu'un grand nombre d'opérations syntaxiques d'une langue sont fortement liées à des conditions lexicales.

La construction d'un lexique-grammaire [M. Gross 1981] passe par la description syntaxique de la langue projetée sur la globalité du dictionnaire. La présentation lexico-syntaxique prend la forme de matrices qui donnent les constructions d'un item lexical (verbe, adjectif, nom, etc.) pour un ensemble donné de formes syntaxiques. Par exemple, le lexique-grammaire des verbes du français contient environ 10 000 verbes et 500 propriétés (formes et phrases) répartis en 60 tables [M. Gross 1975], [J.P. Boons, A. Guillet, C. Leclère 1976a, 1976b].

Ces tables ou classes syntaxiques sont présentées sous forme de matrices. Elles sont définies par une propriété, dite définitionnelle, qui est en général une structure de phrase simple. Les items lexicaux qui ont en commun une propriété définitionnelle sont regroupés dans une même table.

Par exemple, la table française *33* a pour propriété définitionnelle *NO V à NI*, ou *NO* et *NI* sont des substantifs et *V* un verbe.

Tous les verbes de cette table peuvent entrer dans des phrases de ce type, par exemple *mentir*:

Max ment à Léa

Par contre le verbe *flirter* n'appartient pas à cette table:

*Max flirte à Léa

(l'astérisque devant la phrase indique qu'elle n'est pas valide).
Un extrait de la table française *33* est montré figure 1.

FIGURE 1:

N_0						N_1									P.A.		
N_0 = N hum	N_0 = N -hum	N_0 = N nr	N_0 = V-n	N_0 = N plur obl		N_0 V	N_0 est V-ant	N_0 est V pp	N_0 V de N^0pc	N_1 = N hum	N_0 V Prép N pc de N_1	N_0 lui V Prép N^1pc	N_1 = N -hum	N_1 = le fait que P	Ppv = lui	Ppv = y	N_1 = V-n	N_1 = N plur obl	N hum V sur ce point	il V N_0 Ω	cette idée V loc son esprit
+	+	+	-	-	en imposer	+	+	-	+	+	-	-	-	-	+	-	-	-	+	-	+
+	+	+	-	-	insulter	-	+	-	-	+	-	-	+	-	-	-	-	-	+	-	+
+	-	-	-	-	manquer	-	-	-	-	-	-	-	+	-	-	+	-	-	+	+	+
+	-	-	-	-	mentir	+	-	-	-	+	-	-	-	-	+	-	-	-	+	+	-
+	-	-	-	-	naître	+	-	-	-	-	-	-	+	+	-	+	-	-	-	+	-
+	-	-	-	-	obéir	+	+	-	-	+	-	-	-	-	+	-	-	-	-	+	-
+	-	-	-	-	obtempérer	+	-	-	-	-	-	-	+	-	+	+	-	-	-	+	-
+	+	+	-	-	parer	-	-	-	-	-	-	-	+	+	-	+	-	-	-	+	-
+	-	+	-	-	pourvoir	-	-	-	-	-	-	-	+	-	-		-	-	+	+	-
+	+	+	-	-	présider	+	-	-	-	-	-	-	+	+	-	+	-	-	-	+	-
+	-	-	-	-	procéder		-	-	-	-	-	-	+	-	-	+	-	-	-	+	-
+	+	-	-	-	réagir	+	-	-	+	+	+	-	+	-	-	+	-	-	+	+	-
-	+	+	-	-	remonter	-	-	-		-	+	-	+	+	-	-	-	-	-	-	-
+	-	-	-	-	se rendre	+	-	-	-	+	-	-	+	+	-	+	-	-	+	+	-
+	+	+	-	-	résister	+	+	-	+	+	+	-	+	+	+	+	-	-	+	+	-

Les entrées lexicales sont en lignes, les propriétés syntaxiques en colonnes.

Un + à l'intersection d'une ligne et d'une colonne indique que l'item lexical accepte la propriété, un - qu'il ne l'accepte pas.

Comparaison italien-français

En vue de la comparaison d'une partie des lexiques-grammaires de l'italien et du français, un sous-ensemble des tables italiennes possède des informations supplémentaires: à chaque item lexical est associé l'item français correspondant et la table à laquelle appartient ce dernier [A. Elia 1984a, 1984b]. Ce sous-ensemble des tables italiennes concerne 9 tables de verbes à un complément.

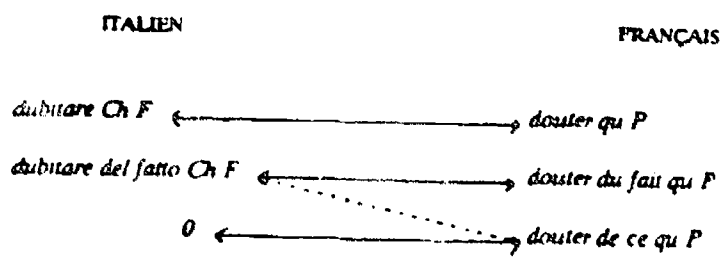
La figure 2 est un extrait de la table italienne *it42* [A. Elia 1984c].

FIGURE 2:

N_0=:Nhum	N_0=:Nnr	N_0=:ChF	N_0=:V[1]	Italien	Prep.	N_1=:Nhum	Ppv=:gli	N_1=:Nhum	N_1=: il fatto ChF	aN_1q	inN_1q	$(di+da)N_1q$	Ppv=:(ci+vi)	dal fatto ChF	N_0V	N_1div^1	$NeVN_0$	Française	Table française
-	+	-	+	accadere	a	+	+	-	-	-	-	-	-	-	-	+	+	arriver	5
-	+	+	+	addirsi	a	+	+	+	+	+	-	-	-	-		-	-	convenir	5
-	+	+	+	arriorare	loc	+	+	+	+	-	+	+	+	+	-	+	+	affleurer	35ST
+	+	+	+	aggradare	a	+	-	-	-	-	-	-	-	-	-	-	-	agréer	5
+	+	+	+	agire	su	+	-	+	-	-	-	-	-	-	-	-	+	agir	5
-	+	+	-	albergare	loc	+	-	+	-	-	+	-	+	-	-	-	+	demeurer	5
-	+	+	-	aleggiare	loc	+	+	+	+	-	+	-	+	+	-	-	+	voleter	35L
-	+	+	+	andare	a	+	+	-	-	-	-	-	+	-	-	-	-	aller	5
-	+	+	+	andare Avvm	a	+	+	+	-	-	-	-	+	-	-	-	+	aller Advm	31R
+	+	+	-	apparire	a	+	+	-	-	-	+	-	+	-	-	-	-	apparaître	17
+	+	+	-	apparire	loc	+	+	+	+	-	+	-	+	-	-	-	+	apparaître	5
+	+	+	+	appartenere	a	+	+	+	-	+	-	-	+	-	-	-	+	appartenir	5
-	+	+	-	arridere	a	+	+	+	+	+	-	-	+	-	-	-	-	favoriser	6
-	+	+	-	attecchire	loc	+	+	+	-	-	+	-	+	-	-	-	+	pousser	35ST
-	+	+	+	attenere	a	+	+	+	-	+	-	-	-	-	-	-	-	revenir	5
-	+	+	+	avvenire	a	+	+	-	-	-	-	-	-	-	+	+	-	arriver	5
-	+	+	-	balenare	loc	+	+	-	-	-	+	-	+	-	-	-	+	étinceler	34LO
+	+	+	-	baluginare	loc	-	-	-	-	-	+	-	+	-	-	-	+	apparaître	5
-	+	+	+	bastare	a	+	+	-	-	+	-	-	+	-	-	-	+	suffire	16
-	+	+	+	bisognare	a	+	+	+	-	+	-	-	+	-	+	+	+	falloir	17
-	+	+	+	bisognare	per	+	-	-	-	-	-	-	-	-	-	-	-	falloir	17
-	+	+	-	brillare	loc	+	+	+	-	-	-	-	+	-	-	-	+	briller	34LO
-	+	+		cadere Avvm	per	+	-	-	-	-	-	-	-	-	-	-	-	tomber Advm	5
-	+	+	+	calere	a	+	+	-	-	-	-		-	-	-	-	-	importer	5
-	+	+	+	capitare	a	+	+	-	-	-	-	-	-	-	+	+	-	advenir	5
-	+	+	-	circolare	loc	-	-	+	-	-	-	-	-	-	-	-	-	circuler	5
-	+	+	-	coesistere	con	-	-	+	+	-	-	-	-	-	-	-	-	coexister	35S
-	+	+	-	coincidere	con		-	+	+	-	-	-	-	-	-	-	-	coïncider	5
-	+	+	+	collimare	con	-	-	+	-	-	-	-	-	-	-	-	-	coïncider	5
-	+	+	-	combaciare	con	-	-	+	+	-	-	-	-	-	-	-	-	correspondre	35S
-	+	+	+	competere	a	+	+	-	-	-	-	-	-	-	-	-	-	convenir	5
+	+	+	-	concordare	con	+	+	+	+	-	-	-	+	-	-	-	-	concorder	5
-	+	+	-	constare	a	+	+	-	-	-	-	-	-	-	-	-	-	apparaître	17
+	+	+	+	contare	per	+	-	+	+	-	-	-	+	-	+	-	-	compter	5
-	+	+	+	contrastare	con	+	-	+	+	-	-	-	+	-	-	-	-	s'opposer	16
-	+	+	+	convenire	a	+	+	+	-	+	-	-	+	-	+	-		convenir	5
-	+	+	+	corrispondere	a	-	-	+	+	+	-	-	+	-	-	-	-	correspondre	35S
-	+	+	+	costare	a	+	+	+	-	-	-	-	+	-	-	-	-	coûter	5
-	-	+	+	degenerare	it.	-	-	+	-	-	-	-	-	-	-	-	-	dégénérer	5
-	+	+	+	dispiacere	a	+	+	-	-	-	-	-	-	-	-	-	-	déplaire	5

Des correspondances entre italien et français ont également été définies pour les propriétés et pour les classes. Ces correspondances ne sont pas bijectives. En effet une propriété ou une classe peuvent ne pas avoir de correspondant, ou en avoir plusieurs: lorsqu'une propriété n'a pas de correspondant le système doit proposer des propriétés de "substitution" formellement voisines (figure 3). Les critères de choix de ces dernières doivent pouvoir être modifiés par l'utilisateur.

FIGURE 3:



La ligne -------- indique le choix d'une équivalence par proximité formelle.

PRÉSENTATION DE TRANSLEGS

TRANSLEGS est une station de travail linguistique performante qui permet d'exploiter une partie de la richesse et de la complexité des données que nous venons de décrire. Le travail présenté ici porte sur un sous-ensemble: les verbes à un seul complément. Leur nombre total est actuellement d'environ 2 300 (verbes italiens et verbes français).

Un prototype a été réalisé en PROLOG [A. Elia, Y. Mathieu 1986]. Il a permis de dégager plusieurs impératifs:

- une représentaiton des connaissances qui soit, d'une part, apte à expliciter les relations structurelles implicites des lexiques-grammaires et qui, d'autre part, permette de connaître à tout moment quelles données sont contenues dans la base et quels opérateurs on peut leur appliquer,
- une interrogation interactive rapide et agréable, destinée aux utilisateurs qui n'ont pas à faire intervenir de connaissances en informatique,
- une administration des données aisée (mise à jour, ajout, suppression),
- une utilisation possible sur micro-ordinateur de type compatible,
- une bonne portabilité.

Le respect de ces impératifs nous a amenée à utiliser un Système de Gestion de Bases de Données (SGBDF) de type relationnel dont le langage d'accès est SQL, couplé avec des programmes écrits en langage C.

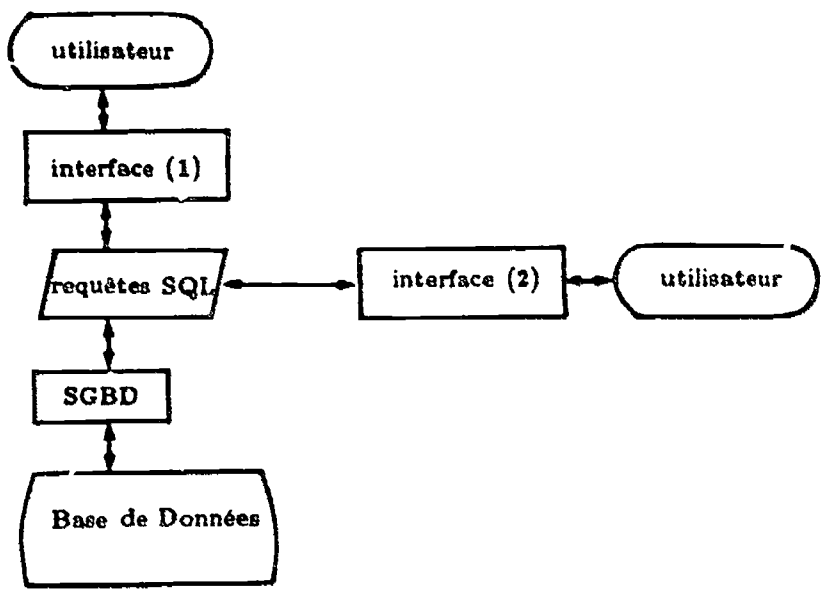
La base de données de **TRANSLEGS** comporte 15 relations.

CONCEPTION ET RÉALISATION

Architecture de TRANSLEGS

TRANSLEGS est composé de différents modules. La figure 4 illustre l'architecture.

FIGURE 4:



Consultation

L'utilisateur dialogue avec une interface (1) écrite en langage C. Le rôle de cette interface est primordial:

- elle affranchit l'utilisateur de toute syntaxe et de toute programmation,
- elle traduit les questions posées en requêtes SQL pour leur traitement par le SGBD.

Cette interface repose sur les concepts de fenêtres et de menus déroulants.

Maintenance

L'interface de maintenance (2), également écrite en langage C, sert, d'une part, au transfert des données de et vers **TRANSLEGS**. La mise à jour des données se fait au niveau des lexiques-grammaires et est répercutée dans **TRANSLEGS**.

Transfert des données

Les données intégrées dans **TRANSLEGS** sont extraites des lexiques-grammaires implémentés sur Vax 780 et Vax 730, ce sont des fichiers séquentiels de type ASCII.

Réciproquement, **TRANSLEGS** peut produire de tels fichiers, utilisables par n'importe quel logiciel sur n'importe quel système.

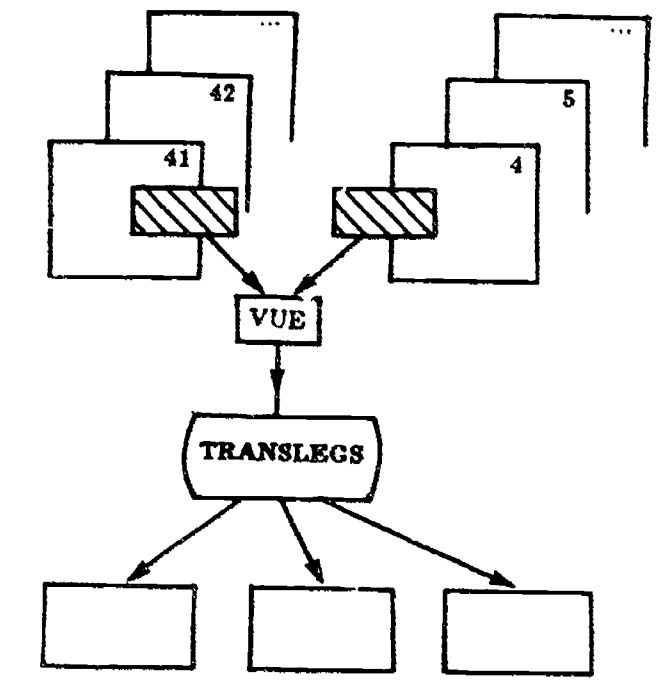
Construction de vues dynamiques

Le linguiste ne travaille pas sur une présentation figée des lexiques-grammaires: il veut étudier une vue temporaire, qui n'existe pas en tant que telle dans les données, et qu'il faut construire dynamiquement au moment de la consultation.

Prenons l'exemple (figure 5) de la comparaison d'un verbe italien et de son correspondant français. Le linguiste s'intéresse à un sous-ensemble du lexique-grammaire italien et à un sous-ensemble du lexique-grammaire français. La juxtaposition de ces deux sous-ensembles constitue une vue partielle et temporaire. **TRANSLEGS** va reconstituer cette vue à partir de ses relations et des liens qui existent entre elle. Cette reconstitution est dynamique et n'existe que le temps de la consultation.

FIGURE 5:

TABLES DES LEXIQUES-GRAMMAIRES

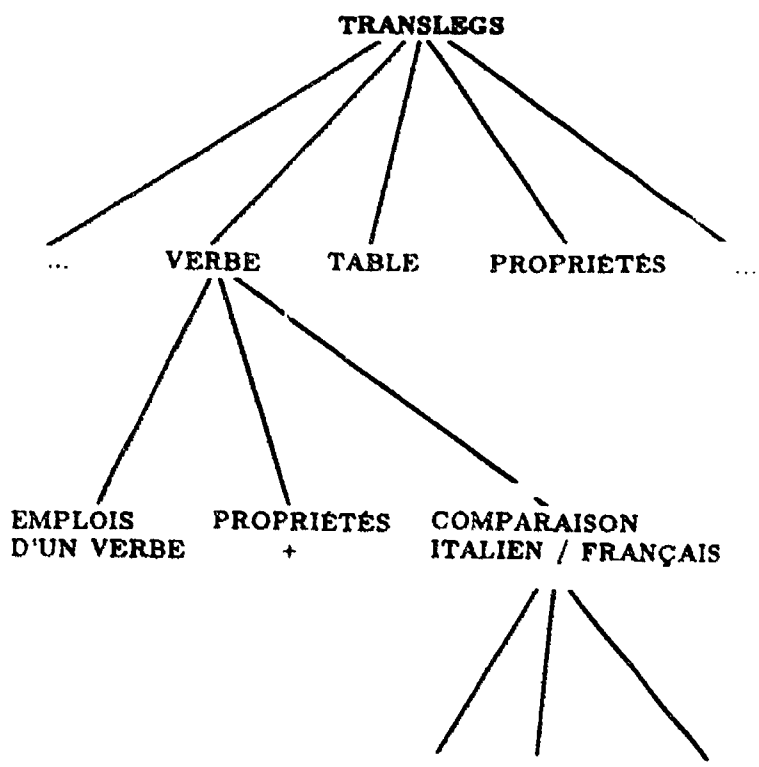


UTILISATION DE T R A N S L E G S

L'utilisation de **TRANSLEGS** repose sur les 2 concepts de menus arborescents et de fenêtres. Dans chaque écran de consultation il y a une fenêtre de dialogue qui permet:

- de se déplacer dans les fenêtres,
- de sauvegarder les résultats affichés,
- de se déplacer dans l'arborescence: A tout moment il peut descendre, remonter, se déplacer latéralement ou arrêter. Un extrait de l'arborescence est donné figure 6.

FIGURE 6:



Des exemples des consultation sont montrés en figure 7 et 8.

La figure 7 concerne la liste des verbes d'une table qui acceptent ou non une sélection de propriétés. Dans cet exemple, les verbes qui possèdent la propriété N0=: Nhum (substantif humain en position sujet) et ne possèdent pas la propriété N1=: N-hum (substantif non humain en position objet) sont affichés dans la partie droite. (Les propriétés concernées sont "marquées" par l'utilisateur dans la liste affichée à gauche).

FIGURE 7:

TABLE IT42

```
+   N0 =: Nhum
    N0 =: Nnr
    N0 =: Ch F
    N0 =: V1 Comp
    N1 =: Nhum
    Ppv =: gli
-   N1 =: N-hum
    N1 =: il fatto Ch F
    a N1q
    in N1q
    (di + da) N1q
    Ppv =: (ci + vi)
    dal fatto Ch F
    N0 V
    N1 di V1 Comp
    N1hum per Aux V1 Comp
```

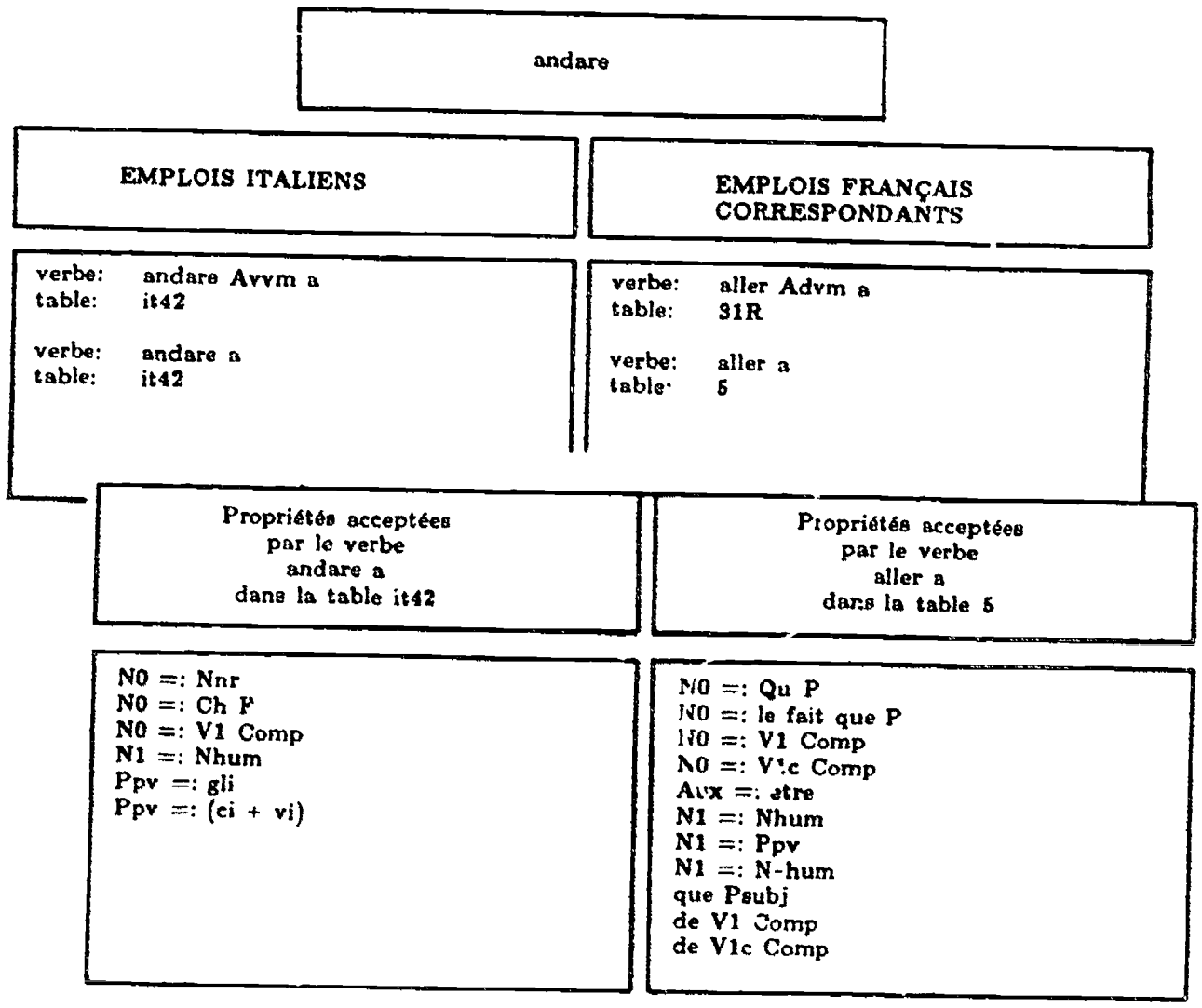
6 VERBES SÉLECTIONNÉS

```
aggradare a
apparire a
baluginare loc
piacere a
repellere a
ributtare a
```

La figure 8 concerne la comparaison italien-français.

Dans cet exemple, l'utilisateur désire étudier les propriétés acceptées par le verbe italien *andare* et par son correspondant français. Un écran lui indique quels sont les différents emplois possibles de ce verbe, ainsi que ceux des traductions françaises correspondantes. Pour obtenir les propriétés d'un emploi particulier, il suffit de le sélectionner et une autre fenêtre s'affiche.

FIGURE 8:



CONCLUSION

TRANSLEGS a été implanté sur un micro-ordinateur PC compatible. Le temps de réponse à une enquête est immédiat.

Etant donnée le nombre total de verbes (environ 2 300) et le nombre de propriétés (qui va de 20 à 36 selon les tables), le nombre d'enregistrements à traiter avoisine 55 000 et le volume actuel de la base est d'environ 7 Mega octets.

Notre premier objectif, la consultation rapide, est atteint et confirme la validité d'une station de travail dédiée à l'exploitation de données linguistiques implantées sur micro ordinateur. Les suggestions et les critiques émises par les utilisateurs de cette base de données devraient nous guider vers la nécessité (ou non):

- d'une interface plus déductive qui prendrait en compte les proximités formelles des propriétés,
- d'une interface en langue naturelle qui permettrait une plus grande convivialité entre l'utilisateur et le système [Y. Mathieu, P. Sabatier 1985].

Ce modèle n'est pas propre au français et à l'italien. Il est généralisable à d'autres langues (pour une utilisation mono, bi ou multilingue), et à d'autres items lexicaux: des études sont en cours pour des applications sur des noms, des noms composés, des expressions figées.

Bibliographie

BOONS, J-P., GUILLET, A., LECLERC, Ch., 1976a, *La structure les phrases simples en français. Constructions intransitives*. Droz, Genève, 377p.

BOONS, J-P., GUILLET, A., LECLERC, Ch., 1976a, *La structure des phrases simples en français. Classes de constructions transitives*, Rapport de recherches de L.A.D.L. n° 6, Université Paris VII, 143 p.

ELIA, A, 1984a, "Sur l'unité du mot et la syntaxe comparée des langues romanes: le morphème INVEST-italien et en français", *Revue Québecoise de linguistique*, Vol. 13, n° 2, Montréal, p.193-216.

ELIA, A, 1984b, "L'étude formelle des différents emplois sémantiques d'un mot: un exemple d'application du Lexique-Grammaire de l'italien et du français", *Cahiers de lexicographie*, Vol. XLIV, n° 1, Paris, p. 51-62.

ELIA, A, 1984c, *Le verbe italien*, Schena-Nizet, Bari-Paris, 300 p.

ELIA, A., MATHIEU, Y.,1986, "Computational Comparatives Studies on Romance language. A linguistic comparison of lexicon-grammars", in *11th International Conference en Computational Linguistics*, Coling'86, Bonn, p.146-150.

GROSS, M., 1981, "Les bases empiriques de la notion de prédicat sémantique", in *Languages* 63, Larousse, Paris, pp 7-52.

GROSS, M., 1975, *Méthodes en syntaxe*, Hermann, Paris, 414p.

Mathieu, Y., à paraître, "**TRANSLEGS**: un outil informatique pour l'étude des lexiques-grammaires", in *Actes du 7ème Colloque Européen sur la Grammaire et le Lexique Comparés des Langues Romanes* (La Croix en Touraine, 21-24 septembre 1988).

Mathieu, Y., Sabatier, P., 1986, "INTERFACILE: a Linguistic coverage and query reformulation", in *11th International Conference on Computational Linguistics*, Coling'86, Bonn, p.46-49.

LA GRAMMAIRE APPLICATIVE UNIVERSELLE

François Rousselot
Scolia

A. INTRODUCTION

Ce texte est un essai de synthèse des divers travaux existants sur la Grammaire Applicative, principalement les deux livres de Shaumyan (1977 et 1987), surtout ceux de 77. Il sera fait mention, bien entendu des travaux de Desclés nombreux (DESCLES 1987, DESCLES 1988) qui ont étendu et précisé l'approche originelle et ceux de Reb (REB 1988) et pour ce qui concerne les subordonnées, des miens (ROUSSELOT 1988). Les projets de la théorie de la Grammaire Applicative Universelle sont ambitieux et de nombreux développements sont encore à faire. Il manquait jusqu'à présent un ouvrage où les fils conducteurs soient visibles et apparents et non noyés dans des techniques de calculs. C'est ce que nous avons tenté de faire ici. L'urgence d'une telle tâche s'imposait: la Grammaire Applicative Universelle (G.A.U. dans la suite) est une théorie formalisée peu connue des linguistes et des logiciens ainsi que des spécialistes de l'Intelligence Artificielle. Or, elle se situe exactement au confluent de ces trois domaines et permet un échange fructueux entre les trois spécialités.

En effet, la G.A.U. s'appuie sur la Logique Combinatoire qui outre l'aspect calculatoire (construction des prédicats, composition, prédicats complexes) procure un environnement apte à la déduction: la Logique Combinatoire est une logique. Quant à la linguistique, la théorie linguistique de la G.A.U. vise à mettre à jour des phénomènes universels permettant d'étudier l'activité langagière i.e. «le langage» et de le ramener à des mécanismes élémentaires utilisés dans toutes les langues. Ces mécanismes appartiennent vraisemblablement au domaine cognitif.

Nous le verrons plus loin, la G.A.U. procure un cadre formel très rigoureux qui permet d'aborder l'étude de la représentation du sens au moyen de «primitives» sémantiques ainsi que la construction du système de primitives. Une telle approche est fort différente de celle bien connue en Intelligence Artificielle (I.A. dans la suite) (Schank 72). Ici la liste des primitives n'est pas fournie par l'intuition, mais induite par le système. Les représentations obtenues sont très fines et ont l'avantage d'être décrites dans un système logique où les inférences sont possibles. Ce dernier aspect ne manquera pas d'intéresser les informaticiens de l'I.A. Dernier point, la logique combinatoire constitue pratiquement un langage de programmation (voisin de LISP). Des recherches sont entreprises d'ailleurs actuellement sur des machines à combinateurs basées sur cette logique (Curien, Robinet etc..). La construction du sens d'un énoncé est très comparable, nous le verrons, à l'exécution d'un programme.

Il convient de dire, ici, que le formalisme de la G.A.U. peut avoir certains côtés techniques qui peuvent sembler arides. Pour réaliser des descriptions avec une certaine finesse, il est nécessaire de décomposer les opérateurs linguistiques en opérateurs abstraits élémentaires. On aura constamment à l'esprit deux niveaux de lecture: le niveau des calculs élémentaires (qu'on pourra qualifier de niveau micro) qui décrit les détails opératoires et un niveau plus élevé (le niveau macro) qui, lui, résumera certaines phases, correspondant à des opérations linguistiques, parfois longues. La pratique de l'enseignement à des étudiants linguistes, pendant plus d'une année, nous a montré que bien qu'un peu contraignant, le formalisme est assimilable par le linguiste qui comprend très vite son intérêt. Sa finesse permet de décrire, ou d'espérer décrire, des phénomènes linguistiques pointus comme par exemple certaines relatives particulières, tout autant que de contribuer à l'ambition centrale de la G.A.U., construire un système d'invariants et bâtir dessus une théorie constructive du sens.

Notre effort actuel est bien entendu tourné vers la réalisation d'applications programmées basées sur la théorie. Il est en effet possible d'envisager d'écrire des programmes de traitement de la langue dans ce paradigme moyennant certaines simplifications linguistiques qui sont alors très *clairement spécifiables*.

B. LA MÉTHODOLOGIE G.A.U. EN GENÉRAL

a) la méthode hypothético-déductive

Dans toutes les sciences modernes, on utilise une telle méthode qui demande la réalisation des quatre étapes suivantes:

1. détermination du problème (simplification)
2. émission d'une hypothèse pour le résoudre
3. déduction des conséquences de l'hypothèse
4. comparaison des conséquences avec les faits réels.

Dans l'approche de la G.A.U., ces quatre étapes peuvent être précisées.

1. considérer un fait linguistique important: par exemple le fait qu'on peut traduire une langue dans une autre.
   Placer l'objet dans des conditions imaginaires idéales (cf. le mouvement sans frottement) sans s'encombrer de certains facteurs. Par exemple, on le verra plus loin, on s'occupera du "sens intrinsèque" d'énoncés: sans aspect, sans détermination.

2. on formule alors des hypothèses: par exemple le fait qu' il existe des invariants langagiers.

3. déduire les conséquences des hypothèses, celles-ci ne doivent pas concerner seulement les faits initiaux, mais doivent en expliquer d'autres inconnus jusque là (pouvoir prédictif). Chaque hypothèse est à la fois un outil d'explication et un outil de prévision.

4. vérifier les hypothèses, les corriger, les rejeter au profit d'autres plus probables.

La théorie de la G.A.U. est une théorie en cours de développement. Elle possède un certain nombre d'hypothèses testées et d'autres éventuellement à trouver ou à remettre en cause. Comme dans toute théorie scientifique, il y a des possibilités de changement. De nombreux exemples dans l'histoire des Sciences montrent que des théories solides peuvent être mises en cause par de nouveaux faits ou le développement de nouveaux points de vue (gravitation de Newton et relativité d'Einstein). Notons que la mouvance des théories dans les Sciences Théoriques abstraites est plutôt signe d'essor, tandis que théorie stable est plutôt synonyme de stagnation.

A l'origine de la construction de toute théorie scientifique, il faut construire un système d'hypothèses. On construit un réseau de concepts définis simultanément, ces concepts sont les éléments, non liés directement à une science empirique, d'un système formel. Comme le dit Einstein:

"Les concepts physiques sont des créations libres de l'esprit humain et ne sont pas, bien qu'on puisse le croire, uniquement déterminés par le monde externe. Dans notre tentative de compréhension de la réalité nous ressemblons à un homme qui essaie de comprendre le fonctionnement d'une montre fermée. Il regarde le cadran et les aiguilles, il écoute même le tic-tac, mais il ne peut ouvrir le boîtier. S'il est ingénieux, il peut former une représentation du mécanisme capable d'expliquer tout ce qu'il observe, mais il ne peut jamais être tout à fait sûr que c'est la seule qui peut expliquer ces observations. Il ne sera jamais capable de la comparer au système réel et même, il ne pourra imaginer la possibilité d'une telle comparaison. Mais, il croit certainement que, au fur et à mesure que sa connaissance croît, sa représentation de la réalité devient plus simple et explique un plus grand nombre d'impressions. Il peut aussi croire en l'existence d'une limite idéale de connaissances approchables par l'esprit humain. Il peut appeler cette limite idéale "réalité objective"".

Il n'y a pas de méthode, de procédure analytique pour fabriquer un système d'hypothèses à partir de données empiriques, il y faut de l'imagination et de l'intuition.

Dans le domaine linguistique, il est important d'observer plusieurs langues, avant d'essayer de procéder par abstraction pour formuler des hypothèses universelles. On échappe ainsi à un certain ethnocentrisme, présent dans bon nombre de théories. On évite de prendre pour catégories universelles des notions induites uniquement par l'étude des langues indo-européennes seules.

b) Abduction

La méthodologie évoquée ici, fait appel au schéma d'inférence dit "d'abduction", fréquemment utilisé par Sherlok Holmes. Il peut se schématiser de la manière suivante: si j'ai une règle d'implication qui me dit A et B ----> C, que j'ai B vrai et C vrai, j'ai de bonnes raisons de croire que A est aussi vrai. Il est clair que cette méthode est à utiliser avec une très grande circonspection et que les hypothèses ainsi formulées doivent être, d'une part, "intuitivement vraies", d'autre part testables ou vérifiables assez rapidement par recoupement.

Exemple: Si une personne a trop bu, elle ne marche pas droit.

Je vois dans la rue une personne qui titube, j'en déduis immédiatement qu' elle a bu: il s'agit d'une inférence plausible. En effet, l'abus de boisson est une cause fréquente de marche non rectiligne.

L'abduction n'est pas toujours liée à la causalité. Si j'ai par exemple la connaissance suivante:

Beaucoup d'habitants du numéro vingt de la rue du Dôme sont alcooliques. Je connais monsieur Dupont, j'apprends qu'il habite à cette adresse, je peux être amené à faire l'inférence (plausible) que monsieur Dupont est alcoolique. Il s'agit ici de l'utilisation d'une constatation et non pas d'une relation de cause.

En effet, l'implication Si Alors qui vient de la logique ne recouvre pas uniquement la causalité.

Finalement, une situation qui confortera une hypothèse formulée après abduction sera l'obtention de "nouvelles" déductions rendant compte de faits nouveaux et en parfait accord avec notre intuition.

C. FONDEMENTS DE LA GAU

a) Sémantique

Après quelques définitions, nous introduirons les raisons qui ont conduit au choix du formalisme.

Un énoncé exprime une pensée complète exprimée avec des mots. Mais qu'est-ce qu'une pensée? Comparons avec la notion de prix d'un objet. Le prix, la valeur d'un objet, la longueur, est un item déterminé par rapport à celui des choses équivalentes. Si on a les moyens de mesurer d'une certaine manière la plus ou moins grande équivalence entre des biens, on peut définir la valeur de ceux-ci, à condition de se fixer en plus un système de référence (étalon).

Dans le domaine des phrases, on va devoir alors décrire la "proximité sémantique" de deux phrases: elles convoyent des messages voisins, elles sont substituables dans une situation donnée.

Pour valeur ou longueur, on peut prendre n'importe quel produit comme valeur de référence. Dans le cas qui nous occupe, on va prendre une certaine classe de phrases. On prendra les phrases qui représentent directement une pensée:

exemple: la phrase de base de "le chien mange un chat" est "chien mange chat".

Il est clair que dans ce cadre, de nombreuses phrases auront le même sens:

le chien mange le chat
le chat est mangé par le chien
ce chat, il est mangé par le chien
le chien mangeait le chat

En d'autres termes, on postule que le sens (profond?) commun à toutes ces phrases est la prédication de laquelle on a supprimé de nombreuses fioritures: thématisation, aspects, temps, détermination...

b) concepts de base Génotype Phénotype

On peut considérer que la théorie va s'occuper de deux ensembles d'entités qui sont toutes deux écrites dans le langage formel:

- l'ensemble des phrases simples correspondant à des pensées: (encore appelées phrases canoniques): le langage génotype primitif
- l'ensemble des autres phrases; ce sont toutes les phrases obtenues à partir des précédentes en combinant divers opérateurs grammaticaux ; appelons-le pour l'instant langage génotype des expressions.

Pour en revenir avec la métaphore "valeur", tout comme l'argent mesure une valeur, la mesure générale du sens consistera en la phrase la plus simple grammaticalement.[1]

Le langage en tant que tel n'existe pas, il faut expliquer son rôle dans le processus de communication linguistique. Pour élaborer la théorie sémantique, il faut être capable de définir le "message" porté par un énoncé. Il est bien clair que différentes phrases peuvent exprimer le même message. Le formalisme postulé est l'appareillage formel de la Grammaire Applicative et le langage appelé Langage Génotype[2] défini par celle-ci.

On postule donc dans chaque langue naturelle l'existence de deux langages:

a) un langage primitif dans lequel le contenu d'un message est représenté sans ambiguïté

b) un langage d'expressions, celles-ci sont produites, par applications successives de différents opérateurs sur les expressions du langage primitif.

Pour modéliser cette hypothèse, le langage génotype est scindé en deux parties: le langage génotype primitif et le langage génotype d'expressions.

Langue Naturelle Langage Génotype	
phrases primitives LG primitif	<-------->
expressions LG d'expressions	<-------->

Nous définirons bien évidemment, dans les chapitres qui suivent, les rapports entre ces différents langages. Le langage génotype d'expressions est obtenu à partir du langage génotype primitif par l'application des règles formelles de notre théorie.

La grammaire phénotype contient les règles spécifiques à une Langue Naturelle particulière qui envoient les phrases de surface dans le langage formel (les flèches horizontales du schéma).

Pour éviter des confusions, il faut mettre l'accent sur les différences essentielles entre certains concepts de la théorie G.A.U et des concepts apparemment similaires ailleurs. En premier lieu, il s'agit d'une théorie de la paraphrase tout à fait nouvelle. Nous présenterons les transformations comme des réductions (et non des équivalences cf Desclés, 88) et les transformations ne resteront pas à un niveau purement syntaxique. En effet, la théorie conduit naturellement à ne plus faire les distinctions habituelles entre syntaxe et sémantique, et à se situer dans un domaine qu'on peut qualifier de "sémantique intrinsèque" (Desclés, 1987).

Le concept de langage génotype ne doit pas être confondu avec celui de structure profonde, ni celui de phénotype avec structure de surface. On peut faire une comparaison entre

---

[1] Tout comme dans les systèmes physiques. En procédant ainsi on fait une approximation: tout le problème est de faire en sorte que celle-ci soit la plus fine possible.

[2] Gödel Escher et Bach: ADN

structure sémantique profonde et langage génotype primitif et une autre entre structure *sémantique de surface* et langage génotype *d'expressions*. Le langage génotype modélise à la fois la structure sémantique profonde, la structure sémantique de surface *et* les opérations qui permettent de passer de la "forme de surface" à la "forme profonde". Attention, nous soulignons qu'il s'agit de structures sémantiques: Chomski travaille uniquement sur le plan syntaxique.

d) Universalité du génotype

L'hypothèse principale faite ici est que les bases sémiotiques du langage naturel sont un système de catégories linguistiques universelles. Pour les conjecturer, il faut d'une part investiguer le plus possible de langages différents (tous) du monde, établir par abstraction des catégories générales. Il faut, d'autre part, définir les propriétés de l'objet langue naturelle et déduire toutes les conséquences de cette définition.

La méthode hypothético-déductive est à la base de la construction. Du point de vue d'une théorie linguistique établie sur les bases postulées ici, un système de catégories universelles n'est ni vrai ni faux, mais est donné par définition. *Le système doit quand même prédire avec succès les catégories linguistiques possibles dans les langages réels*. Si le pouvoir prédicatif du système est fort, cela en fait un outil puissant d'investigation.

Les catégories universelles ne sont bien entendues pas observables, car se situant à un niveau abstrait au-dessus des langues naturelles particulières, on les construit par abduction des niveaux observables.

On peut considérer le langage génotype primitif comme un langage "pivot" entre les langues. C'est, en effet, dans ce langage qu'on exprime les "pensées complètes", il n'y a aucune raison de rattacher ce langage à une langue naturelle particulière.

Reprenons un exemple de S. Shaumyan (Shaumyan, 1977):

> "Doceo pueros grammaticam" et "j'enseigne la grammaire aux enfants" et "Ich lerne den Schulern Grammatik" expriment la même pensée: on peut choisir abitrairement l'une de celles-ci comme pensée standard.

Il est raisonnable de supposer qu'un tel langage "pivot" existe, non ouvert à l'observation directe, mais observable indirectement. Notons que ce langage ne repose pas uniquement sur l'intuition comme par exemple dans certains travaux d'Intelligence Artificielle (SCHANK déjà cité et WILKS), il repose ici sur des bases cognitives et linguistiques. Le schéma précédent devient alors:

LN_1	LN_2
Génotype	Langage
langage primitif --------->	langage génotype
<--------- langage primitif	
primitif	grammaire
grammaire	phénotype I
phénotype II	
langage d'expressions --------->	langage génotype
<--------- langage	d'expressions
d'expressions	

Nous espérons assez rapidement avoir des résultats de recherches exploitant cet aspect du système. Il y a encore eu peu de recherches en ce sens, les principaux travaux se concentrant sur le langage genotype, fondamental il est vrai.

e) Le système formel de la Grammaire Applicative

C'est un systeme spécifié d'objets linguistiques qui est défini par des règles mathématiques de déduction et par les règles permettant de ramener des objets construits complexes à des objets plus simples (règles de réduction). Ces règles sont appelées Grammaire du Génotype ou Grammaire Applicative (la notion d'application joue un rôle essentiel dans la construction des objets du génotype).

Le L.G. comporte un ensemble d'objets initiaux, les atomes et un ensemble fini de règles permettant de ramener des objets à des objets simples construits sur les atomes. Il existe de plus un ensemble de règles qui permet de contrôler qu'un objet complexe est "bien écrit" dans le formalisme (types). La définition, le choix de la taille des atomes peut conduire à faire varier le niveau d'approximation du langage.

Comment obtenir un système formel:

On observe des faits. A ces faits on associe des symboles, des atomes qui serviront de base à la théorie. Pour ce qui est de la théorie de la G.A.U., dans un premier temps les atomes correspondront sensiblement aux entrées lexicales puis dans un second temps à des unités de sens plus fines.

On adjoint ensuite des procédés *constructifs*. On va pouvoir construire des objets symboliques complexes. Par exemple, la théorie des gaz est bâtie sur les atomes et sur les molécules.

Les procédés constructifs doivent permettre de prédire des faits inconnus auparavant mais en principe possibles.

Pour bâtir une théorie "formalisée", il faut se donner un système permettant de déduire des théorèmes.

1. on choisit les axiomes
2. on choisit des règles de constructions qui permettent de construire des objets complexes.
   L'ensemble des axiomes et des objets complexes formera ce qu'on appelle une classe inductive d'objets.
3. on définit sur cette classe inductive des prédicats élémentaires pour pouvoir formuler des assertions sur ces objets.
4. un sous-ensemble des assertions possibles est postulé comme étant vrai (les hypothèses de la base de la théorie).
5. des règles sont postulées pour dériver de ces axiomes d'autres assertions vraies: les théorèmes.

La théorie de la G.A.U. est calquée exactement sur ce modèle. Nous le verrons, pour diverses raisons, nous la présenterons de manière légèrement différente. Plutôt que de privilégier l'aspect constructif, le sens axiome théorème, nous allons privilégier le sens théorème axiome (*"réduction"*), en réduisant des objets complexes à des objets simples, en perdant éventuellement de l'information.

Universaux de la G.A.U.

L'élaboration de la G.A.U. s'est opérée ainsi: la question posée étant de savoir quels sont les universaux les plus simples sans lesquels un langage ne serait pas une langue naturelle. On s'abstient des détails non directement en rapport avec ce problème. Un langage n'existe pas sans énoncés. Ce sont des expressions complètes décrivant une situation ou une action (**type t**). Le flux du discours est coupé en éléments discrets fonctionnant comme les phrases ci-dessus.

Il n'est guère possible d'avoir des langues naturelles sans expressions aptes à dénoter des objets, celles ci seront appelées noms ou termes (**type n**). Avec des phrases et des termes comme éléments universels les plus simples, on peut construire un système montrant comment on peut réduire les éléments complexes à une construction d'éléments simples.

La théorie de la G.A.U. est donc en réalité très simple. Le souci de décomposition en composants élémentaires et abstraits entraînera parfois des chaînes d'explications longues. Deuxième point important, le langage génotype est construit de manière à éliminer tout ce qui est inessentiel par rapport à la communication. Les L.G. modélisent les caractéristiques universelles des Langues Naturelles quand ils codent, transmettent et décodent des messages.

Finalement on peut se poser les questions de savoir quel est le lien d'une théorie abstraite avec l'empirique, quel est le statut ontologique des objets abstraits d'un système. Il faut comparer la théorie de G.A.U. à une carte géographique. Elle représente la réalité sous un certain angle, avec plus ou moins de finesse. La carte a une existence mais très différente de l'existence de ce qu'elle représente. Le terrain appartient (comme la langue naturelle) à la réalité objective. La carte fait partie d'un système symbolique création de l'esprit de l'homme. On l'utilise pour représenter un aspect particulier de la réalité objective.

C. LA G.A.U. ET LA LINGUISTIQUE

a) G.A.U., sémiotique, linguistique et sciences cognitives

Nous l'avons déjà abordé: la théorie de la G.A.U. a des objectifs reliés aux domaines linguistiques, sémiotiques et cognitifs.

1. L'objectif primaire de la G.A.U. est la construction d'un système sémiotique dont les entités de base et les opérations de base seraient les entités et les opérations universelles de représentation

2. Il s'agit ensuite d'étudier formellement les propriétés et la structure de ce système sémiotique.

3. Le langage génotype est l'invariant duquel découlent les diverses langues. L'étude des diverses grammaires phénotypes consiste donc à établir les règles qui projettent le génotype dans les langues observables.

4. Le système étant censé encoder des mécanismes linguistiques universels (auxquels correspondent presque certainement des mécanismes cognitifs), il permet d'aborder une étude formelle et typologique des langues. Le génotype est l'invariant: chaque langue est une projection par des règles différentes (la grammaire génotype) du génotype dans chaque langue particulière. Suivant les axes de travail, on peut donc étudier les évolutions des langues dans le temps (diachronie) ou les comparer structurellement.

5. Plus généralement, les *lois* qui sous-tendent le système sémiotique des langues ont certainement *une portée plus générale*: en musique, par exemple(Jackendoff), ou dans les langages artificiels (Hofstader, 1987, op. cit.). Les invariants dégagés par les recherches de la théorie ont vraisemblablement un statut cognitif. Il s'agit vraisemblablement d'une représentation des opérations élémentaires "précâblées" de notre cerveau qui nous permettent de comprendre et de parler plusieurs langues.

GAU agrée avec l'hypothèse cognitiviste: la langue en tant que produit de l'esprit humain doit manifester les structures abstraites de cet esprit. Ces structures abstraites qui se manifestent dans d'autres domaines de façon analogue, se retrouvent dans le langage et dans la gestion des connaissances dans l'esprit humain (Raccah 86).

b) La GAU et les invariants

La recherche des invariants peut conduire à une réflexion approfondie sur les phénomènes généraux à toutes les langues. Il faut à tout prix se dégager des tentations d'ethnocentrisme. Certaines notions considérées comme fondamentales sont fondées dans certaines théories sur la prédominance des langues indo-européennes: il en est ainsi pour les notions de sujet et d'objet. Certaines structures marginales dans certaines langues sont classiques dans d'autres, "l'ergativité" par exemple (cf. Tchekhoff). Ces structures sont en fait le reflet de mécanismes très généraux dont on décele la présence dans de nombreuses langues. Il faut pouvoir en rendre compte, c'est le propos de la G.A.U. Il en est ainsi, par exemple des opérations de thématisation, réalisées aussi bien par l'ordre des constituants que par des variations phonologiques (intonations), qui tournent uniquement autour de la prédication.

A un niveau plus profond, il est nécessaire de pouvoir prendre en compte des notions plus proches de la sémantique. La communication qui est impliquée par les verbes de mouvement, de changement ou décrivant des situations, répond à des conventions qui permettent de postuler l'existence de "primitives" indépendantes d'une langue particulière. Shaumyan et bien d'autres y ont donné un début de réponse par la théorie des cas (théorie localiste) (Hjemslev et l'école Danoise, Grüber etc) et la théorie sémantique qui en découle. Shaumyan en a entrepris la formalisation dans le cadre de la Logique Combinatoire dès 77.

La GAU langage de communication

Le modèle de la G.A.U. est un cadre formel et général où les problèmes sémiotiques, linguistiques ou de représentations peuvent être décrits de façon rigoureuse. De par ce fait, c'est un excellent moyen de communication. Le modèle de la G.A.U. bien que par son approche particulière nettement différent des modèles existants, ne se pose pas en contradiction des théories existantes. Il permet, on le verra, de prendre en compte des résultats et des idées de la linguistique traditionnelle (Benveniste), ainsi que contemporaine (Cullioli, Pottier, Mel'cuk, Comrie, Langacker, Jackendoff, Portal, etc....).

L'utilisation de G.A.U. sur un problème particulier se fera donc après lecture de la littérature existant sur le sujet, l'éclairage nouveau apporté par le formalisme sera de toute façon intéressant.

La GAU théorie de la construction du sens

Encore un mot sur la principale caractéristique actuelle de G.A.U. Nous l'avons déjà évoqué, la G.A.U. présente un moyen privilégié d'introspection sur la façon dont se construit le sens d'une phrase. Le mécanisme principal va, partant d'un énoncé, produire deux sortes de résultats

- la forme primitive: le sens brut de la phrase
- la série d'opérations grammaticaies qui ont permis de produire la phrase à partir de la forme primitive.

Il est clair que définir précisément la signification grammaticale de ces opérateurs grammaticaux n'est pas toujours chose aisée. Du point de vue du linguiste, la portée exacte de l'opération passivisation, est importante, elle l'est sans doute beaucoup moins, du point de vue de l'informaticien.

On voit donc, ici, le double intérêt de l'approche: mise en évidence des opérateurs grammaticaux à fin d'études dans des conditions facilitant celle-ci, possibilité pour les spécialistes de l'Intelligence Artificielle de préciser, s'ils les ignorent, comment ils les approximent.

Il est à noter que faute d'un langage formel descriptif, il n'est aucun travail d'I.A. qui indique de façon précise quelles sont les simplifications opérées par rapport à la langue naturelle dans son intégralité, d'où la difficulté de formuler d'imaginer même les limitations de la plupart des réalisations programmées.

Sens intrinsèque

Il est bon de souligner pourtant que la notion de sémantique intrinsèque vient de l'idée que le "sens" d'une phrase composée d'un certain nombre de mots est construit à partir du sens de chacun de ces lexèmes et qu'une signification propre existe "dite signification intrinsèque" venant de la composition structurelle de ces lexèmes. La signification réelle dans un contexte ou une situation d'élocution donnée sera calculée par une opération supplémentaire de référentiation à partir du sens intrisèque.

Il reste bien sûr des problèmes à résoudre. La détermination, par exemple, qui est une opération qui se greffe sur la prédication, est encore un problème difficile.

Pour l'instant, la G.A.U. n'a pas de théorie référentielle, il semble que son développement soit inévitable, tant pour la détermination que dans un cadre élargi au discours des références définies, de la deixis et de l'anaphore, car ne traitant que de **sémantique intrinsèque**. Il est

envisageable (et envisagé) de réaliser sur les bases de G.A.U., ou plutôt sur les bases théoriques de Curry (Combinatorial Logic, tome 2) un système comparable à celui de Kamp (Kamp 78); la description sémantique sémantique obtenue serait alors plus fine que celle du modèle de Kamp.

Familles de paraphrases

Pour en finir avec ce chapitre d'introduction, voyons un exemple de familles paraphrastiques.

FAMILLE 1

(1) on a passé de fort beaux disques ce matin

(2) les disques ont été passés

(3) Il se passe beaucoup de beaux disques depuis peu

(4) Ce sont de beaux disque qui ont été passés.

(5) On a passé de beaux disques.

FAMILLE 2

(6) Pierre passe un disque

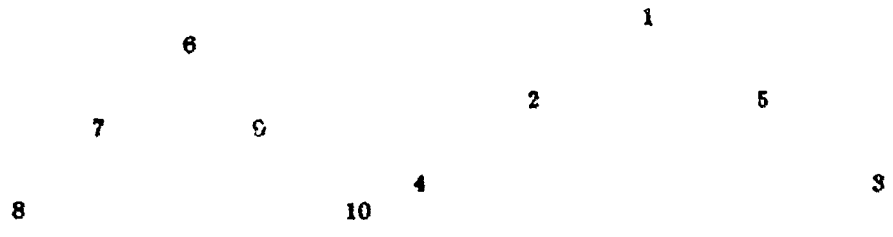
(7) Pierre le passe, le disque

(8) Pierre, il le passe le disque

(9) un disque est passé par Pierre

(10) il est passé par Pierre, le disque

Ces familles se représenteront par des arbres: à la racine 4



On conçoit intuitivement la raison de l'existence de deux branches différentes; l'une décrit la voix active, l'autre la voix passive. Plus on s'éloigne dans les branches, plus la phrase est compliquée. En effet, 8 est obtenue après deux opérations de thématisation, 10, après passivation puis thématisation.

Annexe

Construction d'une théorie scientifique

Une théorie se construit toujours avec deux niveaux de concepts: les concepts correspondant à ce qui est observable, les concepts abstraits. Les concepts élémentaires observables sont des éléments primitifs du domaine: l'électron en physique, le gène en biologie, etc... Les concepts élémentaires sont postulés après *généralisation* des données, ils se situent au niveau empirique.

Pour que la théorie ait un intérêt, il faut qu'elle possede des *lois*. Si celles-ci ne sont que des généralisations de faits observés, elles n'apportent pas grand'chose. On doit donc "construire" un niveau supérieur, c'est là qu'intervient l'*abstraction*.

Les concepts sont une facilité d'écriture, ils permettent d'écrire des généralisations (*abstraction identifiante*).

"un marin aime la mer"

La généralisation, introduction d'un objet abstrait, nécessite l'emploi de "règles d'exclusion d'objet abstrait". Il faut pouvoir déduire "Popeye aime la mer".

L'abstraction relationnelle est fréquemment pratiquée. On en a vu l'exemple pour le concept de valeur qui n'est pas mesurable de façon directe. On fabrique une relation d'équivalence entre les objets.

De façon générale, les "*constructions*" correspondent à de l'inobservable. Pour qu'elles aient une utilité quelconque, il faut spécifier exactement comment elles se rattachent aux objets empiriques, comment établir ces relations pour les objets empiriques (*abstraction systémique*).

Lois

Les lois statistiques n'apportent guère d'informations. Elles généralisent l'état de données empiriques connues actuellement. Elles ne peuvent prendre en compte une évolution temporelle. Les lois statistiques n'aident pas dans la recherche de systèmes potentiels.

Il est beaucoup plus intéressant d'élaborer des lois déductives. Généralement, on se place dans des conditions idéalisées d'expérience et on spécifie la loi construite (exemple: le mouvement sans frottement).

Auteurs **Pierre Plante**
Centre d'ATO, UQAM
Jean Perron
Office de la langue française

Titre **Un projet de recherche et de développement: un système de dépouillement terminologique assisté par ordinateur**

RÉSUMÉ

1. Introduction

Le dépouillement terminologique consiste à recueillir, dans des textes techniques ou scientifiques, la terminologie d'un domaine. Si l'on considère la particularité du terme (qui se présente, très souvent, comme une expression syntagmatique composée d'un nombre variable de mots et construite selon une grande diversité de modèles de formation), peut-on imaginer un système de dépouillement assisté par ordinateur? Quelles seraient les composantes de ce système? Quels types de connaissances faudrait-il emmagasiner dans ce système?

2. Hypothèse et stratégie

En tout premier lieu, une description morphosyntaxique des textes à dépouiller s'impose : catégorisation lexicale et grammaticale, lemmatisation et, enfin, analyse syntaxique. C'est l'analyseur syntaxique qui, en relevant les structures syntagmatiques des textes, produira une première liste de termes virtuels. Toutefois, les unités ainsi dépistées par l'analyseur ne constitueront pas toutes des unités terminologiques bien découpées: plusieurs d'entre elles, bien qu'elles correspondent formellement à des modèles terminologiques, ne sont nullement lexicalisables. Aussi le résultat de l'analyse est-il destiné à être retraité par un ensemble de règles et de critères terminologiques (inscrits dans une base de connaissances) qui rejetteront certaines unités et pondéreront les autres; diverses autres stratégies de traitement des unités, relevant de techniques d'analyse de textes par ordinateur, de la lexicométrie, permettront également de restreindre la liste des unités initiales les plus lexicalisables qui seront soumises, avec leurs contextes, au terminologue pour analyse et validation.

3. Types de connaissances requises

Le succès du développement d'un système de dépouillement repose sur la capacité qu'on aura de décrire et d'exploiter les divers niveaux du fonctionnement discursif: morphologique, syntaxique, lexical et terminologique, sémantique et textuel.

4. Applications

D'une façon générale, le système projeté permettra une exploitation plus appropriée des données comprises dans les textes à portée scientifique, technique ou administrative puisqu'il dépistera les **mots composés** *(termes complexes), ce que ne*

peuvent faire les logiciels habituellement utilisés en lexicométrie. De ce fait, ses applications pourront être variées: si son utilisation en terminologie, en néologie et en lexicologie est évidente, il peut également constituer un outil appréciable pour l'exploitation des connaissances scientifiques, techniques ou administratives elles-mêmes puisque la langue est une représentation de ces connaissances (ex: identification et étude du vocabulaire d'une spécialité à des fins documentaires par les professionnels de cette spécialité ou à des fins de modélisation des données en vue de l'élaboration de systèmes informatisés, etc.).

Auteure **Maria Elisa Macedo**
Centro de linguitisca da Universidade de Lisboa

Titre **Une analyse des prépositions em, a, para, de, du portugais**

RÉSUMÉ

L'analyse proposée donne à ces prépositions le rôle d'élément d'une relation locative et décrit leurs formes sous-jacentes géométriquement marquées. Des verbes supports interviennent dans cette analyse. Ce qui permet de décrire les phrases locatives complexes dans des phrases élémentaires qui traduisent la relation locative entre le lieu et l'argument du lieu.

VARIATIONS DE DÉBIT DANS LA PAROLE DE SYNTHÈSE DE LA SYNTAXE À LA PHONÉTIQUE

Danièle Archambault
Université de Montréal

Cette recherche de base dans le domaine des communications verbales a pour but de permettre à un synthétiseur de parole en français, développé à l'INRS-Télécommunications[1], de parler à différentes vitesses de débit. Cette option supplémentaire sur le synthétiseur permettra à l'usager de choisir parmi trois vitesses de lecture: une vitesse normale pour lecture courante, une vitesse lente lorsque les textes sont difficiles, par exemple, et une vitesse rapide pour *feuilleter* le contenu d'un texte. L'INRS-Télécommunications possède deux systèmes de synthèse par règles qui permettent de générer un texte oral français à partir de n'importe quel texte écrit, chacun mettant à profit une technique différente: un synthétiseur à formants[2] et un synthétiseur par diphones[3]

La première étape de cette recherche consiste à décrire les principales techniques d'accélération et de ralentissement du débit en français, à partir de l'analyse d'échantillons de parole naturelle. Dans la deuxième étape, les règles dégagées lors de l'analyse précédente seront implantées dans le synthétiseur. Le présent article porte sur la première étape de la recherche.

1. COMPOSANTES DU DÉBIT

Les deux principales composantes du débit, généralement défini par le nombre de syllabes prononcées par minute, sont la vitesse d'articulation et les pauses. La plupart des recherches sur le sujet ont eu pour but de cerner les différentes caractéristiques du débit normal à travers différentes épreuves discursives (discours spontané, discours politique, lecture, etc.). On peut citer, entre autres, les travaux de Goldman-Eisler[4] et Lass[5] pour l'anglais, Grosjean et Deschamps[6], et Duez[7] pour le français. Peu de chercheurs, cependant, se sont penchés sur l'examen des variations imposées de débit (Lass[8] et Gilbert[9]). Cette étude est la première à examiner les variations imposées de débit en vue de l'application à la synthèse par règles.

---

[1]L'auteure est également professeure invitée à l'INRS-Télécommunication.

[2]D. O'Shaughnessy (1984) "Design of a real time French text-to-speech system", Speech Communication, vol. 3, p. 233-243.

[3]D. O'Shaughnessy, D. Archambault, D. Bernardi et L. Barbeau et al. (1987) "Diphone Speech Synthesis", Speech Communication. Vol. 7, No 1. 55-65.

[4]F. Goldman-Eisler (1966), "The significance of changes in the rate of articulation.

[5]N. Lass et Deem, J.F. (1971) "Temporal patterns of rate alterations in oral reading", Acta Symbolica, 11, 254-263.

[6]F. Grosjean et Deschamps, A. (1972) "Analyse des variables temporelles du français spontané", Phonetica, 26, 129-156.

[7]D. Duez, (1982), "Silent and non-silent pauses in three speech styles", Language and Speech, Vol. 25, Part 1, 11-28.

[8]N. Lass et Deem, J.F. (1971) "Temporal patterns of rate alterations in oral reading", Acta Symbolica, 11, 254-263.

[9]J.H. Gilbert et W.B. Kenneth (1969) "Rate alterations in oral reading", Language and speech, 12, 192-201.

Dans cette recherche, je cherche à définir le rôle relatif de chacune des composantes du débit afin de dégager un ensemble de règles pour la synthèse de parole. Le jeu des deux composantes (vitesse d'articulation et pauses) est examiné en fonction de trois vitesses de débit et de la longueur du texte. En ce qui concerne les pauses, je cherche à établir leur nombre, leur durée ainsi qu'une hiérarchie dans leur endroit d'occurrences. Au niveau de la vitesse d'articulation, il s'agit de définir les segments qui devront être allongés ou abrégés ainsi que d'établir l'ordre de grandeur de ces modifications de durée.

2. MÉTHODOLOGIE

Dans cette recherche, j'ai eu recours à deux locuteurs ayant l'habitude de parler en public: un professeur (locuteur 2) et un annonceur de radio (locuteur 1). Les enregistrements ont été faits en chambre sourde au laboratoire de phonétique. Les locuteurs avaient à lire une série de textes de différentes longueurs. Pour chaque texte, les locuteurs devaient alterner leur vitesse de débit, c'est-à-dire qu'ils devaient d'abord lire le texte à vitesse normale, ensuite à vitesse lente, relire le texte à vitesse normale, puis le lire à vitesse rapide. Ceci permet aux locuteurs de modifier leur vitesse de lecture en fonction d'une vitesse de référence (vitesse normale).

Pour l'examen des corpus, j'ai eu recours à deux types d'analyse: une analyse perceptuelle et une analyse acoustique. Les différents corpus ont été présentés à des sujets à l'intérieur de deux tests de perception (un pour chaque locuteur). Les sujets avaient pour tâche d'identifier la vitesse à laquelle parlait le locuteur (vitesse lente, normale ou rapide). Cette analyse perceptuelle a pour but de vérifier que le locuteur a réussi à produire la vitesse de débit désirée. Les analyses acoustiques ont été faites au laboratoire de phonétique à l'aide de programmes d'analyse spectrographique implantés sur un ordinateur Zenith AT. Il s'agit principalement de segmentation du signal afin de pouvoir dégager les données nécessaires à l'examen des vitesses d'articulation et des pauses.

3. RÉSULTATS

3.1 Texte long

Le texte long est constitué de cinq paragraphes. Il s'agit d'un article paru dans le journal *Le Devoir*. J'examine ici les données relevées chez le locuteur2. Bien que les trois productions du locuteur (normale, lente et rapide) présentent peu de distinction au niveau de la durée totale du texte, du débit et de la vitesse d'articulation (tableau 1), on peut toutefois remarquer une meilleure distinction entre la vitesse normale et la vitesse lente (durée totale: 105 s et 87,5 s) qu'entre la vitesse normale et la vitesse rapide (87,5 s et 82,8 s). Ceci indique que le locuteur a eu plus de facilité à ralentir qu'à accélérer. Ce phénomène a déjà été relevé par Lass (1971) et peut être relié à des contraintes d'ordre articulatoire. En effet, il y a une limite à la vitesse à laquelle un locuteur peut parler et néanmoins conserver l'intégrité du message. Cependant, il est intéressant de remarquer, au niveau perceptuel, que les auditeurs ont bien distingué le texte rapide du texte lent mais ont perçu le texte lent comme étant de vitesse normale, ce qui semble en contradiction avec les données acoustiques.

FIGURE 1:

Variations de débit. Texte long. Loc. 2

Débit et synthèse

Variations de débit Texte long. Loc. 2

	Lent	Normal	Rapide
Débit (syll./mn)	300	343	360
Durée totale:	105 s	87.5 s	82.8 s
Suites sonores:	50	35	22
Pauses:	49	35	22
Vitesse d'art.	6 3 syll./s	7 syll./s	7.4 syll./s
Rapport phonation/durée totale	77.7 %	82.2 %	81.8 %

Cependant, si les trois productions se distinguent peu au niveau de la vitesse d'articulation, le nombre de pauses par contre est très différent d'une vitesse à l'autre. Le nombre des pauses décroît régulièrement quand on passe du texte lent au texte rapide. On retrouve 49 pauses pour le lent, 35 pour le texte normal et 22 pour le rapide. Le nombre des pauses semble jouer un rôle prépondérant dans l'indication de la vitesse de débit.

La comparaison entre l'analyse acoustique et l'analyse perceptuelle indique qu'un tel texte semble trop long pour être exploitable dans le but qui nous intéresse. En effet, le locuteur a beaucoup de difficultés à tenir un débit constant tout au long du texte et il est difficile de savoir si l'auditeur base son impression de la vitesse de débit sur tout le texte ou sur un des paragraphes en particulier. J'ai donc décidé d'avoir recours à un texte moins long composé d'un seul paragraphe.

3.2 Paragraphe isolé

Le paragraphe comprend trois phrases complexes et un total de 90 mots. On peut remarquer, cette fois-ci, une meilleure distinction entre les trois vitesses de lecture (durée totale: 35,4 s, 28,6 s et 23,6 s) mais les vitesses d'articulation sont toujours assez semblables (6 syll/s, 6,8 syll/s et 7,5 syll/s, tableau 2). Encore une fois, la distinction se fait au niveau des pauses. Ici, il y a trois fois plus de pauses en débit lent qu'en débit rapide et deux fois plus en débit normal qu'en débit rapide. Le rôle privilégié des pauses dans le débit constitue une caractéristique très importante car elle permettrait de modifier le débit du synthétiseur beaucoup plus facilement. En effet, si, pour changer le débit d'un texte, il suffit de modifier le nombre de pauses sans toucher à la vitesse d'articulation, la tâche s'en trouve non seulement facilitée mais il y a ainsi moins de risques de compromettre l'intelligibilité du texte.

FIGURE 2:

Variations de débit. Par. 5. Loc 2

Débit et synthèse

Variations de débit. Par. 5. Loc. 2

	Lent	Normal	Rapide
Débit (syll./mn):	272.9	342	404
Durée totale:	35.4 s	28.6 s	23.6 s
Suites sonores:	20	14	7
Pauses:	19	13	6
Vitesse d'art.	6 syll./s	6.8 syll./s	7.5 syll./s
Rapport phonation/durée totale	76 %	83 %	91 %

Il reste cependant à définir les points d'occurrence des pauses. L'indice le plus important de l'apparition d'une pause est la présence d'un marqueur orthographique (point, virgule, etc.). Il existe une hiérarchie à l'intérieur de cette catégorie: une pause en présence d'un point est obligatoire quelle que soit la vitesse de débit mais, après une virgule, la pause peut tomber lorsqu'on accélère le débit. En l'absence de marqueurs orthographiques, si le texte est trop long, des pauses apparaîtront aux frontières syntaxiques. Cependant, ces pauses ont un statut plus fragile et lorsqu'on accélère le débit, elles sont les premières à tomber.

3.3 Phrases isolées

Pour ce dernier test, j'ai utilisé une série de dix phrases isolées, phonétiquement balancées (Lennig[10]). Le locuteur (Locuteur 1) devait lire chaque phrase dans l'ordre suivant: vitesse normale, lente, normale, rapide. Au moment de l'enregistrement j'ai noté un problème pour les phrases à vitesse lente: elles semblaient déjà trop rapides. Nous avons demandé à la locutrice de les reprendre. Elle a donc relu la série de phrases complètes à vitesse lente seulement. Les résultats de l'analyse des deux séries sont présentés dans le tableau 3 (Lent 1 représente le deuxième essai et Lent 2 le premier). Ce tableau présente une analyse comparative des vitesses d'articulation pour les différentes vitesses de débit. La présence d'une pause dans la phrase est indiquée par une virgule entre les chiffres représentant les vitesses d'articulation des suites sonores de la phrase.

Les différentes vitesses de lecture se distinguent par la présence ou l'absence de pauses. La vitesse lente est caractérisée par la présence d'au moins une pause dans la phrase et ce, pour les deux essais. Seule la phrase 8 ne présente pas de pauses. Au niveau perceptuel, cette phrase

[10] M. Lennig (1981) "Phrases françaises phonétiquement balancées", Revue d'acoustique, 56, 31-42.

a, dans les deux cas, été reconnue par les auditeurs comme étant une phrase à vitesse normale. Du point de vue de la production, la présence d'une pause semble être un élément très important pour distinguer la vitesse lente de la vitesse normale. Cependant, au niveau perceptuel, environ la moitié des phrases à vitesse lente (premier essai, Lent 2) ont été jugées comme étant de vitesse normale dans les tests de perception, ce qui indique que la seule présence d'une pause ne suffit pas à indiquer la vitesse d'articulation. Il n'y a que peu de différence au niveau des pauses entre la vitesse normale et la vitesse rapide. Trois des phrases présentent une pause en vitesse normale et une seule phrase en vitesse rapide. L'auditeur ne peut donc, dans ce cas-ci, baser son jugement sur la vitesse de débit par la présence ou l'absence d'une pause.

La vitesse d'articulation semble donc ici jouer un rôle important. On peut d'ailleurs remarquer une plus grande distinction entre les vitesses d'articulation des trois débits (tableau 3). Les vitesses d'articulation moyennes sont 3,1 syll/s pour la vitesse lente (Lent 1), 5,1 syll/s pour la vitesse normale et 7,5 syll/s pour la vitesse lente (tableau 4). La vitesse d'articulation semble jouer au niveau des phrases isolées un rôle beaucoup plus important que dans les autres types de texte (paragraphes et phrases). En fait, les phrases sont d'une complexité syntaxique assez simple et assez réduite quant au nombre d'éléments. Donc, pour changer la vitesse de débit à l'intérieur de phrases isolées, on ne peut se contenter des pauses. Les modifications de durées segmentales deviennent indispensables.

FIGURE 3:

Phrases isolées. Loc 1

Débit et synthèse

Phrases isolées: Loc. 1

Vitesse d'articulation (syll./s)

Phrase	Lent		Normal	Rapide
	Lent 1	Lent 2		
1	3 4, 3 6	4.9, 4.5	6.0, 5 1	7 5
2	3 4, 4.3	4 3 6 4, 6 1	5 4	7 0
3	3.9, 3.7	2 9, 4.6	5 2	7.0
4	2.2, 2 9	3 0, 3 4, 4 9	4 7	5.9
5	3.5, 2.8, 3 3	4.0, 3 6, 3 9	4.8, 5.2	6.1, 9.6
6	2.3, 4 0	3.6, 4 9	4.0, 5 7	5 9
7	1 4, 3 4	1.8, 4 8, 4.9	5.3	7.0
8	4.6	4 8	5.6	7 2
9	2 9, 3 1	4 5, 3 4, 3 0	4.4	5 2
10	2 5, 5 1	4 3, 4 1	5 14	6 9

4. DISCUSSION

Les résultats des analyses nous amènent à conclure que si les pauses jouent un rôle déterminant dans l'indication de la vitesse de débit pour les textes longs, il n'en va pas de même pour les phrases simples isolées. Dans ce dernier cas, c'est la vitesse d'articulation qui semble l'élément déterminant. Donc, le modèle de modification de débit pour la synthèse de parole devra intégrer les deux composantes.

En ce qui concerne les pauses, le modèle doit prédire le nombre et les points d'occurrence des pauses en fonction des vitesses de débit. Ces pauses sont déterminées par la présence de marqueurs orthographiques ainsi que par des critères d'ordre syntaxique. Il faudra élaborer un corpus permettant de dégager les règles d'apparition des pauses en fonction des frontières syntaxiques et établir une hiérarchie dans l'apparition ou l'abandon de ces pauses. Pour être en mesure d'intégrer ces règles dans le synthétiseur, il sera probablement nécessaire d'élever le niveau de sophistication de l'analyseur syntaxique.

Les modifications de vitesse d'articulation posent plus de problèmes. Idéalement, il faut, à partir de données établies pour la vitesse normale, élaborer un ensemble de règles d'allongement et d'abrègement des segments. Cependant, le modèle de durée actuel (O'Shaughnessy[11]) produit une vitesse de parole souvent jugée trop lente.

FIGURE 4:

Comparaison des vitesses d'articulation

Débit et synthèse

Comparaison des vitesses d'articulation

	Lent	Normal	Rapide
Texte long	6.3	7	7.4
Par. 5 Loc. 2	6	[illegible]	7.5
Par. 5 Loc. 1	5.4	5.8	6.3
Phrases isolees	3.3	5.1	7.5

D'ailleurs, l'examen comparé des vitesses de débit produites par le synthétiseur (Loquax) et celles produites par le locuteur 1 pour les phrases isolées montre que le synthétiseur est très lent (Tableau 5). Les valeurs de vitesse d'articulation en lecture normale sont semblables à celles en vitesse lente chez le locuteur 1. Il faut vérifier à l'aide de tests de perception si ces valeurs sont acceptables comme indicatrices d'une vitesse normale pour une parole de synthèse. Dans l'éventualité d'une réponse négative, il faudra donc refaire un modèle de durée pour vitesse normale à partir duquel nous élaborerons les règles de changement de débit.

[11] D. O'Shaughnessy (1981), "A Study of French Vowels and Consonants Durations", Journal of Phonetics , 9, 385-406.

Il existe un autre modèle de durée pour la parole de synthèse par règles en français (Barkthova et Sorin[12]). Ce modèle comme celui de O'Shaughnessy laisse de côté un aspect linguistique important dans les paramètres de modification de durée: la syllabe. En effet, aucun de ces modèles ne tient compte de la syllabe comme cadre d'application des règles. Pourtant, leur importance est reconnue. Prenons comme exemple le cas de la règle d'allongement des voyelles devant consonnes allongeantes. Cette règle ne s'applique que si la consonne qui suit est dans la même syllabe: *voyage* [vwaja:ʒ] mais *voyageur* [vwaja - ʒœr]. Aucun des modèles actuels ne peut rendre compte de ce phénomène.

FIGURE 5:

Phrases isolées: Loc. 1 et Loquax

Débit et synthèse

Phrases isolées: Loc. 1 et Loquax

Vitesse d'articulation (syll /s)

Phrase	Lent	Normal		Rapide
	Loc. 1	Loc 1	Loquax	Loc 1
1	34, 36	60, 51	44	75
2	34, 43	54	39, 38	7.0
3	39, 37	52	38	7.0
4	22, 29	47	34	5.9
5	35, 28, 33	41 52	35	61, 96
6	23, 40	40 57	39	59
7	14, 34	53	40	7.0
8	46	56	46	7.2
9	29, 31	44	43	52
10	25, 51	514	44	69

De plus, comme on peut le voir dans ce cas-ci, ce ne sont pas tous les éléments de la syllabe qui sont alors modifiés mais seulement ceux de la rime. En effet, quand une voyelle s'allonge sous l'effet d'une telle règle la consonne qui suit doit alors obligatoirement s'abréger[13]. De plus, d'après l'examen sommaire de mes corpus, il semble que l'impression de débit soit principalement véhiculée par les modifications de durée des syllabes accentuées.

Il faudra donc probablement redéfinir un modèle de durée segmentale pour le français. Un autre problème se présentera alors: celui du corpus à utiliser. En effet, les modèles de variation de durée ont généralement pour corpus de base, du moins en partie, une série de mots isolés permettant d'examiner les variations de durée en fonction d'un ensemble complet de contextes phonétiques. Ce type de corpus est inutilisable dans les variations imposées de débit. En effet, plus un mot est réduit quant au nombre de segments et de syllabes, plus il est difficile de faire varier le débit en le prononçant. Un mot isolé laisse peu de place aux indices de débit. Il faudra donc trouver une méthode plus appropriée aux études sur le débit.

---

[12] K. Barkthova et C. Sorin (1987) "A Model of Segmental Duration for Speech Synthesis in French", Speech Communication, 6:3, 245-261.

[13] D. Archambault et al, (1986), "Problème de production ou problème de perception? Le cas du dévoisement dans la désintégration phonétique", Congrès de l'ACFAS, Université de Montréal. Recherche subventionnée par les Fonds FCAR et l'Université de Montréal (CAFIR).

Auteur **Jean-Yves Morin**
Université de Montréal

Titre **Vers le parsage universel**

RÉSUMÉ

Je compte présenter ici une approche nouvelle aux problèmes du parsage (c'est-à-dire de l'analyse morphologique, syntaxique et sémantique de phrases en langues naturelles). Cette approche est essentiellement linguistique, plutôt qu'informatique. Elle encourage des descriptions approfondies, modulaires et transparentes.

On entend par théorie du **parsage universel**, *une approche au problème du parsage caractérisée par les buts suivants :*

a) **Universalité des domaines linguistiques.** *Elle doit être applicable* **mutatis mutandis** *à tous les types de langues:*

- *"configurationnelles" ou non;*
- *strictement ordonnées ou non;*
- *à arguments externes ("sujets") ou non;*
- *ergatives, accusatives ou mixtes;*
- *à relations grammaticales ou rhématiques dominantes;*
- *polysynthétiques, analytiques ou agglutinantes.*

b) **Universalité des objets traités.** *Elle doit également tenir compte de tous les types d'objets et de relations syntaxiques :*

- *catégories mineures aussi bien que majeures,*
- *constructions simples ou complexes.*
- *objets et constructions marqués et non marqués.*

c) **Couverture lexicale.** *Enfin, elle doit permettre une couverture lexicale assez large (de l'ordre de 20 000 mots) pour chaque langue traitée.*

En cela, cette approche s'oppose radicalement à la plupart des approches récentes[1] *:*

a) qui sont orientées vers un type "européocentriste" de langue (configurationnelle, strictement ordonnée, à arguments externes, accusative, à relations grammaticales dominantes et analytiques);

b) qui négligent plusieurs types de catégories (surtout les catégories mineures) et de constructions (en particulier les adjonctions et les appositions) et ne tiennent que peu ou pas compte des relations thématiques et rhématiques et

c) dont la couverture lexicale est insignifiante (de 200 à 2000 mots).

[1] Cf. par exemple Correa (1987), Marcus (1980), Berwick (1985), McCord (1980, 1982, 1985, 1987), Morin (1985), Pereira (1981, 1982), Proudian & Pollard (1985), Weinberg (1987), Wehrli (1983, 1984).

De plus, sur le plan computationnel, la théorie du parsage universel est caractérisée par les buts suivants :

a) **Généricité.** *Les fonctions de parsage doivent être définies à un très haut niveau et être génériques. C'est-à-dire qu'elles doivent s'adapter aux types d'objets (descriptions linguistiques) auxquels elles s'appliquent plutôt que l'inverse.*

b) **Modularité et transparence fonctionnelles.** *Les modules d'analyse doivent être fonctionnellement transparents, c'est-à-dire qu'ils doivent permettre d'accumuler de l'information sur un ou plusieurs objets de façon incrémentale, sans que les résultats d'un module puissent venir contredire ceux d'un autre. Chaque module doit donc traiter des propriétés spécifiques.*

A partir d'exemples concrets, j'essaierai de démontrer l'intérêt et la faisabilité d'une telle approche dans le cadre des grammaires d'unification et de la programmation logique par contraintes.