PAQUIN, L.-C.; BEAUCHEMIN, J. (1989). " Apport de l'ordinateur à l'analyse des données textuelles ". In *Actes du colloque 'La description des langues naturelles en vue d'applications linguistiques'*. Québec: Centre international de recherche sur le bilinguisme, 1989: 21-31.

APPORT de l'ORDINATEUR à l'ANALYSE des DONNÉES TEXTUELLES

par Louis-Claude PAQUIN et Jacques BEAUCHEMIN

Université du Québec à Montréal

0. Préambule

Ce texte n'a pas pour but de présenter des outils ou des méthodes informatiquesà 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 extrêmement 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". 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.

C'est 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.

La pauvreté de certains résultats obtenus par des analyses lexicales imputable à une formalisation insuffisante des données textuelles a fait croire en la primauté du second type d'outil sur le premier. Un tel raisonnement repose sur une définition 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 de l'autre de ces types d'outil, nous proposons de les un 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 des 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 publiques, en sont venus à une production textuelle - faite de rapports, de directives, de projet ou de préprojets - 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 "tissus" et "texte" ont une racine latine (textus) commune. Les réalités désignées se caractérisent par un enchevêtrement, dans le 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 situer 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 et 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 coprésence, 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 discours 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 déploiement de stratégies discursives. Le décodage des stratégies mises en oeuvres 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 dialogisme 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 de 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 et 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'ils puissent 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 pourrons 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 beaucoup de place à l'improvisation. La motivation du lecteur à tendre la main pour prendre une fiche vierge 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 des 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 fondent. La lecture experte souffre donc d'un manque de rigueur rendant sa validation difficile. Par ailleurs, les procédés conventionnels d'analyse de textes interdisent à toutes fins pratiques le traitement de corpus de grande envergure typique des organisations. À 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 auxquels il importe de palier. Elle n'est ni régulière et 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 un 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 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 leurs fréquences 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 soient reflètent la

distribution des mots et non pas leur 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 exemples 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 nom propres désignant des lieux de même que les adverbes de lieu peuvent être regroupés dans une catégories 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 considérés 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 ou 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 à 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 programmations 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) s'est avéré que partiellement opérant puisque les langues naturelles ne constituent pas des systèmes fermés, mais ouverts et que l'ambiguòté 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).

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 (nécessité, 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 grand 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 lexicale 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. À l'instar des systèmes experts qui séparent le savoir exprimé sous forme de règles d'inférences 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'outils 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 répertoriant 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 schématisations 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 dépendances 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 propriétés. 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 technique 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 générée 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 nature, tel l'agglomération d'items lexicaux en certains points stratégiques du texte, les procédés stylistiques, etc.

Une approche interactive à l'analyse textes où la dimension heuristique prime nous semble préférable. L'analyse prend alors la forme d'une démarche cyclique composée 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 appaaÎt importante pour décrire les relations qu'entretiennent les objets de schématisation.

En attendant que les parseurs puissent être appliqués à n'importe quels textes 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.

À paraÎtre dans les actes du colloque: La d scription des langues naturelles en vue d'applications informatiques organisé par le Centre international de recherche sur le bilinguisme (CIRB) de l'Université Laval (Québec), tenu des 7 au 9 décembre 1988.

Chercheur au Centre d'Analyse de Textes par rdinateur, Ph. D en philologie médiévale, âdition critique d'un traité alchimique latin: Liber secretorum.

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

Les auteurs participent depuis janv er 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.

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