SYSTEME D'ANALYSE DE CONTENU ASSISTEE PAR ORDINATEUR (SACAO)[1]

Par Jules Duchastel, Luc Dupuy, Louis-Claude Paquin,
Jacques Beauchemin et François Daoust
Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal

1. Le projet

Le projet SACAO[2] (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 texte 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, ...)[3] 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.

---

[1] Paru dans les Actes du Colloque La description des langues naturelles en vue d'applications linguistiques, en collaboration avec Jacques Beauchemin, Québec Centre international de recherche sur le bilinguisme, 1989.

[2] La conception du projet remonte à 1986. Sa mise en opération effective date de janvier 1988.

[3] Voir Pierre Plante, Jules Duchastel, Lorne H. Bouchard, Potentiel d'applications de Déredec dans le contexte de la burautique, Ministère des Communications du Québec, avril, 1986.

D'un côté, on trouve des usages en traitement informatique de la langue et une quantité croissante de données textuelles déjà disponibles, de l'autre, des procédures diversifiées d'écriture 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[4] 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[5] 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 statistico-morphologique. 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 avancées significatives, mais les contradictions entre diverses approches théoriques ainsi que l'opacité de certains modèles ont peu favorisé l'intégration des connaissances ainsi

---

[4] 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. Grosz et al. Readings in Natural Language Processing, California, Morgan Kaufmann Publishers, inc., 1986, 664 pages.

[5] Op. cit.

produites. La relative courte durée des projets indique l'existence fréquente d'impasses théoriques. La projection très problématique des avancées théoriques dans les applications pratiques a mis en évidence 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. Ceci 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 fondamente 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. Ceci 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 succinte, 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[6] (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 language 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[7], 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

---

[6] 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.

[7] Nous tenons à souligner la contribution importante de messieurs 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.

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 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 faires 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[8].

Nos travaux ou bien s'appuient sur des applications déjà developpées 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.

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.

---

[8] Déredec 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.

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-après 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

À 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 entraîner 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 traitements. 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'utilisteurs, 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[9]. 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)[10].

Finalement, nous disposons d'unités 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)[11], de nature avant tout heuristique, parvient à dépister pour toute propositon, le thème et le propos, des indications sur des compléments verbaux et plusieurs types de

---

[9] 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.

[10] 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.

[11] 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.

détermination nominale. Le second (ALSF)[12], présentement 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 unités 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[13] 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) basées 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.

Dans le cas des unités 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

---

[12] 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.

[13] 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.

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[14] sont des mesures empiriques permettant d'apprécier la difficulté ou la facilité de lecture, de compréhension 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 SACAO

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 matériau textuel, que dans les multiples usages du texte au sein des organisations. Nous avons opté 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 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'assistées, 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 méthodologie pour l'usage intégré de procédures d'aide à la lecture se traduit dans une environnement qui permet la gestion stratégique de ces moyens. L'usager doit pouvoir choisir librement les procédures qu'il retiendra, choisir également les paramètres qui seront activés dans ces dernières. 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,

---

[14] Ces indices sont discutés en détail dans le texte de François Richaudeau, Le langage efficace, Paris, C.E.P.L., 1973, 300 p.

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 équipe dont chaque membre poursuit, par ailleurs, une activité relativement indépendante dans son champ de spécialisation, a dû conçevoir 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 consiste 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 procédures 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 descritions GDSF de la structure thématique des textes d'un corpus de discours politiques. Sur la base de cette validation, certains sous-ensembles de prodédures, enrichis de nouveaux développements, sont utilisés pour établir une description arborescentes 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 destinés à documenter le système et des fiches d'utilisation réservés 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. Systemes-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, pp. 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 possibilités. 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 -Generation 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-Éditions.

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 "Knowledge 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-Étienne, I.U.T. de Saint-Étienne, 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, 67 pages.
Marandin, J.M., (1988) "A propos de la notion de thème de discours. Éléments 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.
Plante, P., (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.