Modélisation informatique de structures dynamiques de segments textuels pour l'analyse de corpus

Thèse de doctorat en Sciences du langage, sous la direction de Jean-Marie VIPREY et la codirection d'Yves MARCOUX

Présentation de la thèse lors de la soutenance du 10 janvier 2011

D'abord merci aux membres du jury de m'accompagner dans cette dernière étape de ce travail de thèse. Un merci particulier à Jean-Marie Viprey et à Yves Marcoux qui ont dirigé la thèse, à Lou Burnard et André Salem qui ont agi comme comme rapporteurs. Et merci à tous ceux et celles qui participent à cette soutenance à Besançon, à Paris et à Montréal.

1- L'origine du projet de recherche

Le titre de la thèse correspond à une problématique de recherche qui remonte à 1993. J'agissais alors comme chef de projet dans une recherche visant à évaluer le potentiel d'un ordinateur à traitement parallèle. À l'époque, le logiciel SATO d'analyse de texte assistée par ordinateur, sur lequel je travaillais depuis les années 1970, tournait sur de *gros PC* fournis par le commanditaire. Ces *gros PC* comportaient 8 méga-octets de mémoire, un processeur Intel 486 et le système d'exploitation DOS.

Dans cet environnement, c'était déjà tout un exploit de disposer d'un programme d'analyse de texte offrant des dispositifs d'annotation. Or, comme nous l'écrivions à l'époque, « la généralisation prévisible des normes de marquage de type SGML va poser avec acuité la question de la représentation informatique, à des fins de traitement et d'analyse, des segments marqués. » Depuis, XML a remplacé SGML, mais la pertinence du propos demeure

la même, plus de 15 ans plus tard.

2- Nos questions de recherche

Si le titre de la thèse en explique l'origine, il est bien insuffisant pour en décrire le contenu. Cette thèse, en effet, a la particularité de s'inscrire au terme d'une trentaine d'années de développement d'outils et de méthodes en analyse de texte assistée par ordinateur. Il est donc apparu essentiel de tracer les grandes lignes de cette évolution afin d'inscrire nos choix théoriques et nos stratégies informatiques dans leur contexte historique.

Les dernières décennies ont été traversées par de multiples influences au gré des changements de paradigme de recherche. Le modèle SATO a traversé ces vagues en maintenant les idées maitresses qui l'ont fondé. Voici les questions de recherche auxquelles la thèse tente de répondre.

- Quelle était la pertinence de ce modèle aux différentes époques de son histoire?
- Jusqu'à quel point ce modèle était-il, ou pas, avant-gardiste, dans le contexte historique, et précurseur de pratiques scientifiques devenues aujourd'hui courantes?
- Quelles sont les insuffisances du modèle par rapport aux nouvelles questions et aux nouvelles possibilités de l'utilisation de l'ordinateur en sciences du langage?
- Dans ce contexte, le modèle existant garde-t-il sa pertinence?
- Et, si c'est le cas, est-il possible de faire évoluer le modèle pour qu'il prenne en compte les tendances nouvelles de recherche dans le domaine?

Pour répondre à ces questions, on devait aborder autant les questions de méthodologie de la recherche en analyse textuelle que les questions de modélisation informatique et d'algorithmique. Un long chapitre de la thèse est consacré à la présentation d'un certain nombre de travaux qui montrent le développement et la méthodologie de SATO dans divers contextes applicatifs et de recherche. Contentons-nous ici de mentionner quelques dates à titre de repères historiques.

Le modèle SATO actuel a été conçu au début des années 1980. Rappelons que la norme SGML date de 1986. Le développement de la version HTML a débuté en 1996 alors que

l'annonce publique du Web date de 1991 et que la norme XML est publiée en 1998. La version serveur déployée en grappe de traitement a été développée au milieu des années 2000.

3- Quels modèles de calcul pour quelles analyses textuelles?

Si notre objectif, au cours de toutes ces années de recherche, a été de produire un modèle informatique, bien fondé du point de vue formel et performant au niveau de son implantation, il est clair que les spécifications à la base du modèle ont été motivées par la nécessité de répondre à des besoins spécifiques manifestés par une certaine pratique d'analyse de texte.

Quels modèles de calcul pour quelles analyses textuelles? C'est cette question que nous posons au début de la thèse. En particulier, nous y reprenons quelques-unes des idées maitresses de l'analyse de discours pour en dégager les implications en termes de modélisation informatique. Guilhaumou, dans *Discours et archive publié* en 1994, écrivait que dans l'analyse de discours, on s'intéresse à la « *manifestation de la langue dans la communication vivante ».* On pourrait aussi dire qu'on s'intéresse à la manifestation du social dans la langue. La notion plus récente de *philologie numérique* actualise le débat sur les traditions fondatrices de l'analyse de discours, en posant la question des rapports entre texte et discours à l'ère du document numérique.

En effet, comme l'écrivait Jean-Marie Viprey dans un article de 2005, l'objet concret de nos analyses, ce sont les artefacts du discours tels qu'ils se manifestent dans des textes. Le « *texte* est un mode opératoire sur le *discours »,* écrit Jean-Marie Viprey, tout en soulignant que « le texte n'est pas un objet en soi, mais une phase vers l'objet fondamental des sciences humaines qu'est le *discours*. » (Viprey, J.-M. 2005). De fait, ce qui sera scruté à la loupe des méthodes informatisées de l'analyse textuelle, notre objet matériel en somme, ce n'est pas LE texte, mais le *corpus,* collection raisonnée de textes établie à des fins d'analyse de la *communication vivante* inscrite dans un contexte historique et social donné.

Comme l'écrivait Michel Pêcheux,

> L'analyse de discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant « le » sens des textes, mais seulement construire des procédures exposant le regard-lecteur à des *niveaux opaques à l'action stratégique d'un sujet*

> [...]. L'enjeu crucial est de *construire des interprétations* sans jamais les neutraliser ni dans le « n'importe quoi » d'un discours sur le discours, ni dans un espace logique stabilisé à prétention universelle. (Pêcheux, 1984: 15,17, cité par Maingueneau 1987:6).

L'objectif de l'analyse de discours, à travers l'analyse textuelle, est donc de construire des *interprétations-explicitations* s'appuyant sur des procédures documentées permettant d'exposer, selon notre interprétation, le fonctionnement du discours tel qu'il se manifeste dans le corpus soumis à l'analyse.

La *polyphonie du texte* et l'espace pluraliste du discours dans lequel il s'inscrit sont invoqués pour soutenir la nécessité d'une pluralité aussi des stratégies de lecture. Ces idées maitresses de l'analyse de discours, reprises depuis dans le domaine de la linguistique textuelle, se sont opposées historiquement à une certaine vision *étapiste* très présente à l'époque dans le domaine du TALN et qui imposait une démarche obligée allant de la morphologie au syntagme jusque, éventuellement, à une grammaire du texte.

Depuis ce temps, les choses ont évolué. Dans *Les linguistiques de corpus,* publié en 1997, Benoît Habert, Adeline Nazarenko et André Salem indiquent qu'on assiste présentement à un *profond changement de paradigme*. Auparavant, la primauté était donnée à la modélisation destinée à formaliser le savoir humain. Le courant dominant était résolument *anti-empirique, anti-numérique et pro-symbolique*. Aujourd'hui, au contraire, on voit de plus en plus la constitution de corpus annotés et de ressources langagières comme une condition au développement de la recherche.

Ces rappels des théories et méthodologies ont pour objectif de nous aider à cerner notre propre objet : l'analyse de texte assistée par ordinateur, l'ATO. Ce qui distingue l'ATO du simple commentaire interprétatif, c'est la construction de dispositifs expérimentaux. Comme l'indique Benoît Habert, le dispositif expérimental est un « montage d'instruments, d'outils et de ressources destinés à produire des « faits » dont la reproductivité et le statut (l'interprétation) font l'objet de controverses » (Habert 2005).

D’un point de vue technique, le dispositif expérimental se matérialise par des procédures de calcul transparentes et reproductibles, et par des procédures assistées de catégorisation dont

la trace doit être explicite. Ainsi, la controverse de l'interprétation pourra s'appuyer sur la discussion serrée des procédures de constitution des faits sur lesquels elle s'appuie.

Le survol dans la thèse des théories et méthodes de l'analyse textuelle est complété par le point de vue informatique sur les *données semi-structurées,* les langages de balisage et des façons de faire pour annoter les corpus de diverses manières.

4- SATO : un modèle informatique pour la construction de dispositifs expérimentaux.

C'est sous l'éclairage de ces théories du texte et de l'analyse textuelle que la thèse présente le modèle à la base du logiciel SATO actuel. Le logiciel est conçu comme une plateforme permettant la mise en place de dispositifs expérimentaux. La nature essentiellement itérative du processus d'analyse-interprétation prend la forme d'un dialogue avec l'artefact textuel dont on gardera une double trace : procédurale et déclarative. L'annotation transforme le matériau textuel par un processus de qualification non destructive alors que le cahier de procédures, le journal et les scénarios qui accompagnent cette qualification, en fournit la clé de lecture.

Le modèle de données utilisé dans le logiciel SATO repose sur la décomposition du flux de caractères, correspondant à la matérialité linéaire première du texte, afin de représenter le texte dans un espace à deux dimensions. D'un point de vue géométrique, le corpus se présente donc sous la forme d'un plan distinguant l'axe lexical (dit paradigmatique), le vocabulaire du texte, de l'axe textuel (dit syntagmatique) traduisant la séquentialité du texte, les contextes d'énonciation.

Sur le plan informatique, cette matrice textuelle rappelle la relation classique classe – instance et l'héritage simple. Chaque mot, dans la séquence textuelle, est l'instance d'une classe lexicale unique et hérite dynamiquement de toutes les propriétés de la classe lexicale. L'intérêt du modèle se mesure à ce qu'il autorise au niveau des fonctions d'annotation. Dans le SATO actuel, ces fonctions d'annotation, appelées propriétés, permettent de donner des valeurs à des variables décrivant les attributs des classes lexicales ou des occurrences (instances).

Pour une version donnée du corpus correspondant à un état donné de l'analyse, le modèle *lexique/occurrences,* couplé au système des propriétés, permet d'appliquer sur le lexique des

ressources lexicales préexistantes et d'enrichir les classes lexicales en condensant sur ces classes des traits provenant des contextes. À l'inverse, il est possible de préciser en contexte des traits lexicaux. Au-delà de l'héritage dynamique issu de la relation lexèmes/occurrences, il est possible de créer de nouvelles propriétés par héritage statique selon chacun des deux axes : texte/lexique, lexique/texte, texte/texte, lexique/lexique.

La construction par SATO du plan *lexique/occurrences* ne repose pas sur une *reconnaissance* de classes conceptuelles associées à une ontologie pré-existante, par exemple. Il s'agit d'un processus de calcul dont l'intrant principal est le fichier corpus complété par des instructions de traitement simple permettant de préciser le statut des diverses séquences de caractères utilisées dans le fichier électronique. Le découpage du flux textuel en unités lexicales n'est pas une opération de lexicalisation au sens linguistique du terme, c'est-à-dire une qualification des chaines de caractères en tant qu'unités terminologiques référant à notre connaissance du monde en tant que communauté agissante à travers le discours. En fait, cette qualification est déjà une opération d'analyse de discours à venir ou, à la limite, déjà inscrite dans le flux textuel suite à un marquage préalable.

Dans notre perspective, rappelons-le, l'objectif est bien de fournir un appareillage, un cadre expérimental pour bâtir des dispositifs de lecture qui vont permettre de qualifier les données selon divers points de vue. La consolidation lexicale est une opération d'analyse qui va pouvoir bénéficier des ressources et procédures informatiques sous gouverne du lecteur-analyste. Le corpus ainsi enrichi pourra donner lieu à une nouvelle version du document numérique et à une mise en forme plus pertinente du plan *lexique/occurrences*. La notion de *pertinence* renvoie, bien sûr, à l'espace social nous rappelant qu'une analyse de discours est aussi une activité de discours.

Comme l'écrit Maingueneau (1987), « Étant donné le statut de l'analyse de discours, on ne peut pas se contenter d'*appliquer* de manière aveugle des protocoles méthodologiques à des corpus. À chaque fois, il faut mener une réflexion spécifique pour construire, de manière interactive, le corpus et son mode d'investigation. »

En conséquence, le modèle de calcul que nous privilégions a des allures de *tableur textuel*, pour reprendre les mots d'un chercheur belge rencontré dans les années 80. L'ergonomie de la plateforme informatique repose sur trois dispositifs qui interagissent.

Il y a d'abord une interface interactive qui, dans sa version actuelle, utilise des formulaires HTML pour guider l'usager dans l'élaboration de ses commandes. Les formes lexicales et les occurrences affichées à l'écran, suite à l'exécution d'une commande SATO, seront des hyperliens donnant accès à un menu de catégorisation qui permet de dévoiler l'annotation et de la compléter. Ce dispositif permet aussi de naviguer entre le lexique et les contextes et d'un contexte à l'autre.

Le deuxième dispositif de l'interface SATO est le langage de commandes qui fait en sorte que toute manipulation interactive produit une commande explicite que l'utilisateur pourra saisir et copier dans des fichiers pour constituer des scénarios de commandes.

Le troisième dispositif caractérisant l'ergonomie de SATO est la production automatique d'un journal dressant un historique complet des interventions sur le corpus. L'examen du journal permet un retour critique sur une démarche analytique effectuée dans les conditions plus spontanées de la phase exploratoire. C'est souvent à partir de l'examen du journal que l'on composera les scénarios qui vont permettre de cristalliser les stratégies de recherche les plus productives.

En conclusion, ce qui signe l'originalité d'une analyse de texte avec SATO, c'est en bonne partie cette démarche itérative qui consiste à explorer le corpus en mode interactif, avec l'assistance des outils textométriques, à revenir sur la démarche suivie à travers l'examen du journal, pour finalement cristalliser les stratégies fécondes sous la forme de scénarios.

5- La dimension documentaire de l'ATO

Le modèle d'annotation sur lequel repose SATO est congruent avec le concept actuel de balisage. Le balisage est le procédé technique qui permet de marquer le corpus en distinguant le texte brut de l'information ajoutée qui permet d'en expliciter les composants et la structure. Ainsi, le format d'entrée du logiciel est le même que le format de sortie, le texte analysé au moyen de SATO se trouvant enrichi en sortie par de nouvelles annotations encadrées de balises.

Depuis la publication de la norme SGML en 1986 et, plus encore, de la norme XML en 1998, le balisage utilise une syntaxe standard augmentée de contraintes qui en précisent la syntaxe pour tenir compte de la sémantique particulière des données échangées dans une

communauté donnée. Un exemple de ces schémas particuliers nous est donné par les propositions de la *Text Encoding Initiative* (TEI), ce consortium qui examine tout particulièrement des façons de faire adaptées aux corpus utilisés en sciences humaines.

Étant donné que le modèle et la syntaxe de SATO ont été élaborés au début des années 1980, le balisage utilisé pour marquer les corpus en format SATO ne suit pas la norme XML. Cependant, la conversion vers la norme XML est facile. Dans le cadre du réseau de collaboration ATONET, nous avons proposé de recourir à un sous-ensemble limité de balises conformes aux propositions de la TEI pour traduire les *syntaxes propriétaires* des logiciels Alceste, DTM, Lexico et SATO. Il s'agit des propositions Sacacomie, du nom du lieu où s'est tenu le séminaire de travail d'ATONET sur les formats d'échange.

La proposition dite avancée proposait le découpage du texte en mots en utilisant la balise TEI <*w*> pour encadrer la chaine de caractères représentant le mot. Ces éléments *w,* accompagnés de l'attribut standard *xml:id,* permettent d'identifier simplement et de façon unique chacune des occurrences du corpus. Pour reprendre le terme suggéré par André Salem, on dispose ainsi d'un texte *tramé* qui fournit un système de pointage permettant de référer précisément et simplement aux unités du texte établies par le lecteur-analyste.

La prochain pas a été de proposer un schéma documentaire pour la constitution de dépôts de données adaptés à la constitution de corpus de recherche. L'objectif est de fournir des lieux pour publier et documenter les corpus et les annotations produites par les chercheurs. Ainsi, d'autres chercheurs pourront réutiliser en tout ou en partie ces ressources pour leurs propres analyses, alimentant ainsi le débat scientifique. Pour faciliter ce mouvement de construction/déconstruction des corpus de recherche, le modèle privilégie, sans l'imposer, des formes de balisage dites *débarquées* qui permettent de publier les annotations sous forme de documents d'annotation dotés de leur propre fiche de métadonnées.

6- La problématique de l'annotation structurelle

En termes généraux, on entend par *segment textuel* une suite continue de mots. Il peut s'agir d'un chapitre, d'une phrase, d'une tirade, d'un syntagme, etc. Le modèle de données de la version actuelle de SATO soutient déjà une forme de *segment dynamique* résultant de la catégorisation de séquences d'occurrences. Les segments sont généralement construits à la

volée pour les besoins du calcul. Lorsque plusieurs mots consécutifs partagent une même valeur de propriété, on peut procéder à une *mise en évidence*, au sens algébrique, faisant ainsi apparaitre une suite de mots comme une entité autonome du point de vue catégoriel. Mais, en fait, dans sa représentation du texte, SATO ne connait pas le segment en tant qu'objet primitif et pérenne. Il s'agit plutôt de *segments virtuels* en ce sens qu'il ne se matérialisent pas dans une structure informatique explicite.

L'absence d'un modèle de données dédié à la gestion des segments textuels a pour conséquence qu'il est difficile de représenter la macrostructure d'un document au sens d'un modèle hiérarchique de type SGML ou XML. Plus encore, la construction d'un texte, on le sait, fait appel à des relations sémantiques, stylistiques, narratives, argumentatives, etc. qui vont bien au-delà de la structure formelle des documents. Ces besoins variés de représentation des structures textuelles requièrent donc que le segment puisse avoir une existence autonome avec ses propres règles.

L'annotation structurelle vise à aller au-delà de l'annotation simple sur les formes lexicales et les occurrences pour marquer des *configurations,* c'est-à-dire des *motifs structurels* qui traversent le texte ou structurent le lexique.

Ce que nous proposons dans la thèse est d'ajouter au modèle existant (lexique/occurrences avec ses propriétés) un nouvel espace d'annotation prenant la forme de graphes représentant les relations entre segments textuels. Allant au-delà des segments dynamiques actuels découlant de la catégorisation des occurrences, l'annotation structurelle vise à organiser ces segments en unités structurelles. Ces unités peuvent servir de nouveaux intrants aux algorithmes de textométrie. Mais, ce sont aussi des clés explicitant le fonctionnement discursif et permettant de multiples navigations interprétatives sur la surface du texte.

L'objectif de la thèse est de proposer des formalismes simples permettant de constituer des documents d'annotation qui réfèrent aux éléments du texte par une utilisation des mécanismes de pointage suggérés par la TEI. Pour illustrer cette proposition, nous montrons diverses mises en forme XML-TEI d'analyses textuelles de discours tirées du livre *La linguistique textuelle* de Jean-Michel Adam. Ces analyses vont du commentaire libre sur un récit de Borges à des analyses précises couvrant, en particulier, une analyse détaillée de la

structurelle compositionnelle du récit de Borges et un exemple d'analyse fonctionnelle de la phrase (relation thème-rhème).

7- Quelques hypothèses pour l'implantation de l'annotation structurelle

Après avoir soutenu qu'il est possible d'utiliser les recommandations de la *Text Encoding Initiative* (TEI) pour marquer de façon formelle, et relativement intuitive, les opérations d'annotation structurelle, il fallait, au moins de manière prospective, aborder la question des *modèles informatiques pour l'exploitation de l'annotation structurelle*. Ce dernier chapitre de la thèse prendra notamment appui sur le principe de l'annotation en couches multiples, sur les recherches autour de l'exploitation informatisée des *arbres linguistiques* et du monde en évolution des outils XML.

Mais, la gestion de cette forêt d'arbres d'annotation pose, dans le contexte de l'ATO, des exigences qui se distinguent de l'exploitation de corpus d'arbres syntaxiques établis.

- D'abord, on doit pouvoir révéler l'annotation structurelle à partir du bas, des occurrences dans un contexte de lecture. On ne peut se contenter du modèle traditionnel de la fouille à partir de la racine des arbres.
- On doit donc soutenir correctement à la fois l'exploration qui donne priorité à la relation de séquentialité dans les énoncés et celle qui donne priorité à la relation de dominance dans les arbres structuraux.
- Dans le cadre d'une annotation dynamique de corpus, la fonction de mise à jour des structures doit aussi être optimisée.
- Finalement, les questions d'interfaces sont très importantes pour éviter que l'utilisateur non informaticien soit découragé par un langage de requête trop complexe. Un modèle possible est inspiré du *Query by Example* (QBE) qui consisterait, dans ce cas-ci, à décrire des arbres ou des graphes partiels correspondant à un prototype de données. Le programme devra compiler ces arbres partiels pour produire une requête dans le formalisme voulu. Bird et Lee parlent de *Query by Annotation* (QBA).

Une question qui nous préoccupait particulièrement dès le début de notre démarche de recherche est la différence entre des langages de requête basés sur l'énoncé de contraintes

algébriques par rapport aux langages, tels *XPath* et *XQuery,* qui s'expriment en termes de chemins dans l'arbre. Il ressort en fait que ces deux formalismes peuvent être équivalents, ce qui permet, par exemple, de représenter les arbres dans des bases de données relationnelles. Le choix du formalisme se pose davantage en termes de stratégies d'implantation qu'en termes de puissance de représentation.

8- Sommaire

Pour terminer, reprenons nos questions de recherche sous la forme, cette fois de propositions que nous croyons être en mesure de soutenir.

1. Nous soutenons que le modèle de données et de traitement qui a présidé à la conception de SATO au début de 1980 était et demeure pertinent.
2. Le modèle de balisage pré-SGML et pré-XML élaboré à l'époque peut être aisément converti vers les formats actuels XML-TEI. Les documents en format TEI peuvent prendre place dans un cadre documentaire permettant de gérer les textes sources, les corpus et les documents d'annotation.
3. L'insuffisance principale du modèle actuel de SATO tient à l'absence de support des modèles hiérarchiques de balisage : structure formelle des documents et annotation structurelle.
4. La prise en charge de l'annotation structurelle peut se superposer au modèle lexique/occurrences sous la forme de documents d'annotation admissibles aux outils XML. L'annotation structurelle offrira une meilleure prise en charge des phénomènes de la *textualité* et offrira de nouvelles possibilités de navigation hypertextuelle.
5. Le code informatique de SATO, malgré ses insuffisances, peut être revu sous l'éclairage de la programmation orientée objet afin de produire un code source documenté. L'hypothèse de l'offrir en code ouvert est à l'ordre du jour.
6. La refonte des interfaces pourrait être facilitée par la refonte du code informatique et l'ajout de modules dédiés à la gestion de l'annotation structurelle.

Merci de votre attention.

Références des citations

Guilhaumou et coll., 1994. Guilhaumou, Jacques; Maldidier, Denise; Robin, Régine. *Discours et archive*, Mariga, Liège, 1994. ISBN 2-87009-520-1.

Habert, 2005. Habert, B. *Instruments et ressources électroniques pour le français* Ophrys Paris ISBN 2-7080-1119-7 p.2., 2005.

Habert et coll., 1997. Habert, Benoît; Nazarenko, Adeline; Salem, André. *Les linguistiques de corpus*. Armand Colin/Masson, Paris 1997, Collection U, série «Linguistique», ISBN 2-200-01775-8, 240p.

Maingueneau, 1987. Maingueneau, Dominique. *Nouvelles tendances en analyse de discours*. Hachette, Paris 1987, ISBN 2-01-012116-3.

Pêcheux, 1994. Pêcheux M. Sur les contextes épistémologiques de l'analyse du discours. *Mots*, no. 9, oct 1984, Presses de la Fondation nationale des sciences politiques.

Viprey, 2005. Viprey, J.-M. Philologie numérique et herméneutique intégrative. In *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité* dir. Jean-Michel Adam & Ute . Slatkine (pp. 51-68).