Cahiers de recherche sociologique

L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une modélisation des procédés analytiques fondamentaux

Computer-assisted textual discourse analysis and textometric softwares: Critical and prospective reflections via a modeling of fundamental analytical procedures

El análisis textual de discursos asistido por ordenador y los softwares textométrico: reflexiones críticas y prospectivas a partir de una modelización de los procedimientos analíticos fundamentales

Élias Rizkallah

Number 54, Winter 2013

Regards croisés sur l'Analyse du discours

URI: https://id.erudit.org/iderudit/1025996ar
DOI: https://doi.org/10.7202/1025996ar

See table of contents

Publisher(s)

Athéna éditions

ISSN

0831-1048 (print)
1923-5771 (digital)

Explore this journal

Cite this article

Rizkallah, É. (2013). L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une modélisation des procédés analytiques fondamentaux. *Cahiers de recherche sociologique*, (54), 141–160. https://doi.org/10.7202/1025996ar

Article abstract

Noting the focus of the French tradition in discourse analysis on textometric softwares, the author attempts to model the fundamental processes underlying the analyst-text interaction distinguishing the modes, operations, dimensions, granularity, contextuality and temporalities of the process, and so, with or without recourse to computer processing. Based on this modeling, textometric softwares show that the assistance of the researcher is often a matter of *bringing to display*, through queries-retrieval, automatic assignments and representations, textual and extra-textual data, but rarely a matter of *assisting* the researcher in his *working of the text* (e.g. customized annotations, multiplicity of reading layers, corpus evolution) to produce and construct meaning by his or her analysis traces in an integrated environment. The origins of this trend are discussed as well as directions for future developments.

L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une modélisation des procédés analytiques fondamentaux

Élias Rizkallah

Introduction

Les chemins de l'Analyse du Discours (AD), du moins l'École française, et de l'Analyse de Texte assistée par Ordinateur (ATO) se sont croisés il y a maintenant plus de 30 ans sous la forme d'une solution logiciel (DEREDEC, UQAM) et depuis plus de 40 ans sous forme de réflexions analytiques et de tentatives d'implantations informatiques ponctuelles, voire collaboratives. Dans le monde du développement logiciel, et non celui des idées, une telle période est assez considérable pour que l'usage actuel en ATO soit encore si divisée en sciences sociales où la majorité des analystes du discours utilisent encore quasi exclusivement soit l'approche textométrique soit l'approche d'annotation sur mesure. Notre propos ici n'est pas de plaider pour une combinaison de ces deux familles, comme d'aucuns le préconisent[1] et

1. Marianne Jørgensen et Louise Phillips, *Discourse analysis as theory and method*, Londres, Sage, 2002. Margaret Wetherell, « Positioning and interpretative repertoires : Conversation analysis and post-structuralism in dialogue », *Discourse & Society*, 1998, vol. 9, n° 3, p. 387-412.

le critiquent[2], mais bien d'examiner de manière critique l'adhésion de l'AD française aux seuls logiciels de textométrie (LT) de la famille de l'ATO. Pour ce faire, notre réflexion partira de l'interaction analyste-texte, avec ou sans assistance informatique, s'inscrit dans le cadre général de l'ATO et vise l'objectif méthodologique de relater, à la lumière des pratiques d'analyses textuelles, quelques limites de l'exploration des textes par les LT libres ou gratuits francophones (Lexico3, DTM, Hyperbase, Iramuteq, TXM, Coocs, Tropes, etc.).

Il serait prudent d'indiquer notre posture dans ce texte par rapport à l'AD et à l'articulation des notions de discours et de texte. Pour nous, le discours comprend deux faces, un objet théorique et un objet empirique. Le premier se situe dans un espace organisé de signes mettant en jeu des locuteurs véhiculant des représentations dans un contexte sociohistorique, alors que le second renvoie à du texte comme trace de l'objet théorique, les deux s'articulant par la schématisation, comme processus et produit, proposée par Grize[3]. En effet, dans l'AD, la recherche et la construction de catégories de description n'est pas une fin en soi mais un préalable à toute analyse, à laquelle s'ajouteraient la place primordiale de la matérialité discursive et les cinq postulats de Grize : l'activité discursive, la situation d'interlocution, les représentations en jeu, les préconstruits culturels et la finalité de l'acte discursif.

Ce texte présentera d'abord une description globale des dimensions et processus fondamentaux pour l'analyste face à ses données discursives, pour ensuite évaluer, à partir de cette description, l'état des LT sur le sujet, et entrevoir les perspectives d'avenir possibles dans les croisements entre l'ATO et l'AD (française ou autre). Plus explicitement, après une brève cartographie de l'ATO, c'est à partir d'une description générique de l'interaction analyste-texte, soit le « travail » de la donnée textuelle, qu'il sera possible de déterminer en quoi l'état actuel de l'ATO dans l'AD française peut ne pas répondre convenablement à la pratique des chercheurs en AD.

Brève cartographie des familles de logiciels d'ATO pour l'AD

Dans l'état actuel de l'ATO pour l'AD, on peut diviser le champ en deux familles d'usages : les logiciels de *Text Mining* (TM) et de Textométrie (Alceste, CooCS, DTM-VIC, Hyperbase, IBM SPSS Text Analytics, Iramu-

2. Reiner Keller, « L'analyse de discours du point de vue de la sociologie de la connaissance. Une perspective nouvelle pour les méthodes qualitatives », *Recherches Qualitatives, hors-série*, 2007, p. 287-306.
3. Jean-Blaise Grize, « Argumentation et logique naturelle », dans Jean-Michel Adam, Jean-Blaise Grize et Magid Ali Bouacha (dir.), *Texte et discours : catégories pour l'analyse*, Dijon, Éditions universitaires de Dijon, 2004, p. 23-27.

teq, Lexico, Rapid Miner, SAS Text Miner, T-lab, TXM, etc.)[4] d'une part, et les logiciels d'annotations « manuelles »[5] d'unités textuelles (AQUAD, ATLAS.ti, CATMA, MaxQDA, NVivo, QDAMiner, Qualrus, RQDA, etc.)[6], d'autre part.

Le justificatif de cette division en deux familles d'usages tient d'abord au fait que l'unité analysée, l'objet ou la trace à la base des interprétations est différente : pour les LT, on utilise surtout l'unité textuelle (mot, segment répété, etc.), alors que pour les logiciels d'annotation, la couche d'écriture (socio-sémantique, énonciative, rhétorique, etc.) au-dessus de l'unité textuelle qui est au fondement des interprétations, au-delà de l'unité textuelle. Deuxièmement, il existe également une différence quant à l'emploi de procédures automatisées autant pour les multiples représentations des données que pour l'assignation des annotations, les logiciels de textométrie étant plus enclins à des traitements automatisés (statistiques multidimensionnelles, traitement automatique des langues [TAL], etc.) que les logiciels d'annotation. Enfin, dans cette deuxième famille de logiciels d'analyse dite « qualitative », d'allégeance « interprétative », mieux connue sous le nom de CAQDAS, il importe de noter que, à la différence de l'AD francophone, les usagers prennent en général leur distance par rapport au statut matériel de la langue et à la dimension linguistique (énonciation, paratextualité, pragmatique, etc.) de leurs données textuelles[7] pour davantage se consacrer à leurs valeurs sociosémantiques.

Cette division en deux familles reste tout de même délicate, car plusieurs passages entre les deux familles se sont récemment[8] effectués. Par exemple, d'un côté T-Lab, réserve de plus en plus de place pour les dictionnaires personnels, et de l'autre, les grandes suites de CAQDAS (QDAMiner, Nvivo, etc.) offrent de plus en plus des traitements statistiques multidimensionnelles

4. Même si nous regroupons le *Text Mining* et la textométrie, désormais désignée par logométrie, il demeure une différence qui se situe principalement sur les plans de la visée et du mode de connaissance : les premiers sont plus orientés vers la confirmation/prédiction, l'apprentissage machine (résumé automatique, extraction des connaissances) et le travail sur des corpus hétérogènes de taille gigantesque alors que les seconds sont plus orientés vers l'exploration de corpus de recherche définis (même en linguistique de corpus) avec retour constant au contexte d'énonciation. Là aussi, il existe aussi une différence, sur laquelle nous ne nous étendrons pas, entre les pratiques anglophones et francophones. Quoi qu'il en soit, sur le plan des unités d'analyse, ils utilisent tous deux l'unité textuelle ou lexème. Comme les premiers sont moins utilisés en AD et en sciences humaines, ils ne seront pas traités dans le présent article.
5. L'emploi des guillemets indique qu'il ne s'agit pas d'annotations automatiques en fonction d'un algorithme ou schème extérieur à la problématique du chercheur (analyseur morpho-syntaxique, désambiguïseur) mais d'annotation sur mesure, car il va sans dire que dans tous les cas qui nous concerne ici, l'annotation se fait par le biais d'une application informatique.
6. Dans le champ des recherches qualitatives cette famille de logiciels est souvent désignée par CAQDAS (*Computer Assisted Qualitative Data Analysis Software*)
7. R. Keller, *op. cit.*
8. Il faut quand même signaler que les deux logiciels (SATO et SEMATO) du centre ATO (UQAM) permettent depuis quasiment 30 ans une articulation entre les deux familles mentionnées.

sur les annotations, voire sur les unités textuelles. Bref, le lecteur aura compris que des usagers (chercheurs du centre d'ATO, UQAM), voire des solutions logicielles (Cassandre, Nooj, SATO, SÉMATO, etc.), peuvent très bien se situer entre ces deux familles, mais qu'en gros cette distinction en ATO reste opérationnelle pour les chercheurs en AD pour la sociologie voire pour plusieurs disciplines connexes (sciences politiques, psychologie, communication, etc.). La question qui se pose c'est l'usage quasi exclusif des LT par les chercheurs en AD, et ce, quand la voie de l'informatique est empruntée.

Esquisse de méta-description des interactions analyste-texte : entre lecture et écriture[9]

Pour situer cette dernière question, il faut d'abord décrire les procédés fondamentaux utilisés lors d'une analyse de texte. C'est pourquoi il s'agira dans cette section de procéder à une méta-description dimensionnelle de la situation d'interaction analyste-texte où l'ATO est en grande partie mise en suspens[10], même si elle inspire quelques-unes des illustrations de cette méta-description. Pour l'exposition de cette méta-description, nous établirons en premier lieu des distinctions/définitions de concepts génériques illustrés par des pratiques communes en analyse textuelle pour ensuite aller dans le cœur des procédés analytiques propres.

D'abord, l'*objet* observé. Nous distinguerons entre le *mode*, l'*opération* et le *moyen*[11] d'appréhender un *objet* où ce dernier se ramène à une trace d'activité discursive qui, en dernière instance, est le matériau d'analyse indépendamment du niveau d'abstraction. Or, cet objet peut être de différentes *natures* et s'appréhender selon différents niveaux de *granularité* et de *contextualité*. En effet, l'objet observé peut être une donnée textuelle brute (un mot ou ensemble de mots dans un corpus), l'une de ses multiples représentations analytiques (tableau de fréquences des mots, projection des mots dans un espace euclidien, etc.), ou bien une de ses annotations (thème, catégorie,

9. Nous incluons aussi tout dispositif d'aide à la lecture ou à l'écriture.
10. De telles initiatives de méta-description se retrouvent dans les documents classiques sur les CAQDAS mais souvent sous une forme de fonctionnalités techniques de logiciels ou bien d'objectifs poursuivis et non suivant une vision dimensionnelle indépendante d'une implantation informatique. Pour les CAQDAS, voir Nigel Fielding et Raymond M Lee, *Computer analysis and qualitative research*, Londres, Sage, 1998 ; Udo Kelle (dir.), *Computer aided qualitative data analysis: theory, methods and practice*, Londres, Sage, 1995 ; Eben A. Weitzman et Matthew B. Miles, *Computer programs for qualitative data analysis*, California, Sage, 1995. Pour une vision globale des pratiques francophones voir Christophe Lejeune, « Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative », *Recherches Qualitatives*, 2010, nº 9, hors-série [en ligne] Consulté le 1er décembre 2012, www.recherche-qualitative.qc.ca/revue/hors_serie/hors_serie_v9/HS9_Lejeune.pdf ; Jacques Jenny, « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 1997, vol. 54, nº 1, p. 64-122.
11. *Cf.* infra pour l'identification de ces trois éléments.

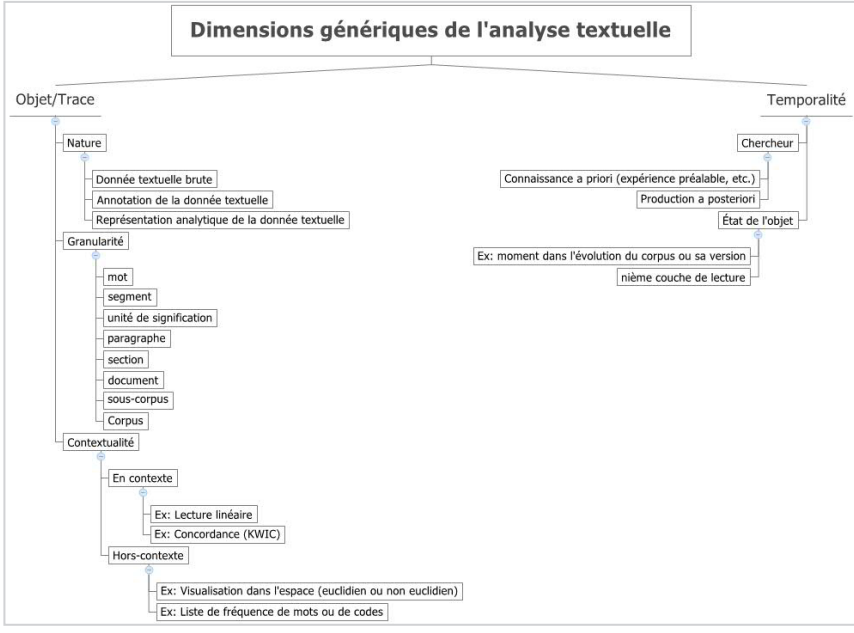
rubrique, code, modalité d'une variable, etc.). L'observation de l'objet varie aussi sur une échelle allant du caractère au corpus[12] lui-même, et ce, suivant un degré de contextualité que nous pouvons grossièrement diviser ainsi : en-contexte v/s hors-contexte. Par exemple, un tableau de fréquence des formes est l'exemple extrême d'un objet pris hors-contexte alors que son pendant en-contexte est la lecture linéaire du texte ; entre les deux, on peut songer à des situations moins extrêmes, comme la concordance (KWIC) d'un mot ou d'un segment ou la poly-cooccurence spécifique de plusieurs mots.

Ensuite, la *dimension temporelle* de l'activité d'analyse qui concerne autant les pré-connaissances du chercheur que l'état de l'objet observé. Les pré-connaissances du chercheur peuvent être d'ordres théoriques (hypothèses), empiriques ou issues de la recherche en cours (pistes de recherche), toujours en fonction de sa situation sociohistorique. Par exemple, le genre d'un texte préexisterait aux opérations de l'analyste, alors qu'une catégorisation (désignée plus loin par annotation) peut être *a priori* (dictionnaire de code préalable, fonctions grammaticales dans une phrase) ou *a posteriori* (i.e. émergeant au fur et à mesure des lectures flottantes du texte). Quant à l'état de l'objet observé, il peut se ramener à un moment déterminé de l'évolution de la constitution du corpus, à la nième itération de sa lecture ou à l'étape de maturation de sa grille d'analyse. En somme, les ingrédients employés par un analyste de texte sont traversés par une dimension temporelle selon le moment où l'acte analytique spécifique prend place, et ce, toujours en fonction de l'expérience (*a priori* v/s *a posteriori*) du chercheur. La figure suivante (figure 1) synthétise les dimensions génériques décrites.

Les dimensions génériques étant explicitées, il s'agit maintenant d'examiner les procédés analytiques propres. Face à un texte à analyser, les activités que déploie un chercheur peuvent être de plusieurs ordres selon qu'il soit en mode lecture d'un objet ou en mode écriture sur un objet. Cette distinction entre lecture et écriture n'est qu'analytique puisqu'au cours de l'action les deux modes participent au même processus itératif (lecture-écriture), celui de « travailler le texte », ou comme l'explicite Duchastel « on lit pour mieux réécrire […] Lire, c'est réécrire un nouveau texte[13]. » D'un côté, est désigné comme mode lecture un acte de consultation ainsi qu'une absence d'intervention sur la trace par le chercheur. D'un autre côté, une opération d'écriture désigne un acte de transformation de cette trace afin de solliciter une autre activité en mode lecture. Cela peut sembler simpliste, mais l'écriture

12. Le continuum peut être le suivant : caractère->mot->segment->unité de signification->paragraphe->section->sous-corpus->document->corpus.
13. Jules Duchastel, « Pour une méthodologie d'aide à la lecture et à l'écriture », Montréal, Office de la langue française/Société des traducteurs du Québec, 1991, vol. 1, p. 584.

Figure 1
Dimensions génériques de l'analyse textuelle



demeure ce qu'il y a de plus complexe dans le processus d'analyse et ce qui sert de base à sa validation concrète. Comme le dit Brooks dans un contexte d'administration publique : « c'est quand l'on écrit que les espaces apparaissent et que les manques de cohérence ressortent[14] ». Par ailleurs, que le chercheur soit en train de consulter (lecture) ou de transformer (ré-écriture), il peut le faire selon différents moyens ; *a priori*, il est possible de penser au triplet suivant : manuelle, automatique ou semi-automatique. Par exemple, la concordance d'un terme, voire multiterme, peut être découpée manuellement ou de manière semi-automatique à partir d'une sélection. De même, représenter les données par une analyse des correspondances à partir des coordonnées peut se faire manuellement ou automatiquement à l'aide d'une application graphique[15].

14. Frederick P. Brooks, *The mythical man-month : essays on software engineering*, Reading, Mass., Addison-Wesley Pub. Co., 1975, p. 111 (notre traduction).
15. Rappelons que le calcul d'une analyse des correspondances, soit la décomposition en valeur singulière, peut très bien se faire manuellement sur des matrices de petite taille, ce qui est très rarement le cas (e.g. des mots associés à des inducteurs) avec des données textuelles.

Dernier élément de la méta-description : les *opérations.* Que le chercheur soit en mode lecture ou écriture (réécriture), quelles sont les *opérations* possibles lors d'une analyse textuelle ? Sans être exhaustif, mentionnons les quatre opérations de base : découpage, annotation, représentation et contraste.

L'opération de *découpage* est un acte de sélection/séparation effectué autant sur le matériel textuel que les données extratextuelles. Dans le matériel textuel, nous incluons des unités d'analyse allant du caractère jusqu'au corpus entier ; bref, autant le texte brut que ces structures formelles *a priori* (section, paragraphes, etc.) ou *a posteriori* (genre/type de discours, etc.). L'extratextuel revient aux concepts de variables ou de métadonnées appliquées aux objets textuels de description (auteur, locuteur, date, époque, contexte sociohistorique, âge, sexe ou profession de l'enquêté-e, etc.). L'une des particularités des propriétés extratextuelles, c'est que les valeurs qui y sont assignées sont *a priori*, exhaustives, non vides et qu'à un même niveau de granularité, l'objet décrit aura la même valeur à travers le corpus. Pour nous l'opération de découpage se déroule surtout en mode écriture qui, de par sa segmentation, permet de servir l'opération de représentation qui fournit des plans de comparaison (différence v/s similitude) potentiels pour l'opération de contraste (*Cf.* infra). Or, elle peut également s'effectuer en mode lecture quand il s'agit de navigation ou de soumettre des requêtes à des données textuelles ou extratextuelles.

Pour ce qui est de l'*annotation*, il s'agit d'une opération qui consiste à enrichir, par définition ou par assignation, la trace par d'autres données (textuelles ou pas), qu'elles soient d'ordre descriptif ou explicatif, c'est-à-dire de lier une donnée par une autre donnée. Ainsi, une annotation peut être autant une définition d'une méta-donnée extra-textuelle que son assignation à une unité textuelle. À ce niveau, l'analyste est forcément en mode écriture, et plus souvent en réécriture. Rappelons que les annotations peuvent être *a priori* ou *a posteriori,* mais elles peuvent aussi suivre une ou plusieurs approches : socio-sémantique, morpho-syntaxique, énonciative, d'argumentation/rhétorique, etc. Comme pour le reste des objets ou traces, il faut donc prendre en considération ici la contextualité de l'annotation et la granularité de l'objet auquel elle est assignée. D'une part, l'annotation peut se faire en contexte[16], soit localisée à partir de sa situation d'énonciation (les moda-

16. Dans la tradition des recherches dites qualitatives, il est d'usage de distinguer deux types d'annotation en contexte, le code d'une catégorie ou d'un thème ; en gros, cette distinction est une question de préséance originelle de l'assignation d'une annotation, à parti du texte v/s à partir du cadre théorique du chercheur. Pour l'annotation en contexte, voir John Seidel et Udo Kelle, « Different Functions of Coding in the Analysis of Textual Data », dans Udo Kelle (dir.), *Computer aided qualitative data analysis : theory, methods and practice,* LondresSage, 1995, p. 52-61.

lisateurs) et tirée des mémos/notes personnels assignés à des occurrences particulières, ou hors contexte, en assignant par exemple une valeur à un uni- ou pluri-terme dans l'ensemble du corpus (catégorie socio-sémantique d'un segment). Dans ce dernier cas, nous incluons la définition d'une catégorie ou d'une variable dans toutes ses assignations dans le corpus, participant ainsi à la constitution du dictionnaire des variables ou des catégories. D'autre part, les annotations peuvent être assignées à des unités de différentes tailles allant du caractère à une chaîne de caractères plus ou moins grande, voire à tout le document. Il est aussi possible, pour rejoindre des phénomènes intertextuels/interdiscursifs, lors d'une même opération d'annotation de chevaucher deux unités d'analyse distinctes, par exemple assigner une même modalité/valeur à plusieurs segments de texte issus de deux documents distincts. Par ailleurs, il arrive qu'un analyste ne se limite pas à annoter des signes adjacents (i.e. un ou plusieurs mots, une phrase, etc.), mais inclut aussi des annotations plus complexes, de type « relationnel », les annotations dites structurelles[17] où une seule annotation réfère à une relation entre plusieurs signes non adjacents (identification des énumérations, des couples thèmes/rhèmes, des constituants d'une parabole, etc.), très utiles en analyse énonciative et argumentative des discours. Il est important finalement de signaler que les annotations peuvent avoir différentes couches d'abstraction, telles qu'observées dans les arbres thématiques[18], et que l'assignation peut se faire de manière manuelle (feutrer par le biais d'un logiciel), automatisée (TreeTaggers) ou semi-automatisée (Cassandre, SATO, etc.).

Pour définir l'opération de *représentation,* nous adoptons la version sémantique de Bunge : « une traduction conceptuelle, visuelle, auditive ou artéfactuelle d'un objet (matériel ou idéel)[19] ». En mode écriture, c'est une mise en relation entre plusieurs signes (une syntaxe) les faisant passer d'un état symbolique à un autre. Cette mise en relation peut être linguistique (textuelle) ou non linguistique. Parmi les non-linguistiques, on retrouve les types non exclusifs et non exhaustifs suivants : les arbres, les graphes, les tables, les espaces (euclidien ou non euclidien, etc.). Par exemple, un treillis peut être représenté par un arbre, les documents et les unités textuelles peuvent être représentés par une matrice réduite dans un espace euclidien, les différents états d'un phénomène peuvent être représentés par une variable (indépendamment de l'échelle de mesure), ou encore un événement peut être représenté par un récit, ou un texte peut être réécrit. Cette opération

17. François Daoust, Yves Marcoux et Jean-Marie Viprey, « L'annotation structurelle », *Journées d'Analyse statistique des Données Textuelles*, 2010, Actes du colloque des 10e JADT, p. 1145-1156.
18. Pierre Paillé et Alex Mucchielli, « L'analyse thématique », dans *L'analyse qualitative en sciences humaines et sociales*, 2e éd., Paris, Armand Colin, 2008, p. 161-207.
19. Mario Bunge, *Philosophical dictionary*, Amherst, Prometheus Books, 2003, p. 251.

comporte aussi une granularité et une contextualité. Par exemple, il est possible de projeter dans l'espace les mots ou les groupes de mots (données textuelles), voire les modalités de variables (données extratextuelles), ou bien de représenter un mot ou groupe de mots dans leur contexte, par exemple avec la procédure de concordance (KWIC).

Enfin l'opération de *contraste*, opération en mode lecture par excellence, en est une qui permet, par le biais d'une représentation commune, d'apprécier les différences et les similitudes entre plusieurs données textuelles, extratextuelles ou leurs diverses couches de représentations pour permettre, le cas échéant, ce qu'on pourrait désigner comme un acte de « découverte », soit le dévoilement d'un élément existant jusque-là inconnu[20] du chercheur. D'ailleurs, en s'inspirant d'une perspective constructiviste piagétienne[21], on peut dire que le processus de découverte se meut, à travers des opérations de contraste, dans une spirale croissante où se côtoient constamment des démarches antagonistes et complémentaires, à savoir l'ancrage dans du familier (une pré-description, une théorie, une recherche antérieure, etc.) et l'objectivation (malgré son résultat provisoire) de la « nouveauté »[22]. En fait, les produits de chacune des trois opérations précédentes (découpage de l'unité d'observation, son annotation et sa représentation) visent ultimement *à faire paraître* des contrastes (différences et similitudes) entre les données. Par exemple, c'est par une opération de contraste qu'un tableau avec des valeurs-tests (représentation commune) permet d'estimer la surutilisation ou la sous-utilisation d'une ou plusieurs unités (mot, groupe de mots, catégorie socio-sémantique, catégorie syntaxique, etc.) entre plusieurs documents.

La figure 2 (page 151) fait la synthèse de ces quatre opérations.

Comme le lecteur l'a peut-être remarqué, les éléments de cette métadescription sont plus modulaires et interdépendants que des scénarï en séquence préfabriquée pour atteindre un objectif à l'instar d'un test d'hypothèse en statistique inférentielle. Ainsi, une unité découpée est susceptible d'être annotée suivant différentes représentations (linguistiques ou non linguistiques) afin de la contraster par rapport à une autre unité. Mais il est aussi possible de garder la même unité découpée et d'en faire varier les représentations. Cela nous ramène à la dimension temporelle des interactions analyste-texte où les différents états d'un corpus interagissent avec

20. *Ibid.*.
21. Jean Piaget (dir.), *Logique et connaissance scientifique*, Paris, Gallimard, 1967, p. 1220-1224.
22. Les concepts d'ancrage et d'objectivation reviennent à Serge Moscovici (*La psychanalyse, son image et son public*, Paris, Presses universitaires de France, 1961) mais font écho à la conception du développement de la science chez Piaget.

plusieurs moments itératifs de lectures/écritures, soit des moments marqués par des allers-retours où aux impasses se succèdent des chemins plus prometteurs ramenant souvent l'analyste à des états antérieurs de son corpus. Ces moments sont autant d'occasions d'enrichir ou de modifier différentes interprétations à travers une multitude d'annotations retraçant les multiples passages.

L'état des LT à la lumière de l'interaction analyste-texte : état d'avancement nécessaire mais non-suffisant

Les LT sont depuis un certain temps sur la scène des sciences sociales. Leur emploi est relativement distribué[23], mais il reste une question importante sur la spécificité et la limite de leur apport au processus du *sense-making*, soit l'activité du chercheur à « faire du sens » avec des données discursives. Pour les enquêtes sociologiques, les LT peuvent être envisagés comme des outils de contraste fournissant, à propos des données discursives à l'étude (surtout des unités lexicales), des représentations qui facilitent les interprétations qu'en donne le chercheur[24]. Pour ce qui est du processus d'interprétation (globale ou locale) lui-même, nous le considérons comme une opération de clôture mobilisant plusieurs traces de lectures et d'écritures de données discursives pour construire un sens en fonction d'une question de recherche.

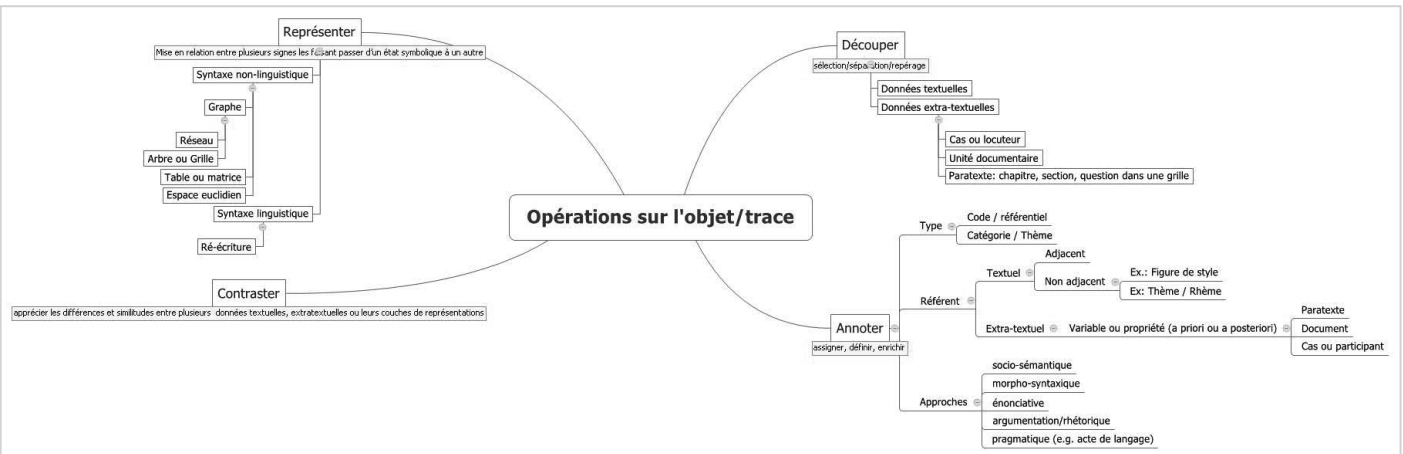
Avant de tenter d'estimer la place des LT dans le champ de l'AD, il semble approprié de situer le contexte général de l'évolution des logiciels d'ATO en sciences sociales. En effet, au moment de son apparition, l'intégration de l'outil informatique se croisait avec deux facteurs importants : 1) l'état d'avancement des disciplines connexes (l'informatique et les statistiques) à l'époque de la manifestation des besoins en analyse textuelle (linguistique, analyse de contenu, analyse automatique des discours, terminologie, etc.) ; 2) les affinités locales et les communautés de pratiques qui prévalaient avant l'arrivée de l'ATO.

Du côté informatique et statistiques, il était plus à la portée des technologies des années 1970-1980 de représenter (dans le logiciel) des structures de données tabulaires voire relationnelles mais non hiérarchiques que de tenter de représenter des relations à structures complexes (graphes ou arbres) nécessitant d'associer plusieurs couches de lectures interdépendantes. En effet, les capacités de calculs étaient moins énergivores avec des analyses

23. François Leimdorfer, « Résumé des réponses au questionnaire sur les usages de la lexicométrie », 2013 [en ligne] Consulté le 1er décembre 2013, www.printemps.uvsq.fr/medias/fichier/resume-enquete_1364992355396-doc?INLINE=FALSE.

24. Didier Demazière, Claire Brossaud, Patrick Trabal et Karl van Meter (dir.), *Analyses textuelles en sociologie : logiciels, méthodes, usages*, Rennes, Presses universitaires de Rennes, 2006.

Figure 2
Les opérations sur l'objet/trace



multidimensionnelles où l'on suppose des relations symétriques entre points d'une matrice projetés dans un espace euclidien[25] que celles de calculs avec des graphes à relations asymétriques[26] ou de calculs de combinaisons effectives/possibles[27] entre plusieurs points d'une matrice.

Du côté des affinités locales et des communautés de pratiques, il faut se rappeler que depuis ses débuts[28], la communauté de l'AD française a gardé un lien constant avec les sciences du langage en général et avec la linguistique en particulier. Ce faisant, le rapport au texte est resté sous l'égide du respect de l'intégrité du texte et de la primauté de ses expressions, ce que Rastier appelle la « dé-ontologie[29] », très bien résumée par Pincemin : « Il s'agit bien d'éviter toute préconception réductrice, on veut surtout rester au plus proche du texte et ne pas commencer par l'étudier à travers le prisme d'une ontologie[30]. » D'ailleurs, ce que préconise Rastier, et qui est préservé par la tradition textométrique, reste la direction allant des expressions aux catégories en passant par les signifiés et les concepts.

Cela dit, une tentative d'articuler notre esquisse de description des interactions analyste-texte avec les LT francophones va permettre de montrer les contributions mais aussi les limites de ces derniers : 1) une concentration quasi exclusive sur l'unité textuelle et l'absence d'annotations locales et complexes sur mesure ; 2) l'architecture des LT n'est pas celle de système d'informations intégrés, soit que les différents traitements pour assister le chercheur dans ses tâches itératives d'analyse demeurent dans le même environnement de travail.

Pourtant, au niveau des pratiques d'analyse, les LT ne manquent pas d'atouts autant en mode lecture qu'écriture ; en proposant plusieurs moyens de repérage et de représentations, ils fournissent des bases essentielles pour contraster les données et en faciliter l'interprétation[31]. Ainsi, il est souvent

25. Ici nous faisons principalement référence à l'analyse géométrique des données (Le Roux & Rouanet, 2004 ; Lebart, Piron, & Morineau, 2006), une approche très française principalement reconnue pour ses analyses en axes principaux (AFC, ACP, ACM, etc.) suivies ou précédées (e.g. ALCESTE) par des classifications automatiques. Pour l'analyse géométrique des données, voir : Brigitte Le Roux et Henry Rouanet, *Geometric data analysis : from correspondence analysis to structured data analysis*, Boston, Kluwer Academic Publishers, 2004 ; Ludovic Lebart, Marie Piron et Alain Morineau, *Statistique exploratoire multidimensionnelle : Visualisation et inférence en fouille de données*, Paris, Dunod, 4e édition, 2006.
26. Régis Gras, Einoshin Suzuki, Fabrice Guillet et Filippo Spagnolo (dir.), *Statistical implicative analysis*, Berlin, Springer Verlag, 2008.
27. Pierre Vergès et Boumedine Bouriche, « L'analyse des données par les graphes de similitude » [en ligne] Consulté le 1er décembre 2013, www.scienceshumaines.com/textesInedits/Bouriche.pdf.
28. Jean Dubois, « Enoncé et énonciation », *Langages*, 1969, vol. 4, nº 13, p. 100-110.
29. François Rastier, « Ontologie (s) », *Revue d'intelligence artificielle*, 2004, vol. 18, nº 1, p. 15-40.
30. Bénédicte Pincemin, « Sémantique interprétative et textométrie », *Texto !*, 2012, vol. 17, nº 3. p. 11. Pour plus d'informations sur l'ontologie de Rastier en linguistique, voir François Rastier, « Les mots sans les choses », dans Adolfo Murguía (dir.), *Sens et références*, Tübingen, G. Narr, 2005, p. 223-255.
31. Damon Mayaffre, *L'analyse du discours assistée par ordinateur*, 2009 [en ligne] Consulté le 1er décembre 2013, http://eprints.aidenligne-francais-universite.auf.org/19/.

possible de découper un corpus en nombre de sous-corpus en qualifiant ces derniers par des données extratextuelles, soit de l'annotation hors contexte ; il est aussi possible de repérer des segments textuels en contexte (concordance) ou hors-contexte. Les LT offrent aussi, quoique de manière hétérogène[32], une panoplie de moyens de représentations allant des techniques synthétiques ou macro (analyses factorielles, classifications automatiques, analyses arborées, analyse de similitude, etc.) aux techniques plus micro (caractérisation du vocabulaire (spécificité), co-occurrences d'unités lexicales particulières, etc.) chacune avec son procédé de visualisation et d'appréciation (valeur-test, test du chi2, écart-réduit, etc.) afin d'esquisser des interprétations à partir des distributions lexicales d'un corpus[33]. Ces fonctionnalités ainsi que ces batteries statistiques ont largement fait bénéficier la communauté des analystes de texte en général (lexicomètre, terminologue, etc.) et ainsi que quelques analystes de discours d'autant plus qu'elles sont rarement ou pas du tout offertes dans les suites logiciels comme les CAQDAS.

Cela dit, nous estimons, et nous tenterons de montrer comment les LT ont de la difficulté à satisfaire les besoins d'une majorité[34] de chercheurs en AD avec cette panoplie de fonctionnalités. Cette remarque s'appuie sur la première limite susmentionnée, qu'en LT, il y a focalisation sur l'unité lexicale au détriment de l'annotation. Or, en sciences sociales, où la majorité des données à analyser s'inscrivent dans du langagier, il semble illusoire de penser que le processus analytique puisse se baser uniquement sur les unités lexicales (i.e. occurrence, co-occurrence locale ou globale, *n-gram*, etc.) vu que toute unité de signification ne se limite pas forcément à la seule forme graphique (uni- ou pluriterme), et que, comme démontré précédemment, d'autres couches d'*écritures* locales ou globales sont susceptibles d'intervenir. Pourtant, dans les LT, le sacro-saint « retour au texte » ne se traduit malheureusement que par les expressions et leurs différentes représentations en contexte et hors contexte.

Pour revenir à la perspective de Rastier, sur le plan analytique, c'est comme si le passage aux signifiés en direction des catégories, c'est-à-dire le début d'un parcours interprétatif, n'est pas à inscrire dans une trace. À moins de craindre que cette trace ne cache les expressions originelles, il semble que

32. En effet, il y a autant chevauchement que dispersion des traitements et chaîne de traitements à travers les différents LT francophones. Pensons simplement, à la concordance, à l'AFC ou à la caractérisation du vocabulaire qui sont présents dans la grande majorité des LT mais suivant différentes appellations, représentations graphiques et possibilité de paramétrage.

33. Pour plus de détails, voir Fidelia Ibekwe-SanJuan, *Fouille de textes : méthodes, outils et applications*, Paris, Lavoisier, 2007 ; Ludovic Lebart, André Salem et Lisette Berry, *Exploring textual data*, Boston, Springer, 1998.

34. D'ailleurs, si l'on affine l'enquête de Leimdorfer (2013), il est probable que ceux et celles qui ont recours aux LT se limitent souvent à des opérations de découpages/repérages et se montrent méfiants quant à l'usage des analyses bi- et multidimensionnelles.

les interprétations locales, qui sont pourtant aussi des lectures de données, soient à maintenir à l'écart. Or, comme en pratique, selon les dires de Rastier, on est condamné à interpréter, il s'ensuit que d'une part les chercheurs en AD vont de toute façon marquer leurs interprétations ailleurs que dans l'environnement de lecture, et d'autre part durant un parcours interprétatif, il est difficile de nier que les données sont *déjà* à l'intérieur du prisme d'une lecture dirigée, justement, par ledit parcours interprétatif. En fait, pour un chercheur en AD, le fait de marquer le passage à même les données n'affecte pas le rapport de l'analyste avec ses données, l'historique du parcours restant toujours à disposition. De plus, la majorité des logiciels privilégiant l'écriture des interprétations (annotation en contexte ou hors contexte) maintiennent scrupuleusement l'intégrité du texte et séparent constitutionnellement les données textuelles de leurs annotations. Nous estimons que cette vision limitative du travail de l'analyste par les LT le freine dans ses analyses des couches de lecture, et ce faisant, l'empêche de comparer ses différentes lectures des données, comparaison souvent heuristique car pouvant mener à une autre compréhension du texte.

Il faut souligner que dans quelques LT, il existe des procédés analytiques qui dépassent l'unité textuelle[35], mais le plus souvent c'est pour aller vers des annotations de type linguistique, toutes très utiles en analyse de discours, comme l'étiquetage morpho-syntaxique ou le codage grammatical, sans parler de la lemmatisation qui est plutôt un regroupement de formes lexicales. Toutefois, dans ces contextes, l'analyste se sert d'une annotation préfabriquée qui, même si quelques LT offrent la possibilité de paramétrer partiellement son automatisation, reste dans bien des cas insuffisante car *a posteriori* et relativement étrangère[36] aux processus sémiotiques de l'analyste en sciences sociales en général[37]. Il ne s'agit pas de bannir cet étiquetage, essentiel à plusieurs égards, mais de montrer que de s'y limiter réduit massivement le processus de *sensemaking*. Ce dernier a autant besoin d'un type d'annotation hors contexte et sur mesure – ce que Lejeune[38] appelle les registres – que d'annotations en contexte (CAQDAS), souvent d'ordre socio-sémantique[39], et qu'il est tout à fait possible, dans les deux cas, de les rendre systématique et, ce faisant, de

35. D. Mayaffre, *op. cit.*
36. D'ailleurs, il suffit pour un chercheur hors sciences du langage de commencer à s'immiscer dans des analyses morphosyntaxiques ou des algorithmes de linguistique computationnelle pour se rendre compte qu'il est assez amateur dans ce genre de problématique set conséquemment, l'acte de faire du sens avec ce matériau s'en trouve encore plus délicat et hasardeux.
37. À cet égard, le logiciel Tropes fait figure à part car ses sorties se fonde une théorie cognitivo-discursive. Voir Rodolphe Ghiglione, Christiane Kekenbosch et Agnès Landré, *L'analyse cognitivo-discursive*, Grenoble, Presses universitaires de Grenoble, 1995.
38. C. Lejeune, *op. cit.*
39. Jules Duchastel et Victor Armony, « La catégorisation socio-sémantique », », *Journées d'Analyse statistique des Données Textuelles*, 1995, Actes du colloque des 3e JADT, p. 193-200.

les soumettre aux mêmes techniques statistiques. Bien sûr, ces annotations préfabriquées ont une visée attrayante pour l'inférence statistique, soit celle d'être exhaustive dans leur assignation à *tout* le matériel textuel. Or, on sait pertinemment que ce n'est pas cette exhaustivité de l'assignation (dont une grande partie reste sans interprétation possible) qui garantit la production du sens recherché chez un analyste du discours.

Ajoutons que la question des annotations dans les LT ne s'arrête pas au niveau de leur type ou fondement, mais s'étend aussi à leur fonction de trace de constructions progressives de sens par le chercheur. Par exemple, face à une matrice résultat (un espace factoriel, un dendrogramme ou un arbre de similitude), les LT (surtout payants) offrent à l'analyste, comme procédure de marquage de son passage par ce chemin, des fonctions de stylage, de déplacement ou de suppression des unités représentées, bref, des utilitaires d'impressions. De telles indications ne favorisent en rien le déclenchement d'algorithmes[40] ou de traces locales situées à même le socle des signes de l'acte herméneutique de l'analyste.

Ce dernier constat permet de passer à une deuxième limite importante des LT, le fait qu'ils ne sont pas des systèmes intégrés et par là faillissent en partie à leur tâche d'assister le chercheur dans son analyse. Pour illustrer ce point, considérons la dimension temporelle de la constitution d'un corpus et surtout aux cycles de vie de ce dernier (C1 au temps 1, C2 au temps 2, etc.). Ainsi, allant des prétraitements jusqu'aux différentes facettes des interprétations locales et globales[41], l'analyste ne cesse d'emprunter et de réemprunter des chemins et de recommencer en peaufinant des chaînes de traitement. Or, ce que les LT[42] lui permettent à cet égard c'est de recommencer la chaîne depuis le début à chaque nouvelle mouture ou tour de roue. Prenons un exemple simple, supposons qu'en cours d'analyse, un chercheur ait apporté des modifications à son corpus (marquage des majuscules, lemmatisation, balisage des sections, regroupement de synonymes, etc.) et qu'il veuille faire « rouler » de nouveau une chaîne de traitement (un thémascope[43]), il n'a

40. Il y a ici des exceptions locales, par exemple avec les sorties graphiques de Lexico3.
41. Jules Duchastel et Danielle Laberge, « Des interprétations locales aux interprétations globales : combler le hiatus », dans Nicole Ramognino et Gilles Houle (dir.), *Sociologie et normativité scientifique*, Toulouse, Presses universitaires du Mirail, 1999, p. 51-72.
42. En fait, ce phénomène se retrouve même dans des grandes suites d'ingénierie documentaire comme GATE (*General Architecture for Text Engineering*)
43. Il s'agit d'une chaîne de traitement standard dans SPAD-T et DTM-VIC où une analyse en axes principaux (AFC, ACP, etc.) est suivie d'abord par une classification automatique sur les coordonnées des axes principaux et ensuite par une description automatique des classes avec des variables supplémentaires. Pour les chaînes de traitement, voir Lebart *et al.*, *op. cit.* ; Ludovic Lebart, « Stratégies du traitement des données d'enquêtes », *La revue de Modulad*, 1989, vol. 3, p. 21-29.

souvent pas le choix que de réindexer le corpus[44], sans parler des multiples recommencements de toute la chaîne de traitement même si on ne fait que changer un simple paramètre (pondération). Autre exemple, supposons qu'un chercheur a soumis son corpus à un étiqueteur morpho-syntaxique, tel Treetagger, et qu'il veuille analyser, voire juste conserver, les formes graphiques *et* chacune de leurs fonctions grammaticales à l'intérieur de son LT, donc les identifier comme un seul et même projet de travail. Or, il n'est pas rare que dans les LT les deux versions soient représentées et stockées comme deux éléments distincts, ce qui signifie que dans la représentation interne de la majorité des LT, une version enrichie d'un corpus, voire un sous-corpus est envisagé comme un autre corpus et non comme la poursuite du travail sur un même corpus.

Finalement, la plupart des LT ont des interfaces qui, à bien des égards, sont des « coquilles » disponibles pour déclencher des algorithmes indépendants menant à des représentations (graphes, tableaux, etc.) que le chercheur interprète et annote ailleurs, hors de l'application, dans des documents annexes. Il en résulte que l'environnement de travail pèche sur le plan de la réutilisabilité et de la cohérence. Insistons sur le fait que notre propos n'est pas une question de convivialité, mais d'architecture du système[45] d'information. En somme, si nous tentons de représenter en image le fonctionnement non intégré des LT, la métaphore[46] d'un ensemble de tiroirs qui ne communiquent pas entre eux et dont la poignée de chacun des tiroirs représente le bouton qui enclenche un paquet de traitement dans l'interface d'utilisation apparaît opportune. En un mot, sans faire l'apologie des logiciels d'annotation (*Cf. supra*)[47] ni dénigrer l'apport indéniable des LT à l'AD, nous dirons que ces derniers ne peuvent que *servir* de manière limitée le processus d'analyse, mais ils ne peuvent s'y *adapter* et encore moins l'*accompagner*. Au contraire, c'est bien souvent à l'usager d'y adapter sa conception du matériel d'analyse, voire ses procédures analytiques. Nous pensons que les LT manquent clairement d'un espace de travail intégré pour donner tout son sens à l'expression « analyse de texte *assistée* par ordinateur (ATO) ». C'est ce que nous tenterons de développer dans la dernière section.

44. Il y a bien sûr une exception dans Hyperbase. Ce dernier offre effectivement la possibilité de visualiser l'unité textuelle, son lemme, son code grammatical et sa structure syntaxique. Il n'en demeure pas moins qu'il ne s'agit que de lemmatisation ou de dimensions morpho-syntaxiques, soit des types linguistique de regroupement de formes graphiques.
45. D'ailleurs comme le lecteur l'aurait remarqué, nous avons mentionné plusieurs exceptions ajoutées hélas a posteriori *sur* la structure originelle du logiciel, probablement pour des raisons d'accroître l'utilisabilité par un plus grand bassin d'usagers et non une conception différente de l'interaction analyste-texte.
46. Cette métaphore est largement inspirée d'une discussion avec François Daoust, le concepteur du logiciel SATO.
47. Rappelons d'ailleurs que dans la communauté des CAQDAS par exemple, il est très rare que la matérialité discursive, chère à la communauté de l'AD, ait une importance.

Prospectives sur des directions potentielles pour l'ATO en AD : du sacro-saint « retour au texte » à l'activité du chercheur

Depuis les années 1990, plusieurs initiatives du centre d'ATO à l'UQAM pointent vers une direction plus intégrée de l'ATO, et ce, bien avant l'apparition du *cloud computing*. Même si l'architecture entière d'une application informatique est loin d'être implantée, l'intuition du plan de conception est assez bien esquissée[48]. À défaut d'avoir une telle architecture intégrée, il convient ici d'en exposer l'esprit général et de soutenir ensuite l'intérêt de poursuivre entretemps des initiatives d'échanges et de collaboration inter-LT.

Comment décrire l'esprit général de ce que nous envisageons comme une plateforme en ligne, libre et conçue à des fins de recherche ? Signalons d'abord que si nous avons associé les LT à des paquets de traitements ou d'algorithmes, c'est entre autres pour faire valoir une distinction entre les chaînes de traitements pré-coordonnées et le « travail du texte » par le chercheur. Bien que les LT fournissent un ensemble de traitements statistiques et linguistiques très appréciable, à notre avis, sur le plan de l'interaction analyste-texte, ces opérations d'assignation, de représentation automatisée et de contraste des données se distinguent de celles de navigation, de découpage et d'annotation sur mesure. Ce dernier ensemble, qu'on appellerait *gestion des données textuelles* (traitées ou brutes), appartiendrait à l'espace de travail de l'analyste qui occuperait une place centrale dans une application à l'architecture orientée services/ressources où ces derniers représentent la périphérie de l'espace de travail. Ainsi, d'un côté, l'espace de travail comporterait des fonctions liées aux prétraitements (édition, construction-déconstruction de corpus et sous corpus, etc.), à l'annotation en contexte, à la navigation (i.e. les fonctions documentaires dans Hyperbase), à la constitution d'une représentation interopérable (tabulaire ou hiérarchique) des unités textuelles, et surtout, au relevé des opérations paramétrables et réutilisables (i.e. pas juste un fichier journal), à l'instar de celui disponible dans le logiciel SATO[49], pour optimiser leur traçabilité.

D'un autre côté, par services nous entendons des services Web, soit des requêtes soumises par le noyau susmentionné pour faire appel à des algorithmes exécutés dans des applications externes. Dans ce cas, il peut s'agir autant de lemmatisation, d'étiquetage morphosyntaxique ou de codage grammatical que de traitements statistiques (classification automatique, analyse des correspondances, co-occurrences, etc.), où il serait plus que plausible

48. Jules Duchastel, Francis J. Lacoste, et François Pizarro Noël, « Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD », 2004, Actes du colloque des 7e JADT, p. 364-372.
49. Pour plus d'informations voir ici : http://corpus.ato.uqam.ca/sato/

d'employer des packages du projet R[50]. Derrière une telle application intégrée, devrait aussi être prévues une place pour la collaboration entre collectivités ainsi que la division des tâches inter-utilisateurs dans un même projet.

Si on continue dans cette direction, la question de faisabilité devient au cœur des préoccupations. Y a-t-il déjà des initiatives qui vont dans le sens d'une application intégrée ? D'un côté, dans les CAQDAS, le concept d'espace de travail intégrée existe déjà, mais manque sérieusement la possibilité d'avoir un relevé des opérations paramétrables et réutilisables, sans parler de la fermeture de leurs formats d'échanges interlogiciels dans un contexte où le logiciel libre est encore très embryonnaire. D'un autre côté, des initiatives à plus grande échelle comme le logiciel libre *RapidMiner*[51], une application *Text Mining* qui offre un véritable espace graphique de construction fine de chaînes de traitements paramétrables et réutilisables faisant appel à une longue liste de procédures automatisées (fouilles de textes, traitement automatique des langues) intégrées dans les logiciels. Comme les LT, ces logiciels ne traitent malheureusement que les métadonnées et les unités lexicales sans aucune place pour de l'annotation en contexte et sur mesure, option qui se trouve par contre très bien implantée dans GATE[52], une autre grande suite d'ingénierie documentaire. Une question se pose alors : pourquoi viser un tel degré d'intégration si ce n'est pas encore disponible dans les grands logiciels d'analyse de données ? Notre modeste réponse se résume au fait que le matériel d'étude (i.e. discours comme objet théorique et empirique), la visée et la démarche du chercheur en sciences sociales ne sont pas forcément préstructurés comme dans les pratiques de *Text Mining* ou d'ingénierie documentaire.

Pour terminer, en attendant la concrétisation d'une telle initiative, il est possible de songer à un éventail de pratiques probantes qui permettent d'accroître les capacités actuelles des LT et d'autres types d'applications en ATO. D'ailleurs, le projet même de la plateforme textométrique, incarnée présentement par le logiciel TXM, part déjà d'un constat d'éparpillement des forces et de démultiplication contre-productive des LT. C'est pourquoi, parmi les idées clés de telles pratiques, on retrouve l'interopérabilité et la collaboration.

Nous nous limiterons ici à une description brève de quelques initiatives, allant dans cette direction, effectuées par le centre d'ATO (UQAM) durant ces 15 dernières années. D'abord, le passage à des applications Web de SÉMATO et SATO. Outre le fait de surmonter les problèmes de compatibilité entre systèmes d'exploitation (Windows, Linux, Mac) ce dernier logiciel

50. Pour plus d'informations voir ici : http://www.r-project.org/
51. Voir ici : http://sourceforge.net/projects/rapidminer/
52. Voir ici : http://gate.ac.uk/

a, dans sa version Web, entamé une vision de plus en plus intégrée des opérations sur les corpus par la possibilité d'établir des sous-corpus contextuels. Aussi, quelques services Web y sont présentement intégrés à même la chaîne de traitement de SATO, par exemple les analyses par le biais de Treetagger peuvent être réintégrées dans les couches antérieures de la vie du corpus à l'étude pour être articulées avec d'autres types de catégorisations dans la suite du processus d'analyse. Aussi, sur le plan de la représentation dans le système, il s'agit d'un corpus où chaque unité d'analyse encapsule l'ensemble des informations qui l'identifient (métadonnées et traitements effectuées). Par ailleurs, un format d'échange de corpus annotés (XML-TEI), hélas peu utilisé hors du centre ATO, a été développé et implanté dans SATO pour permettre des échanges avec Lexico3 (et donc Coocs aussi), DTM et Alceste. Enfin, à partir d'un quasi-modèle formel[53], le centre ATO implante depuis deux ans une plateforme en ligne de dépôt des données de recherche qui accueillera à la fois les corpus des chercheurs et leurs grilles d'annotations. Comme l'application (Fedora Commons) sur laquelle se base la plateforme représente non seulement les données textuelles et extra-textuelles, mais aussi les relations entre les objets déposés (X *annote* Y, Z *est une partie de* Y, etc.), elle permettra à la communauté de chercheurs par le biais des requêtes fines de construire-déconstruire des corpus à partir d'autres corpus et d'y joindre des grilles d'annotations pour des analyses comparatives ultérieures.

En guise de conclusion

Pour terminer, un retour sur l'essentiel de notre propos est résumé en esquissant une réflexion sur les attitudes de l'AD française envers l'ATO. Après avoir commencé par un constat d'une focalisation de l'AD française sur les LT, nous en avons montré certaines limites à partir de l'esquisse de modélisation de l'interaction analyste-texte, surtout dans ce qui est compris *par* et *dans* l'assistance du processus d'analyse du texte. En effet, ce qui ressort de cette analyse, c'est que l'assistance au chercheur dans les LT est souvent une question de *donner à voir* par le biais des procédés d'interrogation, parfois d'assignation automatique et souvent de représentation de données textuelles, mais très rarement une base d'accompagnement du chercheur dans son travail du texte pour produire et construire du sens par ses traces d'analyses. Conséquemment, il est possible de dégager dans les LT des présupposés sur le texte et sur l'analyste. Pour ce qui est du texte, il est considéré comme un objet empirique (cf. introduction) que l'analyste ne fait

53. François Daoust, Jules Duchastel, Yves Marcoux, et Élias Rizkallah, « Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche », 2008, Actes du colloque des 9e JADT, vol. 1, p. 355-367.

qu'observer, découper et projeter sans aucune tentative de le caractériser, le modeler, voir le ré-écrire à travers sa démarche d'interprétation. Cela laisse entrevoir une vision « empiriciste » de la posture et de l'action de l'analyste sur ses données, c'est-à-dire une version radicale de l'empirisme limitant les données à des extractions de l'expérience sensible et de la connaissance à une généralisation inductive à partir des données extraites. Si tel était le cas, on ne devrait pas s'étonner des attitudes envers l'ATO, limitant les pratiques aux seuls LT dans la communauté d'AD française. En effet, selon notre expérience, le continuum des attitudes va de l'emballement, parfois relativement aveugle, basé souvent sur une méconnaissance des procédés de calcul sous-jacent, à l'opposition, voire à la méfiance de procédés techniques sophistiqués amenant à préférer effectuer des analyses manuelles. Ce dernier propos ne veut pas dénigrer les analyses manuelles, bien au contraire, plusieurs des études qui y recourent sont souvent plus pénétrantes que certaines recourant aux traitements informatiques, mais pour lancer la question insistante : est-ce parce que les LT c'est tout ce que l'ATO peut offrir au chercheur en AD ?