Revitalisation du Centre d'Analyse de Texte [assistée] par Ordinateur (ATO)
Projet de colloque – Québec – été 2014

Élias Rizkallah,
Professeur de sociologie
Directeur (futur) du centre d'ATO
UQAM, Montréal

L'esprit qui a toujours animé le Centre d'ATO

Le centre d'ATO est depuis sa naissance (depuis 1982) un lieu d'expertise, de formation et de consultation en analyse de textes par ordinateur rattaché à la Faculté des Sciences Humaines de l'Université du Québec à Montréal. Il a été à travers son histoire un des pionniers de ce champs particulièrement par le développement constant de deux logiciels web (accès gratuits), SATO et SÉMATO (l'ancien DEREDEC) qui ont servis à la réalisation de plusieurs dizaines de projets universitaires, privés et publics en sociologie, sciences politique, sciences du langage et sciences cognitives. Une tradition d'écoles d'été en ATO s'est aussi installée entre 2004 et 2012 réunissant plusieurs formateurs nord-américains et européens. Le Centre d'ATO a toujours eu une posture méthodologique privilégiant la rigueur des procédures et l'assistance aux tâches analytiques des chercheurs universitaires tout en restant à l'affût des avancements technologiques.

Des questions persistantes

Les départs à la retraite de ses fondateurs, les coupures dans les financements de la recherche, l'évolution technologique et son pendant de prolifération dans le marché de logiciels d'ATO, le manque de relève, etc. font en sorte que le centre a besoin d'un repositionnement dans la continuité de ses origines. Quatre interrogations persistantes sont donc à l'origine du projet de renouvellement de l'ATO.

1) Comment se fait-il que depuis plus d'une trentaine d'années que l'ATO prend autant d'expansion, entre autres à cause du développement technologique, il y a encore des chercheurs chevronnés en analyse de discours (AD) qui font des analyses de discours très pénétrantes sur des corpus de grande taille sans avoir recours à de l'ATO alors qu'on trouve par contre de plus en plus de recherches via des logiciels d'ATO[1] qui manquent clairement de profondeurs et de jugement épistémologique? La réponse ne peut être uniquement une question d'utilisabilité de logiciels, il y a aussi que ces derniers ne répondent que partiellement aux besoins des analystes[2]. En effet, plusieurs des solutions logiciels semblent plus être des projets d'informaticiens soucieux de la langue naturelle et offrant des scenarii de traitements applicables, qui ultimement finissent par regrouper des communautés de praticiens intéressées, que des initiatives de réponses à des besoins et des problèmes rencontrés par les chercheurs en analyse de texte. Ce n'est qu'ensuite, après que

[1] On peut grossièrement diviser les logiciels d'ATO entre des logiciels de *Text Mining* ou de textométrie (e.g. Alceste, CooCS, DTM, Hyperbase, IBM SPSS Text Analytics, Lexico, Rapid Miner, SAS Text Miner, T-lab, TXM, etc.) et les logiciels d'annotations d'unités textuelles (e.g. ATLAS.ti, GATE, MaxQDA, NVivo, QDAMiner, RQDA, SATO, etc.)
[2] Sans parler du poids de l'ensemble des contraintes qui accompagnent les suites logicielles proposées.

l'architecture de l'application (libre ou propriétaire) est peu modifiable, qu'on tente autant que faire se peut de répondre à quelques besoins[3].

2) Comment se fait-il que la majorité des logiciels de textométrie (pas les logiciels d'annotations) restent bon gré malgré une somme de traitements ou de chaînes de traitements (automatisés ou semi-automatisés) sur des unités textuelles et non des systèmes intégrés assistant le chercheur dans sa tâche itérative de construction de sens? Autrement dit, la plupart des solutions ont bien des interfaces mais c'est plutôt pour déclencher des algorithmes indépendants menant à des produits (graphes, tableaux, etc.) qu'ensuite le chercheur interprète et note ailleurs, dans des documents annexes, hors de l'application. Pourtant, nous savons très bien que pour un chercheur un corpus a un ou des cycles de vies et surtout une multitude de couches de descriptions qui l'enrichissent et qui constituent autant de représentations qui sont à la base des interprétations du chercheur. Métaphoriquement, on peut rapprocher un paquet de traitement à un tiroir et un logiciel de textométrie à un ensemble de tiroirs qui ne communiquent pas entre eux et où la poignée de chacun représente l'interface d'utilisation qui enclenche le paquet de traitement.

3) Comment se fait-il que les initiatives pour des formats d'échanges (e.g. Corti & Gregory, 2010; Daoust & Marcoux, 2006) entre les données de recherche (les corpus ainsi que leurs grilles de lectures) soit encore si peu utilisée alors que la communauté de chercheurs peut grandement en bénéficier pour des collaborations et des validations? Il y a plusieurs réponses à cette question mais principalement : a) la grande majorité des logiciels d'annotation ou de textométrie (à usage hors linguistique computationnelle) sont des gratuiciels ou des propriétaires (donc fermés), le monde du logiciel libre étant encore à ses balbutiements dans ce champ; 2) les initiatives ne sont efficaces que pour un seul type d'architecture de logiciels d'ATO n'atteignant ainsi qu'une une poignée de logiciels.

4) Comment se fait-il que plusieurs cadres théoriques suffisamment formalisés et ayant parfois même des implications directes dans des méthodes de recherche n'ont pas encore été intégrés dans le cadre de logiciels d'ATO? Il suffit de penser à la logique naturelle de Grize, à la linguistique textuelle d'Adam et à la sémantique interprétative de Rastier. Dans quelques cas, il y a les limites techniques, mais dans la grande majorité il y a une commodité à se limiter à employer de l'ATO avec ce que le marché des logiciels fournit à la communauté et non de rendre opérationnels des cadres théoriques déjà solides.

Toutes ces questions mènent à des orientations majeures en ATO : plus d'applications simples, intégrées (avec des services externes pour des traitements automatisées), orientées par les tâches des utilisateurs et échangeables par la communauté des chercheurs.

Pourquoi l'analyse du discours (AD) comme cadre fédérateur?

L'analyse du discours (au sens large) est une approche complexe et transdisciplinaire qui ne consiste pas en un ensemble spécifique de méthodes et ne mène pas à un parcours interprétatif unique. L'AD, si elle est considérée comme démarche méthodologique, peut être nettement plus fédératrice que les méthodes dites qualitatives ou mixtes. Il y a encore respectivement d'un côté,

[3] Cela sans parler du chevauchement entre les traitements offerts par les différents logiciels, pensons simplement aux logiciels francophones de textométrie.

une méconnaissance et une méfiance à l'égard de l'usage de la quantification, de l'autre une posture "agnostique" sur le plan épistémologique, alors que plusieurs recherches en AD font souvent appel à de la quantification et des chercheurs valorisent la pré-définition de la nature des données discursives et l'élaboration de critères de validité.

Les axes et les activités à viser par le centre ATO.

Pour répondre à ces questions, les activités du centre d'ATO tourneront autour de trois types (incluant quelques activités courantes et prévues) :

1) Centre de recherche et développement en méthodologie de l'ATO: autour d'une équipe de chercheurs en analyse de discours, familiers et surtout pas familiers avec l'ATO, il s'agit de développer des savoir-faire, d'applications informatique, des formats d'échanges de données interopérables, des plateformes de collaboration de données de recherches, des nouvelles techniques à arrimer avec des questions en ATO, etc.

- À partir d'un modèle formelle (François Daoust, Duchastel, Marcoux, & Rizkallah, 2008), le centre ATO implante depuis 2 ans une plateforme de dépôt des données de la recherche qui accueillera et les corpus des chercheurs et leurs grilles d'annotations; comme l'application (Fedora Commons) sur laquelle se base la plateforme représente les relations entre les objets déposés (e.g., X annote Y, X est une partie de Y, etc.), elle permettra 'a la communauté via des requêtes fines de construire des corpus à partir d'autres corpus et de joindre des grilles d'annotations pour des analyses comparatives ultérieures. L'application sera livrée en hiver 2014
- La structure du format d'échanges XML-TEI est en phase de perfectionnement via un méta-modèle inclusif pour une plus grande interopérabilité permettant de faire communiquer une plus grande panoplie de logiciels
- Le logiciel SATO, déjà un logiciel suffisamment intégré (cf. question 2), subi plusieurs améliorations : documentation améliorée pour le déposer en dans la communauté des logiciels libres; tests de plusieurs services webs externes pour des traitements statistiques pouvant se réintégrer ensuite aux couches antérieures de la vie du corpus à l'étude.

2) Centre de formation à l'application des méthodes d'ATO via des écoles d'été, des cours universitaire (présentielle ou virtuelle), des séminaires, des colloques, etc.

- Après l'école d'été 2012, dépendant du financement, on envisage d'en faire une école tous les 2 ans.
- Pour le printemps 2014, dans le cadre de l'ACFAS, le centre organisera un colloque réunissant des chercheurs de la francophonie en AD sur les usages et les besoins en ATO en adressant les questions 1) et 2) susmentionnées pour des usagers et surtout des non usagers d'ATO.

3) Expertise en solution logiciel pour l'ATO: offrir du support logiciels à la communauté des chercheurs francophones (au Québec et ailleurs), et ce, autant avec les outils locaux (SATO et Sémato) que des logiciels libres (RQDA, TXM, Iramuteq, Rapidminer), gratuiciels (DTM, Lexico, Hyperbase, etc.), voire payants (Alceste, T-Lab, Nvivo, MaxQDA, QDAMiner, etc.)