L'analyse de texte assistée par ordinateur dans une perspective d'analyse de discours
(ou l'analyse textuelle des discours)

Présentation au colloque *L'analyse du discours comme approche disciplinaire, ACFAS 2012*

François Daoust, Centre ATO,*Université du Québec à Montréal*

Objectif de la présentation

Cette intervention est d'abord de nature méthodologique et vise à soulever un certain nombre de questions concernant la pratique de l'analyse de discours sur corpus. En particulier, nous voulons aborder la question des outils d'analyse de texte assistée par ordinateur (ATO) aptes à construire des dispositifs expérimentaux supportant l'interprétation. Finalement, nous poserons la question, comme point à débattre, de l'avenir de l'analyse textuelle des discours assistée par ordinateur.

L'analyse textuelle des discours

Nous reprenons ce terme d'*analyse textuelle des discours* de notre collègue Jean-Marie Viprey de l'Université de Franche-Comté. Il résume en quelques mots notre tradition d'analyse textuelle assistée par ordinateur dans une perspective d'analyse de discours.

> À propos de la notion de texte, écrit Jean-Marie Viprey, notons tout d'abord que l'ambition d'une *analyse textuelle des discours* souligne bien, et définitivement, non seulement que *texte* et *discours* ne sont pas des objets du même plan, mais encore quelle est leur relation: le *texte* est un mode opératoire sur le *discours*. Ne pas parler d'*analyse de textes*, mais d'*analyse textuelle*, indique 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. 2005:52)

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*[1] inscrite dans un contexte historique et social donné. La pratique de l'analyse de discours, en tant qu'elle aspire au statut de pratique scientifique, construit donc ses données et ses dispositifs expérimentaux destinés à révéler le fonctionnement des discours et à en interpréter la portée.

1- Dans *Discours et archive*, Guilhaumou et coll. définissent l'analyse de discours comme la *manifestation de la langue dans la communication vivante* (Guilhaumou et coll., 1994: 194).

Un processus itératif d'interprétation-explicitation

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).

Pour nous, l'analyse de texte assistée par ordinateur (ATO) doit être conçue comme un processus itératif allant de l'analyse exploratoire à la conception de dispositifs de lecture qui tendent à modéliser des pratiques discursives concrètes inscrites dans des réalités sociales données.

L'établissement du corpus, en tant qu'échantillon théorique, implique aussi des aspects plus statistiques permettant d'autoriser des mesures valides au sein du corpus et une interprétation légitime de la position des variables externes qui rendent compte du contexte d'énonciation et des acteurs sociaux dont on a saisi les discours sous forme d'artefacts textuels.

La question de la qualification des données fait aussi partie de la *construction des faits* sur lesquels s'appuie l'interprétation. Un corpus n'est pas qu'un paquet de mots. L'interprétation des analyses statistiques sera brouillée si les unités prises en compte sont ambigües, si on confond le discours direct et le discours rapporté, etc. Les premières phases de l'analyse du corpus viseront justement cette qualification en fonction du plan d'analyse envisagé.

Des statisticiens, tels Ludovic Lebart, ont travaillé sur les questions de validation statistique permettant d'établir des zones de confiance, par exemple dans les analyses factorielles, pour valider les proximités et les contrastes entre les sous-corpus.

La question de la validation ne s'arrête pas, cependant, à la rigueur statistique. Les grilles catégorielles, inspirées des résultats statistiques ou issues de modèles préalables que l'on veut vérifier, doivent se traduire par une opération d'annotation (catégorisation) qui explicite les données en y investissant nos modèles. Ce corpus enrichi devrait de nouveau être soumis aux dispositifs d'analyse pour confirmer ou infirmer le caractère explicatif de nos constructions.

D’un point de vue technique, ce 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.

Implications pour les outils logiciels

Cette perspective impose un certain nombre d'exigences aux outils de calcul dont nous avons tenté de tenir compte, au Centre ATO de l'UQAM, dans le développement du logiciel SATO[2]. Ce logiciel est conçu comme une plateforme interactive permettant la mise au point de protocoles expérimentaux d'analyse textuelle. Voici quelques spécifications qui traduisent ces exigences.

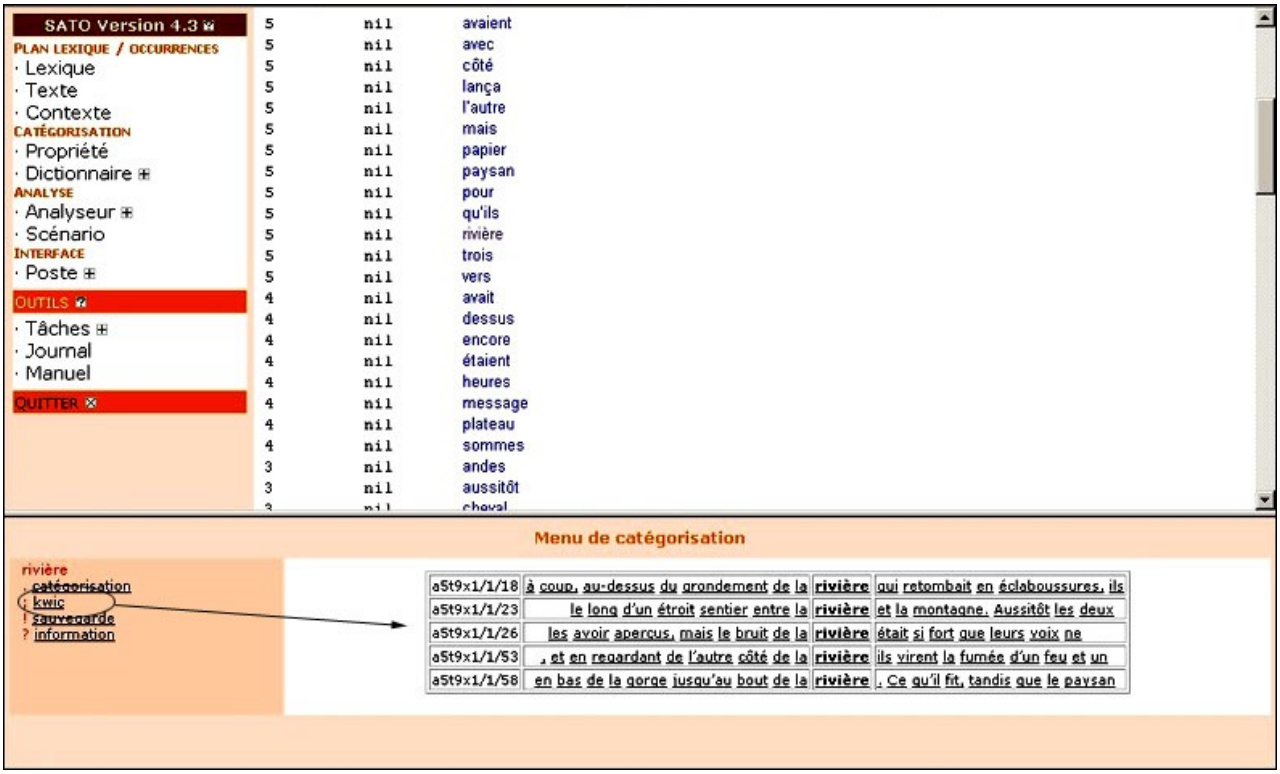
- Il y a d'abord la question du modèle informatique de données requis pour nous donner une emprise sur le corpus. On a opté pour une reconfiguration du texte linéaire sous la forme d'un plan qui fait ressortir le lexique du corpus et les mots en contexte comme des occurrences des entrées lexicales. Sur ce plan, on peut définir des systèmes indépendants de propriétés permettant d'annoter les formes lexicales ou les occurrences de différents points de vue. Des mécanismes d'héritage permettent de faire basculer des annotations du lexique au texte et du texte au lexique.

Fréqtot	**Gramr**							
1	Con	**donc**			x			
2	Proper	**je**	x					
1	Vconj	**pense**		x		x		
1	Vconj	**suis**					x	
			1	2	3	4	5	**NoOcc**
			maj	nil	cap	nil	nil	**Édition**
			prém	prém	conn	conc	conc	**Partie**

*partie=prém Je pense *partie=conn DONC *partie=conc je suis

- La démarche itérative de l'analyse implique un dialogue avec le corpus et donc une interactivité permettant des parcours variés entre le lexique et les contextes, entre les synthèses statistiques et les énoncés du discours. Tous les mots du texte et les entrées du lexique doivent être cliquables pour révéler les couches d'annotation qui les qualifient et pour servir de point d'ancrage à la navigation hypertextuelle.

2 SATO, Système d'analyse de texte par ordinateur, François Daoust, Manuel évolutif en ligne : http://www.ling.uqam.ca/sato/satoman-fr.html



- L'annotation peut être réalisée de façon directe sur une entrée lexicale ou une occurrence en cliquant sur un mot. Elle peut aussi résulter de l'application de commandes et d'analyseurs appelés directement ou indirectement à travers des scénarios. L'annotation lexicale peut être conservée dans des dictionnaires qui pourront être appliqués sur divers corpus.
- La traçabilité des opérations s'effectue par l'enregistrement dans un journal cumulatif des commandes générées par l'interface Web;
- Les commandes peuvent être repiquées dans un scénario de commande permettant de conserver des stratégies d'analyse sous la forme de scénarios pouvant être appliqués sur les corpus.
- Les annotations qui enrichissent le corpus sont autant de ressources qui permettent de caractériser et de reconfigurer les données pour produire, par exemple, des sous-textes construits à la volée et qui pourront être contrastés grâce aux analyseurs statistiques. C'est ainsi que seront établies les frontières des discours attestées sur le corpus et marquées par l'annotation qui traduit les interprétations des résultats de l'application de nos procédures.
- Un corpus enrichi par les annotations peut être exporté et soumis de nouveau au

logiciel pour produire de nouveaux états du corpus, par exemple pour tenir compte d'un raffinement des unités linguistiques : noms propres, locutions figées, mots catégorisés, etc. On peut aussi produire des états du corpus qui sont des réductions (par exemple, en remplaçant les mots par des catégories) afin de vérifier si ces synthèses respectent la configuration du corpus intégral.

- Une exportation en format XML-TEI permet de pérenniser des états stabilisés du corpus pouvant être conservés à long terme dans un dépôt de données. L'exportation peut aussi concerner des résultats construits sur le corpus et pouvant servir d'entrée à des analyseurs externes et à divers logiciels de textométrie.
- Un mécanisme de travail collaboratif permet aux individus de partager un état du corpus et de le faire évoluer dans un espace de travail personnel.

Ces caractéristiques logicielles permettent de combiner de façon itérative des phases de découverte, visant à révéler les caractéristiques du corpus, des phases d'interprétation, qui visent à formuler des hypothèses d'interprétation et à les marquer sous forme de catégories, et des phases de validation permettant de confirmer, infirmer ou préciser ces hypothèses.

Quel avenir pour l'analyse textuelle des discours assistée par ordinateur ?

Ces considérations méthodologiques nous rappellent avec évidence que l'analyse de discours n'est pas qu'un ensemble de théories sur le discours en général, mais une pratique d'analyse à prétention scientifique qui s'appuie sur des artefacts concrets inscrits dans le temps et l'espace social. Les corpus de textes, productions écrites ou transcriptions de l'oral, voire descriptions du geste et de l'image, sont cette matière concrète. L'analyse de texte assistée par ordinateur et les approches quantitatives regroupées sous le vocable de textométrie nous fournissent des instruments pour mener ces analyses textuelles des discours.

Il faut donc considérer l'analyse textuelle des discours comme une véritable pratique scientifique avec ses fondements théoriques, ses objets empiriques, une méthodologie et un savoir-faire technique et procédural.

Mais, quel est l'état de cette pratique, en particulier chez nous en Amérique francophone ? Cela nous amène à une série de questions.

- On a l'impression d'un hiatus entre la communauté de l'AD et la communauté de la textométrie ? Est-ce fondé ? Les approches mixtes que nous pratiquons (approches statistiques, navigation hypertextuelle et annotations) peuvent-elles contribuer à diminuer cette distance ?

- Nos formations universitaires préparent-elles adéquatement nos étudiants et futurs chercheurs à ces procédures d'analyse de corpus qui combinent les méthodes statistiques et l'annotation dans une démarche itérative de qualification des données ?

Débattre de ces questions est certainement de première importance pour donner à l'analyse textuelle des discours la place quelle mérite.

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.

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).