L'analyse de texte assistée par ordinateur

Lunettes de lecture des textes électroniques

Parcours d'une métaphore…

La métaphore de la lunette de lecture peut être prise dans son sens médical de béquille visuelle. Cependant, ce n'est pas dans ce sens-là qu'on l'aborde ici.

La lunette, en sciences naturelles, c'est ce qui a permis de dépasser les limites de notre vision biologique pour atteindre l'infiniment petit et l'infiniment grand. On fait donc référence au microscope et au télescope, ou même à la vision à travers des spectres qu'on ne perçoit pas directement comme l'infrarouge ou les ondes radio.

En passant du support papier au support électronique, le texte devient perceptible par des outils artificiels, c'est-à-dire construits par la technologie sur la base de théories scientifiques. On songe principalement ici à ces outils de calcul logique que sont les ordinateurs.

On connaît les conséquences évidentes du passage du support papier des textes au support électronique: plus grandes capacités de stockage et diminution des coûts du stockage. Mais, c'est surtout au niveau de la diffusion des textes qu'on mesure l'importance de l'édition électronique couplée à sa distribution à travers le réseau Internet. Aussi, même si l'édifice social et économique de l'édition conventionnelle résiste encore au changement, il reste qu'on a maintenant accès a une quantité astronomique de productions écrites sous forme électronique.

Ces changements posent, de façon dramatique, la question de nos capacités de lecture. Tant qu'on en était à l'édition papier, il était difficile d'envisager des solutions en dehors de la sélection miraculeuse, parmi la masse des documents, de ceux qui répondraient à nos attentes. Mais, le miracle, on le sait, ne fait pas vraiment partie de la démarche scientifique…

Cependant, avec l'édition électronique couplée aux capacités de calcul de l'ordinateur moderne, la quantité peut commencer à être perçue comme un allié plutôt que comme un obstacle. C'est-à-dire que la mesure de la masse documentaire peut elle-même devenir source d'information. Ça, c'est pour le côté “macro” ou “astronomique” pour reprendre notre métaphore télescopique…

Mais, l'ordinateur peut aussi nous fournir un accès nouveau au côté “micro” en sachant reconnaître à la surface des textes des régularités difficilement saisissables dans le parcours d'une lecture conventionnelle. On peut donc envisager la conception d'agents, sortes de lunettes de lecture électroniques, capables d'élargir nos capacités de lecture. Ces lunettes, est-il nécessaire de le préciser, doivent être conçus comme des dispositifs au service de nos finalités interprétatives. En d'autres mots, ils doivent être taillés selon notre prescription.

Stratégies de lecture, stratégies d'énonciation

Comme lecteur, quand on aborde des textes, on a des attentes. On lit avec nos questions, nos

connaissances et nos objectifs du moment. Pour répondre à nos attentes, on développe des stratégies de lecture. Ces stratégies de lecture vont devoir se confronter à la stratégie d'écriture du texte par son auteur. Le processus de lecture est donc essentiellement une (re)construction du sens qui suit les prescriptions de lectures suggérées par le texte lui-même mais aussi beaucoup nos propres besoins et dispositions.

Plus encore, la stratégie d'écriture de l'auteur est elle-même largement redevable au contexte général du discours dans lequel elle s'inscrit! Au-delà de lecteur réel, il y a un lecteur fictif auquel l'auteur s'adresse dans un contexte historique précis tel que perçu par l'auteur.

On est donc très loin de la transparence du texte comme simple support de contenus qu'il suffirait de cueillir comme fleurs de printemps!

Examinons une démarche typique d'un lecteur qui recherche des textes dans une démarche informative et analytique.

Dans le contexte de l'édition électronique son premier niveau de stratégie de lecture consistera probablement à rechercher des textes en utilisant, par exemple, les moteurs de recherche sur le web.

Mettons que notre lecteur s'intéresse à la question de la lecture électronique des textes. Donc, il sait que le sujet existe ou il vient de le découvrir. Dépendant de son niveau de connaissance préalable, il devra d'abord trouver les termes utilisés pour aborder le sujet. C'est le problème classique de la recherche documentaire.

Une recherche simple dans les moteurs de recherche risque de nous donner des milliers de références. Notre lecteur devra donc préciser de quel point de vue il veut aborder la question. Recherche-t-il une nouvelle occasion d'affaire? Veut-il savoir comment réagissent les éditeurs à cette nouvelle réalité? Se pose-t-il des question sur l'impact social de l'édition électronique. Est-ce que ça va inciter à lire davantage, par exemple? Est-ce que ça changera nos habitudes de lecture? Etc.

Au terme de cette première étape de repérage, notre lecteur, qui veut s'adonner à une lecture de nature professionnelle ou analytique, aura sélectionné son corpus de référence rassemblant les textes jugés pertinents. Maintenant, il va vouloir répondre de façon plus précise à nos questions.

- “Quels sont les différents points de vue qui s'expriment?
- Quels sont les termes du débat?
- Comment les divers acteurs sociaux se positionnent-ils?
- Est-ce que l'âge, le sexe ou l'origine sociale ou géographique distinguent les diverses positions?
- Est-ce qu'il y a évolution des points de vue dans le temps?
- Si je reprends ma recherche sur Internet dans trois mois, est-ce que je vais pouvoir vérifier facilement si le débat a évolué?”

Pour répondre à ces questions, il faut aussi s'interroger sur la nature des textes, sur leur *genre*. Le texte, on le sait, s'inscrit dans un processus de communication. Il participe à un genre qui en définit la structure générale et que le lecteur se doit de reconnaître pour développer sa stratégie de lecture. La notion de genre dépasse ici l'idée classique de genre littéraire et renvoie plutôt à des conventions sociales plus ou moins explicites. Ces conventions peuvent d'ailleurs appartenir à des groupes sociaux spécifiques. Elles peuvent, ou pas, être sanctionnées par des réseaux institutionnels comme des revues scientifiques par exemple.

D'après François Rastier,

> "Un discours s'articule en divers genres, qui correspondent à autant de pratiques sociales différenciées à l'intérieur d'un même champ. Si bien qu'un *genre* est ce qui rattache un *texte* à un *discours*. (...) L'origine des genres se trouve donc dans la différenciation des pratiques sociales" (1989:40 Sens et textualité, Paris, Hachette, cité p. 22).

On n'abordera donc pas de la même façon un forum de discussion grand public, des écrits dans une revue informatique donnée, un compte-rendu de colloque à l'ACFAS ou un essai philosophique sur l'écrit à l'ère de l'édition électronique!

L'analyse de texte assistée par ordinateur : dispositifs de lecture électronique

Cela nous amène donc au temps deux de la lecture: l'analyse des textes jugés pertinents. On conçoit aisément que la lecture séquentielle des textes à l'écran, ou des textes transférés dans leur format imprimé traditionnel, est le goulot d'étranglement dans notre stratégie de lecture.

C'est là qu'entre en jeux l'analyse de texte assistée par ordinateur.

L'analyse de texte vise à faire ressortir les multiples procédés qui structurent les énoncés et positionnent le texte dans le contexte auquel il participe. Ce n'est pas très différent de la pratique scolaire de l'analyse de texte.

Dans cette tâche, on n'utilise pas l'ordinateur pour *mimer* la lecture humaine. Le radiotélescope ou le microscope électronique ne ressemblent guère à un œil humain pas plus que l'ordinateur ne ressemble à un cerveau humain. Mais, la construction du télescope, son utilisation dans un cadre expérimental et l'interprétation des lectures qu'il nous donne sont l'extension technologique d'une démarche scientifique dirigée par l'humain.

On peut distinguer trois phases dans notre stratégie d'analyse de texte par ordinateur. La <u>première phase</u> consiste à faire parler les données. On a notre corpus électronique et on veut en révéler les caractéristiques générales avant de déployer des stratégies de lecture qui vont dépendre du genre des textes, de leur homogénéité et de leur hétérogénéité et du type de questions posées au texte

Cette première phase inductive n'est pourtant pas empirique. Elle s'appuie sur des hypothèses générales sur la statistique lexicale, sur les divers procédés de la langue et de la pragmatique textuelle. On tiendra compte également de ce que l'on connaît au préalable du processus d'énonciation et de la structure du corpus. On procède aussi dans un va et vient entre le quantitatif et le qualitatif, c'est-à-dire qu'on a toujours le texte au bout des doigts pour vérifier la pertinence des régularités et singularités qu'on dépiste.

Cela nous conduit, de proche en proche, à une deuxième phase d'analyse qui est davantage de nature hypothético-déductive. Il s'agit de construire un dispositif expérimental, une *lunette de lecture* composée de scénarios de commandes et, possiblement d'opérations de catégorisation ou validation manuelles bien définies.

Par exemple, si on pense que les points de vue exprimés sur la lecture électronique diffèrent qu'ils s'expriment à partir du milieu universitaire ou à partir du milieu de l'édition ou des vendeurs de produits, on segmentera le corpus en fonction de leur d'origine. On calculera un lexique pour chacun des segments et on déploiera un analyseur lexicométrique sur certaines catégories de mots qualifiés grammaticalement ou sémantiquement. Voilà comment on peut se tailler sur mesure une petite lunnette de lecture électronique susceptible de valider ou d'invalider notre hypothèse. On verra peut-être que le lexique de départ est insuffisant et qu'il faut aussi s'intéresser aux mots qui sont occurrent avec les termes pivots du débat.

Enfin, on peut imaginer une troisième phase qui consisterait à appliquer nos dispositifs de lecture sur de nouvelles données. La réutilisation et l'adaptation de nos dispositifs de lecture ont alors une double fonction: gagner de la puissance en termes de capacité de lecture, en terme quantitatif, et aussi mesurer l'évolution du discours afin d'ajuster nos modèles de lecture.

C'est ainsi qu'on se constitue nos propres outils de lecture analytique. Ces outils exploitent des mécanismes généraux de la langue et de la pragmatique textuelle. Mais aussi, ils font appel à notre propre connaissance du monde et matérialisent nos théories explicatives. Ils permettent une certaine reproductibilité de nos lectures ou, à tout le moins, une explicitation de nos procédures analytiques.

De grands défis

Manipuler un microscope ou un télescope, ce n'est pas vraiment un jeu d'enfants. L'instrument lui-même doit être compris pour qu'on puisse le lire, c'est-à-dire interpréter sa “vision” intimement reliée à la théorie scientifique qu'il matérialise. L'analyse de texte par ordinateur, en nous obligeant à rompre avec une vision naïve de la lecture, requiert donc un élargissement de notre culture scientifique. La lecture devient objet de science comme le montage d'un dispositif scientifique dans un laboratoire de sciences naturelles. Plus encore, l'analyse de discours qui est un peu la base théorique de l'analyse de texte par ordinateur est essentiellement pluridisciplinaire.

Notre premier défi en est donc un de formation. De la même façon qu'on peut difficilement aujourd'hui ignorer l'intérêt des modèles mathématiques en sciences humaines, sera-t-il possible demain d'ignorer demain l'apport de l'analyse de texte par ordinateur?

Il faut bien constater aussi les limites de nos théories sur le texte et le discours. Ce que l'on maîtrise le mieux, c'est la dimension lexicale et la lexicométrie de même que les l'analyse syntaxique. Mais, au-delà de la phrase ou même de la proposition, on manque de modèles. Le troisième défi est donc théorique.

Également, on affronte les limites de nos outils de calcul. On investit encore très peu dans ce domaine probablement en conséquence de la faiblesse de nos théories et de la formation dans ce domaine. Il faut aussi noter les limites au niveau des formats du texte électronique lui-même qui

est encore le plus souvent une simple image de son équivalent papier ou graphique. La production de textes sur la base de leur marquage logique plutôt qu'éditique est encore à généraliser. C'est le troisième défi.

Mais, à court terme, c'est vraiment le problème de la formation qui bloque l'utilisation à plus large échelle de l'analyse de texte par ordinateur et de la lecture électronique des textes.

François Daoust,

Informaticien au Centre d'analyse de texte par ordinateur, Faculté des sciences humaines, Université du Québec à Montréal.

Étudiant au doctorat à l'École de bibliothéconomie et des sciences de l'information, Université de Montréal.