POUR COMPRENDRE CE QUI VA SUIVRE…

> « Un texte ne saurait être assimilé à une masse de connaissances directement exploitable par la machine. Il faut dans un premier temps prévoir des traitements complexes pour identifier l'information pertinente, la normaliser, la catégoriser et éventuellement la mettre en contexte. Alors seulement l'ordinateur ou l'expert sera capable d'en tirer parti pour mener à bien ses analyses. Mais comment procéder pour extraire l'information pertinente de la masse de données textuelles ? Quels outils utiliser ? Pour quelle pertinence ? […] Plusieurs études ont pointé la frustration des chercheurs en sciences sociales face à ce problème : les textes sont effectivement là, présents et disponibles sur la Toile, mais leur exploitation reste difficile. **Elle exige la collaboration de spécialistes de différents horizons, capables de traiter les données, de fournir les outils pour extraire l'information pertinente et d'ajuster de manière collaborative les traitements** (nous soulignons). »
>
> Thierry Poibeau, « Le traitement automatique des langues pour les sciences sociales : quelques éléments de réflexion à partir d'expériences récentes », *Méthodes digitales – Réseaux*, n° 188, vol. 32, La Découverte, 2014.

Comme le souligne Thierry Poibeau, les chercheurs en Sciences humaines et sociales (SHS) disposent aujourd'hui d'une masse de données textuelles immédiatement disponibles *via* l'internet, et le traitement de ces données les confronte à des questions nouvelles : quels outils statistiques et informatiques utiliser, quelles méthodes mettre en œuvre (comment organiser ces données en corpus par exemple), mais aussi, dans quel cadre théorique ?

C'est à ces questions que cet ouvrage voudrait répondre, en les abordant sous un angle particulier, qui est celui de l'analyse du discours (AD). En

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

effet de nombreux chercheurs ont recours aujourd'hui à des outils de traitement automatique pour « analyser des discours » – que ces discours soient collectés parmi des ressources disponibles ou qu'ils soient suscités, dans des entretiens par exemple. C'est là une pratique de recherche courante dans des disciplines comme les sciences de l'information et de la communication, l'histoire, la science politique, la sociologie, les sciences de gestion… D'un autre côté, si l'on trouve beaucoup d'analyses exclusivement « qualitatives » en AD, le recours à l'informatique et à la statistique fait aussi partie de la discipline, depuis ses débuts. Or l'analyse du discours, comme domaine de recherche ancré dans les sciences du langage, repose sur un certain nombre de postulats – sur la langue, le discours, le sens – et ces postulats théoriques ont des répercussions sur les choix méthodologiques et pratiques : définition des questions de recherche, constitution du corpus, choix des outils informatiques. Ces postulats, nous les rappelons rapidement dans cette introduction, car ils vont guider notre démarche.

Analyser des données textuelles

Les travaux qui analysent les textes et les discours avec l'aide d'outils informatiques ne se situent pas tous ou pas exclusivement en analyse du discours. Il peut être utile de permettre au lecteur de s'orienter parmi toutes les disciplines du texte qui ont recours à l'outillage informatique, en dressant à grands traits la cartographie d'un paysage mouvant aux frontières souvent floues, dont les différentes régions se distinguent tantôt clairement par leurs objectifs et leurs méthodes, et tantôt plutôt par leur histoire ou celle de leur fondateur…

On peut d'abord citer les recherches en **statistique linguistique et lexicale,** auxquelles sont associés les noms de Pierre Guiraud, Charles Muller, Étienne Brunet… L'objet de la statistique lexicale, au départ, ce sont les textes littéraires : elle s'intéresse surtout à la *structure* du vocabulaire de tel ou tel ensemble de textes (voir les travaux de P. Guiraud et de C. Muller). Elle fait intervenir différentes lois concernant les fréquences des mots dans un texte (loi Estoup-Zipf par exemple, chap. 5) ainsi que des tests probabilistes.

L'analyse du discours entre surtout en dialogue avec la lexicométrie, née à la fin des années 1960 autour de Maurice Tournier. Proche de la statistique lexicale, la lexicométrie s'en différencie par le fait qu'elle s'intéresse non pas aux particularités du style d'un auteur mais aux régularités d'un discours, en les mettant en relation avec des déterminations idéologiques ou des positionnements sociaux : c'est ainsi que le Laboratoire de lexicologie politique de l'ENS de Saint-Cloud (ci-après Laboratoire de Saint-Cloud) étudie les tracts de Mai 68 ou les résolutions des congrès syndicaux. La lexicométrie est ainsi très proche de l'AD, mais les deux disciplines se distinguent par le choix des observables : à ses débuts la lexicométrie s'intéresse avant tout

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

au lexique tandis que l'AD, sans pour autant négliger l'unité lexicale, faisait de la syntaxe (les relatives, les nominalisations) son objet d'observation privilégié. À l'heure actuelle, les logiciels de lexicométrie évoluent de plus en plus vers la prise en compte du texte et l'on parle alors plus volontiers de **textométrie**.

De nombreux travaux se situent également à l'intersection de l'AD et de **la sémantique interprétative**, développée par François Rastier, qui vise à rendre compte du sens des textes. Dans ce cadre, l'interprétation se fait de façon différentielle, par un jeu d'oppositions, comme en sémantique structurale (on se rappelle peut-être comment Bernard Pottier définit le terme « chaise » par une combinaison de traits qui le distinguent de « fauteuil » ou de « tabouret »…). Mais pour F. Rastier, le sens d'un mot ne peut être dissocié du texte dans lequel il apparaît, du genre dont relève ce texte et du discours (de la pratique sociale) dont ce genre émane. Sans entrer dans le détail, on peut dire que la sémantique interprétative se distingue voire s'oppose à l'AD sur des points de clivage théorique, en particulier la part de la dimension historique et des contraintes de langue dans la production du sens. Mais étant donné que la sémantique interprétative d'une part travaille sur des corpus représentant des genres de discours et d'autre part a recours à l'outillage et en particulier aux logiciels de textométrie, il arrive fréquemment que des travaux se situent à l'intersection de la sémantique interprétative et de l'analyse du discours (voir aussi « Zoom », p. 49-50 et chap. 6).

La plupart des courants/disciplines listés ci-dessus se retrouvent globalement dans **l'Analyse de Données Textuelles (ADT)**, dénomination englobante et consensuelle qui regroupe une communauté de chercheurs se retrouvant régulièrement dans le cadre des « Journées d'Analyse de Données Textuelles » (JADT). Les chercheurs en ADT analysent des corpus constitués de textes entiers considérés comme représentatifs (d'un genre, d'une pratique, d'une sphère sociale, d'un locuteur, voir chap. 2), dans une visée herméneutique (en relation avec le sens des textes), et ont recours à l'informatique comme outil et comme méthode. Sur tous ces points l'ADT se distingue ainsi du **Traitement automatique des langues (TAL)**, qui vise généralement à produire des outils informatiques permettant de traiter automatiquement des données langagières, dans une visée le plus souvent applicative (industrielle). L'ADT comme l'AD se distingue aussi de la **linguistique de corpus**, qui, dans la tradition de la linguistique empirique britannique, vise à décrire les usages de la langue à travers l'étude de « très grands corpus ». Même si les outils mis en place par la linguistique de corpus, ainsi que l'attention prêtée au contexte dans l'interprétation du sens, peuvent concerner des chercheurs en AD, on voit que l'objectif est ici différent : il s'agit de décrire la langue dont le corpus représente un échantillon.

Par rapport à ces différents domaines de recherche, essayons maintenant de voir quelle peut être la spécificité d'un positionnement en AD.

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

Un positionnement en analyse du discours

Il existe différentes façons de concevoir le discours et l'analyse du discours : citons entre autres l'analyse interactionnelle, l'ethnométhodologie, ou encore l'analyse critique des discours (Critical Discourse Analysis, CDA). Le point commun entre toutes ces approches, **c'est d'envisager les messages ou les textes, qu'ils soient oraux ou écrits, non pas en eux-mêmes mais en relation avec ce qui les entoure.** Certains chercheurs focalisent leur attention sur la situation précise dans laquelle est produit le discours et l'influence que les locuteurs exercent les uns sur les autres ; d'autres plutôt sur la sphère sociale dont émanent les discours, en les reliant à des pratiques sociales ; d'autres encore sur le contexte historique ou politique.

Cet ouvrage se situe globalement dans le courant d'analyse du discours qui s'est développé en France à partir des travaux de Jean Dubois, avec l'accent mis sur l'ancrage social et politique du mot et sur le retour, dans un discours, de séquences posées comme équivalentes, et à partir des théorisations élaborées par Michel Pêcheux et, dans une certaine mesure, par Michel Foucault. Si on en rappelle ici les principales options théoriques, c'est parce que ces travaux sous-tendent la démarche d'ensemble et les choix méthodologiques et pratiques exposés dans les différents chapitres de l'ouvrage.

Pour l'AD, la situation, le contexte, l'environnement du discours ne sont pas ou pas seulement conçus comme matériels : ces extérieurs au discours, ce sont aussi des discours, qui le conditionnent en partie (notion d'« interdiscours »), avec lesquels il « dialogue » (notion de « dialogisme », forgée par Mikhaïl Bakhtine/Valentin Volochinov), ou entre en relation d'une façon ou d'une autre. Par exemple, pour analyser des rapports éducatifs écrits par des travailleurs sociaux, on prendra en compte 1) les discours produits dans ou sur le secteur éducatif et si possible l'institution dans laquelle les rédacteurs travaillent (textes institutionnels, produits en formation, ou entretiens) ; 2) les discours que les médias, l'administration ou les politiques tiennent sur la maltraitance ou l'enfance en danger (articles de presse, rapports officiels) ; 3) les discours juridiques sur ces questions (textes de lois, arrêtés et décrets ; livres écrits par des juges. Voir Cislaru et Sitri, 2012). On va ainsi tenter de saisir ce qui circule d'un discours à un autre, ce qui est repris, reformulé, ce qui répond à un autre discours ou bien anticipe une réponse, une objection. On s'intéresse aux « résonances » (terme employé par André Salem) d'un discours à un autre – et donc d'un corpus à un autre (cf. chap. 6).

De fait, ce qui constitue l'objet de l'AD, c'est la façon dont cet « extérieur » discursif se manifeste dans le discours que l'on étudie et dont il le détermine. Ainsi on ne peut pas comprendre l'emploi de l'expression « être en danger » dans les rapports éducatifs si l'on ne rapproche pas ces rapports du texte de loi qui encadre la protection de l'enfance, où la notion de « danger » ou de

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

« risque de danger » est le motif qui conditionne l'intervention de l'État, ainsi que la levée du secret professionnel pour les professions qui y sont tenues. Ou encore, on peut difficilement interpréter la fréquence d'une formulation figée telle que « X est dans le déni/demande/conflit… » si on ne met pas en relation le discours des éducateurs avec le discours des psychologues ou des psychanalystes dont c'est le lexique professionnel. On peut ainsi rendre compte de la présence/absence/fréquence d'une forme ou encore du choix de cette forme plutôt qu'une autre, ce qui nous conduit au point suivant.

L'AD est une **discipline interprétative** : elle vise à rendre compte de la façon dont se construit le sens dans le discours étudié. Or, pour l'AD, le sens ne se construit pas « en dehors » de la langue, mais dans l'interaction entre les formes (mots, constructions syntaxiques, ponctuation…) et les déterminations extérieures dont on vient de parler. Par exemple, M. Pêcheux montre avec plusieurs exemples comment l'interprétation d'une relative met en jeu le contexte socio-historique et les discours par lesquels il se manifeste : contexte politique et idéologique dans « c'est le devoir du parti et des communistes d'apporter leur soutien au développement **des luttes qui montrent la volonté des salariés de contrecarrer les attaques antisociales et antidémocratiques du gouvernement et du patronat** » (déclaration du bureau politique du PCF, juin 1978), phrase dont l'interprétation est constitutivement ambiguë, selon que l'on considère que « le PCF soutient toutes les luttes car ces luttes montrent... » ou bien que « le PCF soutient les luttes, à condition qu'elles montrent… » (Pêcheux, 1979, *in* Maldidier 1990, p. 273-280). On citera aussi un exemple analysé par André Salem (repris dans le chap. 6) : il montre comment, dans les résolutions des congrès de la CFDT, la forme *travailleurs* disparaît au profit de la forme *salariés*, sans que pour autant on puisse parler de simple substitution. C'est une vision différente du monde du travail (plus sociale d'un côté, plus juridique de l'autre) que chacun des mots engage, et ce sont aussi des contextes linguistiques différents (on ne parle pas de « salariés immigrés »). Pour résumer, ce qui caractérise l'AD, c'est la place centrale accordée aux formes langagières, **le postulat que l'on ne dit pas la même chose en le disant autrement**, et que l'on n'accède pas « directement » au sens d'un discours, que le discours n'est pas le simple « reflet » d'une pensée ou d'une idéologie. De ce point de vue, **l'analyse du discours ne doit pas être confondue avec ce qu'on appelle « l'analyse de contenu »**, qui se propose d'« accéder au sens d'un segment de texte, en traversant sa structure linguistique » (Pêcheux, 1969, p. 4).

Analyse du discours et informatique

Si l'outillage informatique de l'AD s'inscrit dans un mouvement général lié au développement des humanités numériques et des approches quantitatives en SHS, on aurait tort de croire qu'il s'agit là d'une alliance contre nature.

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

Pour M. Pêcheux, l'informatique est une sorte de garant méthodologique : elle « exige des analystes de discours une *construction explicite* de leurs procédures de description, ce qui est *la pierre de touche de leur consistance d'objets théoriques* » (Pêcheux et Marandin [1984] cité par Jacqueline Authier-Revuz, à paraître, chap. 10).

M. Tournier, de son côté, parlait de « refroidir le corpus » (entretien avec Pierre Fiala du 6 avril 2016) pour expliquer comment l'outil informatique permettait de garantir une lecture objective du texte. C'était aussi un moyen de hisser les sciences humaines au même rang que les sciences exactes. L'automatisation des analyses s'est imposée en raison de la complexité des calculs et du volume toujours plus important des données à traiter. On imagine mal par exemple des concordanciers ou des décomptes effectués à la main sur plusieurs milliers de textes. Le perfectionnement technologique a rendu possible cette automatisation.

De fait, un des objectifs de l'AD, dès le début, est d'échapper à une pratique de lecture allant directement au sens pour, au contraire, privilégier des méthodes qui, comme le dit M. Pêcheux, relèvent d'« un parti-pris pour l'imbécillité », c'est-à-dire qui suspendent la compréhension immédiate d'un texte.

C'est ainsi que l'AD s'intéresse à ce qui se répète dans un texte, aux séquences qui reviennent plus ou moins à l'identique dans différents endroits du corpus. Or la mise en série, qui est au cœur de la lexicométrie, entraîne de fait la délinéarisation du texte : on ne l'étudie plus phrase après phrase, paragraphe après paragraphe, mais on utilise des outils qui parcourent l'ensemble du corpus pour produire, le plus souvent, des résultats sous formes de listes. De tels procédés font voir des récurrences ou des rapprochements souvent impossibles à détecter à l'œil nu, à plus forte raison sur de grandes masses de documents. L'observation de l'enchaînement linéaire – ce qu'on appelle le « fil du discours » –, *via* des formes comme les reprises anaphoriques, le discours rapporté…, constitue cependant toujours un objet pour l'AD. Cette dimension linéaire du texte trouve aujourd'hui des débouchés techniques, car les logiciels de traitement des données développent de plus en plus des fonctionnalités qui permettent, au cours de la recherche, de revenir au texte ou de visualiser une portion de texte.

Si la démarche globale présentée dans cet ouvrage relève de l'AD, il est clair que, au sein des disciplines qui explorent les textes avec des outils informatiques et statistiques, se nouent des dialogues, des échanges et des collaborations qui rendent parfois difficile « l'étiquetage » disciplinaire de telle ou telle recherche. Le développement même des outils et de fonctionnalités nouvelles, qui se fait sur un mode largement empirique souvent en relation avec tel ou tel projet, tel ou tel laboratoire, tel ou tel corpus, n'est pas sans influer sur les méthodes, et provoque des « bougés » dans un cadre théorique. Ainsi, à l'heure actuelle, les débats sur l'annotation (cf. chap. 3) semblent largement dépassés

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

et il est admis que l'annotation constitue une ressource qui, à un moment de la recherche, peut se révéler pertinente et produire des résultats intéressants. Plus profondément, on voit que la contradiction entre des observables qui, du côté de la lexicométrie, sont de l'ordre du « mot » (la forme graphique) et, du côté de l'AD, plutôt de l'ordre de la syntaxe, n'est pas sans conséquence sur l'orientation actuelle des recherches en AD outillées par l'informatique.

Pourquoi un ouvrage sur les méthodes automatisées en sciences humaines et sociales ?

S'il existe aujourd'hui plusieurs ouvrages de référence sur la statistique textuelle, la linguistique de corpus ou encore la linguistique « instrumentée » (Lebart et Salem, 1994, Habert *et al.*, 1997, Habert, 2005, Poudat et Landragin, 2017, par exemple) – souvent écrits par des statisticiens et des informaticiens eux-mêmes, la spécificité du présent ouvrage est de **questionner ces outils et leurs méthodes en les articulant directement à des problématiques d'analyse du discours**.

Alors que les données numériques sont innombrables et que les ressources et les outils sont immédiatement disponibles, la tentation est grande pour les chercheurs ou pour les étudiants de s'en « remettre » à la machine ou au logiciel. Or, comme on le verra tout au long de cet ouvrage, l'instrumentation informatique nécessite de prendre des précautions **avant** (préparation des données), **pendant** (analyse des données) et **après** le recours à l'outillage (interprétation des résultats). La machine ne dispense pas le chercheur, bien au contraire, d'un questionnement sur la validité de sa démarche et sur la pertinence de ses données. Les premiers programmes signifiaient au chercheur ses limites – certains se rappelleront ce contrariant message envoyé par la machine : « syntax error » – ce qui l'invitait à s'interroger sur son cheminement, le choix et la préparation de ses données. Les outils contemporains ne favorisent pas d'emblée un tel questionnement dans la mesure où ils fournissent immédiatement des résultats au chercheur. Mais comment ces résultats sont-ils produits ? Que veulent-ils dire ? Que nous apprennent-ils sur le discours que l'on veut analyser ? Il n'est pas toujours aisé de répondre à ces questions. Par ailleurs, la multiplication des logiciels et la diversité des méthodes proposées exigent un minimum de connaissances sur leurs principes et leurs fonctionnalités. C'est à cette condition que l'on peut s'orienter librement vers l'outil le plus adapté à la problématique posée et aux données collectées. Il s'agit également de ne pas mésestimer la dimension heuristique de ces instruments, dont l'usage enrichit généralement les questions de recherche, en conduisant parfois vers de nouvelles pistes et des hypothèses que le volume des corpus ou une analyse « à la main » ne permettait pas d'envisager.

Cet ouvrage a aussi pour ambition de rendre ces méthodes abordables et applicables pour des étudiants et des chercheurs en sciences humaines

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr

et sociales qui s'intéressent au discours. Les auteurs de cet ouvrage, venant de différentes disciplines, utilisateurs voire concepteurs d'outils informatiques et statistiques, ont en effet constaté, dans leurs pratiques de recherche et d'enseignement, un véritable manque tant au niveau épistémologique (**pourquoi compter les mots ?**) que méthodologique (**qu'est-ce qu'on compte et qu'est-ce qui compte ?**) ainsi qu'un besoin de vulgarisation et d'explicitation.

Il s'agit donc de répondre à différentes attentes dans un contexte où le traitement informatique des textes se généralise avec la numérisation d'un plus grand nombre de documents et la production abondante de données numériques, et où compter les mots, mesurer les discours semble aujourd'hui devenu une étape incontournable pour de nombreuses recherches en sciences humaines et sociales. Dans ce contexte, l'AD, dans sa dimension interdisciplinaire, dans sa visée interprétative et dans son dialogue continu avec les autres approches de l'ADT, est une démarche parmi d'autres (stylistique, linguistique de corpus, sémantique…), qui peut tantôt guider tantôt informer une analyse informatisée de données textuelles, de la constitution du corpus à l'interprétation des phénomènes jugés pertinents, en passant par le choix des outils et des fonctionnalités.

Frédérique Sitri, Christine Barats.

Méthodes et outils informatiques pour l'analyse des discours – Émilie Née (dir.)
ISBN 978-2-7535-5499-3 — Presses universitaires de Rennes, 2017, www.pur-editions.fr