Chapitre 9

ANALYSER LES CONTENUS, LES DISCOURS OU LES VÉCUS ? À CHAQUE MÉTHODE SES LOGICIELS ![1]

1. Par **Christophe Lejeune**.

Sommaire

1. Introduction

La recherche qualitative regroupe des méthodes aussi diverses que l'ethnographie, l'analyse de contenu, les parcours de vie, les études narratives, l'analyse de discours, l'approche compréhensive, l'analyse par théorisation ancrée, la méthode d'analyse en groupe ou l'analyse phénoménologique interprétative[1]. Toutes ces méthodes donnent lieu à des recherches recourant à des logiciels d'analyse de textes. Toutefois, seules trois d'entre elles ont spécifiquement motivé le développement de programmes informatiques. À l'origine, chaque logiciel a donc été conçu pour répondre à une question de recherche particulière, issue d'une de ces trois méthodes.

Ce chapitre nourrit des ambitions qui dépassent l'inventaire ou le banc d'essai. Présenter ce que proposent les différents logiciels permet tout d'abord au lecteur qui le souhaite d'opérer un choix en connaissance de cause. Rappeler l'origine méthodologique des différents logiciels évite ensuite l'amnésie d'une littérature scientifique exclusivement consacrée aux logiciels d'analyse de textes (Gilbert, Jackson et di Gregorio, 2014) et stimule la réflexion méthodologique que tout recours à un outil devrait susciter. Distinguer trois formalisations de la recherche qualitative et en identifier la logique interne tempèrent aussi les invitations à croiser, combiner ou conjuguer des méthodes aux présupposés inconciliables. Circonscrire la pertinence de chaque programme désamorce enfin l'illusion de la « boîte magique », dénonce le mensonge de l'outil a-théorique et le mirage du logiciel universel.

Ce chapitre situe les différents logiciels selon deux paramètres : leur prise en charge de l'interprétation des textes et leur verbosité. Certains logiciels se contentent de découper, réorganiser ou réagencer les textes analysés. Poursuivant une visée essentiellement descriptive, leur interface ne comporte que des matériaux ; elle est dépourvue d'éléments interprétatifs supplémentaires. Ces logiciels figurent dans la colonne de gauche du tableau 9.1. À l'inverse, d'autres logiciels intègrent une « couche » interprétative. Leur interface appose (voire substitue) au texte analysé des mots choisis par l'ordinateur, par le concepteur du logiciel ou par le chercheur. Ces logiciels apparaissent dans la colonne de droite du tableau 9.1.

1. Ce chapitre a bénéficié de la lecture serrée de deux évaluateurs anonymes, des remarques avisées de Julie Gilles de la Londe et de Jean-Sébastien Cadwallader ainsi que de l'accompagnement bienveillant de Marie Santiago-Delefosse et de Maria del Rio Carral. Je remercie chacune de ces personnes et reste bien entendu seul responsable des imprécisions qui subsisteraient dans ce chapitre.

Le deuxième paramètre caractérise la verbosité ou le mutisme du logiciel. Certains logiciels calculent des résultats dès leur lancement et les affichent d'entrée de jeu, sans attendre d'intervention de l'utilisateur. Je les qualifie de « bavards » (première ligne du tableau 9.1). À l'inverse, les logiciels muets attendent une interrogation de l'utilisateur pour afficher des résultats (deuxième ligne du tableau 9.1). La direction empruntée dépend de l'initiative de l'utilisateur.

Ces deux paramètres ne relèvent pas (seulement) de l'ergonomie des interfaces. Ils supposent des positionnements épistémologiques distincts.

Tableau 9.1 – *Verbosité et représentation de l'interprétation*

	L'interface réagence le texte sans ajouter d'éléments d'interprétation	**L'interface représente des éléments d'interprétation**
Bavard	Lexicométrie	Dictionnaires, automates, registres
Muet	Concordanciers	Logiciels d'étiquetage réflexif

Suivant un ordre historique, ce chapitre débute avec la première rencontre de la recherche qualitative et de l'informatique. Les premières analyses de contenu portaient sur le relevé des thématiques abordées dans la presse quotidienne. Or l'identification de ces thématiques variait d'une personne à l'autre (section 2.1). Pour en objectiver le codage, les analystes de contenu définirent des dictionnaires, chacun doté d'un nom et composé d'une liste de mots ou d'expressions (section 2.2). Un ordinateur repérait les éléments de la liste et superposait le nom du dictionnaire concerné au texte d'origine. Ainsi fut résolu le problème de la fiabilité du codage.

La deuxième section concerne un courant spécifiquement francophone d'analyse de discours. Réagissant notamment à la réduction des textes à leur contenu, cette méthode se propose de caractériser les formes du discours (section 3.1). Se réclamant de cette tradition méthodologique, les logiciels de lexicométrie (section 3.2.1) transforment automatiquement les textes en listes de mots, dont ils calculent la fréquence. S'appuyant sur les mêmes calculs, les automates suggèrent des pistes d'interprétation générées automatiquement (section 3.2.2). Lexicométrie et automates s'adjoignent des concordanciers, permettant l'affichage en contexte de toutes les occurrences d'un mot donné (section 3.2.3).

La troisième section montre comment la formalisation de l'analyse par théorisation ancrée a inspiré la conception des logiciels d'étiquetage réflexif (section 4.2.1) et des registres (section 4.2.2).

2. De la subjectivité des codeurs à l'objectivité des machines

2.1 Méthode : l'analyse de contenu

Les logiciels d'analyse de textes sont apparus dès les années 1950 comme une réponse aux problèmes logistiques et épistémologiques que rencontraient les praticiens de l'analyse de contenu.

Dans l'entre-deux-guerres, des chercheurs américains s'intéressent au rôle grandissant de la presse et de la radio. Sensibilisée par le déroulement de la Première Guerre mondiale puis par la montée du nazisme en Allemagne, l'équipe d'Harold Lasswell décide d'étudier les phénomènes de propagande. À cette fin, ces chercheurs mettent au point une technique adaptée à leur question de recherche. Entre eux, ils identifient tout d'abord les positions ou arguments sur lesquels les valeurs américaines et nazies s'opposent. Puis, ils en recensent et dénombrent les occurrences dans la presse américaine et allemande. Cette logique comptable découle d'une définition de la propagande, certes rapide, mais pertinente. La propagande est envisagée comme une façon de communiquer consistant à répéter sans relâche le message qu'elle entend diffuser dans l'opinion publique. Puisque la répétition appelle la fréquence, Harold Lasswell et ses collègues baptisent la technique mise au point « sémantique quantitative ».

Bernard Berelson (1952) formalise ensuite cette technique et donne à l'analyse de contenu la forme que nous lui connaissons aujourd'hui. La scientificité de cette méthode dépend de la définition des catégories utilisées. Celles-ci doivent impérativement être exhaustives, mutuellement exclusives, objectives et pertinentes.

- L'exhaustivité renvoie au fait qu'aucune portion du matériau ne peut être écartée. Vu que tout doit être codé, le jeu de catégories doit prévoir tous les cas de figure (ce qui n'est pas nécessairement aisé).
- L'exclusivité implique que les catégories ne peuvent tolérer aucun recouvrement. Pour assurer que les catégories soient mutuellement exclusives, les chercheurs les définissent au moyen d'exclusions logiques, comme les modalités d'une variable nominale en statistique. L'analyse de contenu mobilise dès lors des oppositions classiques. Une portion de matériau donnée peut en effet difficilement se voir qualifiée à la fois de sentiment et de raison, d'argument à charge et à décharge, de compliment et de critique ou encore

de plaidoyer et de réquisitoire. Certes quelque peu réductrices, ces oppositions permettent d'assurer que les catégories ne se recouvrent pas.

- L'objectivité implique une définition claire, univoque et sans ambiguïté. Cette rigueur permet non seulement d'assurer l'exclusivité (décrite au point précédent), mais elle garantit également la scientificité de la méthode. Une définition objective se comprend de la même manière, quelle que soit la personne. Cette univocité détermine la fiabilité de la méthode.
- Enfin, la pertinence désigne l'adéquation entre les catégories, le matériau analysé et la question de recherche. Le jeu de catégories doit être adapté au type de contenu du matériau. Les catégories doivent également recenser des éléments qui permettront, une fois quantifiés, de répondre à la question de recherche. Si elle semble triviale à la lecture, la pertinence n'est cependant pas si évidente à atteindre. Définir des catégories permettant réellement de répondre à la question de recherche est une tâche toujours exigeante, voire parfois complexe.

Une analyse de contenu se réalise en quatre étapes : la définition du livre de codes, le découpage du matériau, son codage et l'analyse des résultats.

Comme l'équipe d'Harold Lasswell, le chercheur commence par définir les catégories qui devront être recensées dans le matériau. Cette définition, particulièrement rigoureuse, respecte les quatre principes susmentionnés. Le document accueillant la définition des catégories (exhaustives, mutuellement exclusives, objectives et pertinentes) s'appelle le « livre de codes ».

Parallèlement à la définition de ce livre de codes, le chercheur procède au découpage des différentes portions du matériau à analyser (ces portions s'appellent des « unités d'enregistrement »). Il identifie clairement les éléments à exclure de l'analyse (par exemple, les photographies, les tableaux et les titres) et segmente le matériau retenu. Ce découpage conditionne le reste de la démarche : une unité d'enregistrement qui comporterait deux idées différentes compromettrait l'exclusivité des catégories.

Dans un troisième temps, le chercheur passe le relais à une équipe composée de personnes n'ayant pas pris part à la définition du livre de codes. Il leur transmet (1) le livre de codes, (2) le corpus découpé et (3) des consignes indiquant comment attribuer les catégories du premier aux portions du deuxième. Ces consignes mentionnent, par exemple,

s'il faut prendre en considération la phrase, le paragraphe ou l'ensemble du texte pour comprendre une unité d'enregistrement donnée.

Une fois ces instructions transmises et comprises, les personnes procèdent individuellement au codage, sans se concerter. Le chercheur récupère le résultat du codage, procède à son dépouillement, à son encodage et à son analyse statistique.

Le premier traitement consiste à mesurer la fidélité inter-codeurs : il s'agit d'identifier si tous les codeurs ont codé le matériau de la même manière. Au moins trois types de divergences peuvent survenir. Si une portion donnée se voit attribuer différentes catégories, c'est l'indice d'un problème de découpage ou d'une ambiguïté inhérente au texte. Quand deux catégories ont tendance à être utilisées l'une pour l'autre, leur définition – et en particulier leur exclusivité – est problématique. Lorsqu'un des codeurs se démarque systématiquement des autres, le chercheur considère que le travail de cette personne n'est pas fiable ; il l'écarte alors de son analyse, ce qui améliore automatiquement la fidélité inter-codeurs.

En analyse de contenu, la convergence des différents codages indique l'univocité (et donc l'objectivité) des catégories du livre de codes. Si différentes personnes comprennent et appliquent une définition de la même manière, c'est qu'elle est objective. La mesure de la divergence entre les codeurs n'est donc valide qu'avec des personnes n'ayant pas participé à la maturation du livre de codes. C'est pour cette raison que le chercheur ne peut procéder lui-même au codage et doit le confier à d'autres personnes : connaissant l'intention ayant présidé à la définition des catégories, il mobiliserait des informations implicites, non reportées dans le livre de codes ; son codage ne permettrait dès lors pas de mesurer l'objectivité des catégories.

L'analyse de contenu est une démarche hypothético-déductive et quantitative. Lors de leur définition, les catégories sont *déduites* des hypothèses de recherche, en amont de tout contact avec le matériau. Les étapes suivantes entendent les éprouver et les *quantifier* : les résultats d'une analyse de contenu s'expriment en chiffres.

2.2 Logiciels : les dictionnaires informatiques

La rigueur de l'analyse de contenu n'empêche pas certaines difficultés pratiques, notamment discutées lors du premier symposium consacré à cette méthode en 1955 (De Sola Pool, 1959). Une de ces difficultés concerne ce que l'analyse de contenu attend du codeur. Idéalement,

celui-ci suit scrupuleusement les consignes et n'en dévie à aucun moment. Sa compréhension de la définition des catégories exclut toute sensibilité personnelle. Il code invariablement de la même manière, consensuellement, sans réfléchir. Étudiants ou professionnels, ces opérateurs particuliers se révèlent difficiles à recruter, même contre rémunération.

Ce problème logistique fut le terreau de l'introduction de l'informatique en analyse de contenu. L'ordinateur constitue bien cet opérateur décérébré se conformant sans relâche aux instructions qui lui sont soumises, sans fatigue, ni sensibilité, ni variation d'humeur.

Pour que le codage puisse être automatisé, le chercheur arrête la liste des termes ou des expressions se rapportant à chacune des catégories du livre de codes. Les catégories sont alors conçues comme des dictionnaires, dont le nom sera automatiquement attribué aux unités d'enregistrement dans lesquelles apparaissent les termes ou les expressions prédéfinies. Dès 1961, le codage automatique et systématique de textes (transcrits sur cartes perforées) est opéré par le *General Inquirer*[1], que Philip Stone (1966) développe avec l'aide d'IBM.

Contrairement à d'autres programmes informatiques (présentés plus bas), les dictionnaires ne se contentent pas de réagencer le corpus de textes, mais offrent une valeur ajoutée directement visible. La présence ou l'absence d'un dictionnaire, l'affectation de son nom à un ou plusieurs passages et le dénombrement de sa fréquence dans l'ensemble du corpus ajoutent une « couche » interprétative. En outre, le logiciel réalise ces opérations sans intervention de l'utilisateur ; il est bavard (tableau 9.1).

Les dictionnaires héritent des soubassements épistémologiques de l'analyse de contenu. Définis en amont du codage, ils procèdent d'une démarche hypothético-déductive. Classés en fonction de leur fréquence, ils s'inscrivent également dans une épistémologie quantitative. Ils s'avèrent donc tout à fait adéquats pour mesurer des phénomènes quantitatifs, répétitifs ou sériels, comme la propagande ou les rumeurs. *A contrario*, les dictionnaires sont moins pertinents pour identifier, dans des entretiens, l'expression de sentiments ou de ressentis personnels, dont l'éventuel caractère répétitif n'est pas démontré.

1. Les adresses de tous les logiciels cités dans ce chapitre sont disponibles sur le site Internet de l'auteur : *http://www.squash.ulg.ac.be/logiciels/*.

Exemples de dictionnaires

Développés à Harvard, les premiers dictionnaires entendaient contribuer à une théorie générale de l'action sociale. Ils avaient donc vocation à convenir à n'importe quelle recherche en psychologie et en sociologie, sans que leur pertinence s'étende pour autant aux disciplines connexes, comme l'anthropologie et les sciences politiques (Stone *et al.*, 1966). Bien que leur extension soit très large, ils ne prétendaient donc pas à l'universalité.

Comme les catégories exclusives de l'analyse de contenu, les dictionnaires fonctionnent souvent par deux (la retenue émotionnelle contre l'extraversion) ou trois (les points de vue favorables, critiques ou neutres). Même si ces oppositions semblent sommaires, leur opérationnalisation autorise une certaine finesse.

Chaque dictionnaire comporte plusieurs dizaines (voire centaines) d'expressions ou de mots-clés. Il n'est donc pas possible d'en reproduire un intégralement ici. À titre illustratif, le dictionnaire de la coopération comprend plus de cent mots-clés dont : harmonieux, se mettre d'accord, aligné, arbitrer, se coordonner, collègue, collectif. Un autre recense les lieux de socialisation : aéroport, église, école, entrepôt, ferme, rue, usine…

Utilisés dans les analyses de presse, ces dictionnaires ont (notamment) permis d'attester la prolifération des faits divers sous différentes rubriques journalistiques, de montrer l'évolution éditoriale (sur dix ans) d'un journal américain (le *New York Times*) comparé à ses homologues européens ou d'identifier les périodes électorales lors desquelles les démocrates et les républicains insistent sur leurs propres arguments ou critiquent ceux du camp adverse.

Les dictionnaires ont joué un rôle précurseur dans l'histoire des logiciels d'analyse de textes. La plupart des développements ultérieurs semblent cependant les avoir oubliés, à part en linguistique, où des dictionnaires (de pronoms, de noms ou de verbes) permettent d'identifier les catégories syntaxiques. Deux exceptions méritent d'être mentionnées : les « scénarios » de *Tropes* et les « êtres fictifs » de *Prospéro* présentent les caractéristiques des dictionnaires. Ces logiciels sont les derniers à mobiliser les dictionnaires pour leur portée interprétative.

Malgré leur potentiel toujours actuel, les dictionnaires de l'analyse de contenu sont largement éclipsés par d'autres traitements informatiques, inspirés par d'autres traditions méthodologiques.

3. Les formes du dire

3.1 Méthode : l'analyse de discours

En France ou au Québec, de nombreux logiciels d'analyse de textes se revendiquent de l'analyse de discours. Cette méthode émerge dans les années 1960. Elle reproche à l'analyse de contenu de se focaliser sur les informations manifestes, au mépris du contexte de leur élaboration. À l'objectivité des livres de codes, l'analyse de discours oppose une tradition littéraire. Elle n'emprunte pas la voie de la quantification, mais se nourrit des répertoires de la philologie et des sciences du langage. Contrairement à l'analyse de contenu, elle ne connaît pas une formalisation unique, mais rassemble plutôt des recherches très hétérogènes. En somme, l'analyse de discours n'existe pas comme une tradition unique, mais comme une bannière rassemblant des recherches très différentes. Identifier le dénominateur commun à ces recherches foisonnantes serait ardu. Tel n'est d'ailleurs pas mon propos. Dans le cadre de ce chapitre, je m'intéresse exclusivement à l'analyse de discours que se proposent d'outiller les logiciels d'analyse de textes. Ce qui suit ne concerne donc qu'un courant parmi d'autres : l'analyse de discours « à la française ».

Pour cette analyse de discours particulière, chaque situation induit une mise en mots déterminée. Partant de ce constat, l'analyse de discours postule que tout texte comporte des traces de son contexte de production. L'analyse de ces traces informe dès lors sur son origine et permet d'élaborer une typologie des discours : chaque genre, chaque style, chaque discours se caractérise par des propriétés formelles particulières, comme les connecteurs argumentatifs, les temps verbaux, la longueur des phrases ou le registre de langue. Ainsi définie, l'analyse de discours vise à caractériser précisément et finement chaque discours, ce qui implique d'en recenser les caractéristiques formelles. Contrairement aux deux autres méthodes présentées dans ce chapitre, cette analyse ne vise pas à découvrir ce qui précède ou sous-tend les discours, mais elle s'y intéresse *pour eux-mêmes*. Sa visée est essentiellement descriptive et typologique.

3.2 Logiciels

Les logiciels qui se réclament de ce courant particulier combinent deux types de traitement, complémentaires l'un de l'autre. Bavard, le comptage de mots propose d'autorité une vision d'ensemble du corpus ;

muet, le concordancier attend que l'utilisateur lui soumette un mot pour afficher ses contextes d'apparition. Le *General Inquirer* intégrait déjà ces deux traitements. Toutefois, réagençant les textes sans participer à leur codage, ces traitements ne trouvaient guère de justification méthodologique dans l'analyse de contenu. Le courant descriptif et typologique de l'analyse de discours à la française la leur fournit.

3.2.1 La lexicométrie

Pour faire (très) simple, la lexicométrie ou la textométrie consiste à appliquer des calculs statistiques à un corpus de textes. Procédant au comptage de mots, les traitements proposés produisent des résultats relevant de la statistique.

Chaque forme est associée à une fréquence. Présentées en liste, ces fréquences peuvent être triées : en tête, les formes, mots ou signes de ponctuation les plus fréquents ; en bas, les mots rares (qui n'apparaissent qu'une fois – on parle d'*hapax*). La fréquence simple peut-être pondérée : répéter « extraordinaire » à trois reprises dans un texte de quatre lignes ou de quatre pages ne revient pas au même. En outre, la distribution des fréquences varie également d'un corpus à l'autre. Le calcul des spécificités permet d'identifier la sur- ou sous-représentation, par texte, de chaque forme par rapport à sa distribution dans l'ensemble du corpus.

Des statistiques simples permettent également d'identifier que deux, trois ou quatre mots se suivent de manière répétée et forment ainsi une expression. On parle de collocation. Les logiciels de lexicométrie repèrent également les mots qui apparaissent régulièrement ensemble sans nécessairement se succéder. On parle cette fois de cooccurrences.

Des logiciels comme *Hyperbase*, *Lexico*, *Sato*, *T-Lab, TXM* ou *Iramuteq* proposent ces différents calculs. Les listes d'occurrences, les spécificités, les collocations et les cooccurrences constituent des indicateurs utiles, faciles à mettre en œuvre et peu sensibles à l'augmentation de la taille du corpus. Autrement dit, une liste ne change pas de nature en s'allongeant ; elle reste une liste. En outre, au fur et à mesure que le corpus s'étoffe, le lexique tend à se répéter ; la liste des occurrences s'allonge donc moins vite.

Comme les dictionnaires de l'analyse de contenu, la lexicométrie est bavarde (tableau 9.1) : le logiciel calcule en une fois les occurrences, les spécificités, les collocations et les cooccurrences pour tout le corpus. Les résultats s'affichent sans intervention de l'utilisateur. Ce fonctionnement convient particulièrement à l'approche descriptive de l'analyse

de discours. Cette « verbosité » présente en outre des vertus heuristiques, des résultats inattendus étant susceptibles de s'imposer à l'utilisateur.

La lexicométrie présente une double assise, descriptive et quantitative. Les listes et les autres modes de visualisation disponibles procèdent d'un réagencement des parties du discours. L'interface des logiciels de lexicométrie ne superpose donc pas aux textes une « couche » interprétative comparable aux noms des dictionnaires de l'analyse de contenu (tableau 9.1). Prolongeant la visée descriptive de l'analyse de discours présentée plus haut, elles visent à caractériser le(s) discours. Certains utilisateurs de ces logiciels revendiquent d'ailleurs un certain ascétisme interprétatif, affirmant qu'il ne leur appartient pas d'aller au-delà de cette caractérisation. Cet argument ne rencontre cependant pas l'unanimité ; des débats récurrents divisent la communauté des utilisateurs (Bonoli, 2014).

Les logiciels lexicométriques s'adossent également au paradigme statistique. Ce faisant, ils s'écartent de la tradition d'analyse de discours dont ils se réclament. Pour rappel, à l'origine, l'analyse de discours oppose une méthode *qualitative* à l'analyse de contenu. Or la lexicométrie procède d'une analyse quantitative du discours. À première vue, cette proposition peut sembler contradictoire avec le positionnement originel et historique de l'analyse de discours. À tout bien considérer, la quantification s'inscrit toutefois dans le programme épistémologique initial. Au fond, les discours subsument leurs manifestations : ce concept typologique vise la façon dont on parle, plus que ce dont on parle (le contenu) ou ce que l'on exprime (le vécu). Agréger de grandes quantités de textes permet de caractériser des discours, rencontrant ainsi le projet initial de l'analyse du discours.

3.2.2 Un cas particulier : les automates

Les logiciels de lexicométrie offrent différents types de calculs. Ceux-ci multiplient les parcours du corpus de textes, ainsi que les croisements qu'ils autorisent. Issus des mêmes statistiques textuelles que la lexicométrie, certains outils se proposent pour leur part d'exploiter un seul type de calcul. Afin d'en simplifier l'exécution, différentes opérations sont automatiquement enchaînées et génèrent une représentation en deux dimensions, comme un plan factoriel, un dendrogramme, un diagramme stratégique ou une carte de cooccurrences[1].

1. Le chapitre de Valérie Capdevielle-Mougnibas illustre l'usage croisé de la lexicométrie et d'un de ces automates.

Avec cette mise en boîte noire, les automates prennent leur distance avec la tradition de l'analyse de discours. Principe au cœur des logiciels *Leximappe*, *Candide*, *RéseauLu* et *Calliopé*, la cooccurrence se voit rebaptisée en « réseau de mots associés » et ainsi inscrite dans la théorie des acteurs réseaux, développée en anthropologie des sciences et des techniques (Lejeune, 2004). Le logiciel *Alceste* enchaîne, quant à lui, des traitements issus des statistiques de Jean-Paul Benzecri (connu en France pour ses analyses factorielles). Son concepteur, Max Reinert, se revendique de la sémiologie de Charles Sanders Peirce. Enfin, seul logiciel libre de cette famille, *Iramuteq* propose à la fois les procédures automatisées par *Alceste* et les traitements classiques de la lexicométrie.

3.2.3 Les concordances

La lexicométrie et les automates tirent parti de la force du comptage. Cet atout présente cependant un revers : traiter le texte comme un « sac de mots », c'est risquer d'amalgamer des homonymes aux référents bien distincts (Rastier, 2011). Pour contrer ce problème, les logiciels de statistique textuelle (comme *Hyperbase*, *Lexico* et *TXM*) se sont dotés d'une fonctionnalité complémentaire : la concordance.

Fruits des controverses sur l'interprétation des textes sacrés, les concordances émergent au XIII^e^ siècle[1]. Avec leur informatisation, dès la fin des années 1950, elles alignent verticalement les différentes occurrences de l'expression ou du mot recherché, présentées dans leur cotexte originel (Luhn, 1960). Ce mode de présentation permet de se faire une idée des significations en présence d'un coup d'œil.

Exemple de concordances

Après vingt ans d'enseignement secondaire, Carine Willemsen (2015) reprend des études en sciences de l'éducation et décide de réaliser un mémoire sur la mobilité des enseignants. En entretien, une enseignante belge lui raconte son séjour dans une école du sud de la France pratiquant la pédagogie Freinet. Lors de la lecture de la transcription de l'entretien, la mémorante est intriguée par le rôle joué par la directrice de l'établissement belge. Afin d'en avoir le cœur net, elle utilise un concordancier pour afficher en contexte toutes les occurrences de la « directrice » et de la « direction ».

☞

1. Sur l'histoire passionnante des concordances, voir Pincemin (2006) et Lejeune (2007).

☞

```
 1 la barre haut comme ça. Et la direction aussi parce que là ça fait
 2 créé deux classes-pilotes. La direction de l'école nous a donné l'a
 3  quand on a compris, quand la direction nous a montré toutes les po
 4  fortement négatives. Donc la direction nous a suggéré de nous mett
 5 C'est l'année passée donc, la direction nous dit : « Écoutez, vous
 6  venir qui viennent. C'est la direction qui a invité et l'associati
 7 rs Erasmus comme ça. C'est la direction qui parce qu'on était déjà
 8 c'est vraiment le fait que la direction soit venue avec un dossier
 9  rencontre entre enseignants, direction, PMS, élèves. Et ce groupe
10 ié d'abord en interne sans la direction. Donc les profs concernés o
11 e-là, donc l'année passée, la directrice a été informée d'opportuni
12 vendredi. Donc voilà. Donc la directrice est aussi en attente de re
13 re. » Mais je pense que si la directrice n'était pas venue avec le
14 dire que, elle ne nous… Si la directrice nous avait proposé cette a
15 tte année de bilan, il y a la directrice qui a été informée d'Erasm
16 Informatrice : Mais ça, notre directrice, elle l'a vécu. Je pense q
```

D'un coup d'œil, la mémorante identifie les passages attribuant à la directrice une qualité d'empathie envers les enseignants (ligne 16), un rôle de transmission des possibilités de séjour (ligne 11) et une fonction d'instigatrice du projet (lignes 3 à 5, notamment).

Comme la statistique textuelle, les concordances procèdent du réagencement des textes ; elles ne leur superposent pas une « couche » interprétative, comme le faisaient les dictionnaires de l'analyse de contenu. Ce faisant, même si son origine diffère, la concordance s'insère sans difficulté dans la tradition d'analyse de discours à laquelle se réfère la lexicométrie. Elle la complète d'autant mieux que, contrairement à la statistique textuelle, la concordance est une fonctionnalité muette, qui ne répond qu'aux sollicitations de l'utilisateur : pas de résultats sans question (tableau 9.1). Si le logiciel a la main lors des calculs lexicométriques, l'utilisateur la reprend lorsqu'il demande l'affichage de telle concordance particulière.

Utiliser un logiciel de lexicométrie, un concordancier ou un automate

Sans installer aucun logiciel sur son ordinateur, le lecteur peut tester quelques traitements lexicométriques ainsi qu'un concordancier sur la plateforme en ligne *Voyant Tools*. Il lui suffit de s'y rendre, sur Internet, et d'y copier-coller le texte qu'il souhaite analyser. Si ce premier contact le convainc, le lecteur peut ensuite se procurer un des logiciels libres suivants.

- *AntConc* constitue un concordancier très complet.
- Les plateformes en ligne *Sato* et *TXM* intègrent traitements lexicométriques et concordances. *TXM* peut également être installé sur un ordinateur personnel.
- *Iramuteq* propose également des outils de statistiques textuelles, un concordancier et les automatismes popularisés par *Alceste*.

Pour débuter avec ces logiciels, le lecteur convertit préalablement son matériau textuel en simples fichiers textes, dépourvus de mise en forme. Pour un usage avancé, les fichiers devront, par contre, suivre des règles de formatage spécifiques.

4. Des données qualitatives à l'analyse qualitative

4.1 Méthode : l'analyse par théorisation ancrée

Hors francophonie, les logiciels d'analyse de textes se réfèrent systématiquement à l'analyse par théorisation ancrée (*Grounded Theory Method*). Barney Glaser et Anselm Strauss (2010) formalisent cette méthode à la fin des années 1960. Contrairement à l'analyse de contenu, qui avait été créée pour une fin particulière (identifier et mesurer la propagande dans les journaux), la formalisation de l'analyse par théorisation ancrée procède moins d'une innovation méthodologique que de la mise en cohérence d'une série de pratiques en vigueur dans les recherches qualitatives de l'époque.

L'analyse par théorisation ancrée propose une interaction continue entre la collecte, l'analyse et l'écriture (Paillé, 1994). Elle s'écarte de la succession d'étapes de l'analyse de contenu et préconise de constamment aller et venir entre les différentes activités de la recherche. Certes, les analyses de contenu et de discours préconisent opportunément de constamment rapporter les analyses du chercheur au matériau, mais ces allers-retours ne concernent que la phase analytique. Pour sa part, l'analyse par théorisation ancrée étend les itérations à l'ensemble de la démarche de recherche.

Le présent chapitre se focalise sur les opérations qui ont inspiré des développements informatiques[1]. Celles-ci sont au nombre de trois : l'étiquetage du matériau, l'écriture et l'intégration d'une conceptualisation.

- L'analyse par théorisation ancrée est célèbre pour son rapport au codage du matériau empirique. Pour bien distinguer cette opération du codage systématique et contradictoire de l'analyse de contenu, je la qualifie d'étiquetage (Lejeune, 2014, p. 13). Étiqueter consiste à apposer des mots dans la marge des transcriptions d'entretiens ou en regard des notes d'observation. Cette opération se veut conceptualisante, originale et créative. Il ne saurait donc être question, comme en analyse de contenu, ni de la déléguer à d'autres ni d'éviter la subjectivité individuelle, bien au contraire.
- Qu'il s'agisse d'observations, de documents ou de témoignages, les matériaux collectés sur le terrain sont analysés au fur et à mesure de leur obtention. Chaque séance d'analyse fournit des résultats provisoires. Elle soulève également de nouvelles questions, qui sont explorées lors du contact suivant avec le terrain. Ces résultats et ces questions intermédiaires, le chercheur les consigne au jour le jour dans son journal de bord, à la fois pour conserver la trace du déroulement de la recherche, pour en organiser la mémoire et pour catalyser sa conceptualisation.
- Au fur et à mesure qu'il collecte son matériau, l'analyste organise les étiquettes produites et entend les intégrer à un système conceptuel. Ce travail d'articulation passe par le dressage de tableaux synoptiques et de représentations graphiques. Celles-ci prennent typiquement la forme de diagrammes reliant entre elles les étiquettes. Tracer ces schémas accompagne la conceptualisation ainsi que la tenue du journal de bord.

4.2 Logiciels

4.2.1 Les logiciels d'étiquetage réflexif

Dès les années 1980, des logiciels ont proposé un soutien informatique à l'analyse par théorisation ancrée (citée notamment dans l'introduction de leur manuel d'utilisation). Ils en outillent les trois opérations présentées dans la section précédente, reprises ci-dessous dans l'ordre de leur informatisation.

1. Pour une présentation complète, le lecteur est invité à se reporter au chapitre que Pierre Paillé consacre à l'analyse par théorisation ancrée.

- Dès 1984, alors que les PC n'étaient pas encore dotés d'interface graphique, le logiciel *The Ethnograph* permettait de sélectionner du texte, de lui attribuer une étiquette (dont le nom était choisi par l'utilisateur) et d'afficher, à l'écran, les étiquettes attribuées dans la marge du texte analysé. Aujourd'hui, ces activités d'étiquetage s'opèrent à la souris au sein d'interfaces graphiques rappelant un traitement de textes simplifié.
- À la même époque, *Nud*Ist*, l'ancêtre de *NVivo*, propose différents espaces de rédaction pour prendre les notes documentant, accompagnant ou conceptualisant l'étiquetage. D'un fonctionnement limité à leurs débuts, ces espaces de rédaction sont devenus essentiels aux logiciels d'étiquetage réflexif.
- Au début des années 1990, *Atlas.ti* intègre l'étiquetage à un outil d'articulation graphique (Gilbert, Jackson et di Gregorio, 2014). Depuis lors, les fonctionnalités d'édition de diagrammes permettent d'organiser graphiquement les étiquettes[1].

Les anglophones parlent de logiciels d'aide à l'analyse de données qualitatives (*computer assisted qualitative data analysis*, expression parfois contractée sous le sigle CAQDAS). Cette qualification peut paraître ambiguë, étant donné que les logiciels des autres familles permettent également l'analyse de données qualitatives. Cette ambiguïté ne se comprend que si l'on se souvient que cette famille est la seule recensée dans la littérature internationale : comme déjà mentionné, les dictionnaires appartiennent au passé ; la lexicométrie, les concordanciers et les automates constituent une spécificité francophone. La littérature anglophone ignore la diversité des logiciels présentés dans ce chapitre. Afin d'éviter, en français, l'ambiguïté d'une traduction littérale, je qualifie la présente famille de « logiciels d'étiquetage réflexif » (Lejeune, 2010).

Contrairement à l'ascétisme interprétatif des logiciels se revendiquant de l'analyse de discours, les fonctions d'étiquetage, de rédaction (journal) et de schématisation offrent, au sein du logiciel, différents supports à l'interprétation. Cette prise en charge informatique de l'interprétation ne s'accompagne cependant d'aucune verbosité. Comme les concordanciers, les logiciels d'étiquetage réflexif sont muets (tableau 9.1). Le chercheur garde la main de bout en bout, sans que l'ordinateur lui signale des phénomènes d'intérêt (ce que fait la lexicométrie) ni ne lui propose des

1. De nombreux autres logiciels proposent aujourd'hui ces fonctionnalités, notamment *Dedoose*, *HyperResearch*, *Kwalitan*, *LaSuli*, *MaxQDA*, *Saturate*, *Sonal*, *TamsAnalyser*, *Transana* et *WeftQDA*.

éléments de conceptualisation (comme le font les automates). Comme le logiciel n'opère en somme aucun traitement de lui-même, il n'est pas sensible à ce qui se passe à la surface du texte. Invité par le logiciel à lire et à analyser son matériau de lui-même, le chercheur analyse et interprète le vécu des acteurs rencontrés, plus que les mots employés. Contrairement aux autres familles, ces logiciels se révèlent dès lors adéquats pour « lire entre les lignes ».

Du point de vue de la conception informatique, les traitements nécessaires à une analyse par théorisation ancrée sont relativement simples à mettre en œuvre. Aujourd'hui, des dizaines de logiciels les proposent, ce qui génère une certaine concurrence. Les éditeurs de logiciels d'étiquetage réflexif y répondent en développant régulièrement des fonctionnalités additionnelles. Cette stratégie leur permet à la fois de se positionner par rapport à la concurrence et de justifier la sortie (et la vente) de nouvelles versions de leurs outils. En assistant le travail collaboratif (en équipe) ou en permettant d'étiqueter des images et de la vidéo en plus des textes, certains de ces ajouts ont bénéficié à la communauté scientifique.

Mais tous les ajouts n'entraînent pas des avancées. Comme les ordinateurs ont été développés, dès leurs origines, en tant que machines à calculer, les traitements les plus aisés à développer relèvent du comptage. Alors même qu'ils se réfèrent explicitement à une méthode qui exclut les comptages (Strauss et Corbin, 2004, p. 28), les logiciels d'étiquetage réflexif se sont progressivement dotés de feuilles de calcul rappelant les tableurs, de générateurs de camemberts ou d'histogrammes permettant de représenter des proportions, voire d'outils de mesure de la fidélité inter-codeurs. La stratégie des sociétés éditrices n'est pas absurde : les dictionnaires étant tombés dans l'oubli, pratiquement plus aucun outil ne se propose d'outiller l'analyse de contenu. Les logiciels d'étiquetage réflexif se profilent dès lors comme des outils polyvalents, permettant d'outiller l'analyse par théorisation ancrée et l'analyse de contenu.

La polyvalence des logiciels d'étiquetage réflexif comporte cependant un risque : les utilisateurs sont susceptibles de mobiliser, dans une même analyse, des traitements issus de deux méthodes que presque tout oppose. Conscients du risque, les éditeurs de *MaxQDA* ont programmé une fenêtre mettant en garde l'utilisateur qui appelle une procédure statistique. Malgré ce genre de mise en garde, un usage répandu des logiciels d'étiquetage réflexif consiste à quantifier les étiquetages thématiques opérés par le chercheur. Cet usage ne trouve de justification ni en analyse de contenu ni en analyse par théorisation ancrée. Pourtant, ceux qui procèdent ainsi se revendiquent de l'une et de l'autre, vraisemblablement

sans évaluer la tension épistémologique inhérente à cette position. Pire, il arrive que la section méthodologique se contente de citer le nom du logiciel utilisé, témoignant une confusion manifeste entre l'outil et la méthode (Paillé, 2006).

Utiliser un logiciel d'étiquetage réflexif

Le lecteur désireux de se familiariser avec un logiciel d'étiquetage réflexif peut télécharger le logiciel libre *WeftQDA*. Très aisé à prendre en main, ce logiciel permet d'étiqueter quelques textes et d'opérer des croisements. Il dispose d'un espace de rédaction, accessible via l'onglet « détails », mais est dépourvu de fonctions de schématisation graphique. La sobriété de ce logiciel convient particulièrement à un premier contact. Si l'utilisateur souhaite aller plus loin, il pourra acquérir un logiciel d'étiquetage réflexif plus élaboré[1].

4.2.2 Les registres

Parmi les fonctionnalités ajoutées aux opérations centrales de l'analyse par théorisation ancrée, il en est une qui tire opportunément parti, de manière créative, de la concurrence entre logiciels d'étiquetage réflexif. Son originalité m'a incité à lui consacrer une sixième famille : celle des registres.

À l'origine des registres, il y a l'idée qu'un étiquetage réflexif peut bénéficier de la recherche d'expressions ou de mots-clés. Les logiciels *Atlas. ti*, *HyperResearch* et *Cassandre* permettent à l'utilisateur de rassembler tous les passages correspondants à une (voire à plusieurs) expression(s) donnée(s) et de leur attribuer, à tous, une étiquette temporaire. Les registres désignent cet étiquetage à la fois semi-automatique et réflexif (Lejeune, 2008 ; Bénel, Lejeune et Zhou, 2010). Leur dimension semi-automatique rappelle le fonctionnement (simple et efficace) des dictionnaires. Les registres ne sont toutefois pas définis en amont, mais au fil de l'analyse. Leur nature temporaire et transitoire les éloigne des dictionnaires, dont l'objectivité et la stabilité conféraient sa scientificité à l'analyse de contenu. L'utilisateur crée les registres les uns après les autres, au gré d'où ses analyses le conduisent. Il les rebaptise, en modifie les expressions ou les mots-clés, ou les abandonne rapidement. Les registres s'insèrent donc dans une logique de la découverte et ne relèvent pas, comme les dictionnaires, de l'administration de la preuve.

1. Les liens vers les différents logiciels d'étiquetage réflexif figurent sur le site Internet de l'auteur.

5. Discussion : assumer ses choix

Les analyses de contenu, de discours et par théorisation ancrée n'épuisent pas les façons d'analyser des matériaux qualitatifs. D'autres méthodes existent. Il n'est pas question ici de réduire la diversité des traditions méthodologiques, mais de rappeler que, découlant de questions de recherche particulières, les logiciels d'analyse de textes souscrivent à des postulats théoriques distincts.

Les différents utilisateurs d'un même logiciel ne poursuivent pas pour autant des recherches similaires. Cette diversité préside à l'inventivité, à la fois méthodologique et théorique, qu'il n'est pas non plus question de restreindre ici.

La présente histoire intellectuelle de l'émergence des différents types de logiciels entend par contre sensibiliser les praticiens et les novices aux implications de leurs choix techniques. Bien sûr, un logiciel n'est jamais « qu' » un outil mais, en éclairant certaines caractéristiques du matériau, chaque outil en occulte d'autres.

Aucun logiciel ne se démarque si l'on ne tient pas compte des investigations à conduire. Chacun outille une démarche singulière. Pour choisir l'outil adapté, il importe donc d'être au clair sur sa question de recherche et sur le type de méthode que l'on souhaite emprunter. Le dictionnaire permet d'appréhender des phénomènes comptables, répétitifs ou sériels, comme la propagande ou les rumeurs. La lexicométrie permet de cerner des questions de style et de caractériser des genres de discours. Les logiciels d'étiquetage réflexifs ont été conçus pour assister l'analyse de l'expérience intime et subjective des acteurs[1].

Les trois traditions et les six types de logiciels présentés dans ce chapitre assistent l'analyse de matériaux qualitatifs. Mais les analyses et les traitements engagés ne sont pas tous qualitatifs. Évidemment, les logiciels d'étiquetage réflexif s'imposent comme les outils conçus pour assister le cœur de ce que recouvre typiquement une démarche qualitative, alternant typiquement l'écriture, la collecte et l'étiquetage du matériau. L'usage des outils de cette famille n'est cependant pas obligatoire. D'autres fonctionnalités peuvent également convenir. S'agissant de technologies muettes, laissant la main à l'analyste, les concordanciers respectent également l'essence de la démarche qualitative. Les outils

1. Lejeune (2016) recense les questions à se poser avant de choisir.

bavards ou verbeux conviennent par contre plus difficilement, surtout s'ils reposent sur des fonctions de comptage. Il est dès lors difficilement tenable, épistémologiquement et méthodologiquement, de se revendiquer de la recherche qualitative et de recourir à la lexicométrie, aux logiciels automatiques ou aux dictionnaires[1]. Ces trois familles ne sont ni moins légitimes ni moins scientifiques pour autant : leurs fondements méthodologiques et épistémologiques sont simplement distincts et en circonscrivent à la fois la pertinence et la validité.

6. Conclusion : frugalité informatique et audace intellectuelle

J'assiste souvent à des démonstrations de logiciels et à des formations à leur usage. Immanquablement, l'auditoire interroge l'auteur ou le formateur sur la présence de telle ou telle fonctionnalité. Oubliant les considérations de méthode, ces questions trahissent le fantasme de l'outil complet, remplaçant définitivement tous ses prédécesseurs. Or les logiciels de statistique textuelle sont dépourvus d'outils d'étiquetage et les logiciels d'étiquetage réflexif ne permettent pas de compter les mots, non parce que leurs concepteurs sont paresseux ou en retard. Regretter l'absence de ces fonctionnalités, les envisager comme un manque ou espérer leur intégration dans la prochaine mise à jour est absurde. De la même manière, il serait injuste d'affirmer que les logiciels de lexicométrie et de concordances ne permettent « que » d'explorer et d'organiser les textes (tableau 9.1). Chaque outil s'inscrit dans un paradigme de recherche et permet de répondre à une question particulière.

Désireux de se former en une fois à un outil utile pour toutes leurs recherches, les utilisateurs recherchent souvent le logiciel le plus complet ou le plus polyvalent. À l'inverse, je préconise d'opter pour l'outil le plus simple et le plus adapté au projet en cours. Cette option permet de se former et d'être opérationnel rapidement. Elle évite également d'être tenté par des fonctionnalités peu compatibles avec ledit projet.

Prôner la spécificité plutôt que la polyvalence, en matière de logiciels d'analyse de textes, présente la vertu de recentrer l'échange scientifique sur des questions de méthode. Focaliser son attention sur la démarche

1. Pour autant que l'on soit conscient du paradoxe, il est possible de le dépasser (Lejeune et Bénel, 2012).

de recherche fait prendre conscience que les écueils attribués aux outils proviennent souvent d'un usage inadéquat ou d'attentes inopportunes. Au moment où se pose la question du choix du logiciel, il est donc également pertinent de se rappeler que l'usage d'un logiciel ne garantit pas la qualité des résultats. Le chercheur qui maîtrise le fonctionnement d'un outil n'est pas nécessairement avisé dans son interprétation. Travailler sans logiciel devrait donc toujours rester une option. Choisir une technique (informatisée ou non) s'inscrit dans une démarche de recherche, réfléchie, délibérément construite et sans cesse évaluée.

Lectures conseillées

Sur l'analyse de contenu :

Le lecteur désireux de réaliser ses propres analyses de contenu peut consulter le guide très didactique de Roger Mucchielli (2006).

Sur l'analyse de discours :

Organisées tous les deux ans, les journées d'analyse statistique des données textuelles (JADT) rassemblent les concepteurs et les utilisateurs de logiciels issus de l'analyse de discours. Leurs actes sont disponibles en ligne : *http://www.jadt.org/*.

Sur l'analyse par théorisation ancrée :

La présentation originale de l'analyse par théorisation ancrée est disponible en français (Glaser et Strauss, 2010). J'ai également rédigé un manuel facilitant sa mise en œuvre pratique (Lejeune, 2014).