CAROLINE MOTHE
ESTELLE DELFOSSE
ANNE MARIE BOCQUET
Université Savoie Mont Blanc, IREGE

L'Analyse de Données Textuelles assistée par ordinateur pour les chercheurs et praticiens : Les discours sur les réseaux de chaleur[1]

L'analyse de données textuelles (ADT), courant qui s'est développé grâce aux techniques de la linguistique informatique, a évolué de l'analyse lexicale à l'analyse sémantique. Cette approche particulière est l'occasion d'une expérience amplifiée par les possibilités graphiques d'interaction et de partage du web. En mobilisant cette approche pour l'analyse d'un corpus de près de 200 documents sur les réseaux de chaleur en France, nous montrons ainsi l'utilité de l'ADT assistée par ordinateur pour les chercheurs, les praticiens et tous les acteurs qui se trouvent face au défi de devoir traiter des données massives.

Computer-Assisted Textual Data Analysis for researchers and practitioners: Speeches on district heating

The use of Textual Data Analysis (TDA) has spread thanks to information science techniques applied to linguistics, and has evolved from lexical analysis to semantic analysis. This specific approach allows an amplified experience thanks to the possibilities provided by the web of graphical interactions and sharing. Using this approach to analyze a corpus of about 200 documents on district heating networks in France, we show the interest of automated Textual Data Analysis methods for researchers, managers, and all those who have to analyze a wide corpus of documents.

[1] Cette étude a été réalisée dans le cadre du projet PACs-CAD qui est soutenu par le programme européen de coopération transfrontalière Interreg France-Suisse 2014-2020 et a bénéficié à ce titre d'un soutien financier du Fonds européen de développement régional (FEDER) et de fonds fédéraux Interreg suisses.

La « recherche qualitative » implique des méthodes d'analyse basées sur le recueil d'informations documentaires, la réalisation d'entretiens ou l'observation participante. Ces méthodes consistent à prendre connaissance des matériaux ainsi rassemblés pour proposer une interprétation argumentée par des citations de terrain et des références aux théories. Plus ou moins formalisé, ce travail d'analyse s'inscrit dans la tradition herméneutique et exclut le plus souvent le recours aux méthodes quantitatives (Dumez, 2013). Il existe trois grandes manières de faire de la recherche qualitative et d'aborder les textes : dans la tradition littéraire, les textes sont lus et commentés par le lecteur ; l'analyse de contenu classique « à la Krippendorff » permet de systématiser et d'industrialiser la lecture ; l'ADT est une analyse assistée par ordinateur, une automatisation qui s'appuie sur la formalisation.

L'ADT se définit donc par les procédés contemporains qui s'appuient sur la statistique, l'ingénierie linguistique et sémantique dans des applications informatiques[2]. Elle se développe dans la deuxième moitié du XXème siècle au confluent des études littéraires, de la statistique, de la linguistique et de l'informatique (Benzecri, 1973a, 1973b ; Lebart et al., 2020). Sa mise en œuvre repose sur des outils logiciels qui ont beaucoup évolué avec les technologies informatiques. Cependant leur usage par les chercheurs en sciences humaines reste assez limité, car les oppositions entre approche qualitative et quantitative restent fortes.

L'objectif de cet article est de montrer comment les méthodes de l'ADT peuvent contribuer à l'analyse des corpus de documents qui présentent un double défi pour le chercheur : celui de la charge de travail (lié à la lecture de corpus très volumineux) et celui de l'objectivité (afin de respecter les ambitions d'un travail scientifique). Nous cherchons à présenter la démarche ADT et sa pertinence pour les chercheurs (et les praticiens) dans le cas d'une recherche (ou analyse) documentaire en illustrant, exemple des réseaux de chaleur à l'appui, ce qu'il est possible de faire avec l'ADT. Notre objectif est double : il consiste d'une part à montrer la contribution potentielle de l'ADT dans un monde où la masse de données textuelles explose. La recherche documentaire n'est ici qu'un exemple de données textuelles ; nous aurions également pu prendre d'autres types de données textuelles comme celles issues du web, d'entretiens, d'articles académiques (comme pour les revues de littérature systématiques) ou de réponses libres dans des questionnaires. D'autre part, notre objectif est de montrer comment faire une analyse de textes de manière automatique, soit de manière inductive par exploration avec l'analyse lexicale, soit de manière déductive avec l'analyse ciblée de contenu

[2] Dans ce qui suit, nous utilisons le terme « ADT » exclusivement pour caractériser cette analyse de données textuelles assistée par ordinateur, formalisée, statistique et systématique ; celle-ci se distingue de la lecture d'un texte (et non de « données textuelles »), dans l'analyse de contenu par exemple).

sémantique. Ainsi, l'ADT permet de remplacer la lecture de textes par l'observation neutre du texte, pris comme objet d'expérience. L'exploration faite par la statistique lexicale permet de caractériser le texte de manière statistique sans en donner le sens et, pour aller plus loin, de lire ces substituts et d'interpréter. Dans ce contexte, l'ADT est très proche de l'analyse qualitative « pure » (ou ancrée, cf. Dumez, 2013) et d'une démarche inductive. Par ailleurs, l'ADT permet une démarche plus déductive avec un usage lié à la statistique sémantique pour caractériser les contenus. Dans ce cadre, le chercheur déduit des codes de la théorie et les fait rechercher dans le texte par l'ordinateur, utilise un thésaurus standard ou en construit un, et applique une lecture automatique pour chercher à le caractériser, à déterminer la fréquence d'occurrence des concepts et à formaliser les idées.

Ainsi, pour le courant de l'ADT, le texte est vu comme une donnée, et on l'analyse de manière statistique et « froide », sans intervention et interprétation. Cette vision s'oppose totalement à celle du texte vu comme un corpus qui a du sens et sur lequel on mène des analyses de contenu, approche la plus mobilisée à l'heure actuelle en France en sciences de gestion[3].

Dans une première partie nous présenterons les enjeux de l'analyse de corpus de documents et les principes et méthodes de l'ADT. Une seconde partie permettra de montrer l'apport et la pertinence de l'ADT, à partir d'un corpus de près de 200 documents qui portent sur l'étude des réseaux de chaleur. L'analyse, menée avec les logiciels Sphinx IQ2 et DataViv, sera illustrée par des liens qui mettent en évidence la « vision ADT » de ce corpus. Le lecteur pourra ainsi se faire sa propre opinion sur cette approche, dont la portée et les limites seront discutées en conclusion.

I. Documents et analyse de données textuelles

1. Les méthodes traditionnelles d'analyse de corpus de documents

Le recours aux documents se distingue de l'utilisation des entretiens non directifs, individuels ou de groupe, un usage plus largement répandu en sciences sociales. Cette approche rejoint la tradition de recherche en histoire (Bogaert et al., 2018) : « Trouver les documents, les archiver mais aussi les qualifier et les critiquer, tel est le travail de l'historien. A la différence des entretiens, les archives ne sont pas suspectes d'une quelconque influence du chercheur mais

[3] En anglais, on parlerait de « *computer-aided textual data analysis* » (Mossholder et al., 1995), de « *computer-assisted discourse analysis* » (Buhler et Lethier, 2020), de « *automated text analysis* » (Humphreys et Jen-Hui Wang, 2018) ou, plus généralement, mais surtout en informatique, de « *text mining* » (Kobayashi et al., 2018) pour mettre l'accent sur le fait que l'analyse de données textuelle est automatisée.

doivent être situées dans le contexte de leur production et par les intentions de leur auteur ou des institutions qui les conservent. Un rapport de conseil d'administration ne se lit pas comme un article de presse ou un tract syndical. L'historien doit connaître les contextes et les finalités pour être capable d'apprécier les contenus et en faire une critique autant externe qu'interne. En histoire contemporaine, cela peut donner un avantage au chercheur immergé dans le terrain de sa recherche.

Dans cette démarche de recherche, le chercheur se méfie de ses a priori et, s'il est guidé par un modèle, il cherche davantage à le préciser ou à le faire évoluer qu'à le vérifier d'une manière systématique. Le recours aux documents a pour objectif de refléter la variété des situations possibles plutôt que de représenter fidèlement une population comme on cherche à le faire en administrant un questionnaire.

Cette approche implique donc l'utilisation de documents nombreux et/ou volumineux, souvent plus d'un millier de pages pour une première approche documentaire fondée sur une analyse systématique. Cette démarche de recherche applique des procédures statistiques à des données qualitatives, tout en s'appuyant sur des procédés de visualisation interactive (Moscarola, 2018b). Cet usage innovant de l'analyse de données textuelle confère son originalité à cette recherche dont les apports sont également méthodologiques : conduire au développement de recherches documentaires considérées comme un aboutissement, plutôt que comme un préalable inachevé au recueil d'interviews.

Recherche qualitative « pure »

L'analyse de ces corpus documentaires renvoie à la tradition des recherches littéraires ou historiques qui consiste à lire, comprendre et interpréter le matériau rassemblé dans le but d'en rendre compte de manière descriptive (ce dont il est question), explicative (influence des personnes, circonstances ou périodes) ou compréhensive (articulations internes à ce qui est dit). L'analyste, comme l'historien ou le journaliste, construit ainsi l'histoire qu'il écrit, à partir des informations contenues dans les documents et des connaissances et théories mobilisées pour organiser et présenter ces informations. Le travail de synthèse mené par le chercheur s'appuie sur des citations extraites du corpus, mobilisées pour appuyer les interprétations proposées ; cette démarche correspond à une tradition de recherche qualitative « pure »[4] (au sens de Dumez, 2013, qui fait référence à un codage émanant du matériau

[4] Une recherche qualitative dite « pure » (Dumez, 2013 ; Moscarola, 2018a) est donc celle qui est menée dans la tradition des humanités par un travail intellectuel de lecture, de réflexion et d'écriture purement littéraire et non instrumenté ; la critique littéraire, la dissertation, l'essai ou le reportage journalistique en sont des exemples.

indépendamment des cadres théoriques - « à la Glaser et Strauss », et de Moscarola, 2018a). Ce travail s'envisage ainsi dans la perspective de la théorie ancrée (Glaser et Strauss, 1967) car il applique une démarche purement inductive pour progresser sans idée ou modèle a priori, de la découverte du corpus à la construction d'une nouvelle théorie. Dans leur ouvrage fondateur, Miles et Huberman (1994) élargissent le point de vue et décrivent la recherche qualitative comme une synthèse progressive qui organise la confrontation entre le terrain et les idées du chercheur (théories, modèles).

Analyse thématique du contenu

Le courant de l'analyse thématique du contenu (Drisko et Maschi, 2016) propose une approche plus formalisée et systématique de cette prise de connaissance. L'analyse permet d'identifier des passages du corpus, que le chercheur distingue en fonction de leur signification, qui est précisée par des codes. Selon la granularité du découpage en unités de significations et le niveau de détail des codes, ce travail peut s'effectuer au niveau global du document (considéré comme un tout) ou à un niveau plus fin (les phrases). De nombreux travaux décrivent cette approche selon le niveau de détail du code et son intention (Saldaña, 2013) ou selon la manière de faire évoluer la codification (Gioia et al., 2012). Le travail d'interprétation du chercheur se trouve ainsi encadré et systématisé (Krippendorff, 2013) et sa présentation est objectivée (Bardin, 1977). Il reste alors deux difficultés : la subjectivité de la lecture ou de la codification, et la lourdeur du processus quand le corpus est très volumineux.

Analyse du contenu assistée par ordinateur

Longtemps effectué à la main, ce travail de codage, de repérage et d'annotation peut être assisté par ordinateur grâce à des logiciels, appelés CAQDAS (Computer Assisted Qualitative Data Analysis Software), qui sont apparus dans les années 80. Les plus répandus correspondent à l'approche de Miles et Huberman (2003) ; c'est le cas de NVivo et d'ATLAS.ti, dont le fonctionnement consiste à repérer des extraits significatifs et à les affecter à des systèmes de codification évolutifs, ce qui aide le chercheur à organiser ses idées et à rendre compte des thèmes qu'il découvre. D'autres logiciels comme QDminer, MaxQDA, Hyper Base, Sphinx, Alceste ou Iramuteq sont des outils statistiques d'exploration ou de *text mining* qui permettent essentiellement de faire de l'exploration lexicale. Ces logiciels utilisent la statistique pour rendre compte des contenus, de leur organisation interne et de l'influence des contextes. Ils ont comme fonction première de lister, compter et catégoriser les mots

présents dans un texte de grande taille. Certains (Alceste, Iramuteq) se concentrent sur l'analyse lexicale et la classification hiérarchique descendante (séquence d'analyse factorielle des correspondances multiples), qui permet de construire des typologies de manière automatique. D'autres (SphinxIQ) sont plus généralistes et intègrent des thésaurus, permettent de faire de l'analyse sémantique des contenus et de la visualisation de données. Enfin, à partir de la tradition des thésaurus informatisés (Wordnet, Troppes), l'ingénierie linguistique a ouvert la voie à l'analyse sémantique et à la reconnaissance automatique des contenus, des opinions ou des sentiments (Goddard, 2011).

Malgré les progrès rapides des logiciels d'ADT de plus en plus puissants, leur usage reste assez limité, et le recours à Nvivo ou Atlas TI, les logiciels les plus fréquemment mobilisés dans l'analyse des corpus de documents, ne s'accompagne pas toujours d'une contribution visible sur les résultats obtenus dans l'analyse des corpus de documents[5]. On peut observer que les méthodes de recherche qualitative ne sont que très marginalement affectées par l'évolution des techniques logicielles et par les contributions de l'ADT. Toutefois, nous le verrons dans la section suivante, l'ADT tend à être de plus en plus utilisée par les chercheurs, y compris en sciences de gestion.

2. L'analyse de données textuelles (ADT)

2.1. Historique et utilisation en sciences de gestion

A l'origine développée pour l'analyse de grands corpus littéraires ou historiques, l'analyse de données textuelles vise à aborder un texte comme un ensemble de données constituées de mots et structurées selon les règles de la syntaxe et de l'organisation du discours (parties, paragraphes et phrases). La répétition des mots utilisés fait sens (Lebart et al., 2020), et donne une première indication sur les contenus, précisée ensuite par la manière dont ils sont associés les uns aux autres pour constituer des configurations thématiques. Les fréquences d'utilisation des termes du corpus et les associations permettent de préciser le contenu des documents. Enfin, les éléments de contexte, l'origine des documents, la période et les différences dans l'usage des mots permettent de qualifier leur influence. Ainsi, en traitant statistiquement le texte comme un ensemble de mots, on peut mettre en évidence des mots clés, des configurations thématiques ou des spécificités contextuelles.

[5] En effet, ces outils servent plus à gérer la multiplicité des sources impliquées et à aider le chercheur dans la structuration de sa vision, tout au long de la recherche, qu'à analyser des corpus particuliers de façon systématique. Ce sont des logiciels d'aide à l'analyse de contenu mais sans automatismes, le codage et l'analyse restant très largement manuels et humains - et non systématiques et statistiques.

L’ADT en tant que telle s’inscrit dans une épistémologie positive : elle aborde le texte comme un objet de langue composé de formes graphiques, dont les décomptes et analyses statistiques révèlent les propriétés et structures lexicales. Ceci étant, l’usage de l’ADT dans une démarche de recherche qualitative s’apparente aussi à l’épistémologie constructiviste, car elle donne du sens aux structures indicatrices des contenus. En complément, ou à l’appui de l’approche interprétative subjectivement contrôlée décrite par Dumez (2013), l’ADT appuie la construction du sens sur le texte, qui est abordé comme un ensemble de données objectives.

L’ADT, à l’origine plutôt utilisée par les linguistes et les historiens, est une méthode de plus en plus répandue en sciences de gestion, notamment dans les pays anglophones où l’intérêt de l’ADT est démontré depuis près de 30 ans (cf. article de l’*Academy of Management Journal* de Gephart, 1993, qui mobilise l'ADT pour générer des informations sur la manière dont l'organisation donne un sens au risque et au blâme). Dans *Organizational Research Methods*, Kobayashi et al. (2018) observent que, malgré l'omniprésence des données textuelles, peu de chercheurs ont jusqu’ici appliqué l'exploration de texte en tant que technique pour répondre à des questions de recherche liées aux organisations. L’analyse de données textuelles (données généralement volumineuses, ce qui implique une approche quantitative), aide à accélérer la découverte des connaissances en augmentant les quantités de données qui peuvent être analysées. L’article vise à familiariser les chercheurs avec la logique fondamentale qui sous-tend l'exploration de texte, les étapes analytiques impliquées et les différentes techniques contemporaines qui peuvent être utilisées en fonction des objectifs visés. Mossholder et al. (1995) présentent l’utilisation de l’ADT comme une technique de recherche qualitative qui permet d’évaluer le contenu émotionnel de réponses ouvertes à une enquête (cf. le courant de l’analyse des sentiments), et montrent comment l’ADT peut être utilisée en complément d’une méthodologie quantitative. C’est aussi ce que font Raich et al. (2014), dont l’objectif est de fournir des preuves tangibles des phénomènes dans l'organisation. Les ensembles de données textuelles se composent de deux blocs différents, qualitatifs et quantitatifs. Les chercheurs combinent souvent plusieurs méthodes pour exploiter ces deux aspects, et ils le font fréquemment de manière comparative, convergente ou séquentielle. Raich et al. (2014) préconisent une méthode d'analyse hybride, qui utilise analyses qualitatives et quantitatives de manière simultanée sur le même ensemble de données textuelles (méthode illustrée par le cas d’une banque autrichienne).

En marketing, un article récent du *Journal of Consumer Research* (Humphreys et Jen-Hui Wang, 2018) réalise une excellente synthèse sur l’utilisation de l’analyse de texte automatique

pour les recherches sur le consommateur, notamment pour examiner des tendances et structures que des chercheurs ne pourraient pas identifier dans les textes sans aide logicielle. Plus récemment encore, Buhler et Lethier (2020) utilisent la textométrie (ou ADT), i.e. l'analyse systématique assistée par ordinateur de données textuelles, pour analyser un corpus de documents de planification et d'urbanisme (contexte proche de notre thématique des réseaux de chaleur). En intégrant des dizaines de documents à la fois, cette méthode étend l'échelle d'analyse et permet d'identifier des transitions massives du discours dans le temps, et des contrastes majeurs entre les discours qui émanent de groupes d'acteurs spécifiques.

Des chercheurs ont également analysé la manière dont l'ADT est utile pour mesurer des construits : Tsao et al. (2020) s'appuient sur cette méthode d'analyse systématique pour générer automatiquement des dictionnaires, qui peuvent permettre de mesurer des opinions ou des contenus, si on les utilise de manière analogue aux échelles de Likert. Nous développerons dans la suite de l'article les différents usages possibles de l'ADT, en les appliquant à l'exemple des réseaux de chaleur.

En France, l'ADT commence à se répandre : 34 articles en sciences de gestion ont été identifiés dans la base Cairn avec les mots clés « Analyse de Données Textuelles » (toutefois, certains ne se réfèrent pas à l'analyse automatisée et assistée par ordinateur, ce qui révèle un certain flou sémantique). Aujourd'hui, la plupart des logiciels d'analyses lexicales disponibles en France sont inspirés de la démarche du mathématicien Jean-Paul Benzécri (1973a, b) : Alceste, Iramuteq (logiciel libre) et SphinxIQ (qui a l'avantage de fournir un spectre plus large avec la possibilité de faire également de l'analyse sémantique). On parle souvent d'analyse à la française»[6] (Jenny, 1999 ; Gauzente et Peyrat-Guillard, 2007).

2.2. Les différentes techniques et utilisations de l'ADT

Comme nous l'avons vu, les utilisations de l'ADT sont diverses :

* L'exploration lexicale permet de caractériser un grand volume d'information par la production de substituts lexicaux ou sémantiques (nuages de mots, cartographie de concepts) de faible volume, obtenus de manière automatique et reproductible ;

* La classification hiérarchique, qui se situe entre l'exploratoire et l'analyse ciblée, peut être vue comme une première étape vers la création de sens, car elle permet d'interpréter les classes obtenues par l'exploration sémantique avec l'utilisation d'un thésaurus standard. Un

[6] La principale différence réside dans le fait que les logiciels anglo-saxons s'appuient sur l'analyse en composantes principales, les logiciels francophones comme Alceste sur la classification hiérarchique.

thésaurus est une modélisation de la connaissance et de l'analyse sémantique, pour les informaticiens, qui permet de faire une analyse systématique et automatique de contenu. Cette analyse peut aussi se faire avec un thésaurus spécifique créé par le chercheur.

Notons que ces analyses peuvent toutes être réalisées de manière contingente, puisque la statistique permet d'analyser le texte en fonction de certaines caractéristiques contextuelles prédéfinies. Par ailleurs, les outils de visualisation proposés par certains logiciels - comme SphinxIQ - permettent de partager avec les destinataires de la recherche la fouille des corpus par entrées lexicales ou sémantiques, en leur apportant les mêmes ressources que celles utilisées par le chercheur.

ADT et approche exploratoire

Au cours des trente dernières années, les logiciels ont beaucoup évolué grâce aux progrès des techniques de classification, de l'ingénierie linguistique et de l'analyse sémantique (Goddard, 2011). Les algorithmes de classification (Reinert, 1983) ont contribué à la popularité des logiciels de l'ADT[7]. Leur puissance s'est enrichie des apports de la lemmatisation, des thésaurus et de l'analyse sémantique (Baulac et al., 2014). Ces algorithmes permettent d'identifier non seulement les mots clés, les expressions et les catégories thématiques révélées par les associations, mais également les concepts auxquels ces mots et expressions renvoient. Sur ces bases, la description des contenus peut être complétée par l'analyse de l'articulation interne des contenus et des influences du contexte.

L'ADT permet d'obtenir une analyse réalisée de manière automatique et reproductible, sans aucune lecture ou codification préalable. Les résultats obtenus sont le reflet statistique des structures lexicales et sémantiques, présentées sous forme de nuages de mots, de cartes de configurations lexicales ou sémantiques, ou de graphes d'influence, qui permettent de résumer en quelques pages l'intégralité du corpus. Le travail de l'analyste consiste alors à interpréter et à donner du sens à ces résultats. Il faut être conscient que ces « substituts lexicaux ou sémantiques » (Moscarola, 2018a) constituent une approximation qui peut se révéler trompeuse, et amener le chercheur à projeter sur ce qu'il voit des interprétations erronées. Le retour au texte est donc indispensable pour vérifier, par la lecture des documents, ce que les substituts semblent indiquer.

[7] Initialement implantés dans Alceste puis Iramuteq et Sphinx.

Visualisation de données et infographie dynamique

Les logiciels de dernière génération permettent une expérience de lecture hypertextuelle, par la visualisation des substituts lexicaux produits par l'ADT. La meilleure manière d'en saisir la portée pour l'analyse de documents est de l'expérimenter grâce au lien suivant. Dans la suite de l'article, d'autres liens pointent sur des pages qui contiennent des nuages de mots clés ou des catégorisations thématiques. En complément de ces substituts lexicaux et sémantiques figure le verbatim des documents, découpés en phrases. En cliquant sur un des éléments des substituts, on lit le verbatim correspondant, ce qui permet de vérifier le sens attribué aux mots clés, catégories thématiques ou concepts sélectionnés. De la même manière, en sélectionnant la source du document, on peut constater l'incidence sur les contenus. Cet accès au corpus en donne une vision différente, à la fois très synthétique et détaillée, et constitue une autre façon de découvrir le contenu des documents.

Classification automatique

Il s'agit de repérer statistiquement les similitudes entre les documents ou les phrases en se basant uniquement sur les mots et/ou concepts qui les composent. Deux documents ou phrases se ressemblent s'ils utilisent les mêmes mots ou renvoient aux mêmes concepts. On définit ainsi des classes en associant les mots ou concepts qu'elles ont en commun, classes qui se distinguent les unes des autres par l'absence d'éléments communs. Chaque classe est caractérisée par les mots et concepts qui lui sont spécifiques et dont la lecture conduit à une interprétation thématique. Ainsi la construction purement statistique et objective d'une typologie débouche sur une interprétation subjective, fondée sur la signification qui émerge des éléments spécifiques et des verbatim correspondants.

Par ce procédé, on peut obtenir un résultat similaire à celui qu'on pourrait obtenir grâce à une analyse de contenu « classique », codes à l'appui, qui se situe au plus près du texte par une codification mot à mot, codification « pure » selon Dumez (2013), qui débouche, par agrégation et interprétation de deuxième ordre, à la formulation de thèmes en référence aux théories et connaissances sur le domaine (e.g. Gioia et al., 2012 ; Saldaña, 2013). Grâce à l'ADT, cette première 'lecture', automatisée, est instantanée, parfaitement objective et reproductible (elle correspond aux « catégories de premier ordre » dans le vocabulaire de Gioia). Elle est complétée par une deuxième lecture par le chercheur, qui doit alors lire les résultats de la typologie en se fondant sur les mots et concepts spécifiques qui caractérisent ses éléments. Comme le radiologue qui doit connaître les propriétés physiques du système utilisé, le chercheur doit comprendre les principes statistiques de l'algorithme de

classification. Il doit aussi avoir une bonne culture du domaine, comparable aux connaissances anatomiques et médicales du radiologue.

Ajoutons à ces compétences celles qui permettent de définir :
- l'objet à observer : les documents dans leur globalité ou dans le détail de leurs phrases ?
- le spectre dans lequel mener l'observation : les mots, les expressions, les concepts ?

Les logiciels disponibles permettent désormais de faire varier ces paramètres avec une grande facilité, et de trouver la meilleure combinaison pour révéler les structures du corpus.

ADT ciblée

L'exploration peut être poursuivie en ciblant plus précisément la recherche pour caractériser de façon statistique la forme des documents, leur longueur, style, contenu, et ce qui les distingue selon le contexte de leur production (date, auteur, statut…). La structuration des documents (parties, paragraphes, phrases) permet de les différencier (par exemple, les articles de presse sont plus concis que les rapports d'études). D'autres indicateurs lexicaux permettent de qualifier le style des documents. Les travaux sur l'énonciation (e.g. Gavard-Perret et Moscarola, 1996) fournissent des références utiles pour caractériser la forme des documents. Par exemple, les auteurs scientifiques évitent systématiquement le 'je', privilégient les articles définis et peuvent construire de très longues phrases. Au contraire, la communication d'un dirigeant ou d'un homme politique recourt abondamment aux pronoms personnels (je, vous, nous), utilisés dans des phrases plutôt courtes.

Sur le fond, la nature des mots et expressions utilisées peut varier beaucoup d'un document à l'autre. L'usage de tests statistiques permet ainsi de mettre en évidence les « zones de langage », caractéristiques de différents types de documents selon les auteurs ou les institutions. Plus globalement, en croisant le contexte de production des documents avec les catégories thématiques révélées par les algorithmes de classification ou avec les concepts dégagés par l'analyse sémantique, on peut identifier l'influence des statuts ou des contextes sur le contenu des documents.

Thésaurus

Depuis une dizaine d'années, l'ADT s'est enrichie des apports de l'analyse sémantique, qui consiste à utiliser un thésaurus pour identifier les idées présentes dans le texte. Les documents peuvent être ainsi 'lus' pour chercher, en fonction des mots qu'ils contiennent, des catégories

correspondantes du thésaurus[8]. Les logiciels peuvent ainsi associer à chaque document ou à chaque phrase les idées et concepts qui les caractérisent. Les résultats obtenus dépendent du texte et du thésaurus utilisé, ce qui peut, dans certains cas, justifier la construction d'un thésaurus ad hoc mieux adapté au domaine abordé plutôt que d'utiliser le thésaurus générique présent dans les logiciels d'ADT comme Sphinx IQ2.

Choisir une approche

Exploration ou analyse de contenu ciblée, le choix d'une approche dépend de l'orientation définie par une question de recherche plus ou moins précise. Dans tous les cas, la description des propriétés statistiques et lexicales du corpus, contrôlée par la visualisation de données et le retour au texte, permettent un exposé objectif et critique du matériau analysé. Cependant, le recours à l'ADT n'est véritablement justifié que pour des corpus d'une taille suffisante, dans lesquels les effets de répétition sont statistiquement significatifs. Dans la tradition des études quantitatives, au moins trente documents sont nécessaires, et selon la théorie des actes de langage, c'est le nombre de mots du corpus qu'il faut considérer[9].

Le chercheur doit-il s'engager dans un travail « classique » de lecture et de codification ou recourir à l'ADT ? Cela dépend de la taille du corpus et des catégories conceptuelles auxquelles le modèle de recherche renvoie. La codification peut prendre beaucoup de temps, mais elle est inévitable dans certains cas, selon que les idées ou concepts qu'on cherche à repérer sont plus ou moins abstraits. Si les concepts renvoient à des faits circonscrits et repérables par des listes de mots, on peut construire un thésaurus. Au contraire, si les concepts impliquent une interprétation ou une appréciation, la lecture se révèle incontournable, quitte à mener une ADT sur un échantillon du texte.

II. Illustration de l'apport de l'ADT: les réseaux de chaleur en France

Nous appliquons ces méthodes à un corpus de près de 200 documents sur les réseaux de chaleur. L'opportunisme méthodique de Girin (1989) nous permet de justifier le choix du secteur analysé (les réseaux de chaleur). Ce domaine, concerné par des initiatives privées et des démarches administratives ou publiques, fait l'objet d'une très abondante production de textes professionnels ou administratifs et de discours politiques. Les données documentaires ont été sélectionnées comme suit :

[8] Un thésaurus est une arborescence de concepts définis par des ontologies (listes de mots) et des réseaux sémantiques.
[9] Le logiciel Sphinx met en garde l'utilisateur quand le corpus fait moins de 5 000 mots pleins.

Encadré méthodologique : Identification et sélection des documents analysés

Nous avons effectué une recherche avec le mot clé « réseau de chaleur » (dans le titre), sur les sites des acteurs nationaux des réseaux de chaleur. Notre recherche s'est concentrée sur les documents postérieurs à 2015, année de promulgation de la loi relative à la transition énergétique pour une croissance verte (TECV). Nous avons obtenu 9 documents provenant du Ministère de la Transition Ecologique et Solidaire, 33 des agences de l'Etat (ADEME, CEREMA et Plan Bâtiment durable), 18 des associations de promotion des réseaux de chaleur (Amorce et ViaSèva), 11 de la FNCCR (Associations des collectivités territoriales), 11 des fédérations et syndicats privés (FEDENE et SNCU) et 6 des associations des abonnés et usagers (USH, ARC et CNL).

Concernant les documents qui émanent des collectivités locales, nous avons sélectionné des villes en région Auvergne-Rhône-Alpes qui possèdent des réseaux de chaleur : Annecy, Annemasse, Chambéry, Grenoble, Lucinges et Voreppe. Pour chacune de ces villes, nous avons recherché sur leur site et leur journal communal les articles qui se rapportent aux réseaux de chaleur, et ainsi obtenu 17 documents qui émanent des collectivités territoriales. Pour les opérateurs des réseaux de chaleur, une recherche a été faite selon les mêmes critères sur les sites des 4 grands opérateurs (Dalkia, Cofely, Idex et Veolia). Nous avons retenu pour chacun d'eux 4-5 articles ou pages web qui présentent leurs actions sur les réseaux de chaleur - si possible en rapport avec les villes sélectionnées précédemment, soit au total 15 documents dont 7 communiqués de presse communs avec les collectivités territoriales, qui annoncent la mise en place d'un nouveau réseau de chaleur.

Nous avons également recherché des articles en provenance des médias grand public et spécialisés via les bases de données Europresse et Techniques de l'ingénieur. N'ont été retenus que les articles parus après 2015 dont les mots « réseaux de chaleur » sont présents dans le titre, soit 31 articles de la presse grand public et 131 de la presse spécialisée. Pour limiter le nombre d'articles, et pour plus d'homogénéité, nous n'avons conservé qu'un tiers des articles de la presse spécialisée, soit 38.

Enfin, des articles académiques ont été identifiés via la base de données Cairn avec le même processus de recherche. Sur les 58 articles ainsi obtenus, nous avons conservé ceux pour lesquels les termes « réseau de chaleur » apparaissent dans l'article intégral, soit 14 articles in fine.

Cette recherche a été menée en utilisant les logiciels Sphinx IQ2 et Dataviv (visualisation de données). Les chercheurs ont pu bénéficier d'un budget pour une formation et un accompagnement de la société Sphinx et du Professeur émérite Jean Moscarola, qui leur a apporté ses connaissances et son expertise en analyse de données textuelles.

Les liens présents dans le texte permettent au lecteur de faire l'expérience de ce corpus et des méthodes utilisées.

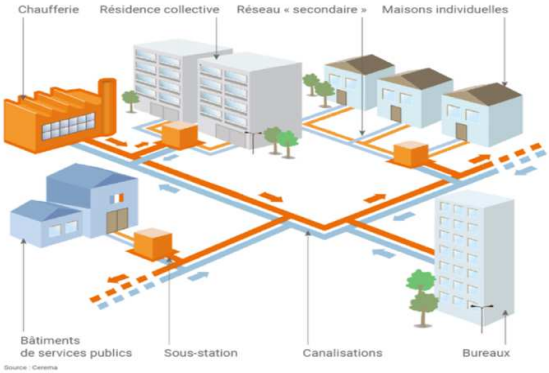


Figure 1 – Schéma de fonctionnement d'un réseau de chaleur (Cerema, 2019)

1. L'étude des discours sur les réseaux de chaleur

Qu'est-ce qu'un réseau de chaleur ?

Un réseau de chaleur permet de chauffer un quartier, une ville ou un territoire. Il est constitué de plusieurs parties : l'unité de production d'énergie ou chaufferie, un système de canalisations pour transporter de l'eau chaude ou de la vapeur, et un point de livraison de la chaleur

appelé sous-station (le système primaire). Au-delà de la sous-station, commence le réseau secondaire. Il s'agit d'un réseau interne à l'immeuble constitué de tuyaux et des radiateurs ou planchers chauffants (cf. Figure 1). Apparus au début du XXe siècle en France dans les villes avec une forte demande en énergie, les réseaux de chaleur connaissent un regain de popularité depuis la loi relative à la transition énergétique pour la croissance verte de 2015, car ils sont perçus comme le moyen de mobiliser massivement des énergies renouvelables et de récupération dans le chauffage urbain pour lutter contre le changement climatique (ADEME, 2017). Créer ou développer un réseau de chaleur implique de nombreux acteurs à tous les échelons : l'État fixe le cadre, les collectivités territoriales déclinent les politiques publiques sur les territoires, la ville décide de créer un réseau, en intégrant l'ensemble des acteurs concernés le plus en amont possible du projet (Amorce, 2017b).

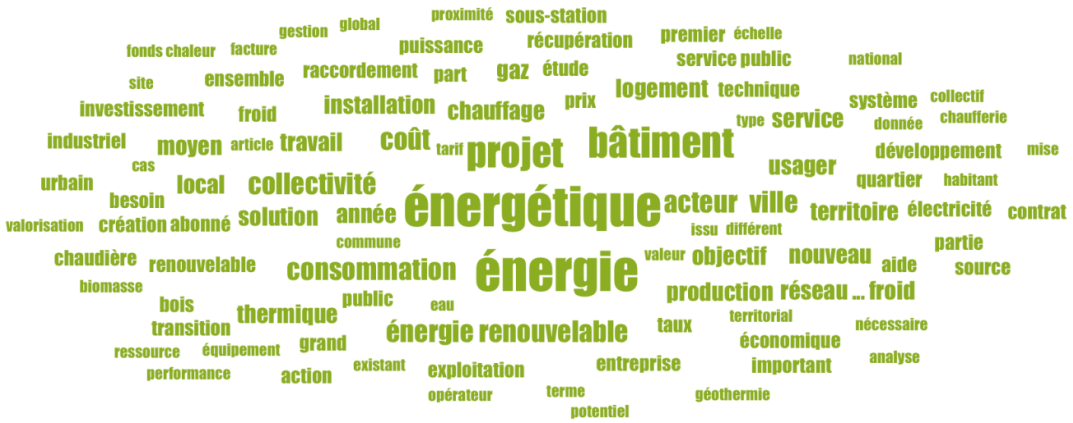
Les documents et le corpus

Chaque document comporte en moyenne 2 500 mots pour un corpus total de 510 530 mots et 16 045 phrases. Une exploration lexicale et sémantique des 196 documents a été réalisée, afin de faire émerger les principaux thèmes, catégories et associations significatives. Les documents proviennent majoritairement des médias et des agences de l'État (cf. Répartition des documents selon son origine). Certains documents sont relativement courts (par exemple les articles des médias) et d'autres beaucoup plus longs (des rapports émanant des collectivités publiques), il est donc pertinent d'analyser les documents en tenant compte du nombre de mots. On observe ainsi que les associations de promotion des réseaux de chaleur (Amorce et ViaSèva), les agences de l'État et les chercheurs sont les acteurs majoritairement représentés dans le corpus.

2. Exploration lexicale et sémantique

Les cinq mots clés les plus fréquents (cf. Figure 2), hors réseau et chaleur (énergétique, énergie, bâtiment, projet, et collectivité) permettent d'identifier les préoccupations des acteurs qui sont à l'origine des documents. Ainsi, le corpus de documents aborde les réseaux de chaleur en tant que systèmes énergétiques, dont l'objectif est de promouvoir les énergies renouvelables. Les réseaux fournissent de la chaleur à des bâtiments. Les projets sont portés par les collectivités et impliquent de nombreux acteurs. Le coût et la consommation d'énergie sont des critères importants.

Figure 2 – Nuage des mots les plus fréquents pour le corpus de phrases (hors réseau/chaleur)

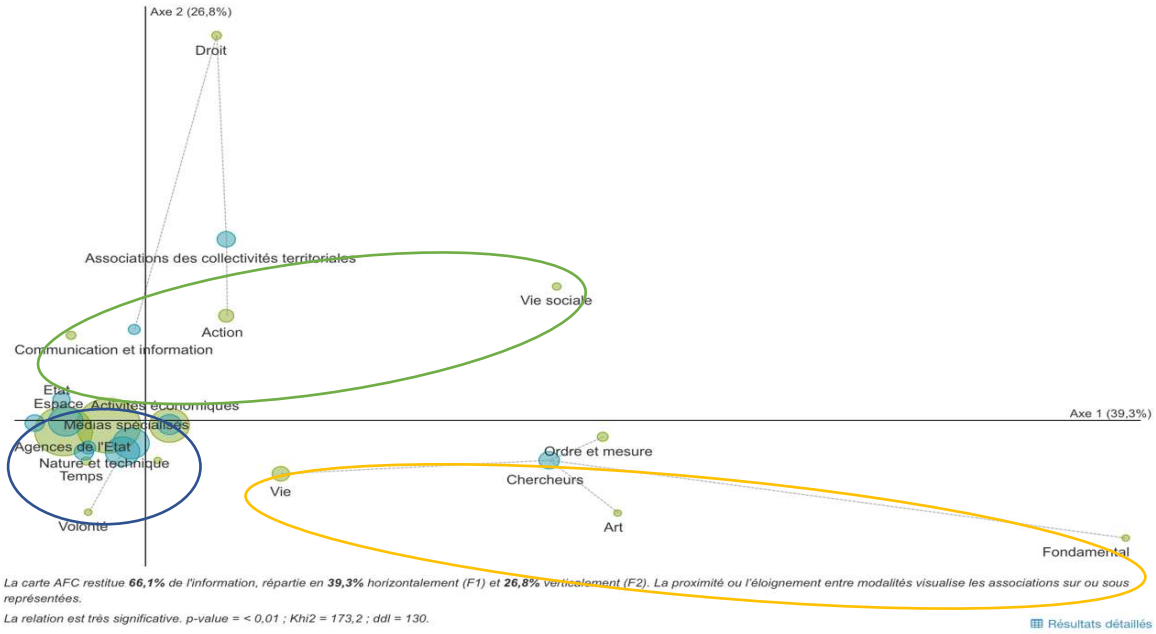


Les préoccupations spécifiques de chaque acteur sont mises en évidence par des mots clés. Ainsi les associations des collectivités territoriales insistent sur la notion de service public, sur l'importance des usagers et des consommateurs ; les chercheurs mettent l'accent sur les notions d'acteurs et de proximité. L'application des catégories d'un thésaurus standard permet de mettre en évidence l'importance de sujets et de concepts liés à la technique et à l'aménagement du territoire, et également dans le domaine économique. Ces résultats d'une première approximation lexicale et sémantique permettent ainsi un premier niveau d'interprétation.

Interprétation

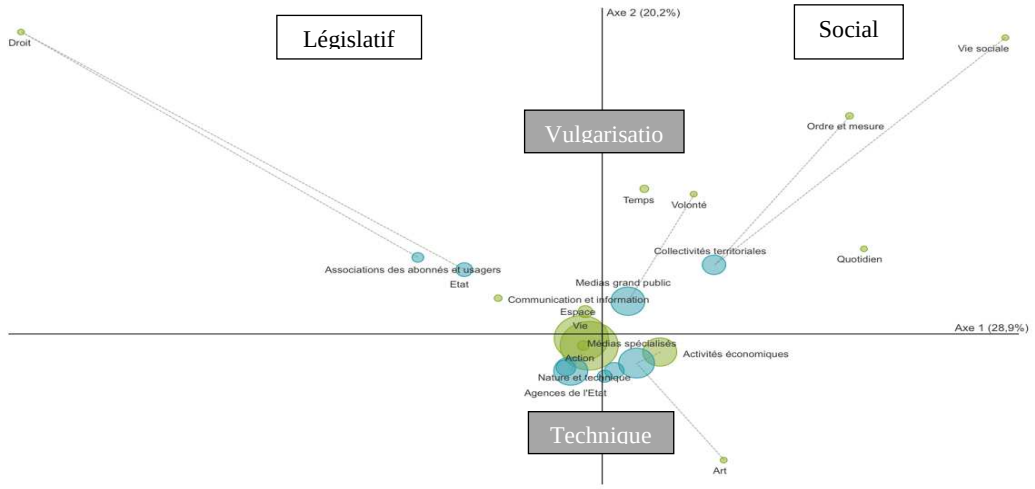
Une analyse factorielle des correspondances (cf. figure 3) permet de relier les concepts identifiés avec les acteurs, analyse qu'on pourrait traduire par « qui parle de quoi ? ». On distingue clairement trois catégories : les associations, qui représentent les collectivités territoriales et les usagers, et dont le discours est tourné vers préoccupations proches du terrain, du service à l'usager, en lien avec le droit et la vie sociale. Les chercheurs forment une catégorie homogène : ils cherchent à donner du sens et à définir des modèles. Enfin, les administrations publiques, les associations de promotion des réseaux de chaleur, les exploitants et leurs fédérations traitent de thèmes similaires (liés à la technique et à l'aménagement du territoire) et semblent avoir le même type de discours. L'État fixe le cap ; les associations de conseils sont là pour relayer la parole de l'État et promouvoir les réseaux de chaleur ; les exploitants des réseaux font la promotion des réseaux de chaleur à partir de leur réussite technique.

Figure 3 – Qui parle de quoi ? Les concepts et acteurs (AFC)

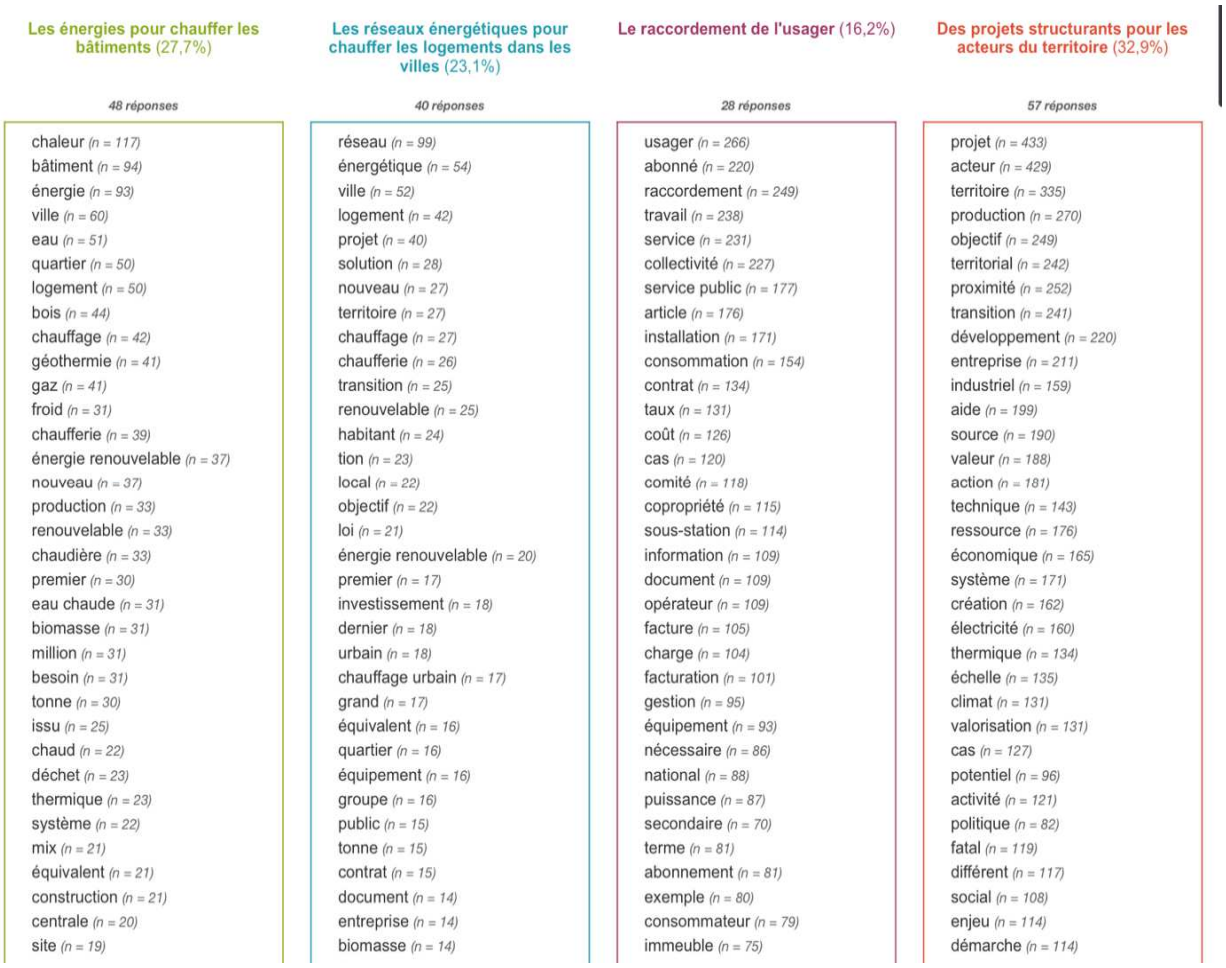


Cependant, un zoom sur la partie centrale (cf. figure 4) met en évidence des différences de discours autour de 4 zones. L'axe vertical distingue la vie sociale et le droit tandis que l'axe horizontal oppose un discours technique à un discours de vulgarisation, accessible au grand public. Ainsi les collectivités territoriales, qui s'expriment dans les journaux communaux, s'adressent à leurs administrés et abordent les réseaux de chaleur à partir de ce qu'ils peuvent apporter aux usagers. Les associations des usagers se focalisent sur les droits des usagers, les recours possibles en cas de litiges (par exemple l'ARC, l'association des représentants des copropriétaires : « Depuis maintenant plus d'un an, l'ARC dénonce la hausse des tarifs du réseau de chauffage urbain de la ville de Paris gérée par la CPCU », ARC, 2015). On retrouve dans cette partie l'État, qui fixe le cadre législatif.

Figure 4 – AFC (zoom sur la partie centrale de la Figure 4)



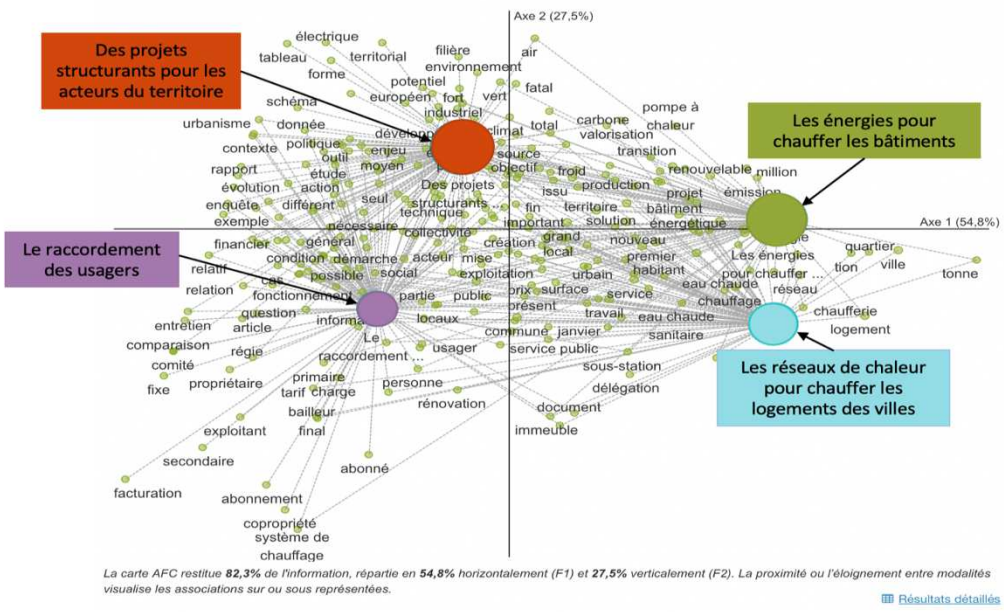
Dans la partie inférieure du graphique, les associations de promotion des réseaux de chaleur et les agences de l'État traitent de la partie technique des réseaux de chaleur. De même, les exploitants et leurs fédérations communiquent sur les performances techniques des réseaux, à l'instar de Dalkia : « Ce réseau, le plus grand construit en France sur les cinq dernières années, s'étendra sur 36 km depuis l'unité de valorisation des déchets du Mirail jusqu'au quartier de Montaudran. Ses sous-stations de livraison sont connectées au Desc de Dalkia (*Dalkia Energy Savings Center*) pour un pilotage numérique en temps réel et une optimisation des consommations. Alimenté par la chaleur issue du centre de valorisation des déchets et par celle d'un data center, son mix énergétique est composé à 70% d'énergies de récupération locales » (Dalkia, 2019). On observe, sans surprise, une distinction entre les médias spécialisés, qui s'adressent à un public averti et qui abordent des thèmes techniques, tandis que la presse grand public a un discours plus accessible, en lien avec l'intérêt des réseaux de chaleur dans le quotidien ou dans le domaine social.



A la lecture de ces résultats, la tentation est grande de conclure que l'ADT n'apprend pas grand-chose de nouveau sur le fond. Néanmoins, il s'agit d'une première étape rassurante pour le chercheur ou le praticien, qui retrouve des résultats attendus grâce à une méthode qui lui permet d'objectiver ces résultats.

Allons plus loin maintenant en recherchant les thèmes traités par ce corpus. La figure 5 ci-dessous situe les quatre groupes sur une carte d'analyse factorielle des correspondances. Elle représente une typologie des réponses, dégagée par une classification hiérarchique descendante de l'ensemble des documents pris dans leur globalité. La proximité ou la distance entre les éléments permet de visualiser les associations sur - ou sous - représentées. Elle permet de définir quatre grands thèmes. Cette classification est obtenue de façon automatique, et donc reproductible et objective. C'est ensuite l'interprétation, par le chercheur ou le praticien, qui permet de donner tout son sens à cette classification, à l'aide de l'ADT qui offre la possibilité de replacer très rapidement chaque mot dans son contexte (cf. figure 5).

Figure 5 : Classification hiérarchique descendante des 196 documents



Ainsi, les réseaux de chaleur sont abordés dans 57 documents comme des projets structurants pour les acteurs du territoire, comme dans cet extrait formulé par la Ministre de la Transition écologique et solidaire, Elisabeth Borne : « Les réseaux de chaleur sont désormais

majoritairement alimentés par des énergies renouvelables. Il s'agit là de projets de territoires, qui constituent un vecteur indispensable pour exploiter massivement les énergies renouvelables et de récupération » (MTES, 2019).

En approfondissant cette analyse, grâce à une typologie détaillée au niveau des phrases, on constate que la thématique des projets est traitée d'une part du point de vue de la relation contractuelle entre la collectivité et l'opérateur, avec des notions légales/juridiques (type de contrats, litiges) et par ailleurs sur le plan des relations de proximité nécessaires entre les acteurs : « La résolution des conflits passe en effet par la capacité des acteurs à mobiliser et/ou à construire une proximité relationnelle susceptible de faciliter la production de règles collectives et d'un projet de développement partagé » (Beaurain et al., 2017). L'importance est mise également sur la notion de projet de territoire pour répondre à des enjeux écologiques et économiques, le recours à certaines énergies renouvelables ou de récupération permettant de contribuer à la protection de l'environnement et à la préservation de l'emploi local (ViaSèva, 2018).

Le thème suivant traite des énergies pour chauffer les bâtiments. Les réseaux de chaleur sont décrits comme les seuls moyens d'utiliser certaines énergies, comme les énergies de récupération par exemple. L'analyse des idées détaillées dans cette thématique fait émerger de nouvelles questions : quelles énergies pour les réseaux de chaleur de demain ? Comment introduire plus d'énergies renouvelables ? Quelle place pour la production d'électricité par les réseaux de chaleur ? Par exemple : « En combinant la production d'électricité et de chaleur, ce procédé, permet, pour une quantité d'énergie produite, d'utiliser moins d'énergie primaire » (Amorce, 2017b). Les documents tendent à montrer que des innovations technologiques vont se développer : « Des améliorations sont à attendre également sur le stockage de l'énergie et sur le pilotage intelligent » (ADEME, 2019).

Le troisième thème abordé par ces documents peut se résumer ainsi : « les réseaux de chaleur pour chauffer les logements dans les villes ». Les réseaux de chaleur y sont traités de façon très générale et globale de façon à montrer leur intérêt, par exemple comme un levier pour la transition énergétique et comme un moyen de lutter contre la précarité énergétique : « Réussir la transition énergétique suppose en effet que chacun suive le mouvement, y compris les plus fragiles » (ViaSèva, 2018).

Le dernier thème est lié au raccordement des usagers, autant sur les volets économique (« le coût du raccordement doit être incitatif », Amorce, 2016a) que juridique (« la responsabilité des installations de chauffage et d'eau chaude sanitaire interne à l'immeuble incombe au

gestionnaire de l'immeuble », ViaSèva, 2018). Les usagers doivent être pris en compte le plus en amont possible dans le processus de décision : « En associant les futurs abonnés et usagers dès le début du projet, le maître d'ouvrage garantit une bonne compréhension des choix, du fonctionnement du futur réseau par tous les acteurs du projet » (Amorce, 2017b). En effet, les usagers restent sceptiques : « si les réseaux de chaleur sont, aujourd'hui, pour la plupart d'entre eux, compétitifs, cette réalité n'est pas toujours perçue par les usagers » (ViaSèva, 2018), d'autant que les prix diffèrent d'un réseau à l'autre : « Les analyses montrent la grande disparité des prix de vente d'un réseau à l'autre, et ce quelles que soient les sources d'énergie majoritaires » (Amorce, 2016b).

L'influence du type d'acteur

La figure 6 met en évidence l'influence de l'acteur à l'origine du document sur les thèmes principaux traités. Sur la carte, les proximités entre acteurs et le thème caractérisent les discours selon l'acteur à l'origine du document. De nouveau, c'est l'interprétation du chercheur ou du praticien qui permet de donner du sens à cette représentation des données. Ainsi, il est possible de distinguer d'une part le discours de projet de celui orienté vers l'usager, d'autre part le discours générique du discours plus spécifique appliqué au territoire.

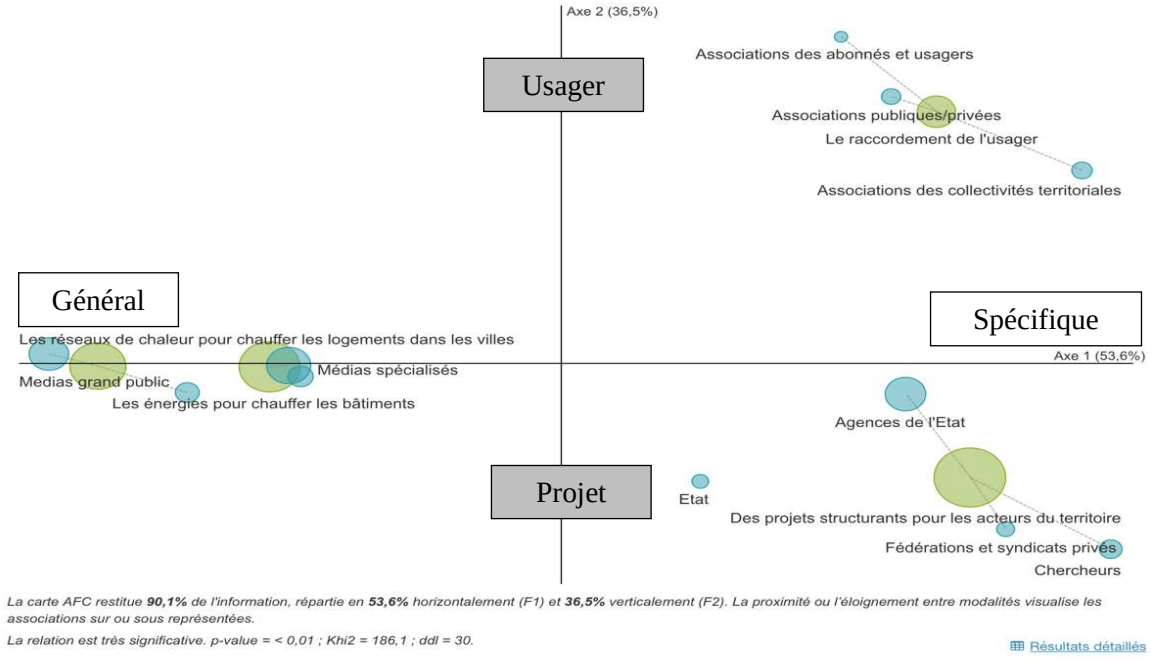
On peut ainsi observer que la première thématique, qui concerne les réseaux de chaleur en tant que projet de territoire avec des objectifs de développement, est abordée par l'État et surtout par les agences de l'État, qui ont pour rôle de déployer les politiques nationales. Les chercheurs sont également présents, car ils étudient les réseaux de chaleur en tant que projet de territoire avec des acteurs de proximité : « Nous adoptons la position qui consiste à distinguer les proximités géographiques, organisationnelles et institutionnelles » (Dehez et Banos, 2017). On retrouve ici les fédérations des entreprises privées, avec un discours orienté projet et développement.

Les associations des collectivités territoriales, les associations des abonnés et des usagers ou encore les associations de promotion des réseaux de chaleur sont les acteurs qui abordent le plus la thématique du « raccordement de l'usager ». Ce discours s'adresse aux collectivités territoriales, aux propriétaires des réseaux, aux exploitants qui vendent la chaleur et aux usagers, pour expliquer, rassurer sur les consommations d'énergie, les charges, le coût de l'énergie : « Le calcul du R1 qui correspond à la facturation d'énergie consommée et du R2 qui est une redevance fixe liée à l'abonnement pour prendre en compte les coûts liés à l'exploitation et l'entretien du réseau primaire » (ARC, 2019).

Le dernier type de discours est un discours plus généraliste, essentiellement issu de la presse, que ce soit la presse spécialisée ou « grand public », qui présente les réseaux de chaleur comme un outil énergétique pour chauffer les logements dans les villes. Les médias s'adressent à un public large et traitent le sujet de façon très générale, avec l'objectif de promouvoir les réseaux de chaleur et ses atouts ; on y trouve rarement d'articles polémiques : par exemple « Un financement participatif a été lancé pour la rénovation et le verdissement du réseau de chaleur et de froid de la ville de Courbevoie » (Aujourd'hui, 2018).

L'analyse plus poussée de ces quatre grands thèmes et des discours des différents acteurs permet de faire ressortir un élément plus surprenant : la problématique de la transition énergétique ne ressort pas ou peu. Alors qu'au niveau national, les réseaux de chaleur sont présentés comme : « un des leviers essentiels pour atteindre la neutralité carbone à l'horizon 2050 » (MTES, 2019), cette thématique n'est pas, ou très peu, abordée dans les discours des différents acteurs, y compris ceux qui émanent de l'État ou des agences de l'État. L'accent est davantage mis sur les aspects techniques et financiers des réseaux de chaleur, au détriment d'un discours motivant qui mettrait l'accent sur leurs avantages environnementaux, les gains en matière d'émissions carbone, de qualité de l'air, les avantages pour l'économie locale… au risque de renforcer l'idée que les réseaux de chaleur sont des systèmes techniques compliqués.

Figure 6 - Thèmes abordés selon les acteurs



III - Discussion et contributions

Les méthodes qualitatives traditionnelles peuvent être classifiées en deux catégories :

* La recherche qualitative « pure », dans la tradition des humanités (selon Dumez, 2013 et Miles et Huberman, 1994);

* L'analyse de contenu formalisée (selon Berelson, 1952 et Krippendorff, 2013), fondée sur un travail de codification systématique.

Les justifications du recours à l'ADT comme alternative à ces méthodes pour analyser du texte sont les suivantes :

1/ Le volume des informations à analyser est allégé par le recours à l'ADT, ce qui évite une surcharge de travail pour le chercheur et des tris arbitraires de données.

2/ L'objectivité est mieux garantie avec l'ADT, qui évite la récursivité de l'analyse qualitative « pure » ou de l'analyse de contenu (Dumez, 2013).

3/ L'ADT favorise la transparence de l'analyse et l'exposition à la critique, souvent identifiées comme des points faibles des approches qualitatives.

Notre article permet ainsi d'avoir une vision plus complète des apports, des différentes utilisations et des techniques de l'ADT.

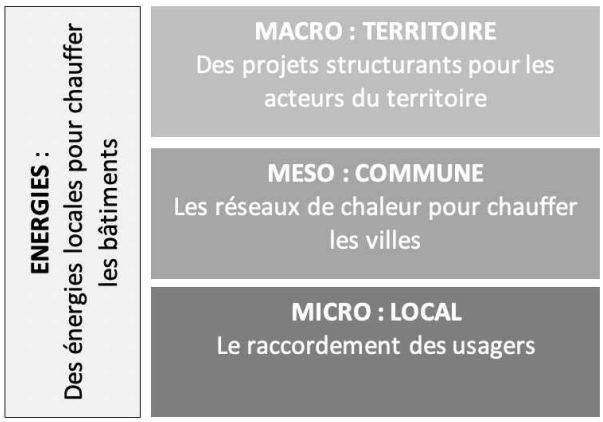
Au niveau du fond

Telle qu'elle est menée dans cet exemple, la recherche documentaire peut être considérée comme un aboutissement en soi, dans la mesure où elle apporte des résultats très intéressants, tant pour les chercheurs que pour les praticiens (par exemple les acteurs des réseaux de chaleur). L'analyse des documents par l'analyse de données textuelles (ADT) permet d'identifier les grands thèmes abordés par les différents acteurs et de faire ressortir de manière objective les dimensions du discours, les perceptions des acteurs (thèmes) et leur importance (fréquence). Les réseaux de chaleur sont perçus comme des projets structurants pour les acteurs, car ils permettent de chauffer les bâtiments urbains avec un mix énergétique moins polluant, mobilisent des énergies locales et créent des emplois pérennes et locaux. Néanmoins, plusieurs éléments ressortent de l'analyse menée dans ce vaste corpus de documents, qui permettent de nuancer ce constat. D'une part la question de la transition énergétique n'apparaît pas au premier plan, et d'autre part, l'usager final et son confort énergétique ne sont que très peu pris en compte dans le processus de décision concernant les réseaux de chaleur, s'agissant de projets qui impliquent pourtant de multiples parties prenantes.

Grâce à l'ADT menée, on peut faire le constat suivant : bien que l'Etat inscrive la transition énergétique comme une de ses préoccupations principales en lien avec les réseaux de chaleur (Loi TECV, 2015[10]), ce discours n'est pas repris par les acteurs dans leur communication sur les réseaux de chaleur. On constate que dans ce corpus, le réseau de chaleur est avant tout abordé comme un objet technique, instrumentalisé par les différents acteurs, avec leurs intérêts et intentions propres. La problématique territoriale est davantage présente et mise en avant que le réseau lui-même en tant qu'instrument de la transition énergétique, et le discours sur le réseau de chaleur semble ainsi avant tout l'occasion pour les acteurs d'affirmer des actions territoriales et locales.

On observe aussi un décalage entre les intentions des pouvoirs publics et la mise en œuvre au niveau du territoire (la ville, le quartier et le bâtiment). Il n'y a pas nécessairement une convergence entre les stratégies des territoires et la logique de transition énergétique, qui semble finalement moins prise en compte que les problématiques spécifiques de chacun des acteurs. Cette opposition entre des logiques territoriales et leur mise en œuvre est liée au statut des acteurs. L'ADT permet au chercheur d'écouter la force du « bruit » et des répétitions dans le langage utilisé par les différents acteurs. L'observation des discours et d'un ensemble de « langages » fait ressortir dans cette analyse quatre principales catégories et idées clés véhiculées par les acteurs, qui révèlent trois niveaux de prise en compte des réseaux (territoire, ville et bâtiment), ainsi que la thématique énergétique, qui est transversale à ces trois catégories (Figure 7).

Figure 7 – Représentation des quatre thèmes issus du corpus de documents



[10] Loi n°2015_992 du 17 août 2015 relative à la transition énergétique pour la croissance verte.

Au niveau de la méthode

Nous avons montré avec cet exemple que les corpus de documents peuvent être abordés de manière exploratoire, par les techniques de l'analyse de données textuelle, lexicale et sémantique. Cette première approche du corpus mobilise l'analyse statistique comme préalable à la lecture et à l'interprétation des substituts lexicaux produits. Le retour au texte et la visualisation de données permettent de contrôler cette lecture qui, in fine, donne un sens aux structures du texte révélées par l'ADT.

Cette étude exploratoire permet au praticien et au chercheur d'avoir rapidement une vision des principaux thèmes traités dans un nombre élevé de documents, d'identifier qui parle de quoi et de voir les thèmes privilégiés selon l'acteur à l'origine du document. Cette analyse rapide, facilitée par les représentations graphiques, permet également d'identifier les thèmes non abordés et d'en déduire des problématiques qui pourront être étudiées et approfondies, par l'analyse d'autres documents ou le recours à des entretiens. Cette analyse documentaire favorise la mise en évidence des positions officielles, des informations et des intentions présentes dans le contenu des documents analysés, et permet ainsi de s'affranchir des discours subjectifs recueillis lors d'un entretien. L'analyse des données textuelles permet aussi de quantifier, ce qui garantit l'objectivité de la description des contenus étudiés. Les résultats peuvent dans un premier temps se substituer à une lecture « classique », ce qui évite au chercheur un travail fastidieux et subjectif. Dans un deuxième temps, le chercheur peut mener une lecture avertie, à l'aide des lexiques, concepts ou typologies mises en évidence par l'ADT, tout en exposant à la critique des lecteurs le travail d'interprétation mené.

La mise en œuvre du logiciel utilisé (Sphinx IQ2) et d'un outil de visualisation (Dataviv) favorise une expérience renouvelée de la prise de connaissance des corpus. La visualisation de données et l'infographie dynamique permettent d'échapper à la linéarité du texte en stimulant une recherche créative et la découverte. L'application web, qui autorise le partage de corpus dématérialisés, offre au lecteur de les découvrir par lui-même et surtout de discuter les interprétations proposées par l'auteur dans l'esprit critique du débat scientifique. Nous espérons que les liens proposés dans cet article permettent de percevoir cette expérience cognitive particulière. Plusieurs obstacles doivent être surmontés pour favoriser un recours plus fréquent à ces méthodes d'analyse : en premier lieu celui des habitudes, à commencer par les normes de la communication écrite, encore peu ouvertes aux extensions web. Par ailleurs, le clivage qualitatif /quantitatif constitue également un obstacle, ainsi que la technicité des notions et méthodes de l'ADT, qui exige des compétences spécifiques. Enfin, ces techniques,

par leur efficacité de premier abord et l'aisance qu'elles donnent à ceux qui les maîtrisent, peuvent comporter un risque d'usages abusifs ou trompeurs, si la construction du sens se fait au détriment du temps, long, de la lecture et de la culture.

CONCLUSION

Notre travail répond à un objectif d'informer, de porter à connaissance, de vulgariser l'ADT et d'en montrer les utilisations principales. Les chercheurs et les consultants, *data scientists* ou *community managers* qui informent les décideurs, praticiens et managers, devraient s'intéresser à l'ADT pour répondre au défi des données massives. Lorsque le volume des données est réduit, la recherche qualitative « pure » reste la meilleure méthode pour les analyser, et potentiellement faire de l'induction et permettre la théorisation. Il y a donc bien trois manières d'analyser des textes et de faire de la recherche qualitative : la recherche qualitative « pure », dans laquelle le chercheur commente les textes ; l'analyse de contenu, qui permet la systématisation et l'industrialisation de la lecture (cf. Krippendorff, 2013) ; enfin, l'ADT, ou l'analyse assistée par ordinateur, qui permet l'automatisation et s'appuie sur la formalisation.

Dans la tradition de recherche qualitative pure, le chercheur lit plusieurs dizaines de documents, sélectionnés de manière plus ou moins éclairée, ou analyse statistiquement les milliers de références relative à son sujet en faisant lire "titres, résumés, mots clés" par des automates linguistiques, pour obtenir, grâce à la statistique et à l'analyse de données, une idée des thématiques, des spécialisations, des évolutions et des impacts.

La pertinence de l'ADT est liée au fait qu'elle permet une analyse rapide et automatisée de l'intégralité des documents. Appliquée à des corpus volumineux, la méthode de l'ADT permet de révéler des structures, des généralités et des contrastes qui ne peuvent pas être facilement détectés par une lecture humaine non instrumentée des textes (Comby, 2015). C'est ce nous observons dans ce travail, en appliquant cette méthode d'analyse à un corpus de documents sur les réseaux de chaleur. L'ADT nous a permis de détecter, au-delà des langages propres à chaque type d'acteur, des différences importantes de discours à différents niveaux (macro, méso et micro), ainsi que des différences de préoccupations et de perceptions selon les acteurs. Nous exploitons ici un avantage indéniable de ces outils pour l'analyse systématique d'un corpus très important, irréalisable sans outil.

Cet atout est contrebalancé par une première limite, liée au fait que les structures textuelles révèlent fatalement des évidences. Ces évidences permettent de prouver la pertinence de l'ADT, qui permet ainsi de conforter l'analyse, de l'étayer par des preuves objectives, et de

partager les analyses avec le lecteur. Il incombe au chercheur d'approfondir ces premiers éléments d'analyse. Comme dans toute démarche qualitative, la recherche consiste à construire le sens au-delà des premières impressions en déplaçant le focus, en ciblant ou en approfondissant, à la recherche de structures moins triviales. L'ADT permet de réaliser ce travail d'approfondissement en s'appuyant sur des données objectives et non pas uniquement sur l'analyse subjective du chercheur. L'ADT permet ainsi de recueillir des éléments inattendus et de réaliser une interprétation plus approfondie, ce qui constitue une exigence pour toute méthode de recherche. Nous avons montré dans cet article comment la recherche menée peut produire des éléments inattendus, concernant les niveaux d'analyse et de perception de l'objet « réseaux de chaleur » observé ici.

Une deuxième limite des outils de l'ADT est liée à l'utilisation d'un thésaurus standard pour identifier les idées et les concepts. Dans notre travail, l'élaboration d'un thésaurus plus technique, spécifique aux réseaux de chaleur permettrait une analyse plus pointue encore, ce qui peut faire l'objet d'un prolongement du travail réalisé ici, dans le cadre d'une étude ultérieure. Une dernière limite, la principale sans doute, provient du fait que l'ADT nécessite de maîtriser l'usage d'un logiciel spécifique. Cette méthode, à cheval entre méthodes qualitatives et quantitatives, nécessite donc à la fois des connaissances en traitement statistique et des compétences littéraires de création de sens, de conceptualisation et d'abstraction, qui permettent d'éviter les paraphrases et d'approfondir la lecture des analyses effectuées.

Concernant la thématique des réseaux de chaleur, étudiée ici grâce à l'ADT, la recherche menée permet d'apporter une double contribution, méthodologique et théorique, et d'envisager des pistes de recherche futures. L'analyse réalisée montre par exemple que la question de la transition énergétique, pourtant à la base du déploiement des réseaux de chaleur en France (MTES, 2019), n'est que peu abordée dans les discours des acteurs ; ceux-ci perçoivent le réseau de chaleur comme un objet structurant pour les acteurs du territoire, avant d'être un outil au service de la transition énergétique. Plusieurs pistes peuvent être avancées pour expliquer ce paradoxe : est-ce le discours porté par l'Etat qui ne porte pas ? Est-ce que cela révèle un manque d'intérêt la part des acteurs des territoires ? Ou bien est-ce qu'il manque un outil qui permettrait aux territoires de s'emparer des réseaux de chaleur comme vecteur de la transition énergétique ? Selon le scénario NegaWatt (2015), la transition énergétique peut être appréhendée par les trois volets : sobriété, efficacité et développement des énergies renouvelables, mais on observe que ces questions sont assez peu abordées dans

ce corpus de documents sur les réseaux de chaleur. De même, la gouvernance entre les multiples acteurs, ainsi que la coopération inter-organisationnelle, sont des thématiques peu évoquées, alors même que l'analyse des discours est loin de manifester une convergence spontanée.

Ces interrogations ouvrent des perspectives sur des pistes de recherche futures, et bien d'autres pistes peuvent être proposées. Le développement des réseaux de chaleur implique qu'ils puissent répondre à un certain nombre de défis écologiques, organisationnels, technologiques et économiques. Par exemple, les enjeux liés au réchauffement climatique impliquent de mobiliser de nouvelles énergies renouvelables, et également de répondre à un besoin de production de froid qui va augmenter. De nouvelles technologies vont voir le jour pour répondre à ces nouveaux besoins. Comment ces technologies innovantes vont-elles trouver leur place au sein des réseaux de chaleur ? Quels acteurs décideront de leur adoption, et selon quels critères : écologiques, économiques, politiques, etc.? Quel sera le système de gouvernance dans des réseaux où les usagers pourront également être producteurs d'énergie ou co-financeurs ? Face à un prix de pétrole au plus bas, comment les réseaux de chaleur peuvent-ils rester compétitifs ? De nouveaux modèles économiques sont-ils envisageables ?

BIBLIOGRAPHIE

ADEME (2019). « Les réseaux de chaleur et de froid, Etat des lieux de la filière, marchés, emplois, coûts », Expertises, Rapport, 79 p.

ADEME (2017). « Les réseaux de chaleur et de froid. Etat des lieux de la filière. Marchés, Emplois, Coûts », Expertises, Rapport, mai, 89 p.

Amorce (2017a). « Guide de création d'un réseau de chaleur, Éléments clés pour le maître d'ouvrage », Série Technique, RCT 46, mars, 55 p.

Amorce (2017b). « L'élu et les réseaux de chaleur », Collection guide de l'élu, 72 p.

Amorce (2016a). « Coût de raccordement et dispositifs incitatifs », Série économique, RCE 22, 23 p.

Amorce (2016b). « Compétitivité des réseaux de chaleur en 2015. Comparaison des modes de chauffage et prix de vente moyen de la chaleur », Série économique, RCE 26, 67 p.

ARC (2019). « Des nouveaux outils à destination des copropriétés concernant le raccordement à un réseau de chaleur », https://arc-copro.fr/documentation/de-nouveaux-outils-destination-des-coproprietes-concernant-le-raccordement-un-reseau.

ARC (2015). « L'arc vient en aide aux abonnés », https://arc-copro.fr/documentation/larc-vient-en-aide-aux-abonnes-dun-reseau-de-chauffage-urbain.

Aujourd'hui (2018). « Courbevoie : déjà 316 prêteurs pour le réseau de chaleur », Aujourd'hui en France (site web), 14 décembre.

Bardin L. (1977). L'analyse de contenu, PUF, Paris.

Baulac Y., Ganassali S., Moscarola J. (2006). « 40000 pages pour un livre, le cas du Débat National sur l'Ecole Publique », Journées de l'Analyse de Données Textuelles, Avril.

Beaurain C., Maillefert M., Lenoir Varlet D. (2017). « La proximité au cœur des synergies éco-industrielles dunkerquoises », *Flux*, 2017/3, n°109-110, 23-35.

Benzecri J.P. (1973a). L'Analyse des données. Tome 1 : la taxinomie, Dunod, 615 p.

Benzecri J.P. (1973b). L'Analyse des données. Tome 2 : l'analyse des correspondances, Dunod, 619 p.

Berelson B. (1952). Content Analysis in Communication Research, Glencoe: Free Press.

Bogaert J.C., Moscarola J., Mothe C. (2018). « Recherche historique, narration et documents d'archives ». Dans Chevalier F., Cloutier L.M., Mitev N. (Eds.), Les méthodes de recherche du DBA, chapitre 14, Editions EMS Management et Société, collection BSI, France.

Buhler T., Lethier V. (2020). « Analysing urban policy discourses using textometry: An application to French urban transport plans (2000–2015) », *Urban Studies*, 57(10), 2181-2197.

Cerema (2019). Raccordement des copropriétés aux réseaux de chaleur. Guide méthodologique, Juillet, 28 p.

Comby E. (2015). « L'analyse de données textuelles et l'acceptation sociale », Dans Depraz S., Cornec U., Grabski-Kieron U. (Eds), Acceptation sociale et développement des territoires, Lyon: ENS Editions, 131-136.

Dalkia (2019). Inauguration du réseau de chaleur et de froid Plaine Campus, https://www.dalkia.fr/fr/espace-presse/communique-de-presse/dalkia-reseau-chaleur-froid-toulouse.

Dehez J. et Banos V. (2017). « Le développement territorial à l'épreuve de la transition énergétique, le cas du bois énergie », *Géographie, économie, société*, 1, vol. 19, 109 – 131.

Drisko J.W., Maschi T. (2016). Content analysis, Oxford University Press, Oxford.

Dumez H. (2013). Méthodologie de la recherche qualitative, Vuibert, Paris.

Gauzente C., Peyrat-Guillard D. (2007). Analyse statistique de données textuelles en sciences de gestion : concepts, méthodes et applications, EMS.

Gavard-Perret M.-L., Moscarola J. (1996). « Lexical analysis in marketing: discovering the contents of the message or recognizing the models of enunciation », French-German Workshop on New Developments and Approaches in Consumer Behavior Research, University of Potsdam, Germany, 40 - 59.

Gephart, J.R.P. (1993). « The Textual Approach: Risk and Blame in Disaster Sensemaking », Academy of Management Journal, 36(6), 1465–1514.

Gioia D.A., Corley K.G., Hamilton A.L. (2012). « Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology », *Organizational Research Methods*, vol. 16, n°1, 15-31.

Girin J. (1989). « L'opportunisme méthodique dans les recherches sur la gestion des organisations », Communication à la journée d'étude la recherche-action en action et en question, AFCET, Collège de systémique, École Centrale de Paris, 10 mars.

Glaser B., Strauss A. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research, Mill Valley, CA: Sociology Press.

Goddard C. (2011). Semantic Analysis: A Practical Introduction, Oxford University Press.

Humphreys A., Jen-Hui Wang R. (2018). « Automated Text Analysis for Consumer Research", *Journal of Consumer Research*, 44(6), 1274–1306.

Jenny J. (1999). « Pour engager un débat avec Max Reinert, à propos des fondements théoriques et des présupposés des logiciels d'analyse textuelle », *Langage & société*, 90(1), 73‑85.

Kobayashi B., Mol S.T., Berkers H.A., Kismihok G., Hartog Den D.N. (2018). « Text Classification for Organizational Researchers: A Tutorial », *Organizational Research Methods*, Vol. 21 (3), 766-799.

Krippendorff K. (2013). Content Analysis. An Introduction to Its Methodology (3rd edition), Sage, Thousand Oaks, California.

Lebart L., Pincemin D., Poudart C. (2020). Analyse des données textuelles, Presses de l'Université du Québec, Canada.

Miles M. B., Huberman M. A. (2003). *Analyse des données qualitatives* (2ème édition), De Boeck, Paris.

Miles M. B., Huberman M. A. (1994). Qualitative Data Analysis, Thousand Oaks, CA: Sage.

Moscarola J. (2018a). Faire parler les données, Editions EMS, France.

Moscarola J. (2018b). « Visualisation de données et infographie dynamique : le logiciel Sphinx », Dans F. Chevalier, M. Cloutier, N. Mitev (eds), Méthodes de recherche pour le DBA. EMS, Paris, 340 - 357.

Mossholder K.W., Settoon R.P., Harris S.G., Armenakis A.A. (1995). Measuring Emotion in Open-Ended Survey Responses: An Application of Textual Data Analysis, *Journal of Management*. 21(2), 335.

MTES (2019). Réseaux de chaleur et de froid. Ministère de la transition écologique et solidaire – une filière d'avenir, Dossier de Presse. Octobre.

NegaWatt (2015). Manifeste NegaWatt – En route pour la transition énergétique !, Actes Sud / Association NegaWatt, Coll. Babel, 400 p.

Raich M., Müller J., Abfalter D. (2014). Hybrid analysis of textual data. *Management Decision*, 52(5), 737-754.

Reinert A. (1983). « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », Les cahiers de l'analyse des données, tome 8, n°2, 187-198.

Saldaña J. (2013). The coding manual for qualitative researchers (2ème ed.), Sage, London.

Tsao H., Campbell C.L., Sands S., Ferraro C., Mavrommatis A., Lu S. (2020). A machine-learning based approach to measuring constructs through text analysis. *European Journal of Marketing*, 54(3):511-524.

ViaSèva (2018). « Comment agissent les réseaux de chaleur pour lutter contre la précarité énergétique ? », 28 p.