Sphinx iQ3 : L'alliance de l'ADT, de la Statistique et de la Datavisualisation

Younès Boughzala, Jean Moscarola

Abstract 1

The purpose of this paper is to present a new Textual Data Analysis software: Sphinx iQ3, which aims to be a tool that blends analysis approaches and resources, mixing the strength of statistics with the power of data visualization. It is a complete set of tools that enables the analysis of large corpora from various origins (open questions, online reviews, scientific or press articles, free or semi-directive interviews, focus groups, forums, social network pages, Open Data...), and the automatic combination of synthesis, content analysis (manual or by dictionary) and text mining. All of this with more visual, dynamic, and interactive graphical restitutions (classic report or online reporting) makes the corpora speak for themselves. The notice highlights the functionalities of the tool by using examples to better assess its scope and its limits.

Keywords: Textual Data Analysis, Lexical Analysis, Semantic Analysis, Data visualization, Sphinx iQ3

Résumé

L'objectif de cette communication est de présenter un nouveau logiciel d'ADT : Sphinx iQ3, qui se veut un outil qui mélange les approches et les ressources d'analyse, en mixant la force de la statistique avec la force de la visualisation des données. C'est un ensemble complet de fonctionnalités permettant d'analyser des corpus volumineux et de diverses origines (questions ouvertes, avis en ligne, articles scientifiques ou de presse, entretiens libres ou semi-directifs, focus-group, forums, pages de réseaux sociaux, Open Data…), et de combiner des synthèses automatiques, des analyses de contenu (manuelles ou par dictionnaire) et des fouilles de textes. Le tout avec des restitutions graphiques (rapport classique ou reporting en ligne) plus visuelles, dynamiques et interactives, pour bien faire parler les corpus. La communication met en évidence les fonctionnalités de l'outil en s'appuyant sur des exemples permettant d'apprécier sa portée et ses limites.

Mots clés : Analyses des Données Textuelles, Analyse lexicale, Analyse sémantique, Datavisualisation, Sphinx iQ3

1. Introduction

L'objectif de cette communication est de présenter un nouveau logiciel d'Analyses des Données Textuelles (ADT) : Sphinx iQ3. Le dernier né des logiciels Sphinx d'enquêtes et d'analyses quantitatives et qualitatives. Cet outil mélange les approches et les ressources d'analyse, en mixant la puissance de la statistique avec la force de la visualisation des données. C'est un ensemble complet de fonctionnalités permettant d'analyser des corpus volumineux, structurés ou non structurés, et de diverses origines (questions ouvertes, avis en ligne, articles scientifiques ou de presse, entretiens libres ou semi-directifs, focus-group, forums, pages de réseaux sociaux, Open Data…), et de combiner des synthèses automatiques, des analyses de contenu (manuelles ou par dictionnaire) et des fouilles de textes. En capitalisant sur les approches traditionnelles et pour répondre aux nouvelles exigences en termes de volume et de structure des données collectées, tous les éditeurs multiplient les efforts en R&D afin de proposer des outils de plus en plus performants qui répondent aux nouveaux besoins des chercheurs, entreprises et cabinets d'études.

En se référant à trois courants, les CAQDAS, les outils de Traitement Automatique des Langages (TAL) et les moteurs de recherche Web, pour l'analyse des données textuelles, Sphinx iQ3 intègre trois types d'approches complémentaires : lexicales, sémantiques et statistiques. Le tout avec des restitutions graphiques visuelles et/ou interactives, pour faire parler les corpus, et les partager en ligne sous forme de vues infographiques.

Le papier met en évidence les fonctionnalités de l'outil pour l'ADT : l'utilisation des moteurs sémantiques (thésaurus, ontologies, analyse des sentiments), les moteurs lexicaux et statistiques (classification hiérarchique descendante, exploration de verbatim) et l'analyse de contenu (grille thématique, dictionnaires et analyse automatique en cours de collecte). La première partie rappelle les différentes approches de l'ADT, les cas d'usages et les enjeux. La deuxième présente la notion de visualisation et ses usages pour partager des données. La troisième partie expose les fonctionnalités du logiciel pour l'ADT. La présentation orale s'appuiera sur des exemples permettant d'apprécier la portée et les limites de ce nouveau logiciel.

2. L'ADT

L'ADT s'est développée au cours de la deuxième moitié du 20ème siècle. Elle vise à découvrir l'information contenue dans un corpus textuel en qualifiant ses éléments au moyen de catégories lexicales et/ou sémantiques et en quantifiant leur fréquence et répartition statistique (Boughzala et al., 2014). C'est un courant qui s'est étendu grâce aux avancées des techniques de la linguistique informatique et du Web 2.0. Sa mise en œuvre repose sur des outils logiciels qui ont beaucoup évolué avec les technologies informatiques, linguistiques, et de l'intelligence artificielle (Mothe et al., 2021). Elle permet aujourd'hui de traiter des données massives contenues dans les interactions et partagées sur le Web.

Dans le domaine de la recherche qualitative en sciences humaines et sociales, l'ADT permet une analyse assistée et automatisée, complétant la tradition littéraire consistant à lire et commenter les textes (Moscarola, 2018). Elle recouvre des pratiques très variées aussi bien dans le monde académique que professionnel. A l'ère du Big data ou encore du Big Quali (Bô, 2022), où la masse des données textuelles explose, l'ADT permet d'étudier des données de différents types et de diverses sources. Ainsi, le corpus peut être disponible (littérature, compte-rendu, correspondances, articles scientifiques ou de presse, sites Web, forums, avis en ligne, pages réseaux sociaux...) ou volontairement généré ou produit pour le besoin d'une étude au moyen d'entretiens libres ou semi-directifs, focus-group, observations du terrain, etc…

L'ADT consiste à prendre connaissance du corpus, à le lire et le fouiller pour en extraire les mots clés, classer les fragments spécifiques, ou encore les coder manuellement sur la base d'une grille d'analyse pour dénombrer les principaux thèmes (Boughzala et al., 2014). Les CAQDAS (*Computer-Aided Qualitative Data Analysis Software*), ont ainsi contribué à une extension méthodologique de l'approche qualitative traditionnelle donnant jour à des analyses plus au moins automatisée grâce à des outils quantitatifs. Pour analyser un corpus textuel, trois approches se distinguent (Moscarola, 2018).

La première appartient à la tradition littéraire, il s'agit de construire un nouveau texte pour rendre compte des texte étudiés. Lire et écrire pour produire une synthèse, un résumé ou un commentaire critique dont le but est de défendre ou de contredire un point de vue. Dans tous les cas, il s'agit d'articuler une pensée autour d'idées ou de concepts illustrés par des citations intelligemment choisies. La qualité du résultat dépend alors des aptitudes du rédacteur à convaincre par la clarté de son exposé et la pertinence de ses citations.

La deuxième, manifeste l'ambition des sciences humaines et sociales qui cherchent à remplacer la subjectivité de l'auteur par l'utilisation de méthodes objectives. Il s'agit alors d'expliciter les méthodes et d'exposer les modalités de construction du sens. C'est la démarche de l'analyse thématique ou de contenu. Elle consiste à coder le texte en utilisant une grille de lecture construite par le chercheur ou le chargé d'études.

La troisième approche est apparue avec le traitement informatique et statistique des textes. Lexicale (dénombrement des mots), puis sémantique (identification automatique des contenus : concepts). Pour les tenants de l'ingénierie linguistique et de l'intelligence artificielle elle pourrait même complètement remplacer le lecteur. Pour les chercheurs et chargés d'études, elle amplifie la capacité de prise de connaissance des corpus en produisant des substituts du texte qui révèlent ses structures lexicales et sémantiques.

Ces trois approches ne doivent pas être considérées comme alternatives mais plutôt comme complémentaires (Moscarola, 2022), notamment avec la montée en puissance des méthodes mixtes de recherche (*Mixed Methods Research*) qui combinent les approches qualitatives et les approches quantitatives (Molina-Azorin, 2011) pour contrebalancer les faiblesses de l'une par les forces de l'autre et produire ainsi des résultats plus pertinents (Moscarola, 2018).

3. La Datavisualisation

La datavisualisation ou visualisation des données, appelée aussi *Dataviz*, est une branche de la statistique descriptive et un outil de stimulation dans l'esprit de la théorie enracinée (Walsh, 2015). Son but est de transmettre des idées et des indicateurs de manière efficace, sous forme statique et/ou dynamique, en leur associant des formes esthétiques et expressives. Ces aperçus de données éparses et/ou complexes ont pour but de faire percevoir l'essentiel des contenus. Ils contribuent à faire parler les données et raconter une histoire (*Storytelling*) ancrée sur les représentations graphiques des résultats obtenus (Moscarola, 2022). Elle fait appel aux apports de la statistique exploratoire et de l'analyse de données multidimensionnelle (Tukey, 1977) qui permettent de réaliser des sélections et synthèses, préalablement à la mise en scène des données. L'exploration et l'analyse des données par le lecteur peuvent s'en trouver facilitée.

L'intégration de la datavisualisation (et/ou de l'infographie) dans les logiciels d'analyses des données présente une réelle opportunité pour les chercheurs et les chargés d'études, mais aussi pour le lecteur des rapports d'études. Elles permettent de prolonger le travail de recherche et d'analyse par la mise en forme d'une communication en ligne plus efficace, accessible, moderne et largement diffusable. Le lecteur bénéficiera d'une représentation visuelle précise et complète pour synthétiser et donner du sens à la masse des données brutes. Les représentations peuvent prendre différentes formes : graphiques, diagrammes, cartographies, vidéos, etc. La finalité étant de rendre intelligibles données et indicateurs afin de les transformer en outils décisionnels.

Dans le domaine du management et de la recherche en sciences humaines et sociales, la datavisualisation donne au lecteur de l'étude l'occasion de découvrir les corpus en étant guidé par les indicateurs lexicaux ou sémantiques proposés par le chercheur ou le chargé d'études, mais il peut aussi suivre sa propre curiosité en découvrant les verbatim et en explorant d'autres aspects du corpus, ce qui apporte une meilleure compréhension et répond à l'intérêt pour de description qui distingue la recherche qualitative pure (Moscarola, 2018). Ainsi, l'usage de l'interactivité permet de vivre une véritable expérience de lecture et de navigation (cliquer, filtrer et illustrer) et de raconter une histoire en construisant des scénarii de lecture (avec une intrigue, des acteurs et des décors).

4. Sphinx iQ3

4.1. Présentation du logiciel

Lancé le 15 novembre 2021, Sphinx iQ3 est, à la fois, un logiciel de conception d'enquêtes, de collecte multicanal, d'analyses statistiques, de datavisualisation et de partage de données. Développé sur la plateforme OS Windows, il permet l'analyse des données quantitatives collectées à travers des enquêtes et sondages et l'analyse qualitative de corpus textuels collectés ou importés.

Figure 1 : Les menus du logiciel

Sphinx iQ3 s'installe en monoposte ou en réseau[1]. Disponible en cinq langues (Français, Anglais, Allemand, Espagnol et Portugais), il est organisé en quatre grands menus : Conception du questionnaire, Diffusion et Collecte, Gestion de données et Analyse des résultats (Figure 1). La conception du questionnaire revient à définir et à paramétrer l'outil de collecte des données (questionnaire, guide d'entretien, grille de saisie d'observations...). Elle se déroule en trois grandes étapes : la rédaction du questionnaire, le paramétrage des scénarios de déroulement et la définition des options de présentation et de mise en forme des formulaires. Le questionnaire peut être diffusé sur plusieurs supports de collecte (papier, en ligne, SMS, téléphone, site web, bornes, réseaux sociaux…) avec un ensemble de fonctionnalités de suivi et de relance. Une fois les données collectées ou importées, la gestion de données permet, entre autres, de les qualifier et nettoyer, de les redresser, de les échantillonner ou encore de fusionner des données et calculer de nouvelles variables. Le 4ème menu regroupe les analyses quantitatives et qualitatives. Les analyses statistiques sont multiples : analyses à plat, analyses croisées (avec tests de significativité) ou encore analyses multivariées et avancées (analyses factorielles, régression linéaire ou logistique, analyses de classification, matrice importance/performance ou de Llosa, équations structurelles PLS…). Pour les corpus textuels, le menu « Analyse des résultats » propose un ensemble d'analyses textuelles et sémantiques[2].

Sphinx iQ3 est une solution hybride, dont l'environnement d'analyse est accessible en local pour le chargé d'étude ou en ligne pour les différents destinataires de l'étude.

[1] Configuration minimale requise pour l'installation de Sphinx iQ 3 : Windows 10 en 64 bits à jour avec .Net. Framework version 4.7.2 fonctionnel. Processeur à 2.5 Ghz . Mémoire vive RAM : 8 Go. Espace libre disque dur : 3 Go. Carte graphique : DirectX 9c compatible 3D avec 256 MB de mémoire dédiée. Résolution écran : 1366x768 pixels.
[2] Un didacticiel détaillant l'ensemble des fonctionnalités est disponible en ligne : https://infos.lesphinx.eu/DOC/Sphinx_iQ3/fr/Home.htm

4.2. L'ADT avec Sphinx iQ3

Pour l'analyse des données textuelles, Sphinx iQ3 mélange les approches et les ressources d'analyse (lexicales, sémantiques et statistiques) permettant d'analyser des corpus très volumineux de diverses origines. Pour ce faire, il intègre comme composant les moteurs d'analyse sémantique « Synapse », société spécialisée en ingénierie linguistique. Ces moteurs donnent accès à un dictionnaire morphosyntaxique d'environ 158 000 lemmes, un thésaurus à 4 niveaux de 3781 feuilles documentées par autant d'ontologies et un module d'analyse des sentiments. Le thésaurus est construit à partir du thésaurus Larousse.

Les corpus qui peuvent être automatiquement importés sont des fichiers textes (.txt, .doc), des fichiers structurés (.xls, .csv, .mdb, etc.), des données Web ou encore des saisies manuelles. Le logiciel assiste l'utilisateur dans cette phase d'importation des données en lui demandant de spécifier le contexte de collecte de données (entretiens dirigés ou non-dirigés, focus groupes, articles ou document) et lui donnant les consignes de préparation du corpus et de conversion du texte en base de données Sphinx (variables de contexte ou signatures, observations, possibilités de découpage…). Le logiciel n'impose aucune limitation de volumétrie et de taille de corpus, la seule limitation est le temps de traitement et d'analyse.

Pour l'ADT, Sphinx iQ3 restitue les textes sous plusieurs formes : Verbatim, mots, regroupement de réponses, orientations, concepts, expressions, classification thématique…Les différentes analyses sont regroupées sous quatre catégories : Exploration lexicale, Analyse sémantique, Spécificités par contexte, Analyse de contenu (Figure 2).

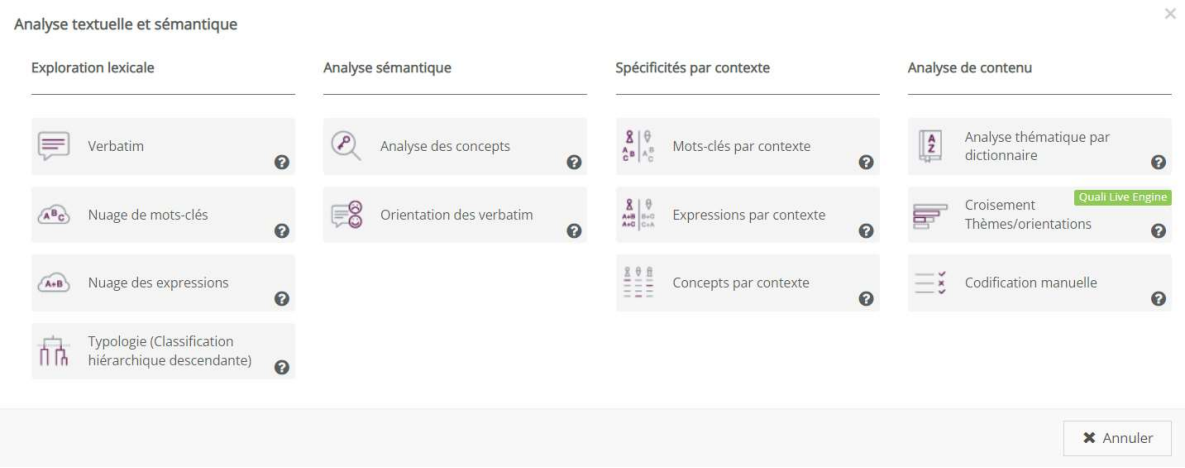


Figure 2 : Les différentes analyses textuelles et sémantiques

4.2.1. L'exploration lexicale

Dans une perspective de « fouille de texte », le logiciel permet de prendre connaissance du corpus à partir des mots et expressions qu'il contient :

- Identifier les principaux mots et expressions, sous leur forme lemmatisée (nombre d'occurrences, nombre d'observations) ;
- Différencier les mots selon leur statut grammatical (noms, verbes, adjectifs…) ;
- Naviguer dans le corpus par entrée lexicale et de rechercher les verbatim ;
- Marquer les mots par couleur ou classer les citations par contexte ou signature (genre, CSP…) ;
- Découper les observations en phrases ou paragraphes ;
- Etablir une typologie des unités de significations par classification hiérarchique ascendante ;
- Regrouper les mots et construire des dictionnaires ad hoc.

L'utilisateur peut ainsi se faire une première idée du texte et la documenter à partir de verbatim judicieusement choisis et filtrés (Figure 3).

Figure 3 : Exemples de restitution pour l'exploration lexicale

4.2.2. L'analyse sémantique

L'analyse lexicale peut être complétée par une analyse sémantique qui consiste à déterminer les concepts auxquels les unités de significations renvoient en application du thésaurus générique intégré au logiciel, ou d'un thésaurus ad'hoc construit par l'utilisateur. Un thésaurus définit un ensemble de significations, idées concepts organisés suivant une nomenclature arborescente qui va du général au particulier. Chaque concept est décrit par la liste des mots qui le définissent (ontologie, dictionnaire).

Il est alors possible de :

- Identifier les thématiques présentes dans le texte et les principaux concepts avec un niveau de détail choisi (seuil de sévérité) par l'utilisateur ;
- Illustrer les concepts par les verbatim correspondants ;
- Adapter les terminologies ;
- Créer des variables fermées sur les concepts.

4.2.3. L'analyse de l'orientation des opinions

L'analyse de l'orientation (ou de sentiments) est une technique très utile pour automatiser la synthèse de multiples commentaires afin d'obtenir efficacement une vue d'ensemble des

opinions sur un sujet donné. En effet, les sources de données textuelles porteuses d'opinion disponibles sur le Web se multiplient : avis *Google*, avis *TripAdvisor*, forums, réseaux sociaux…

L'analyseur de sentiment détermine dans l'ensemble du corpus les opinions exprimant un sentiment, un jugement ou une évaluation. Il précise la tonalité du texte en situant la nature et l'intensité des opinions émises par rapport à un répertoire de sentiments. Le moteur renvoie, pour chaque texte analysé, aux éléments du thésaurus qui lui correspondent.

Il est alors possible de :

- Identifier l'orientation du corpus selon les opinions ;
- Marquer les opinions positives, partagées, négatives ou neutre avec des couleurs ;
- Repérer les passages des fragments de corpus exprimant une opinion grâce aux marqueurs d'expressions subjectives ;
- Déterminer la valence ou l'orientation de l'opinion grâce aux champs lexicaux des opinions positives et négatives ;
- Dégager la synthèse de l'orientation globale du fragment analysé (algorithme d'agrégation rhétorique ou majoritaire).

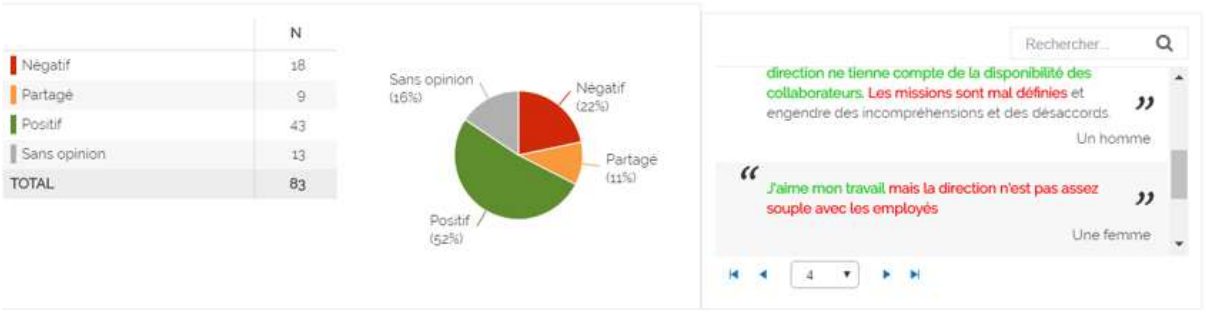


Figure 4 : Exemples de restitution pour l'analyse de l'orientation

4.2.4. L'analyse de contenu

Avec la confluence du TAL et de l'ADT, Sphinx iQ3 permet de :

- Analyser le corpus par dictionnaire pour une analyse automatique, et au fil de l'eau (le logiciel intègre plusieurs dictionnaires thématiques : expérience client, climat social, évaluation enseignement…). Ces dictionnaires peuvent être enrichis selon le contexte ;
- Créer un Code book mono ou multi grille pour une analyse manuelle ;
- Croiser les thèmes et les orientations ;
- Contrôler le défilement du corpus par taille ou contenu ;
- Recueillir et marquer des extraits significatifs ;
- Choisir le niveau du thésaurus pour mener l'analyse ;
- Vérifier la stabilité de la codification par vision des éléments déjà codés ;
- Réviser la grille et de produire des résultats.

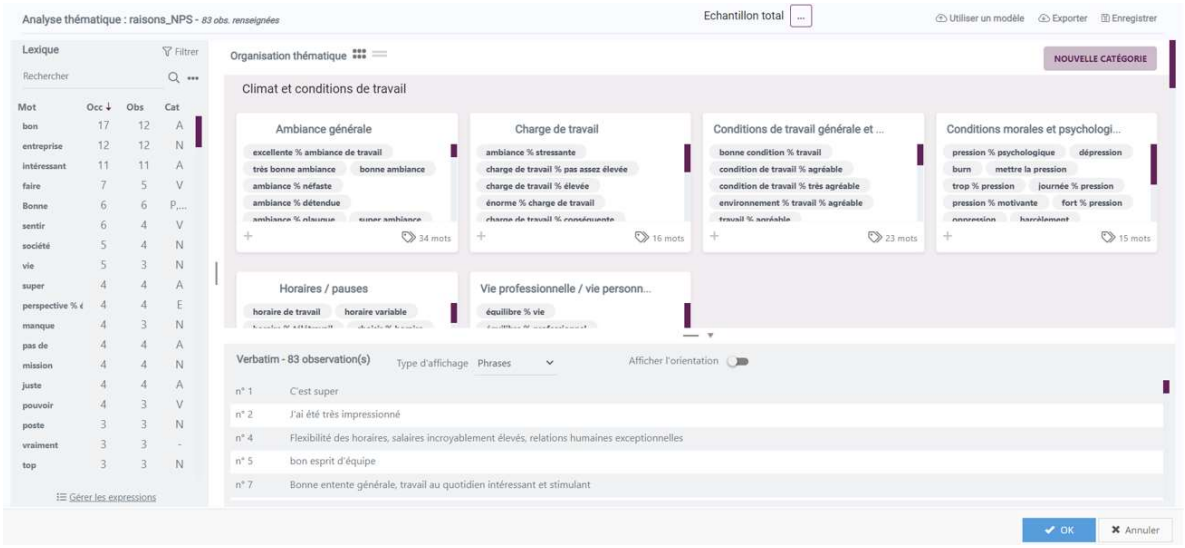


Figure 5 : Exemple d'analyse de contenu par dictionnaire

4.2.5. Le calcul des spécificités

Les calculs de spécificités (Lebart et Salem, 1994) consistent à répondre aux questions suivantes : Qu'est-ce qui différencie les contenus provenant de tel contexte, de telle catégorie de locuteurs, ou de contenus ? Il s'agit de caractériser les observations correspondantes à un sous ensemble d'observations : classes de la classification thématique, contextes, orientations des réponses. Pour cela, la procédure utilise un test qui met en évidence les éléments lexicaux et/ou sémantiques sur-représentés dans des sous-ensembles. Les algorithmes de spécificité sont fondés sur des tests statistiques (rapport de fréquence ou comparaison de fréquence) et permettent de trouver automatiquement les mots, concepts, ou phrases les plus révélateurs. Les résultats de ces calculs déterminent alors les éléments affichés dans les nuages de mots ou dans les tableaux de caractéristiques, et servent à identifier les influences du contexte, à interpréter les classes thématiques, à contrôler la codification manuelle et à sélectionner les verbatim spécifiques ou les plus pertinents.

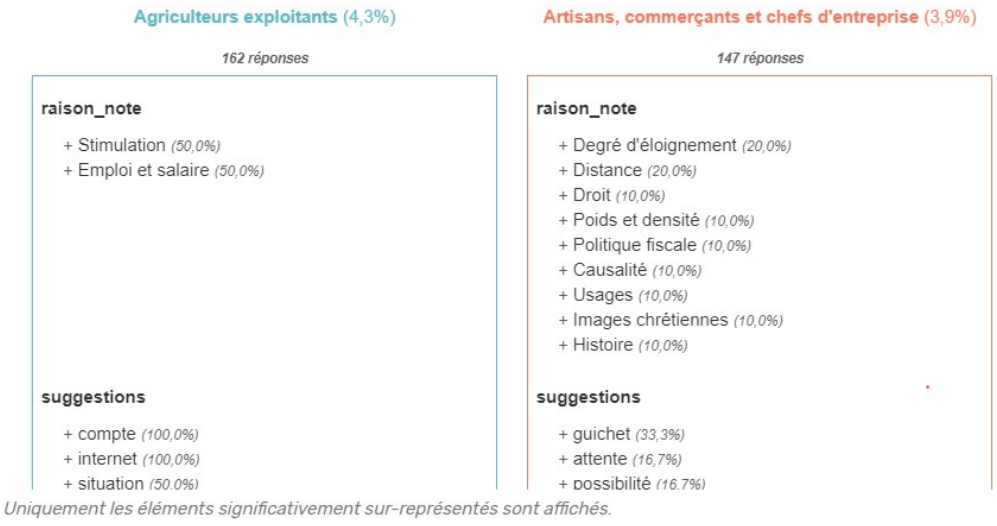


Figure 6 : Exemple d'une présentation des spécificités par contexte

4.2.6. Les analyses statistiques

Pour restituer les résultats de l'ADT, notamment à la suite de la création de variables fermées sur les lexiques, les concepts clés ou les classes, il est possible de mettre en place toutes les fonctions d'analyse statistique disponibles dans le logiciel.

5. Illustrations et conclusion

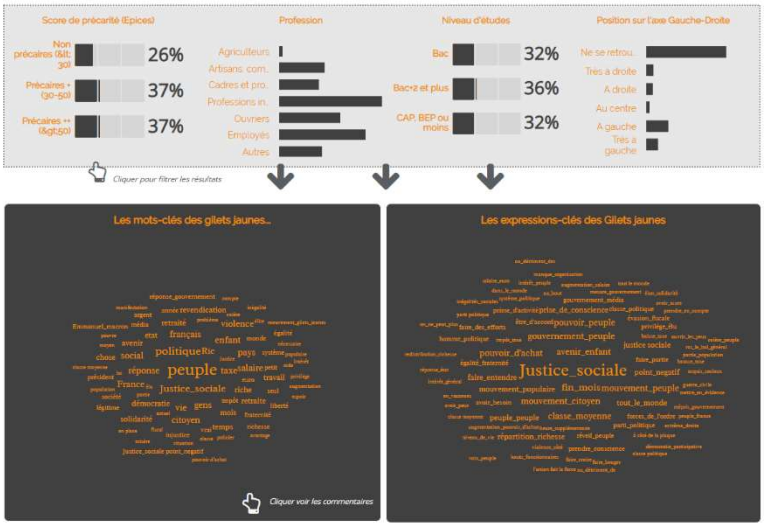
La présentation orale permettra de voir en détails, à travers plusieurs exemples, les usages de l'ADT avec Sphinx IQ3. Pour les lecteurs, les liens suivants pointent sur des exemples de rapports en ligne illustrant les restitutions graphiques et interactives construites avec le logiciel.

Exemple 1 : La fouille et l'analyse d'avis en ligne. Quatre étapes comme alternatives à l'enquête de satisfaction.

https://app.dataviv.net/reporting/report/662030e0-49c0-48d0-75bd-08d9058f5303

Exemple 2 : Questionnaire en ligne, recherche universitaire effectuée à partir de 300 groupes Facebook au début du mouvement des gilets jaunes. La classification automatique et l'analyse thématique mettent en évidence les caractéristiques de ce mouvement.

https://app.dataviv.net/reporting/report/3fd1178b-904c-49bb-421b-08d7681a6cc4



Malgré la richesse de ses fonctionnalités, Sphinx iQ3 présente certaines limites. L'« intelligence » de l'outil ne permet pas de remplacer celle d'une lecture critique et attentive. En effet, la reconnaissance des significations n'est pas sans faille. Elle dépend de la rectitude orthographique et syntaxique du texte. Elle est plus délicate pour les niveaux bas du thésaurus (concepts détaillés ou très détaillés). Ainsi, plus les textes sont courts, plus pauvre est l'analyse automatique, ce qui peut conduire à des méprises évidentes pour le lecteur informé, par ailleurs, de la nature des textes analysés. Pour l'analyse de contenu, la procédure du double codage n'est pas disponible.

Mais ces limites peuvent être compensées par la mise en œuvre de stratégies de recherche adéquates : construction de thésaurus ad hoc, triangulation par exemple. On l'aura compris, même si le logiciel apporte efficacité et rigueur, il ne saurait remplacer l'attention critique du chargé d'étude ou du chercheur.

6. Bibliographie

Bô D. (2022). Big Quali : La puissance des études qualitatives à l'ère du Big Data. Dunod.

Boughzala Y., Moscarola J. and Hervé M. (2014). Sphinx Quali : un nouvel outil d'analyses textuelles et sémantiques., *12ème Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014)*, Paris.

Lebart L. and Salem A. (1994), Statistiques textuelles. Dunod, Paris.

Molina-Azorin, J.F., (2011), "The use and added Value of mixed Methods in Management research", Journal of Mixed Methods Research, 5, 7-24.

Mothe C., Delfosse E. and Bocquet A M. (2021). L'analyse de données textuelles assistée par ordinateur., *Revue Française de Gestion*, 2021/2 (n°295) – pp11-37.

Moscarola J. (2018). Faire parler les données. Edition EMS, Coll. Business Science Institute.

Moscarola J. (2022). Let Your Data Speak. Edition EMS, Coll. Business Science Institute.

Tashakkori, A., Teddlie, C., (2003), "Handbook of mixed methods in social and behavioral research", Thousand Oaks, CA: Sage.

Tukey J. W. (1977). Exploratory Data Analysis. Addison Wesley, Reading, Mass.

Walsh I. (2015). Découvrir de nouvelles theories. Collection Business Science Institute, Editions EMS.