Sphinx Quali : un nouvel outil d'analyses textuelles et sémantiques

Younès Boughzala[1], Jean Moscarola[2], Mathilde Hervé[3]

[1] Le Sphinx/Université de Savoie – yboughzala@lesphinx.eu
[2] Université de Savoie – jean.moscarola@univ-savoie.fr
[3] Sphinx Institute – mherve@lesphinx.eu

Abstract

This paper focuses on "Sphinx Quali" presentation, a new Software specialized in Textual Data Analysis (TDA), which combines approaches and resources analysis. It is a complete set of tools used to analyze very large corpus of various origins (opened questions, scientific or press articles, historical writings, free or semi-structured interviews, websites, forums, social networking pages...) and to assemble automated synthesis, content analysis and text mining.
Referring to three different streams, which are Computer Assisted/Aided Qualitative Data AnalysiS (CAQDAS), tools for language automatic processing and web search engines, Sphinx Quali integrates three types of approaches, increasingly complementary: lexical, semantic and statistical methods.
This paper highlights the main innovations: semantic engines (thesauri, ontologies, and sentiment analysis), lexical and statistical engines (descending hierarchical classification, verbatim selection), learning and automatic extension of the content coding. Examples will be used to introduce the main innovations and thus to appreciate the scope and limitations of this new software.

Résumé

L'objectif de cette communication est de présenter un nouveau logiciel d'Analyses des Données Textuelles (ADT) : « Sphinx Quali ». Le dernier né des logiciels d'analyse des données textuelles, il se veut un outil qui mélange les approches et les ressources d'analyse. C'est un ensemble complet d'outils permettant d'analyser des corpus très volumineux de diverses origines (questions ouvertes, articles scientifiques ou de presse, écrits historiques, entretiens libres ou semi-directifs, sites Web, forums, pages réseaux sociaux…), et de combiner des synthèses automatiques, des analyses de contenu et des fouilles de texte. En effet, en se référant à trois courants différents, à savoir les CAQDAS, les outils de traitement automatique des langues et les moteurs de recherche Web, cet outil intègre trois types d'approches de plus en plus complémentaires : lexicales, sémantiques et statistiques.
La communication met en évidence les principales innovations : l'utilisation des moteurs sémantiques (thésaurus, ontologies, analyse des sentiments), les moteurs lexicaux et statistiques (classification hiérarchique descendante, sélection de verbatim) et l'apprentissage et l'extension automatique de codification de contenu. La présentation des principales innovations s'appuiera sur des exemples permettant d'apprécier la portée et les limites de ce nouveau logiciel.

Mots-clés : Analyses des Données Textuelles (ADT), logiciel pour l'analyse textuelle, Sphinx Quali, analyse lexicale, analyse sémantique

1. Introduction

L'Analyse des Données Textuelles (ADT), appelée aussi l'Analyse Qualitative des Données (AQD), est l'ensemble des approches, méthodes et outils informatiques qui visent à découvrir l'information contenue dans un corpus textuelle. Son objectif est de qualifier les éléments essentiels d'un corpus à l'aide de catégories lexicales et/ou sémantiques et à les quantifier en analysant la répartition statistique des éléments de ce corpus. L'usage de l'ADT est très ancien et varié en sciences humaines et sociales (Lebart et Salem, 1994). Avec le

développement du Web 2.0, des réseaux sociaux et du Big Data, la grande quantité de textes disponibles sur le Web a rendu très laborieux l'usage des approches traditionnelles de l'ADT. Ainsi, en capitalisant sur les acquis de ces approches, le recours à des logiciels et outils linguistiques et informatiques répond à des nouvelles exigences en termes de volume, de complexité, de sources, de moyens humains et financiers, de temps, etc. De ce fait, les éditeurs de logiciels multiplient leurs efforts en termes de Recherche et Développement (R&D) afin de proposer des outils de plus en plus performants et qui répondent aux nouveaux besoins des chercheurs, entreprises et cabinets d'études.

L'objectif de cette communication est de présenter un nouveau logiciel d'ADT : « Sphinx Quali ». Lancé le 17 octobre 2013 à Paris, c'est un outil qui se veut complet et qui mélange les approches et les ressources d'analyse, permettant d'analyser de manière rapide des corpus très volumineux et de diverses origines, et de combiner des synthèses automatiques, des analyses de contenu, des fouilles de texte et des analyses statistiques. La première partie de cette communication rappelle les différentes approches de l'ADT, les cas d'usages et notamment les nouveaux enjeux. La seconde partie présente l'organisation du logiciel Sphinx Quali. Elle met en évidence les principales innovations : l'utilisation des moteurs sémantiques (thésaurus, ontologies, analyse des sentiments), les moteurs lexicaux et statistiques (classification hiérarchique descendante, sélection de verbatim) et l'apprentissage et l'extension automatique de codification de contenu. Enfin, pour apprécier les apports et les limites du logiciel, la troisième partie est consacrée à la présentation d'un exemple de corpus analysé avec le Sphinx Quali.

2. L'ADT : tradition et nouveaux enjeux

La notion d'études qualitatives ou d'ADT recouvre des pratiques très variées dans le monde académique et dans le monde professionnel. Les corpus peuvent être des documents de différentes natures. Des documents disponibles (littérature, compte-rendu, correspondances, articles scientifiques ou de presse, sites Web, forums, pages réseaux sociaux, etc.) ou des documents produits par le chercheur ou le chargé d'études pour les besoins de son étude (entretiens, focus-group, observations du terrain, réponses à des questions ouvertes, etc.). Traditionnellement, l'ADT consiste à prendre connaissance d'un corpus textuel, généralement un discours, le lire et le fouiller pour en sortir les mots clés, de classer les fragments spécifiques, ou encore de les coder manuellement sur la base d'une grille d'analyse pour dénombrer les principaux thèmes. Avec l'apparition des CAQDAS (*Computer-Aided Qualitative Data Analysis Software*), nous avons alors assisté à une extension méthodologique de l'approche qualitative traditionnelle et plusieurs analyses plus au moins automatiques ont vu le jour grâce à des outils quantitatifs (Moscarola, 2001 ; Jenny, 1997).

Pour analyser un corpus textuel, trois approches se distinguent. La première appartient à la tradition littéraire, il s'agit de construire un nouveau texte pour rendre compte des textes analysés. C'est la production d'une synthèse, d'un résumé ou d'un commentaire critique dont le but est de défendre ou de contredire un point de vue. Dans tous les cas, il s'agît d'articuler une pensée autour d'idées ou de concepts illustrés par des citations judicieusement 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 la démarche critique du chercheur. Il s'agit alors d'expliciter les méthodes et d'exposer les modalités de prise de connaissance. C'est la démarche de l'analyse thématique ou de contenu. Elle consiste à situer le texte par rapport à une grille de lecture (*Code Book*) explicitement construite par le

chercheur, et utilisée pour coder la présence des différentes catégories de contenus (thématiques) et les dénombrer. La troisième approche est apparue avec le traitement informatique des textes. Lexicale (compter les mots), puis sémantique (identifier automatiquement les contenus : concepts), qui bénéficie des progrès des outils de l'ingénierie linguistique. Pour les tenants de l'intelligence artificielle, elle pourrait même complètement remplacer le lecteur. Pour les chercheurs et les chargés d'études, elle amplifie la capacité de prise de connaissance du corpus en produisant des substituts du texte qui révèlent ses structures lexicales et sémantiques. Ces trois approches ne doivent plus être considérées comme alternatives mais plutôt complémentaires (Moscarola, 2013). Cela dit, pour analyser un corpus textuel, il est possible de recourir à plusieurs types d'analyses. Nous pouvons citer les analyses lexicales (lemmatisation, calcul des occurrences des mots, proximité des mots, dictionnaires, classes, associations, etc.), les analyses linguistiques (progression thématique, analyse des marqueurs de forme et les connecteurs dans le discours), les analyses thématiques ou de contenu (fréquences des thèmes de la grille d'analyse ou modèle par l'affection des fragments à catégories thématiques), les analyses sémantiques (fréquences des concepts et ontologies), ou encore les analyses cognitives (chaînages cognitifs, niveau d'abstraction des concepts et types de liens). Selon (Fallery et Rodhain, 2007), plusieurs facteurs déterminent le choix du type d'analyses à utiliser, notamment le cadre méthodologique (exploratoire ou test de modèle d'hypothèses), l'implication du chercheur, l'axe temporel (analyse instantanée ou longitudinale), l'objet de l'analyse (un groupe ou un individu), la taille et la lisibilité du corpus (qualité), l'homogénéité du corpus (discours d'une seule personne ou d'un groupe), la structuration du langage et le moment de l'analyse statistique (ex-ante ou ex-post).

En parallèle, nous assistons, depuis plusieurs années, de manière exponentielle, à la combinaison des approches qualitatives avec les approches quantitatives, ce qu'on appelle les méthodes mixtes de recherche (*Mixed Methods Research*). Ces dernières sont définies comme «*l'ensemble des procédures de collecte et d'analyse de données quantitatives et qualitatives menées dans le cadre d'une même étude*» (Tashakkori et Teddlie, 2003). Les méthodes de recherche mixtes sont nombreuses (nous pouvons citer par exemple la méthode du mur d'images), elles offrent une multitude d'opportunités et avantages pour la collecte et l'analyse des données (Boughzala et Moscarola, 2013). En fait, selon ses adeptes, recourir à une approche «multi-facettes» permet de contrebalancer les faiblesses d'une approche ou technique par les forces d'une autre et produisent des résultats plus pertinents (Molina-Azorin, 2011). Selon (Onwuegbuzi et Teddlie, 2003), mixer les méthodes de recherche et d'analyse s'explique pour au moins deux raisons, à savoir la représentation et la légitimation. Pour ce faire, il est nécessaire alors d'utiliser un ou plusieurs logiciels pour mettre en places des analyses qualitatives et des analyses quantitatives. De ce fait, les éditeurs de logiciels enrichissent de manière continuelle leurs outils en essayant de permettre de coupler, dans un même outil, les ADT avec les analyses quantitatives « conventionnelles ». L'affluence de nouveaux logiciels et outils d'ADT est la résultante de la forte demande du monde professionnel et du monde académique. Sur le Web, les entreprises peuvent désormais collecter de manière très facile et rapide une quantité gigantesque de textes (commentaires sur les pages des réseaux sociaux, tweets, messages sur les forums, e-mails, dépêches, articles, rapports, enquêtes en ligne, etc.). Ces données peuvent être exploitées pour la veille stratégique, la mesure de la notoriété ou encore de l'e-notoriété, la capitalisation des connaissances, la prospection commerciale, l'assistance technique, le service après-vente, etc. Les chercheurs quant à eux, ont de plus en plus besoin d'une autre alternative, soit à l'analyse thématique jugée trop subjective, soit à des simples analyses par mots clés jugées trop pauvres (Bournois et al., 2002). Ces derniers ont étendu « *les méthodologies qualitatives « assistées »*

par des outils quantitatifs (SpadT, Sphinx-Lexica, Alceste, Tropes, Decision Explorer, NVivo... parmi les plus cités en France » (Fallery et Rodhain, 2007). En plus de ces logiciels, nous assistons au développement des moteurs et outils du Web sémantique et du Traitement Automatique des Langues (TAL). L'objectif ultime est, à travers l'application de programmes et techniques informatiques, « *d'organiser automatiquement les contenus, d'extraire de l'information à partir d'un magma hétérogène de textes peu structurés»* (Fallery et Rodhain, 2007). Ceci est de plus en plus possible avec le progrès des ressources technologiques et c'est dans ce mouvement, qu'après Sphinx-Lexica, la société « Le Sphinx Développement » a développé un nouveau logiciel : Sphinx Quali.

3. Sphinx Quali

3.1. Présentation du logiciel

Sphinx Quali se veut un outil qui répond à tous les usages. Il mélange les approches et les ressources d'analyse permettant d'analyser des corpus très volumineux de diverses origines et de combiner des synthèses automatiques, des analyses de contenu et des fouilles de texte. Il intègre les avancées récentes de l'ingénierie de la connaissance (ontologies, réseaux sémantiques…). En effet, en se référant à trois courants différents, à savoir les CAQDAS, les outils de Traitement Automatique des Langues (TAL) et les moteurs de recherche Web, cet outil exploite trois types d'approches de plus en plus complémentaires : lexicales, sémantiques et statistiques. Pour ce faire, ce logiciel a intégré comme composant le moteur d'analyse sémantique « Synapse », société spécialisée en ingénierie linguistique. Ce moteur se base sur un dictionnaire morphosyntaxique de 158 000 lemmes, un thésaurus à 4 niveaux de 3781 feuilles documentées par autant d'ontologies. Le thésaurus est construit à partir du thésaurus Larousse (Larousse, 1994).

Figure 1. Page d'accueil de Sphinx Quali

Développé sur la plateforme OS Windows, il est caractérisé par une ergonomie intuitive et proche des outils MS Office. Pour analyser un corpus, Sphinx Quali est organisé en trois grands menus en fonction de l'objectif, du contexte et du corpus de l'utilisateur (Figue 2). Ce dernier peut aborder son corpus selon sa convenance, aucune obligation de commencer par tel ou tel menu (Figure 3). Pour les utilisateurs qui souhaitent compléter les ADT par des analyses statistiques. Le menu « Reporting » permet d'accéder à l'environnement des tableaux de bord et de mettre en place une multitude d'analyses quantitatives : analyses à plat, analyses croisés avec tests de significativité analyses factorielles, etc.

Figue 2. Les trois menus d'analyse

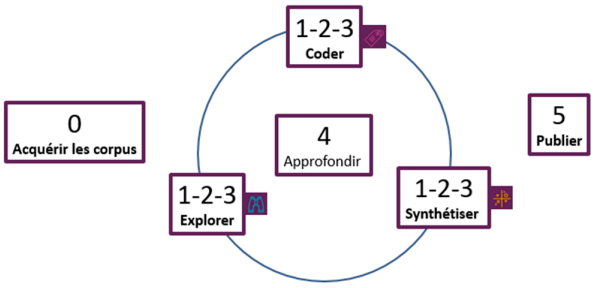


Figure 3. Les différents scénarii selon la problématique, le corpus et les capacités de l'utilisateur

Les corpus qui peuvent être automatiquement importés sont des fichiers textes, des fichiers structurés (.xls, .csv, .mdb, etc.), des données Web et réseaux sociaux (recherche et collecte de données depuis des moteurs de recherche (Google, Bing) ou réseaux sociaux (Twitter, Google+)) 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 (entretien semi-directif, focus-group…) 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…). En effet, à travers le menu « Corpus », il est possible de préparer le corpus, de l'organiser, d'identifier les variables selon le volume de leurs contenus, de déterminer l'influence du contexte sur la taille du corpus (indicateur d'intérêt, de motivation ou d'engagement), etc. Le logiciel n'impose aucune limitation de volumétrie et de taille de corpus, la seule limitation est le temps de traitement et d'analyse[1]. Le Menu « Synthèse » permet d'obtenir automatiquement des indicateurs lexicaux et sémantiques et produire des synthèses (cartes de mots et concepts clés, tableaux synthétiques, classes thématiques, caractérisation des opinions, spécificités lexicales par contexte…). Le Menu « Codification » permet de définir une grille d'analyse, de coder le texte manuellement et de marquer les extraits spécifiques. Il offre aussi, par apprentissage, une extension

[1] Le logiciel n'impose aucune limite de volume, mais pour les très gros corpus, les temps de calcul pour la lemmatisation et les analyses sémantiques peuvent être longs. En revanche, les calculs lexicaux et statistiques sont instantanés. Par exemple, les synthèses du corpus des présidentielles exemples ici-bas (114000 mots, 1194 tours de parole) sont produites avec un délai de 58 secondes. Il faut 2 minutes et 29 secondes pour obtenir les synthèses sémantiques d'un corpus de 730 articles de presse (1275 pages, 613 472 mots). Enfin, le corpus des 106 985 mots formant les 3136 réponses à la question ouverte d'une enquête est analysé en 24 secondes. Ces temps sont établis avec un ordinateur portable de puissance moyenne (Windows 7, 2.5GHz, 4 Go de RAM).

automatique de l'analyse de contenu. Le Menu « Exploration » quant à lui, permet d'explorer les textes et les lexiques du corpus (listes des mots lemmatisés, des concepts, des verbes...).

3.2. Les moteurs du logiciel et les analyses possibles

Dans ce qui suit nous présentons les principales analyses proposées par Sphinx Quali en présentant les moteurs et procédures intégrés dans le logiciel.

3.2.1. L'analyse lexicale

Dans une perspective de « fouille de texte », Sphinx Quali permet de prendre connaissance du corpus à partir des mots qu'il contient à travers une analyse lexicale avec lemmatisation et correction orthographique. L'utilisateur peut ainsi se faire très rapidement une idée du texte et la documenter à partir de verbatim judicieusement choisis. Il est possible alors de :

- Identifier les principaux mots sous leur forme lemmatisée (nombre d'occurrences, nombre d'observations) et les mots composés,
- 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,
- Regrouper les mots et construire des dictionnaires ad' hoc.

3.2.2. L'analyse sémantique

L'analyse lexicale peut être complétée par une analyse sémantique qui consiste à déterminer les concepts auxquels les mots explorés renvoient. Elle permet de dépasser l'écueil en définissant les conditions nécessaires pour passer du lexique au sens (idée, concept...). Elle fait appel aux notions de thésaurus, d'ontologie ou dictionnaire et de réseau sémantique (Figure 4). Un thésaurus définit un ensemble de significations, idées concepts et les organise suivant une nomenclature arborescente qui va du général au particulier. Une ontologie ou un dictionnaire est un ensemble d'éléments qui définissent une notion. Il s'agît d'un ensemble de mots (sous leur forme lemmatisée) qui renvoie à une signification et donc une feuille du thésaurus. Un réseau sémantique est un ensemble de relations entre éléments signifiants (mot d'un corpus ou significations d'un thésaurus) conduisant à préciser le sens de ces éléments en fonctions des éléments auxquels ils se trouvent reliés.

Ainsi, avec Sphinx Quali, il est possible principalement 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.

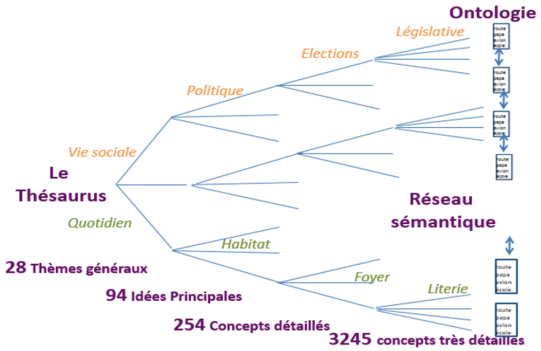


Figure 4. Thésaurus, ontologie et réseau sémantique

3.2.3. L'analyse automatique des sentiments

Le traitement automatique pour l'analyse des sentiments permet d'automatiser la synthèse des multiples avis pour 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 d'internautes, forums, réseaux sociaux…Une expression d'opinion possède une polarité, qui peut être soit positive, soit négative, soit neutre. La valeur neutre correspond à une opinion de polarité ambiguë, qui sera éventuellement désambiguïsée par le contexte. 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, les éléments du thésaurus qui lui correspondent. Il est alors possible de :

- Identifier l'orientation du corpus et les opinions positives et négatives et les sentiments exprimés,
- Marquer les opinions positives ou négatives 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).

3.2.4. L'analyse de contenu

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

- Créer un Code book mono ou multi grille,
- 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,
- Etendre la codification manuelle par une codification automatique et assistée grâce à un apprentissage sur la base des premiers éléments codés et la reconnaissance de contenue (similarité, affectation sur la base d'une recherche des plus proches voisins). Ce procédé est efficace lorsque les textes codés sont courts et mono-focalisés. Si ces conditions ne sont pas remplies, il peut être nécessaire de redécouper le corpus par phrases.
- Réviser la grille et de produire des résultats.

3.2.5. La classification hiérarchique descendante

Inspirée de la méthodologie ALCESTE (Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte) (Reinert, 1990 ; Benzecri, 2007), cette procédure consiste à scinder le corpus en classes homogènes selon les mots et/ou les concepts qu'elles contiennent. Elle permet de révéler les structures thématiques du texte. Celles-ci sont révélées par l'affichage des nuages de mots spécifiques (significativement sur-représentés) de chaque classe. Elle procède par itérations successives à partir d'une analyse factorielle des correspondances multiples (Figue 5). La table des données comporte en colonne les lemmes et/ou concepts des textes décrits par les lignes du tableau. La première itération conduit à une partition selon le premier plan factoriel. La classe la plus nombreuse fait l'objet d'une nouvelle partition. Les itérations se poursuivent par partition de la classe de plus grand effectif, tant que celle-ci est supérieure à un pourcentage de la population totale fixé comme critère d'arrêt.

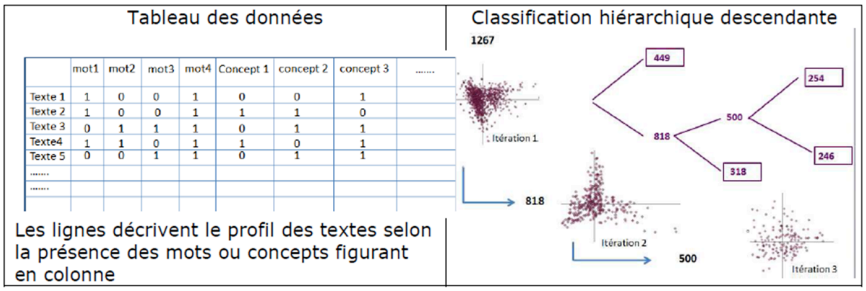


Figure 5. Classification hiérarchique descendante

Cette analyse est particulièrement utile lorsque l'analyse porte sur un grand nombre de textes de nature identique : réponses à une question ouvertes, contributions à des forums, d'interviews non directives, articles de journaux…

3.2.6. Les calculs de 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 correspondants 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.

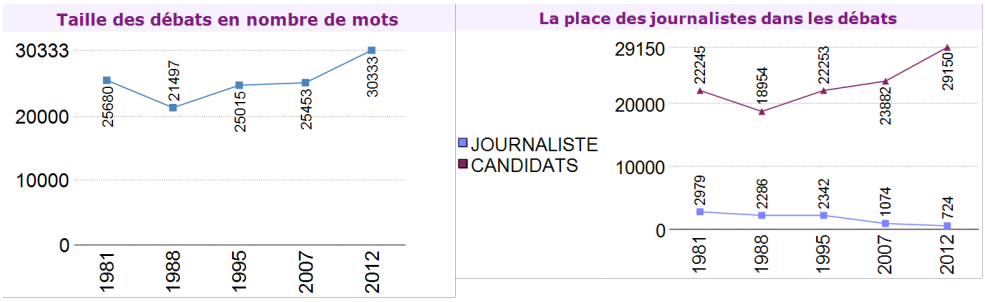
3.2.7. Les analyses statistiques

Pour restituer les résultats de l'ADT, notamment suite à la création de variables fermées sur les lexiques, les concepts clés ou les classes, il est possible de mettre en place plusieurs types d'analyses statistiques : analyses à plat, analyses croisées, analyses factorielles, etc.

4. Exemple

Pour illustrer certaines fonctionnalités de Sphinx Quali, nous présentons un exemple d'ADT effectué sur le corpus de cinq[2] débats de 2ème tour des élections présidentielles en France de 1981 à 2012. C'est un corpus de 127 978 mots, 300 pages.

- *Des débats de plus en plus longs et vifs :* Comme le montre la Figure 6, les débats sont de plus en plus longs et vifs. Les journalistes interviennent moins et les échanges sont plus nombreux et plus brefs. Le débat entre François Hollande et Nicolas Sarkozy est le plus long et le plus intense. Les journalistes interviennent peu, les échanges sont plus nombreux et plus rapides.



Candidats uniquement

Longeur moyenne des prises de parole et nombre d'intervention

	Longueur		
	Moyenne	Somme	Effectif
1981	**142,45**	**22650**	**159**
1988	**126,13**	**19171**	**152**
1995	**90,60**	**22651**	**250**
2007	**69,41**	**24364**	**351**
2012	**62,58**	**29598**	**473**
Total	**85,51**	**118434**	**1385**

Figure 6. Caractéristiques du corpus

- *Les mots clés des cinq débats :* Les mots montrent bien qu'il s'agît d'abord de la France et des Français, de président de la république et des problèmes de gouvernement, etc.

Figure 7. Nuage des mots clés

[2] Cinq débats pour six élections car en 2002 il n'y pas eu de débat pour le 2ème tour entre Chirac et Le Pen.

- *Les mots clés de chaque débat et de chaque candidat :* Il s'agit d'identifier les spécificités des débats (Figue 8) et des candidats (contexte) et notamment les mots communs (Figure 9) : « *premier* », mais surtout « *responsabilité* », « *européen* » et « *débat* ».

Figure 8. Les mots clés pour chaque débat, pour chaque candidat et les mots communs

- *La classification thématique :* La classification hiérarchique descendante effectuée sur l'ensemble des débats met en évidence cinq classes qui répartissent les interventions en cinq catégories (Figure 9). L'examen des termes spécifiques à chaque classe permet de distinguer ainsi cinq types de discours ou thématiques. Le thème de la Gestion, qui renvoie aux actions d'un gouvernement qui gère. Le thème de la Politique parle de : « *majorité* », « *vote* », « *gouvernement* », « *président de la république* ». Celui de la volonté et du Promouvoir : « *vouloir* », « *permettre* », « *donner* », « *projet* », « *responsabilité* ». Celui de la Décision : « *problème* », « *falloir* », « *risque* », « *autorité* »… Enfin, celui du Financier : « *milliard* », « *millions* », « *nombres* », « *payer* », « *euros* »...

Ces thèmes représentent des poids inégaux. En les enregistrant comme une variable fermée supplémentaire dans la base, il est possible ensuite de caractériser leur poids selon les périodes ou les candidats.

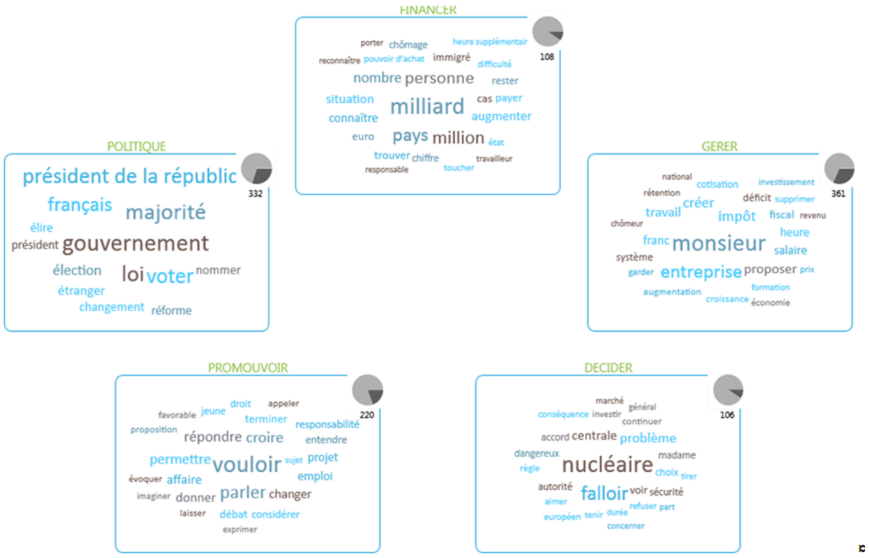


Figure 9. Classification des thèmes

- *Les références sémantiques des discours :* Les trois principaux champs sémantiques sont : Activités économiques, Vie sociale et Ordre et mesure. En bas de la liste, nous retrouvons : Droit, Volonté et Etre humain (Figure 10).

Nom	Effectifs	%
Activités économiques	393	28,4%
Vie sociale	318	22,9%
Ordre et mesure	229	16,5%
Action	162	11,7%
Fondamental	140	10,1%
Communication	139	10%
Hiérarchie	130	9,4%
Esprit	107	7,7%
Matière	89	6,4%
Droit	87	6,3%
Volonté	75	5,4%
Etre humain	66	4,8%
	Total observations : 1386	

Figure 10. Les références sémantiques des débats

- *Les références sémantiques selon la période et les candidats :* La Figure 11 met en évidence la place croissante au cours du temps du discours gestionnaire par rapport au discours politique. De même, l'opposition entre les débatteurs qui accordent de l'importance au thème du Politique, les Mitterrand et Giscard du début de période et ceux qui se positionnent plus sur le Décider, Gérer, Sarkozy et Royale. De manière très remarquable, Chirac et Hollande se situent au centre de graphique ce qui indique qu'ils ne privilégient aucun de ces thèmes, mais les développent tous de manière proportionnée.

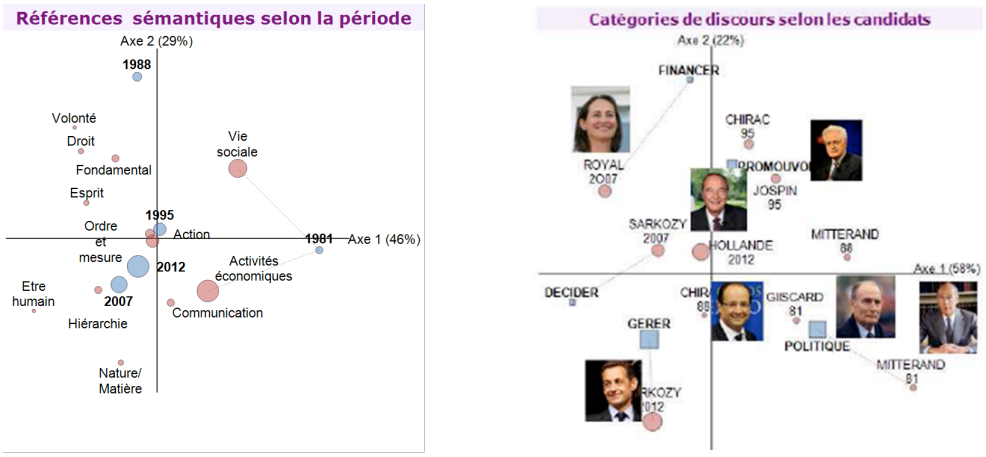


Figure 11. Les références sémantiques selon la période et les candidats

5. Conclusion et limites

L'objectif de cette communication est de présenter aux chercheurs et chargés d'études le nouveau logiciel d'ADT « Sphinx Quali ». Il est le fruit de plusieurs mois de R&D et de plusieurs années d'expertise de l'équipe de Sphinx Institute et de ses partenaires. La présentation orale permettra de voir avec plus de détails, à travers plusieurs autres exemples, les différentes analyses et fonctionnalités proposées dans cet outil. Sa première version présente encore certaines limites, notamment sa dépendance à la langue française, l'unique traitement des textes (et non pas des vidéos) ou encore l'absence de la procédure du double codage. Certes, c'est un outil intelligent mais l'intelligence artificielle ne permet pas encore de remplacer complètement le lecteur. En effet, la reconnaissance des significations n'est pas sans faille. Elle dépend bien sûr de la rectitude orthographique et syntaxique du texte. Elle est plus délicate pour les niveaux bas du thésaurus (concepts détaillés ou très détaillés). Enfin, plus les textes sont courts, plus pauvre est l'analyse des réseaux sémantiques, ce qui peut conduire à des méprises évidentes pour le lecteur informé, par ailleurs, de la nature des textes analysés.

Le logiciel Sphinx Quali est distribué en France par la société Le Sphinx Développement (www.lesphinx.eu). Une documentation détaillée (mode opératoire) est disponible en ligne sur : http://infos.lesphinx.eu/docquali/

Références

Benzecri J P. 2007. Linguistique et lexicologie. Dunod (réédition).

Boughzala. Y. et Moscarola J. (2013). Le mur d'images dans les enquêtes en ligne : comment stimuler pour observer et mesurer ?, *International Marketing Trends Conference*, 17-19 janvier 2013, Paris, France.

Bournois F., Point S. et Voynnet-Fourboul C. (2002). L'analyse de données qualitatives assistée par ordinateur : une évaluation, Revue française de Gestion, 137, janvier-mars 2002.

Fallery B. et Rodhain F. (2007). Quatre approches pour l'analyse des données textuelles : lexicale, linguistique, cognitive, thématique*, 16ème Conférence Internationale de Management Stratégique*, 6-9 juin, Montréal.

Ghiglione R., Landre A., Bromberg M. et Molette P. (1998). L'analyse automatique des contenus, Paris, Dunod, 1998.

Jenny J., (1997). Méthodes et pratiques formalisés d'analyse de contenu et de discours dans la recherche sociologique française contemporaine : état des lieux et essai de classification, Bulletin de méthodologie sociologique (BMS) N° 54.

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.

Moscarola J. (2001). Contributions des méthodes de l'analyse qualitative à la recherche en psychologie interculturelle : Sphinx et MCA, 8ème Congrès International de l'ARIC, Genève 2001.

Moscarola J. (2013). « Du Lexical au Sémantique : La nouvelle version de Sphinx pour les études qualitatives », Working Paper, Le Sphinx.

Onwuegbuzi A.J. et Teddlie C. (2003). "A framework for analyzing data in mixed methods research, in Handbook of mixed methods in social and behavioral research".

Péchoin D. (Sous la direction de) (1994). Thésaurus Larousse : Des idées aux mots, des mots aux idées, 2ème édition.

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, pp. 187-198.

Reinert M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval, Bulletin de méthodologie sociologique, n°26, pp. 24-54.

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