Analyser les corpus d'avis en ligne : Analyse lexicale exploratoire et/ou modélisation sémantique ?

Jean Moscarola[1], Younès Boughzala[2]

[1]Université Savoie Mont-Blanc – Annecy - France

[2]Le Sphinx/Université Savoie Mont-Blanc – Annecy - France

Abstract

The aim of this paper is to present and illustrate an example of Textual Data Analysis (TDA) from the Web by implementing two approaches: exploratory through lexical classifications and confirmatory through semantic modeling. For this, 600 online comments of French tourists evaluating accommodation in Algeria, Morocco and Tunisia are collected on the online reviews site TripAdvisor.com. The aim of this collection is to study the experiential perception of these three tourist destinations depending on type of accommodation chosen (hotel or guest house).

The corpus available on the Web (comments on specialized sites, social networks, etc.) have the advantage of being abundant, independent and spontaneous but unstructured. Their exploration is a new avenue of research in humanities and social science. To analyze them, traditional methods of TDA (lexical properties) should be enriched with semantic ad hoc approaches (thesaurus construction and application) and analysis of sentiment.

Based on this sample of corpus imported from the Web, we will present in a practical and accessible way the contributions and the richness of this combination of approaches for researchers in humanities and social sciences and professionals.

Résumé

L'objectif de ce papier est de présenter et illustrer un exemple d'Analyse de Données Textuelles (ADT) provenant du Web en mettant en place deux approches : exploratoire à travers les classifications lexicales, et confirmatoire à travers la modélisation sémantique. Pour ce faire, 600 commentaires en ligne de touristes français évaluant des hébergements en Algérie, au Maroc et en Tunisie sont collectés sur le site de commentaires en ligne TripAdvisor.com. L'objectif de cette collecte est d'étudier la perception expérientielle de ces 3 destinations touristiques selon le type d'hébergement choisi : hôtel ou maison d'hôtes.

Les corpus disponibles sur le Web (commentaires sur les sites spécialisés, les réseaux sociaux, etc.) présentent l'avantage d'être abondants, indépendants et spontanés mais non structurés. Leur exploration est une nouvelle voie de recherche en sciences humaines et sociales. Pour les analyser, les méthodes traditionnelles des ADT (propriétés lexicales) doivent être enrichies par des approches sémantiques ad hoc (construction et application de thésaurus) et des analyses de sentiments.

En se basant sur cet exemple de corpus importé depuis le Web, nous présenterons de manière pratique et accessible les apports et la richesse de cette combinaison d'approches pour les chercheurs en sciences humaines et sociales et les professionnels des études.

Mots-clés : Analyse de Données Textuelles (ADT), avis en ligne, analyse lexicale, modélisation sémantique

1. Introduction

La conséquence directe du développement du Web 2.0, des réseaux sociaux et du Big Data est de rendre disponible sur le Web une quantité gigantesque de textes numérisés. On dit même que l'humanité a produit ces dix dernières années un corpus textuel plus important que celui des 300 ans antérieurs. Le phénomène le plus marquant est la croissance exponentielle des commentaires disponibles gratuitement sur les forums, les pages des réseaux sociaux (Facebook, Twitter, LinkedIn, Instagram...), les plateformes et sites spécialisés (booking.com, Tripadvisor.com…), les blogs… Leur exploration ouvre une nouvelle voie de recherche en

sciences humaines et sociales (Beauvisage et al., 2013). Ces données peuvent être exploitées dans de très nombreux domaines. L'objectif ultime est alors de faire parler ces données de terrain dont le Web se trouve désormais dépositaire sans passer par de longues et coûteuses enquêtes.

Les corpus d'avis en ligne qui nous intéressent présentent l'avantage d'être abondants, indépendants et spontanés (Ayeh et al., 2013) mais non structurés et hétérogènes. Leur volume, la multiplicité des sources et surtout la qualité des données peuvent rendre leur mobilisation très laborieuse par les méthodes traditionnelle de l'analyse de contenu (Bardin, 1989). Ainsi, l'usage des approches qualitatives traditionnelles est pratiquement impossible. Le recours à des logiciels et outils linguistiques répond à ces nouvelles exigences en termes de volume, de complexité, de moyens humains et financiers, de temps, etc. Les éditeurs de logiciels multiplient leurs efforts en termes de Recherche et Développement afin de proposer des outils de plus en plus performants et qui répondent aux nouveaux besoins des chercheurs, entreprises et cabinets d'études (Boughzala et al., 2014). Ainsi, grâce aux CAQDAS (*Computer-Aided Qualitative Data Analysis Software*) (Fallery et Rodhain, 2007) et aux outils du Web sémantique et du TAL (Traitement Automatique des Langues) (Miller et al., 2010), nous avons 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 (Jenny, 1997 ; Moscarola, 2001 ; Boughzala et Moscarola, 2015).

L'objectif de ce papier est de présenter et illustrer un exemple d'Analyse de Données Textuelles provenant du Web en mettant en place deux approches : exploratoire à travers les classifications lexicales et confirmatoire à travers la modélisation sémantique. Pour ce faire, 600 commentaires en ligne de touristes français évaluant des hébergements en Algérie, au Maroc et en Tunisie sont collectés sur le site de commentaires en ligne TripAdvisor.com. L'objectif de cette collecte est d'étudier la perception expérientielle de ces trois destinations touristiques selon le type d'hébergement choisi (hôtel ou maison d'hôtes).

2. Faire parler les contenus : les différentes approches

2.1. De la tradition littéraire à l'ingénierie linguistique

Exégèse, critique littéraire ou analyse de contenu, toutes ces méthodes visent à «faire parler les textes», c'est-à-dire savoir les lire et les commenter, ou encore extraire leur sens. Cet objectif est très ancien, la tradition littéraire nous y a préparé de longue date (Roger, 1997). Elle nous a appris à lire et construire un nouveau texte pour rendre compte et commenter les corpus analysés. La communication de leur contenu se fait sous forme d'une synthèse, d'un résumé, de citations ou de commentaires critiques dont le but est de défendre ou de contredire un point de vue. Ce travail repose alors sur les talents de l'analyste, ses aptitudes de rédaction et notamment la confiance dont il bénéficie. La subjectivité est, en revanche, toujours redoutée.

L'analyse de contenu vise à systématiser le travail de la tradition littéraire (Bardin, 1989). Elle a l'ambition de remplacer la subjectivité de l'auteur par la démarche critique et objective de l'analyste (Boughzala et al., 2014). Elle se manifeste par :

- la constitution d'une grille thématique (*Code Book*) définissant les catégories de contenu présentes dans le corpus ou susceptibles de s'y trouver,
- la lecture exhaustive et contrôlée du corpus en vue de reconnaître les catégories qui s'y trouve (codage),

- le compte rendu statistique sur la fréquence d'évocation des thèmes et sur la compréhension de leur déterminants et relations.

L'analyse de contenu conduit à la rigueur par la définition précise des catégories de sens et du décompte de leur répétition. Mais, elle est très coûteuse lorsque les corpus sont volumineux et d'autant plus fastidieuse qu'ils sont pauvres en contenu (problème fréquent pour les enquêtes, notamment pour les enquêtes auto-administrées).

Ces deux approches sont certes riches mais nous laissent souvent démunies lorsque les textes sont très volumineux, répétitifs ou lorsqu'ils sont produits dans un contexte utilitariste conduisant directement à l'action. C'est particulièrement le cas des corpus collectés sur le Web (commentaires en ligne, messages sur les forums, les réseaux sociaux, etc., ou encore les réponses aux questions ouvertes dans les enquêtes).

L'ADT est apparue avec le traitement informatique des textes (Lebart et Salem, 1994). Lexicale (compter les mots) puis sémantique (identifier automatiquement les contenus : concepts). L'analyse lexicale se concentre sur la donnée, c'est-à-dire sur les mots composant le texte plutôt que sur son sens. Elle met en évidence les propriétés statistiques lexicales du corpus étudié. Les premières utilisations portaient sur l'attribution d'œuvre à un auteur en fonction des traits stylistiques mis en évidence de manière statistique. Ces méthodes ont vite évolué vers des préoccupations moins littéraires par :

- la production de lexique des formes graphiques ou segments répétés les plus fréquentes,
- le traçage de cartographies lexicales ou la construction de typologies révélant les univers lexicaux caractéristiques du corpus,
- la mise en évidence des expressions récurrentes.

En conséquence, l'analyse lexicale déplace l'effort de lecture vers des substituts du corpus beaucoup moins volumineux (les 50 premiers termes du lexique par exemple, une carte lexicale, ou la liste des termes caractéristiques d'une typologie), communicables et révélateurs de structures qui auraient pu échapper à une lecture de surface. Elle a connu un succès conforté par l'usage de l'analyse morphosyntaxique permettant de passer de la forme graphique au mot et du segment répété à l'expression syntagmatique (Miller et Vandome, 2011). Ainsi, les connaissances linguistiques viennent préciser en amont l'analyse des propriétés statistiques du corpus. Si cette approche réduit l'effort de lecture, elle ne résout pas, en revanche, la question du sens puisqu'il faut toujours un analyste pour interpréter les lexiques, les cartes ou les univers lexicaux. Même si ces substituts sont produits de manière objective et reproductible, leur interprétation demeure subjective et dépendante de la culture et du talent de l'analyste. Mais, son commentaire peut être réfuté par le lecteur sur la base des éléments qui le fondent.

L'analyse sémantique (Rastier, 2009 ; Goddar, 2011) comme son nom l'indique, aborde la question du sens et s'apparente à l'analyse de contenu. Mais, avec les progrès de l'ingénierie linguistique et de l'intelligence artificielle, son ambition est bien plus grande puisqu'elle vise à l'automatisation complète de la lecture et de la reconnaissance des catégories de contenu (Normier, 2007 ; Veronis, 2010). Elle repose sur le modèle ancien de la linguistique générale (Saussure, 1916) selon lequel, lors de la lecture, le sens émerge des interactions entre le signifiant (le mot), le signifié (le concept) et le référant (le contexte). Les termes de cette trilogie sont mis en œuvre avec les thésaurus, les ontologies et les réseaux sémantiques

utilisés pour la lecture automatisée de corpus numérisés. Le thésaurus organise l'arborescence du général au particulier des connaissances ou des concepts repérés par les ontologies (listes de mots définissant les concepts). Au final, les réseaux sémantiques permettent de contrôler, en fonction du contexte, la pertinence des affectations signifiant/signifié. On peut ainsi construire un lecteur artificiel chargé de lire à notre place. A cet égard, il convient de distinguer deux cas selon que l'on utilise un thésaurus standard susceptible de représenter toutes les connaissances pour une langue donnée, ou un corpus spécialisé relatif à une discipline, un métier ou un ensemble de corpus. Ce qu'on désigne par : la linguistique générale *versus* la linguistique de corpus (Rastier, 2002 ; Teubert, 2009). L'usage d'un thésaurus standard peut suffire dans certain cas, mais, le plus souvent, il convient d'en construire un adapté au domaine de l'étude, ce qu'on appelle un thésaurus ad hoc ou ciblé. Ce travail peut être comparé à une analyse de contenu, pour laquelle la construction de l'arborescence du thésaurus correspond à la construction du code book et la documentation de l'ontologie au travail de codification. Dans le cas de corpus très volumineux, le gain de temps peut être considérable, même si la calibration du fonctionnement des réseaux sémantiques requiert un temps d'apprentissage.

L'analyse des sentiments (Turney, 2002 ; Pang et Lee, 2008) est un TAL permettant la synthèse de multiples avis ou commentaires pour obtenir une vue d'ensemble des opinions sur un sujet donné. Une opinion se caractérise par une polarité, pouvant être soit positive, soit négative, soit neutre. L'analyseur des sentiments identifie 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 et des lexiques qui les caractérisent.

2.2. Des corpus à la connaissance : les apports de l'analyse statistique des données

Avec les évolutions méthodologiques et technologiques de l'ADT et du TAL, la tradition littéraire consistant à produire un texte critique pour rendre compte des textes analysés a évolué vers deux directions mobilisant des méthodes statistiques. Il s'agit de :

- mettre à jour, sans à priori et de manière purement statistique, les propriétés lexicales du corpus que l'analyste interprète en cherchant à donner du sens à ce qu'elles révèlent. La statistique précède l'interprétation et la production du sens.
- exploiter statistiquement la distribution et les liens entre catégories de contenu formalisées à priori et identifiées au terme d'une lecture attentive (analyse de contenu) ou automatisée (analyse sémantique ou de sentiment) du corpus. La statistique n'intervient que dans un deuxième temps, elle est consécutive à la reconnaissance des contenus.

Ainsi, la traditionnelle division entre études qualitatives et études quantitatives (Bolden et Moscarola, 2000) perd de son sens dès lors qu'on utilise la statistique pour transformer le texte initial ou pour analyser statistiquement les catégories de contenu qu'il contient. Le vrai clivage est, en fait, celui qui distingue l'approche inductive de l'analyse lexicale à l'approche déductive de l'analyse de contenu fondée sur les modèles sémantiques du code book ou du thésaurus. Pour analyser les corpus collectés, il semble indispensable de mixer les méthodes traditionnelles (propriétés lexicales) avec des approches sémantiques ad hoc (construction et application de thésaurus) et des analyses de sentiments.

L'exemple du corpus des avis déposés sur TripAdvisor.com nous permettra d'illustrer ces deux approches et de discuter leur portée. Il mettra également en évidence qu'on a intérêt à les combiner.

3. Illustration : La place de la destination et des modes d'hébergement dans la relation de l'expérience touristique et de la satisfaction

Cette recherche en science de gestion vise à éclairer les décisions de marketing touristique à partir des avis en ligne. De manière classique, le marketing touristique (Frochot et Legohérel, 2014) s'intéresse à l'offre articulée sur la destination ou le mode d'hébergement et aux attentes qui orientent le choix des touristes. Le marketing expérientiel propose une vision centrée sur l'expérience du séjour comme source de la satisfaction (Batat et Frochot, 2014). Dans quelle mesure les avis reflètent ces différentes visions, et comment apprécier leur poids dans la formation de l'expérience touristique et de la satisfaction ?

Pour répondre à cette question, 600 commentaires en ligne (Figure 1) ont été collectés sur le site d'avis de consommateurs de nuitées hôtelières : TripAdvisor.com. Cette plateforme d'évaluation est devenue un guide de référence pour juger la perception de l'expérience touristique et la qualité des unités hôtelières (Vasquez, 2011 ; Ayeh et al., 2013 ; Limberger et al., 2014 ; Miguéns et al., 2008 ; Tuominen, 2011). Le corpus a été collecté comme un plan d'expérience construit sur 2 facteurs : la destination et le mode d'hébergement. 3 pays ; l'Algérie, la Tunisie et le Maroc et 2 types d'hébergement : les Hôtel/Hôtel club et les Gîtes/B&B (maisons d'hôtes). 100 commentaires par type d'hébergement dans chaque pays.

Ce corpus présente l'avantage d'une expression affranchie des inhibitions résultant d'entretiens en face à face (Hanna Richard et al., 2005) ou des contraintes liées à l'usage d'un questionnaire (Pincott et Branthwaite, 2000 ; Boughzala et Moscarola, 2015). De plus, il est mixte puisque le contenu qualitatif des commentaires est complété par une note d'évaluation globale à 5 niveaux. Mais, ce protocole n'est pas sans biais. Par définition, TripsAdvisor.com est centré sur l'hébergement, ce dont nous devrons tenir compte.


"Riad charmant - Personnel très agréable"

Avis écrit le 16 avril 2015

Nous avons passé un séjour merveilleux dans ce magnifique Riad à 8 minutes de la Place Jemaa El Fna. Aziz nous attendait sur le parking et nous avons immédiatement été charmés par son sourire et sa joie de vivre.Nous nous sommes sentis comme à la maison dès les premières minutes. Tout le personnel est aux petits soins et prêts à répondre à nos désirs. Les petits-déjeuners étaient excellents et nous avons eu le bonheur de déguster un excellent repas le premier soir, dans le patio, sous les orangers. Laurent était présent durant notre séjour et s'était un plaisir de faire sa connaissance et de parler avec lui. Si nous devons retourner à Marrakech (et nous le voulons) nous séjournerons à nouveau dans ce Riad.

Figure 1 : Le corpus : exemple de commentaire

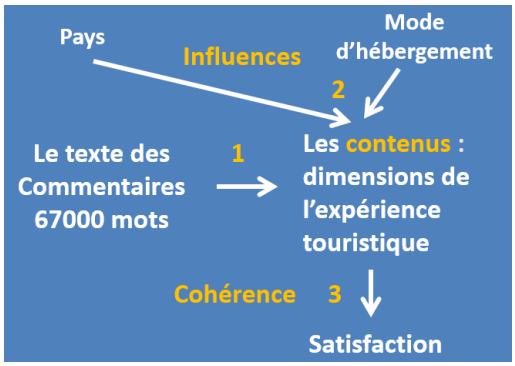


Figure 2 : La problématique de la recherche

Conformément au modèle ci-dessus (Figure 2), la recherche consistera à analyser le contenu des avis par une exploration lexicale et sémantique puis par la construction d'un thésaurus ciblé (1). Cela nous permettra de préciser l'influence des caractéristiques de l'offre (2) et de comprendre la formation de la satsifaction (3).

3.1. Les contenus : la destination en arrière plan

Le corpus collecté est composé de 30 620 mots (16 982 mots pour les hôtels et 13 638 mots pour les maisons d'hôtes), environ 75 pages. Le traitement a été effectué avec le logiciel Sphinx iQ 2 - Quali.

Le premier aperçu lexical et sémantique (Figure 3) montre que la longueur médiane des commentaires est de 37 mots, ce qui atteste d'une certaine richesse. Les 5 principaux champs sémantiques (thésaurus général de la langue française) sont : convivialité, emploi et salaire, domicile, alimentation et restauration et habitat. Cet aperçu donne une idée générale sans surprise ni découverte. Le concept de la destination (pays, villes, lieux) est absent à ce niveau de généralité.

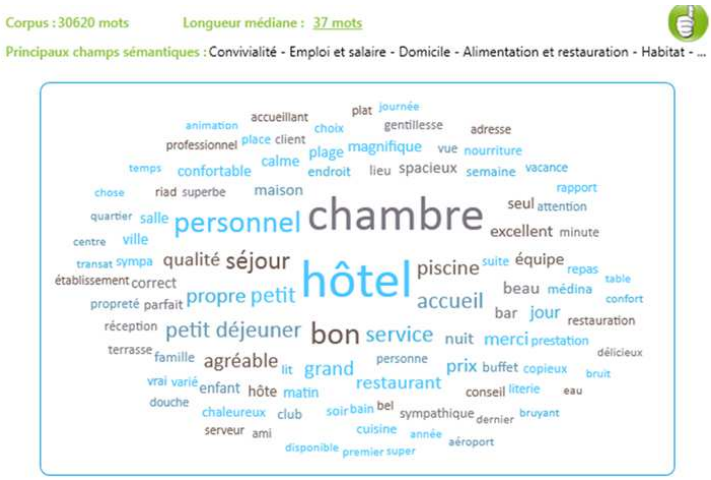


Figure 3 : Synthèse globale lexicale et sémantique (formes lemmatisées et concepts)

Cette première analyse est réduite à l'examen des mots clés les plus fréquents. Pour aller plus loin en restituant les mots dans leur contexte, on identifie les différents univers lexicaux grâce à une classification hiérarchique descendante (Reinert, 1983) qui groupe les avis en classes similaires selon les associations lexicales et sémantiques. 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 en procédant à des itérations successives à partir d'une Analyse Factorielle des Correspondances Multiples (Benzecri, 2007).

L'examen des 7 classes produites par le logiciel et interprétées par l'analyste (Figure 4) montre que les 600 avis sont plutôt centrés sur l'expérience touristique (« grand merci » est le syntagme le plus cité) et les hébergements. Les références à la destination ne sont pas structurantes et ne sont pas identifiées à ce niveau d'analyse qui reflète d'abord la spécificité de TripAdvisor.com : évaluer des hébergements.

Pour repérer les références à la destination, il faut une analyse de contenu plus précise et ciblée sur ce qu'on cherche à identifier. Pour cela, nous mettons en place une modélisation sémantique en construisant un thésaurus ad hoc regroupant les mots du lexique en familles, et en organisant et en nommant ces familles en référence aux connaissances du domaine touristique. La destination y trouve par conséquent sa place. (Figure 5).

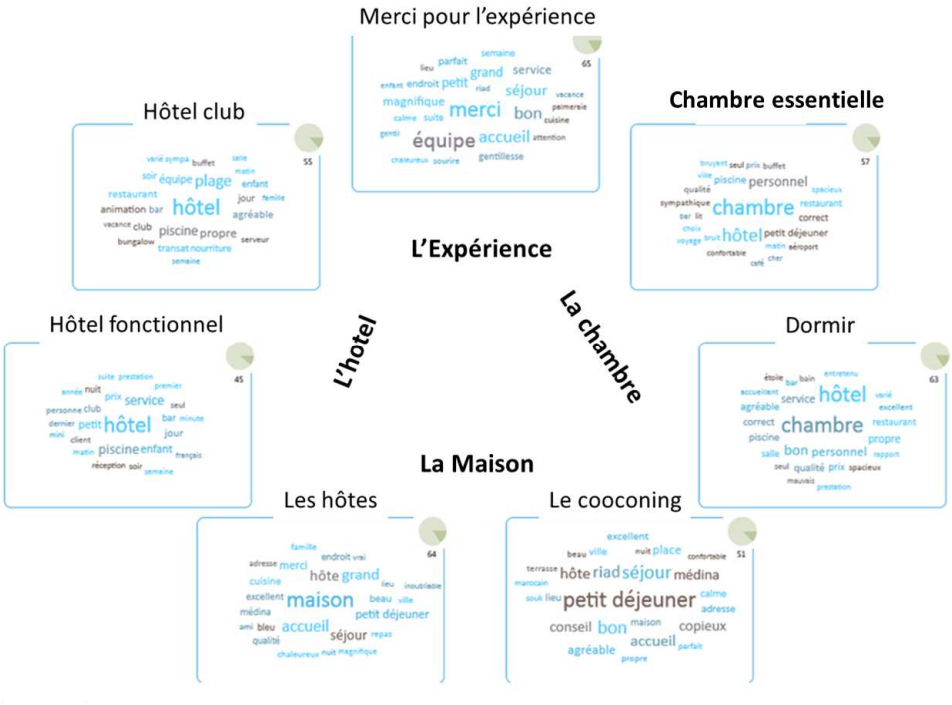


Figure 4 : Classification thématique automatique (classification hiérarchique descendante)

Ce thésaurus correspondant à la grille d'analyse de contenu (Bardin, 1989) qui aurait pu être utilisée pour un codage manuel. Le fait d'associer à chaque feuille une liste de mots, définit une ontologie qui permet de mettre en œuvre une codification automatique. Elle conduit à repérer la présence des thèmes ou sous thèmes par une variable nominale ou par une mesure d'intensité lexicale évaluant le poids de chaque thème / sous thème (Boughzala et al., 2014). Ces résultats sont présentés dans la Figure 5 et confirment celui de l'approche exploratoire quant à l'importance de l'expérience et du mode d'hébergement.

En effet, les évocations du thème « destination » ne viennent qu'en troisième lieu. Elles se manifestent surtout par les références aux ambiances et à l'usage des noms de lieux (Figures 5 et 6). Ce résultat est intéressant puisqu'il permet d'apprécier la présence de la destination, qui n'apparaissait pas lors de l'approche exploratoire. Pour compléter l'analyse, nous effectuons le test de Chi-deux entre les variables pays et mode d'hébergement d'une part, et les thèmes du thésaurus d'autre part. Les relations du test sont très significatives. Le mode d'hébergement et le pays influencent le contenu des avis. L'influence du type d'hébergement est la plus forte, et se manifeste par une sur-représentation très significative du thème « destination » dans les avis provenant du type d'hébergement « maison d'hôte et B&B » (Figure 8).

3.2. Cohérence entre orientation des contenus (analyse des sentiments) et note de satisfaction

L'analyse des sentiments porte sur le repérage des phrases ou propositions comportant des tournures évaluatives et utilise des répertoires de termes positifs ou négatifs. La composition discursive ou sommative (Chardon, 2013) de ces éléments permet d'établir le caractère plus ou moins positif ou négatif d'un avis. Ainsi, parvient-t-on sur la seule base du contenu des commentaires à apprécier leur orientation (Figure 7). Elle est dans la grande majorité des cas

globalement positive et fortement corrélée (r=0,71) avec la note directement donnée par les auteurs. Ceci confirme la cohérence des auteurs.

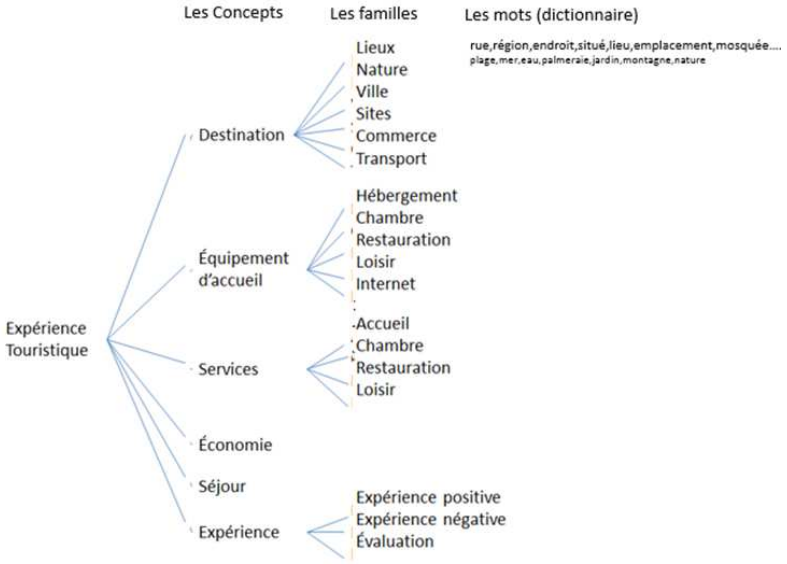


Figure 5 : Le thésaurus ad hoc

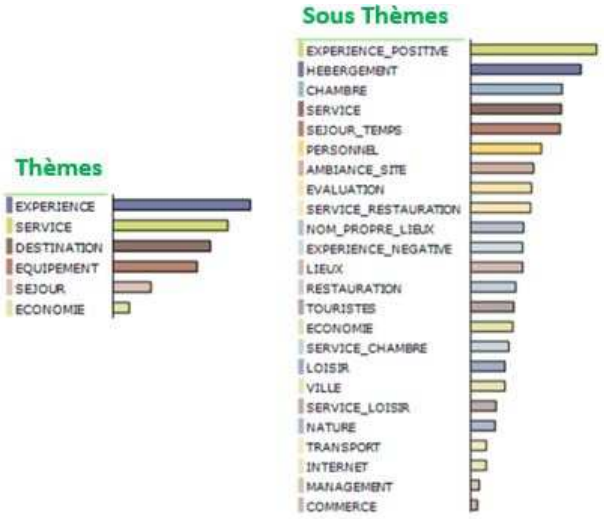


Figure 6 : La présence des thèmes du thésaurus ad hoc

3.3. Influences et facteurs explicatifs

Le pays et le mode d'hébergement affectent les thèmes évoqués de manière significative, quelle que soit la manière dont on les définit. Les cartes ci-dessous mettent également en évidence que le mode d'hébergement est le plus structurant et que les références à la destination sont spécifiques au B&B au Maroc et en Tunisie.

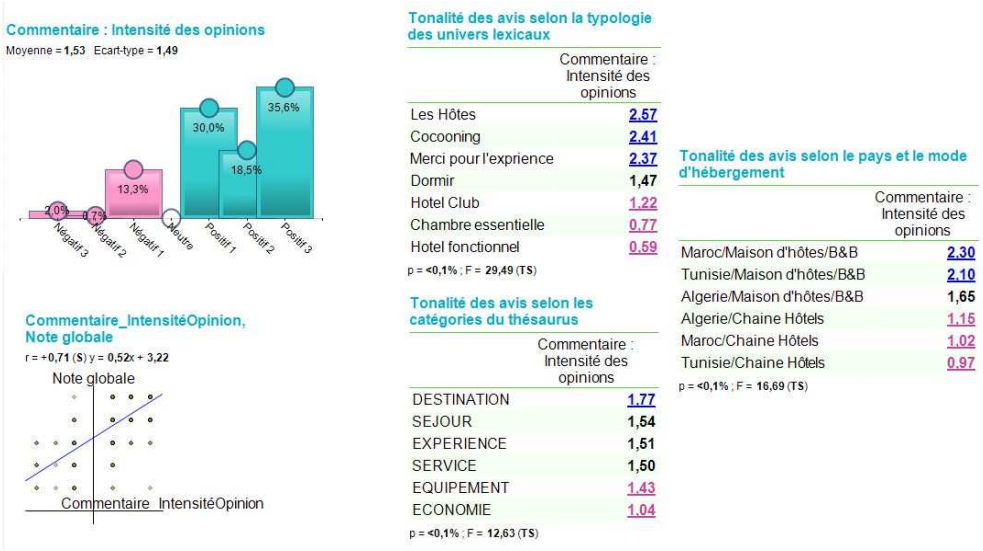
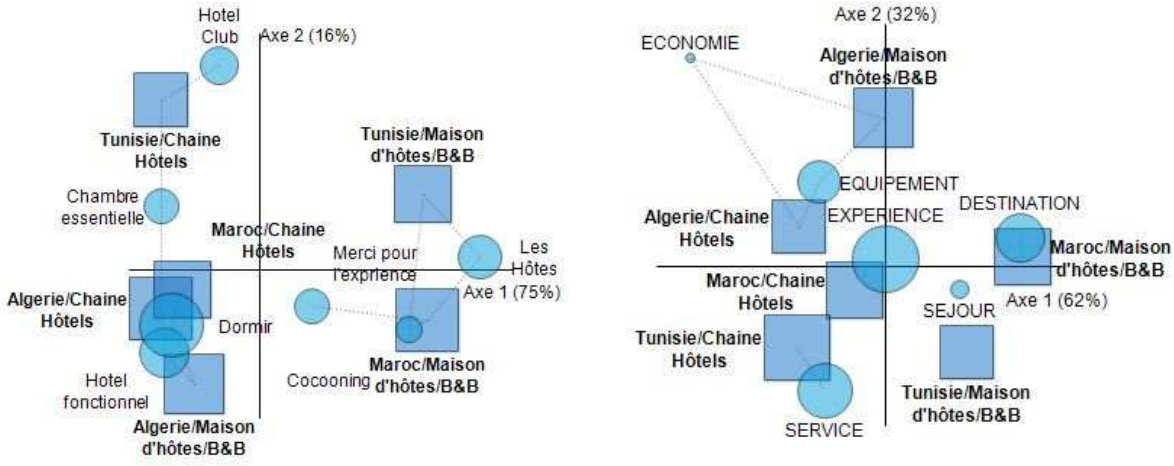


Figure 7 : Orientation des commentaires : cohérence et différences



Thématique classification automatique *Thématique du thésaurus ad'hoc*

Figure 8 : Les thèmes dépendent du pays et du mode d'hébergement

D'autre part, en rapprochant les contenus thématiques et les orientations du commentaire, nous pouvons mettre en évidence les thèmes correspondants aux avis exprimant les sentiments les plus positifs (Figure 7) : ils évoquent « les hôtes » « le cocooning », « l'expérience » et « la destination ».

Ces deux analyses permettent d'établir l'influence du pays et du mode d'hébergement sur le contenu des avis : thèmes abordés et orientation plus ou moins positive.

3.4. Le poids de l'expérience et de la destination dans la formation de la satisfaction globale.

En référence au modèle SERVQAL (Frochot et Legohérel, 2014), nos données nous permettent aussi d'expliquer la formation de la satisfaction globale, définie par la note et l'orientation donnée par l'analyse des sentiments, en fonction des thèmes définis par le thésaurus ad hoc et évalués par les intensités lexicales. Nous construisons pour cela une

modélisation en équations structurelles (Tenenhaus, 1998) de type PLS qui permet d'établir les poids respectifs des différents thèmes (Figure 9).

Le modèle n'explique que 35% de la variance de la satisfaction. Ceci indique que la thématique identifiée par le thésaurus ne couvre qu'une partie des contenus ou plus probablement que les commentaires sont trop focalisés. Néanmoins, le modèle tend à accréditer la vision du marketing expérientiel : le construit « Expérience » pèse plus que les construits « Destination » et « Equipement » correspondant à la vision d'un marketing de l'offre.

Comme tous résultats, ces conclusions méritent bien sûr d'être discutées.

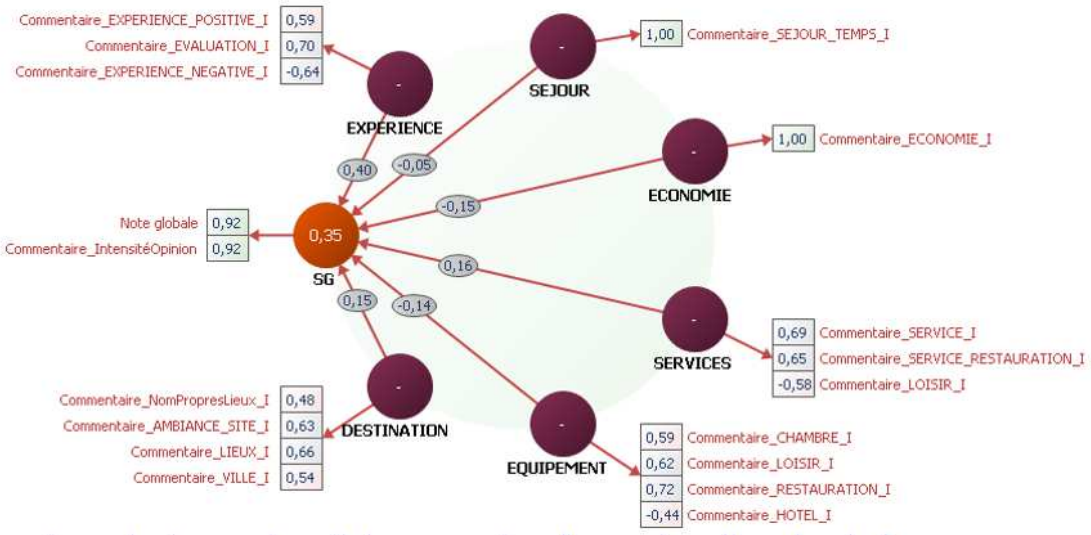


Figure 9 : Le modèle explicatif de la satisfaction (équations structurelles PLS)

5. Conclusion et limites

L'objectif de cette communication était de présenter aux chercheurs et aux chargés d'études un exemple d'analyse de corpus d'avis en ligne. Ces données sont abondantes et assez faciles d'accès. Elles présentent l'avantage d'être spontanées et libres. Leur exploration peut ainsi être une voie complémentaire ou alternative aux sondages d'opinion.

Pour les analyser, les méthodes traditionnelles des ADT (propriétés lexicales) ont été enrichies par des approches sémantiques (thésaurus, analyse de sentiments). Nous cherchions à déterminer l'influence de la destination et du mode d'hébergement dans la perception de l'expérience touristique. L'analyse exploratoire visant à identifier les univers lexicaux n'a pas permis de juger de la place de la destination qui n'a pas été révélée par la classification hiérarchique descendante. La construction d'un thésaurus ad hoc, regroupant les mots du lexique en référence aux connaissances du domaine touristique, a permis de focaliser l'analyse et de faire émerger le thème de la destination. Ceci atteste de l'importance d'enrichir l'analyse lexicale par une analyse sémantique ciblée. Enfin, une modélisation en équations structurelle mettant en relation le poids des thèmes avec la satisfaction a permis de tester un modèle comme on l'aurait fait au terme d'une enquête par questionnaire.

Ainsi, la portée de cette recherche est d'abord méthodologique. Elle met en œuvre des protocoles mixtes (Walsh, 2015) mélangeant données qualitatives (les avis) et quantitatives

(les notes), approches exploratoires (classification en univers lexicaux) et confirmatoires (thésaurus et modélisation structurelle). Elle fait appel à la statistique (calculs de fréquences, d'intensité lexicales, croisements, classifications) et à l'interprétation subjective (interprétation des typologies, construction du thésaurus…).

La complexité des phénomènes étudiés peut ainsi être abordée sans réduction excessive tout en satisfaisant aux exigences de la démarche scientifique : objectivité des traitements quantitatifs, résultats synthétiques et communicables donnant au lecteur la possibilité de critiquer les interprétations qualitatives de l'auteur. Ainsi sur le fond, un examen plus approfondi des résultats obtenus permettrait de les discuter plus précisément pour focaliser encore la recherche ou proposer de nouveaux développements. Par exemple, en ne sélectionnant que des avis évoquant la destination et en les répartissant à part égale selon le niveau de la note associée à l'avis.

6. Références

Ayeh J K., Norman A. and Law R. (2013). Do we believe in TripAdvisor? Examining Credibility Perceptions and Online Travelers' Attitude toward Using User-Generated Content. *Journal of Travel Research*. 52(4) 437-452.

Bardin L. (1989). *L'analyse de contenu*, PUF.

Batat W. and Frochot I. (2014). *Marketing expérientiel*, Collection : Tendances marketing. Dunod.

Beauvisage T., Beuscart J.S., Cardon V. and Trespeuch M. (2013). Notes et avis des consommateurs sur le web, *Réseaux*, 177(1) :131.

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

Bolden, R. and Moscarola, J. (2000). Bridging the quantitative-qualitative divide: the lexical approach to textual data analysis, *Social Science Computer Review*. 18(4), 450-460.

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

Boughzala. Y. and Moscarola J. (2015). Le mur d'images dans les enquêtes en ligne : comment stimuler pour observer et mesurer ?. In Kalika M., Beaulieu P. *La création de connaissance par les managers*. Editions EMS.

Cardon V. (2014). Des chiffres et des lettres : Évaluation, expressions du jugement de qualité et hiérarchies sur le marché de l'hôtellerie. *La Découverte*. n° 183, p. 207-245.

Chardon B. (2013). Chaîne de traitement pour une approche discursive de l'analyse d'opinion. Université Paul Sabatier.

Fallery B. and 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

Frochot I. and Legohérel P. (2014). *Marketing du toursime*, Collection: Marketing sectorial. Dunod.

Goddard C. (2011). Semantic Analysis: A Practical Introduction *Oxford Text Book.*

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.

Limberger P F., Dos Anjos F A., De Souza Meira J V. and Gadotti dos Anjos S J. (2014). Satisfaction in hospitality on TripAdvisor.com. *Tourism & Management Studies*. Vol. 10 Issue 1, p.59.

Miguéns J., Baggio R. and Costa C. (2008). Social media and Tourism Destinations: TripAdvisor Case Study. *IASK ATR2008 (Advances in Tourism Research 2008)*. Aveiro. Portugal. May. 26-28.

Miller F. and Vandome A. (2011). *Analyse Morphosyntaxique*. Alphascript Publishing.

Miller F., Vandome A. and McBrewster J. (2010). *Traitement Automatique du Langage Naturel*. Alphascript Publishing.

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. and Bolden, R. (1998). From the data mine to the knowledge. In Jan M. Żytkow and Mohamed Quafafou. *Principles of Data Mining and Knowledge Discovery*. Springer-Verlag.

Normier B. (2007). *L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle*. ADBS.

Pang B. and Lee L. (2008). Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval.* Vol. 2, No 1-2 (2008) 1–135.

Rastier F. (2009). *Sémantique interprétative*. PUF.

Rastier F. (2002). *Enjeux épistémologiques de la linguistique de corpus*. PUF.

Reinert M. (1983). Une méthode de classification descendante hiérarchique?: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, vol 8. n°2.

Roger J. (1997). *La critique littéraire*. Dunod.

Saussure F. (1916). *Cours de linguistique générale*. Payot.

Tenenhaus M. (1998). *La régression PLS : Théorie et Pratique*. Éditions Technip.

Teubert W. (2009). *La linguistique de corpus : une alternative*. SEMEN.

Tuominen P. (2011). The Influence of TripAdvisor Consumer-Generated Travel Reviews on Hotel Performance. *University of Hertfordshire Business School Working Paper*.

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the Association for Computational Linguistics.*

Vasquez C. (2011). Complaints online: The case of TripAdvisor. *Journal of Pragmatics*. Volume 43, Issue 6, May 2011, Pages 1707–1717.

Veronis J. (2010). *Le Traitement automatique des corpus oraux*. Hermès Science.

Walsh I. (2015). *Découvrir de nouvelles théories*. Editions EMS.