Synergies Tunisie n° 5 - 2022 p. 127-141

GERFLINT
ISSN 2105-1054
ISSN en ligne 2257- 8390

Proposition d'une démarche informatique linguistique pour l'étude d'une problématique managériale : approche discursive

Soumaya Mejri
Université de Tunis, Tunisie
ESSEC-Tunis
soumayamejri@yahoo.fr
https://orcid.org/0000-0001-8823-2145

Reçu le 27-12-2021 / Évalué le 15-01-2022 / Accepté le 16-03-2022

Résumé

La littérature sur le discours en sciences de gestion est foisonnante. Elle cherche à optimiser ce support pour tirer des enseignements utiles aux entreprises. L'objectif de ce travail est de proposer une démarche linguistique informatique afin d'exploiter différemment et d'une manière optimale cet outil de travail. Dans un premier temps, il s'agit d'expliciter nos choix méthodologiques tant sur le plan managérial que sur le plan linguistique et informatique. Dans un second temps, l'objectif est de décrire et d'évaluer les résultats de l'étude quant à l'analyse du discours des clients d'hôtels émis sur un site web. Le tout est dans l'objectif de fournir des recommandations managériales aux entreprises.

Mots-clés : avis clients, hôtel, ingénierie, sciences de gestion, analyse linguistique

Proposal of a linguistic IT approach for the study of a managerial problem: discursive approach

Abstract

The literature on the discourse in management sciences is abundant. She seeks to optimize this support to draw useful lessons for companies. The objective of this work is to propose a computer linguistic approach in order to use this work tool differently and in an optimal way. First, it is a matter of clarifying our methodological choices on a managerial level as well as on a linguistic and IT level. Secondly, the objective is to describe and evaluate the results of the study regarding the analysis of the speech of hotel guests posted on a website. Everything is aimed at providing managerial recommendations to companies.

Keywords: customer reviews, hotel, engineering, management sciences, linguistic analysis

Dans la littérature en sciences de gestion, le discours et son analyse constituent un créneau de recherche fortement mobilisé pour suivre l'évolution des pratiques et des écrits en management. Il s'agit d'une méthode appropriée pour répondre à de diverses

problématiques comme la finance, le marketing, la GRH, la RSE, la communication, la stratégie, ... (Pondy, 1976 ; Girin, 1989 ; Varaa, 2006 ; Knight et Morgan, 1991 ; Thomas, 1998 ; Samra-Fredericks, 2003 ; Lilley, 2001 ; Chanal, Lacroux et Mounoud, 2001 ; Girot et Giordano, 1998 ; Chauzal-Boutonnet, 2002 ; Alvesson et Kärreman, 2000 ; Ashcraft, 2004 ; Piette et Rouleau, 2008 ; Czarniawska, 2005 ; Igalens, 2006 ; Maurel et Pantin, 2017). L'objectif fondamental recherché dans l'étude du discours consiste à optimiser ce support pour en extraire des informations pertinentes et saisir des réalités managériales. Nous nous inscrivons dans cette logique qui considère le discours comme un outil privilégié susceptible de fournir des enseignements utiles en management stratégique. Tout en mobilisant le discours, l'objectif de cet article est de proposer une démarche linguistique informatique pour l'exploitation des données utiles aux entreprises. Dans cette étude, nous nous inscrivons dans une démarche d'ingénierie de la recherche en sciences de gestion (Chanal, Lesca et Martinet, 2015). Pour cela, nous décrirons dans un premier paragraphe, notre problématique relative à un acteur stratégique (clients), la méthodologie adoptée et les outils mobilisés. Un deuxième paragraphe sera consacré à l'analyse des données et à la présentation des résultats. Pour finir, le dernier paragraphe portera sur une évaluation de notre modèle et les enseignements tirés de cette étude.

1. Problématique et méthodologie de recherche

Notre étude soulève la problématique de l'innovation technologique en misant sur la pluridisciplinarité (gestion, linguistique et informatique). En effet, dans le cadre mondial actuel et dans un contexte tunisien post-révolution tout particulièrement, la détention d'informations pertinentes constitue un enjeu majeur à tous les niveaux : économique, social, politique, etc. Pour cela, nous cherchons à concevoir un algorithme qui permet d'extraire des informations du Net relatives à la clientèle d'hôtels tunisiens pour identifier objectivement leur appréciation en misant sur une analyse linguistique (lexicale, syntaxique et sémantique). En effet, notre objectif consiste à proposer un outil d'évaluation du feed-back de la clientèle des hôtels en Tunisie. Par le biais de cet outil, nous répondons à la fois aux besoins de la recherche qui est en quête de données pertinentes à ceux des entreprises dans leur quotidien managérial. Notre projet vise à :

- montrer comment on peut automatiser des données à partir d'un dictionnaire. Pour le chercheur et la recherche, il s'agit d'un excellent outil de travail qui permet de maîtriser et d'exploiter le nombre considérable des données disponibles sur le Net ;
- mettre à la disposition des entreprises (les hôtels) un outil linguistique informatique permettant d'extraire du Net le feed-back de la clientèle (commentaires). Cet outil managérial donnera aux organisations une meilleure visibilité, et par conséquent, il

servira d'outil d'aide à la décision stratégique. Cet outil est d'autant plus important et nécessaire qu'il s'agit d'évaluer les services rendus pour maintenir la reprise d'un secteur d'activité spécifique (tourisme) et la faire durer. Nous retenons les trois articulations fondamentales suivantes :

- L'enjeu économique qui consiste à évaluer le programme de la relance du tourisme en Tunisie (tout particulièrement le retour de la clientèle européenne et francophone tout particulièrement). Sur le plan managérial, il s'agit de proposer la technologie (et ses avancées) pour une meilleure visibilité (une synthèse détaillée) du feed-back des clients d'une entreprise donnée à partir des données du Net ;
- L'enjeu linguistique consiste à exploiter des données textuelles abondantes et disponibles gratuitement. Sur le plan linguistique, il s'agit de suivre une démarche linguistique (analyses lexicale, syntaxique et sémantique) en traitant des données linguistiques en l'occurrence les données textuelles ;
- L'enjeu informatique consiste à concevoir un algorithme permettant l'exploitation d'informations accessibles d'une manière pertinente.

En nous basant sur cette logique tripartite et combinatoire, la problématique peut être énoncée de la manière suivante : « Comment évaluer les services rendus au sein des hôtels tunisiens pour maintenir la reprise et la faire durer ? ».

Pour répondre à cette problématique, notre dispositif méthodologique se décline en trois niveaux respectifs, à savoir :

- le niveau managérial relatif au choix du domaine d'activité et de l'acteur économique cible de l'étude ;
- le niveau linguistique qui se rattache aux choix des données textuelles relatives au site web TripAdvisor et de la méthode d'analyse linguistique qui dépasse l'analyse lexicale, fréquemment utilisée en sciences de gestion, pour intégrer le niveau sémantique ;
- le niveau informatique où nous retenons le Python comme langage de programmation et Unitex comme logiciel d'analyse des données.

2. Méthodologie au niveau managérial : choix des clients d'hôtels tunisiens

Ayant des ressources naturelles limitées par rapport à ses pays voisins (l'Algérie et la Libye), la Tunisie mise beaucoup sur le tourisme. Il s'agit d'une activité qui a été développée depuis les années 60. A partir de 2018, dans un contexte post-révolution et pour assurer la relance du tourisme dans le pays, une attention particulière a été accordée à ce domaine d'activité. Dans le secteur du tourisme, les clients des hôtels bénéficient d'une place privilégiée, dans le sens où l'évaluation du service fourni

permet l'amélioration, voire l'ajustement des pratiques managériales des hôtels. Outre le fait que le client fait partie du micro-environnement de n'importe quelle entreprise et qu'il jouit d'ores et déjà d'une attention spécifique, dans le secteur du tourisme, son rôle devient doublement primordial. Dans les hôtels, la clientèle joue à la fois un rôle commercial, de marketing (vente) et un rôle stratégique managérial (décisions) et ceci par le biais de l'évaluation de la prestation fournie (feed-back). Ce double rôle montre parfaitement comment les clients des hôtels constituent un pivot fondamental et incontournable dans ce type d'activité.

3. Méthodologie au niveau linguistique : choix des données textuelles et de l'analyse linguistique

3.1. Choix des données textuelles du web TripAdvisor

L'analyse des données textuelles (ou ADT) se présente comme une approche des sciences humaines qui définit les textes comme un ensemble de données fournies et organisées. Une fois cet ensemble de discours considérés comme un corpus, il peut être analysé indépendamment de l'énonciateur, voire de l'énonciation. L'analyse des données textuelles a le privilège d'être une approche à la fois qualitative et quantitative. Elle cherche à qualifier les éléments des textes à l'aide de catégories et à les quantifier en analysant leur répartition statistique. Cette approche, très utilisée sur des corpus de textes littéraires ou de textes politiques, ne cesse d'évoluer à partir des années 2000 avec le développement des outils informatiques d'un côté et le progrès de l'ingénierie linguistique et du traitement automatique des langues d'un autre côté. Comme le résume parfaitement Moscarola J. (2018 : 191-217), dans son ouvrage collectif *Faire parler les données. Méthodologies quantitatives et qualitatives,* l'« analyse de données textuelles ou fouille de texte, (…) font appel à la statistique, à l'analyse de données, à l'ingénierie linguistique et à la sémantique. Elles nécessitent l'usage de logiciels appropriés et offrent la possibilité d'aborder de très grands corpus ».

Pour le choix des données textuelles, nous retenons le site web TripAdvisor[1] comme base de données volumineuse à partir de laquelle nous pouvons extraire les informations appropriées à notre étude. Le site web américain TripAdvisor nous permet d'obtenir un corpus d'analyse suffisamment volumineux. Il s'agit d'un site très élaboré qui offre des avis et des conseils touristiques émanant des consommateurs sur des hôtels, des restaurants, des villes, des lieux de loisirs, etc., à l'international. TripAdvisor possède également une unité destinée à faciliter le contact des professionnels du tourisme avec leurs propres visiteurs. Il est entièrement gratuit pour les utilisateurs. Son logo représente une tête de hibou avec un œil rouge et l'autre vert. Présente dans 45 pays, l'entreprise de TripAdvisor accueille plus de 315 millions de visiteurs uniques chaque mois et recueille plus de 500 millions d'avis et d'opinions.

3.2. Choix de l'analyse linguistique : choix du corpus et des thèmes de l'analyse

Comme le dit si bien Neveu (2000 :86), « les corpus sont ainsi des artefacts, c'est-à-dire des objets construits. Leur construction répond à un programme de recherches déterminé par un certain type d'usages langagiers qui sont censés n'offrir qu'une représentation partielle. Aucun corpus ne saurait en effet refléter la langue dans son ensemble, et se poser en référence universelle. Ce que rappelle John Sinclair (1996), qui définit le corpus comme une collection de ressources langagières sélectionnées et organisées à partir de critères linguistiques explicites et destinées à servir d'échantillons représentatifs. On appelle généralement corpus électronique une collection de ressources textuelles réunies suivant ce principe, et encodées de manière standardisée et homogène afin de permettre des extractions non limitées a priori ».

En effet, notre corpus porte sur les commentaires des clients d'un hôtel tunisien sur le site web TripAdvisor. Par le biais de scripts en Python 3, nous allons extraire de ce grand site web tous les avis émis sur l'hôtel choisi pour évaluer la prestation générale positive ou négative. Certes, nous nous limitons à examiner les avis de la clientèle relative à un seul hôtel qui est Hôtel A[2]. Même s'il y a 51 hôtels dans la même zone touristique, 112 hôtels sur l'île de Djerba (Djerba Island) et 230 établissements touristiques (hôtels et autres hébergements) accessibles sur le même site web TripAdvisor, le choix d'un seul hôtel se justifie par de multiples raisons. Cet hôtel possède 2981 avis et se présente parmi les hôtels les plus visités, les plus commentés et les plus notés en langue française. Le choix d'un seul hôtel se justifie surtout par le contenu spécifique de ce corpus (corpus atypique car c'est un discours familier et personnalisé) mais également par la nécessité d'examiner l'intégralité de son contenu pour permettre son analyse linguistique.

Notre corpus, constitué de commentaires individualisés, présente un certain nombre de spécificités et de difficultés. Nous sommes loin d'un discours clair, officiel, bien structuré et correct au niveau lexical, syntaxique et sémantique. Mais il s'agit d'avis individualisés, parfois d'éléments parachutés non structurés où il n'y a pas de phrases. Ce sont des commentaires personnalisés qui utilisent un langage familier avec un nombre important de fautes de frappe, de fautes d'orthographe, de grammaire, de conjugaison, de syntaxe, etc. Ces caractéristiques nous obligent à effectuer un prétraitement sans lequel une analyse linguistique ne serait pas possible. Même si nous avons inclus dans le lexique étudié certaines notions fausses ou inexistantes (comme des mots sans accents ou des mots abrégés comme *sympa* ou des mots comme *sympatoche* pour dire *sympathique* ou encore *supppppper* pour intensifier le *super*), nous avons effectué un nettoyage complet de notre corpus pour nous assurer de la fiabilité de nos résultats. Pour ces raisons, notre corpus est limité en termes de volume et ne porte que sur les avis de la clientèle d'un seul hôtel (Hôtel A). Ces avis figurent sur 511 pages web et chaque page comporte en moyenne 5 avis.

L'analyse linguistique porte à la fois sur le lexique, la syntaxe et la sémantique. Nous établissons, dans un premier temps, une liste du lexique relatif à notre thème général, à savoir le lexique évoqué majoritairement par la clientèle des hôtels. Ce lexique est classé en sept principaux sous-thèmes qui sont systématiquement évoqués par les clients dans les avis émis sur leur séjour à l'hôtel : Accueil, Chambre, Hôtel, Restauration, Animation, Personnel et Services.

Dans une deuxième étape, nous dégageons une liste de mots relatifs à chaque sous-thème. Par exemple, dans 'Personnel', il y a 'animateur', 'réceptionniste', 'directeur', 'direction', 'serveur', 'danseur', 'musicien', cuisinier', 'chef', 'masseur', etc.

Le choix du sous-thème et de ses items (le lexique qui lui est rattaché) n'est pas du tout aléatoire mais suit une structure bien définie. Ce choix découle d'une démarche bien structurée suivie par la majorité de la clientèle, retraçant ainsi la chaîne chronologique des services, de l'arrivée au départ :

- accueil à l'aéroport (ponctualité, disponibilité, transport, trajet,…)
- accueil à l'hôtel (réception, attente, cocktail,…)
- la chambre attribuée (vue, espace, hygiène, lit, toilettes,…)
- le séjour dans l'hôtel qui porte principalement sur :
    - l'hôtel en lui-même en termes d'architecture, hall, espace, jardin … et surtout la piscine (transat, dimension, propreté,…)
    - la restauration : les plats, le buffet, l'hygiène, …
    - le personnel en général (serveurs, femme de ménage,…)
    - l'animation (ambiance, animateurs, soirées,…)
    - les services fournis (mini-club pour enfants, salle de jeux, excursions,…)

En suivant cette démarche globale des internautes, nous dégageons 7 thématiques fondamentales à étudier (Accueil, Chambre, Hôtel, Restauration, Personnel, Animation et Services). Cette première partie nous permet d'identifier la fréquence du lexique général utilisé par les clients de l'hôtel. Une seconde partie consiste à identifier l'usage fait de ce lexique par le biais des graphes syntaxiques utilisés dans le logiciel UNITEX. En élaborant ces graphes, nous effectuons une analyse syntaxique et sémantique pour mieux évaluer les avis émis sur les prestations de l'hôtel. Pour cela, nous avons intégré deux autres thèmes à l'analyse, à savoir les appréciations : Prestation positive ou Prestation négative. Pour chacune de ces thématiques, nous proposons une série de lexiques appropriés. Il s'agit d'identifier l'orientation globale de chaque commentaire et de dégager l'appréciation de la prestation générale des clients, qui est soit positive soit négative. Cela n'est possible que par le biais de l'analyse syntaxique et sémantique des énoncés permettant une analyse binaire des unités lexicales, puisqu'elle étudie la

corrélation entre les concepts étudiés. Dans ce qui suit quelques exemples des thématiques traitées ainsi que des thèmes sous-jacents :

- *Chambre :* chambre.N, suite.N, lit.N, vue.N, balcon.N, hygiène.N, toilette.N, serviette.N, drap.N, climatisation.N, climatiseur.N, meuble.N, propreté.N, …
- *Personnel :* personnel.N, direction.N, directeur.N, responsable.N, hôtesse.N, hôtesse.N, animateur.N, bagagiste.N, serveur.N, réceptionniste.N, financier.N, …

4. Méthodologie au niveau informatique

La méthodologie adoptée au niveau informatique porte sur deux grands axes. Le premier concerne les scripts Python élaborés pour l'extraction et la constitution du corpus. Le second décrit l'outil informatique UNITEX et la démarche suivie pour l'analyse et l'obtention des résultats.

Le premier axe présente deux scripts différents. Par le biais d'un premier script Python, nous effectuons l'extraction du contenu intégral des différentes URLs choisies. Il s'agit là de 511 pages du site TripAdvisor qui portent, entre autres, les commentaires rédigés par des internautes, clients qui ont visité l'Hôtel A. Par le biais du second programme Python, nous nous limitons à extraire uniquement les commentaires des internautes sur l'hôtel concerné. Après quoi, nous réalisons un nettoyage du corpus pour pouvoir le soumettre à l'analyse avec le logiciel UNITEX. Ce nettoyage est loin d'être un choix mais il s'agit d'un impératif incontournable, vue les caractéristiques et les spécificités du corpus choisi (familier, incorrect, parlé, …).

4.1. Les scripts en python3 : constitution du corpus

Tout d'abord, il existe de nombreux langages de programmation. Chaque langage de programmation répond à des objectifs spécifiques. Par exemple, un langage peut être meilleur pour gérer une base de données, tandis qu'un autre sera particulièrement adapté à créer une interface utilisateur … Notre choix du langage de programmation a porté sur un langage populaire (pas le plus populaire) : Python. Il s'agit d'un langage simple, gratuit, portable, orienté objet, dynamique, avec des scripts courts, une syntaxe simple et des types de données évolués (listes, dictionnaires, ...).

Face à des données textuelles de volume important, le Python se présente comme un choix approprié pour exploiter au mieux ce volume d'informations. Pour extraire le contenu intégral des différentes URLs sélectionnées, nous intégrons dans le script nommé (aspirateur.py), la première adresse URL relative à la première page des avis (Figure 1) et qui est différente des autres pages qui suivent (Figure 2).

```
url1=http://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-Djerba_
Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html
```

Figure 1: URL de la première page Web

Pour le reste des pages, nous accédons à une adresse avec des parties invariables et une variable n qui doit être incrémentée par 5. En effet, l'examen des autres URLs révèle une variation au niveau d'un chiffre à l'intérieur de l'adresse. Cette variation se présente comme suit :

```
url2=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-or»+5+»-
Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»

url3=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-or»+10+»-
Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»

url4=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-or»+15+»-
Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»

…

url499=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-
or»+2490+»-Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governo-
rate.html»

urli=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-or»+n+»-
Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»

url1=»http://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-Djerba_
Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»

urli=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-or»+n+»-
Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Governorate.html»
```

Figure 2 : Les différentes URLs

Ainsi pour accéder aux différentes adresses URLs, nous créons une boucle while qui conditionne la variable n < 2490 et qui l'incrémente par 5 tout en initialisant cette variable à 0 (n=0) avant la boucle (Figure 3).

```
Urli=»https://www.tripadvisor.fr/Hotel_Review-g297948-d578174-Reviews-
or»+str(varn)+»-Djerba_Plaza_Hotel_Spa-Midoun_Djerba_Island_Medenine_Go-
vernorate.html»#Reviews

#html=get_url(urli)
```

Figure 3 : URLs des autres pages web des avis émis sur Hôtel A

Pour récupérer le contenu intégral de ces différentes URLs, nous proposons une fonction get url (urli) qui permet d'ouvrir la page, la lire et récupérer son contenu. Pour cela, nous importons urllib.request.urlopen (urli) puis le read (). Le write (urli) est mis à l'intérieur de la boucle pour permettre l'affichage du contenu de chaque page consultée. Le tout est précédé par la création d'un fichier (.txt) qui récupère tous les résultats obtenus, c'est-à-dire le contenu intégral des différentes URLs sélectionnées.

a = codecs.open ('Code_avis_tunisie.txt', 'w+')

Pour permettre la vérification de la validité de notre script Python, nous demandons l'affichage des différentes URLs traitées (a.write (urli)).

Une fois le contenu intégral obtenu dans un fichier (.txt), nous réalisons un autre script Python nommé (corpus.py) pour sélectionner de ce contenu global uniquement les commentaires et les avis des clients. Cela n'est possible que par le biais de la commande re.finditer de l'expression régulière de la Figure 4. Il s'agit de rechercher tout contenu situé entre les balises jaunes (les tirets et le slash étant déspécifiés).

```
'<q class="hotels-reviews\-list\-parts\-ExpandableReview_reviewText\-\-
3oMkH"><span>([a-zA-Z0-9àè……..]+)<\/span>'
```

Figure 4 : L'expression régulière

Etant donné que le corpus extrait du web est un texte illisible, rempli de caractères spéciaux, de fautes d'orthographe, de grammaire, de conjugaison, de syntaxe, d'espaces, de mots créés par les internautes, nous procédons au repérage des imperfections par la commande Rechercher dans Edition et à un remplacement par la commande Remplacer. Le nettoyage Edition/Rechercher/Rechercher et remplacer permet d'obtenir un texte correct dans l'ensemble, prêt à l'analyse linguistique. Quelques exemples des types d'erreurs repérés dans le corpus sont récapitulés dans le tableau qui suit.

Erreurs au niveau du corpus	Types d'erreurs	Correction
\xc3\xa8	caractères spéciaux	è
\xc3\xa9	caractères spéciaux	é
\xc3\xb4	caractères spéciaux	ô
\xc3\xa0	caractères spéciaux	à
etait	absence d'accents et pb de conjugaison	était / étaient
ds / sdb / cpl	Abréviations	dans / salle de bain / complément
ma-gni-fi-que GENIALLISSIME / ttrreess	langage parlé	magnifique génial/ très

Tableau 1 : Synthèse des erreurs repérées dans le corpus d'analyse

La démarche informatique suivie pour l'obtention du corpus d'étude peut être décrite de la manière suivante (description de la chaîne de traitement) :

- accéder à la page web TripAdvisor,
- repérer et récupérer les liens https relatifs à l'hôtel tunisien Hôtel A,
- récupérer les différentes URLs qui concernent notre projet,
- extraire leur contenu (le contenu intégral de toutes les URLs),
- repérer et extraire uniquement le texte relatif aux avis des clients (les distinguer des réponses), c'est-à-dire les données relatives aux avis des clients des hôtels ;
- récupérer ce texte (le corpus) dans un fichier .txt ;

Une fois le corpus constitué, il faut procéder à un nettoyage dans le fichier .txt par Edition/Rechercher et remplacer pour le rendre plus lisible, compréhensible et surtout valable à une analyse linguistique (en éliminant les caractères spéciaux, en corrigeant certaines fautes ou encore en ajoutant de l'espace entre les mots).

4.2. Analyse des données par le logiciel UNITEX

Comme le précise Paumier (2016 : 13-4), « UNITEX est un ensemble de logiciels permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire. (…) Les dictionnaires électroniques décrivent les mots simples et composés d'une langue en leur associant un lemme ainsi qu'une série de codes grammaticaux, sémantiques et flexionnels. (…) Les grammaires sont des représentations de phénomènes linguistiques par réseaux de transitions récursifs (RTN), un formalisme proche de celui des automates à états finis. (…) Ces grammaires sont représentées au moyen de graphes que l'utilisateur peut aisément créer et mettre à jour. (…) Les tables de lexique-grammaire sont des matrices décrivant les propriétés de certains mots. (…) UNITEX permet de construire des grammaires à partir de telles tables. UNITEX est un moteur permettant d'exploiter ces ressources linguistiques. »

Une fois UNITEX installé et la langue française choisie, nous créons les graphes relatifs aux différents thèmes de notre étude. Le corpus constitué est sous forme d'un fichier (.txt). Il s'agit d'un fichier brut qu'il faut prétraiter pour qu'il soit analysé par UNITEX. Une fois le prétraitement effectué, UNITEX génère un fichier (.snt).

5. Résultats de l'étude

Les résultats obtenus pour les neuf thèmes étudiés se présentent comme suit, en termes de nombre d'occurrences et de concordanciers correspondants. Exemple de résultats relatifs au thème chambre :

Result Info
707 matches
2376 recognized units
(2.735% of the text is covered)

Nous retenons le thème *Prestation positive* comme le premier de la liste avec 2286 occurrences suivi par le thème *Animation* et le dernier de la liste est *Prestation négative*. Ce résultat laisse penser que le thème *Prestation positive* et *Animation* sont les plus pertinents de l'étude. D'ailleurs, la littérature en sciences de gestion, qui mobilise principalement l'analyse lexicale, traduit ce résultat par la pertinence de ces thèmes. Or le calcul de f-score, qui est relatif à la performance des modèles en prenant en considération le taux de précision et de rappel, nous permet une nouvelle lecture de ces résultats. Le taux de précision mesure l'efficacité d'un système d'étiquetage établie à partir du ratio entre le nombre d'informations pertinentes trouvées lors de l'étiquetage d'un document et le nombre total d'étiquettes fournies par le système. C'est un indicateur de mesure du bruit. Le bruit est un ensemble d'étiquettes non pertinentes trouvées lors de l'étiquetage d'un document. Le taux de rappel consiste à mesurer l'efficacité d'un système d'étiquetage établie à partir du ratio entre le nombre d'étiquettes pertinentes spécifiées lors de l'étiquetage d'un document et le nombre total d'étiquettes pertinentes du document. C'est un indicateur de mesure du silence. Le silence est l'ensemble d'étiquettes pertinentes non spécifiées lors de l'étiquetage d'un document.

A = Items pertinents trouvés par le script ;
B = Items pertinents qui devaient être trouvés ;
C = Items trouvés par le script ;
Score de précision est A/B
Score de rappel est A/C
Score F-mesure f = (2*Précision*Rappel)/(Précision+Rappel)
Score de F-mesure est f = 2*(((A/B)*(A/C))/((A/B)+(A/C)))

Le calcul de la f-mesure relative à notre corpus d'entraînement nous permet d'élaborer le tableau récapitulatif suivant :

Thèmes/ taux (%)	**Score de précision**=Items pertinents trouvés	**Silence**	**Score de rappel** (silence)=Items qui devraient être trouvés	**Score f-mesure**
Accueil	0.69	0.02	0.9718300986	0.807017544
Chambre	0.83	0	1	0.907103825
hôtel	0.88	0.02	0.977777778	0.926315789

Thèmes/ taux (%)	**Score de précision**=Items pertinents trouvés	**Silence**	**Score de rappel** (silence)=Items qui devraient être trouvés	**Score f-mesure**
Restauration	0.94	0	1	0.969072165
Animation	0.32	0.02	0.941176471	0.47761194
Personnel	1	0	1	1
Services	0.86	0	1	0.924731183
Prestation positive	0.66	0	1	0.795180723
Prestation négative	0.8	0.01	0.987654321	0.883977901

Tableau 2 : Statistiques du corpus d'entraînement (Djerba Hotel A)

Dans le tableau des statistiques (Tableau 2), le thème *Animation* ressort comme un thème complètement défaillant puisqu'il a un score inférieur à 0,5 suivi par le thème *Prestation positive*. Les deux thèmes initialement repérés au niveau des occurrences s'avèrent les moins performants. Cela nous conduit à examiner de plus près leur f-score. Pour le thème *Animation*, nous nous rendons compte rapidement que le taux de précision s'élève uniquement à 32% alors que le bruit est nul. 68% des items retenus constituent des items partagés avec d'autres thèmes de l'étude. Nous avons l'exemple d'« animateurs » qui fait partie à la fois du thème *Animation* et *Personnel*.

Les spécificités du corpus d'un côté (familier, incorrect, pas de phrases, …) et l'intersection des thèmes managériaux dans un certain nombre d'items de l'autre côté peuvent expliquer les statistiques de nos résultats. Ces items partagés affectent directement le score de précision, qui affecte à son tour notre score F-mesure de l'efficacité de nos résultats. Les seconds résultats biaisés par un taux de précision faible à cause des items partagés sont relatifs au thème *Prestation positive*. Ce thème est fortement lié aux items des autres thèmes dans le sens où les internautes portent une évaluation sur un des thèmes évoqués.

6. Évaluation et enseignements

Pour évaluer notre étude, nous avons retenu les avis des clients relatifs à l'Hôtel B qui possède un nombre de commentaires relativement important. En appliquant la même démarche, les résultats qui en découlent font ressortir les thèmes *Prestation positive* et *Animation* comme des thèmes pertinents à l'étude du point de vue nombre d'occurrences. L'étude statistique relative au f-score et au corpus d'évaluation (Tableau 3) retient toujours le thème *Personnel* en termes de performance et montre les défaillances des thèmes *Prestation positive* (0,26) et *Animation* (0,24) avec des taux inférieurs à 0,5.

Thèmes/Taux (%)	**Score de précision**=Items pertinents trouvés	**Silence**	**Score de rappel** (silence)=Items qui devraient être trouvés	**Score f-mesure**
Accueil	0.65	0.01	0.984848485	0.78313253
Chambre	0.9	0.02	0.97826087	0.9375
hôtel	0.9	0	1	0.947368421
Restauration	0.92	0.03	0.968421053	0.943589744
Animation	0.14	0	1	0.245614035
Personnel	1	0	1	1
Services	0.84	0.01	0.988235294	0.908108108
Prestation positive	0.15	0	1	0.260869565
Prestation négative	0.89	0	1	0.941798942

Tableau 3 : Statistiques du corpus d'évaluation (Hôtel B)

L'évaluation a permis de confirmer les résultats de notre étude en retenant le thème *Personnel* comme le thème le plus pertinent et en montrant les lacunes des deux thèmes *Prestation positive* et *Animation*. Pour mieux illustrer notre propos, nous confrontons les deux corpus. À travers ce tableau, nous constatons que les items communs et partagés entre les thèmes affectent considérablement la fiabilité de l'étude (Tableau 4).

Thèmes/Corpus	**Hôtel A / Corpus d'entrainement**			**Hôtel B / Corpus d'évaluation**		
	Précision	Bruit	Partagé	Précision	Bruit	Partagé
Animation	32%	0%	68%	14%	0%	86%
Prestation positive	60%	4%	36%	15%	1%	84%

Tableau 4 : Confrontation des thèmes de deux corpus (d'entrainement (Hôtel A) et d'évaluation (Hôtel B))

En guise de conclusion, nous pouvons dire que cette analyse a permis de valider sept thèmes sur neuf et surtout de montrer l'ambiguïté et le chevauchement entre les thèmes choisis. Sur le plan managérial, ce travail a permis de valoriser l'interdisciplinarité et la nécessité de collaboration entre manager, linguiste et informaticien. Ce travail a le mérite d'avoir évoqué 3 volets spécialisés et non exploités dans ce créneau. Sur le plan linguistique et informatique, en recommandant la collaboration tripartite

entre manager, linguiste et informaticien, la présente étude a essayé de montrer l'intérêt de l'analyse linguistique, l'utilité du TAL et l'efficacité du Python et de Unitex pour servir des intérêts économiques et managériaux. Sur le plan méthodologique, nous avons proposé une démarche innovante d'analyse du discours en dépassant la simple analyse du contenu et en visant une analyse sémantique via la mobilisation des outils linguistiques intégrant des descripteurs syntactico-sémantiques.

Cependant, il est important de relever certaines limites relatives aux thèmes choisis, au corpus, aux graphes proposées et aux scripts informatiques et d'ouvrir ainsi des voies fécondes d'investigations futures. D'abord, un travail plus élaboré sur la problématique et les items choisis ne peut qu'enrichir les conclusions retenues. Nous recommandons par ailleurs, un corpus plus volumineux et, par conséquent, plus appropriée à la démarche suivie en étudiant l'intégralité des hôtels tunisiens. De plus, au niveau linguistique, l'élaboration de graphes plus fournies quant aux thèmes étudiés appuiera certainement les résultats de l'analyse discursive. Pour finir, au-delà du niveau managérial et linguistique, la présente étude gagnerait en intérêt en intégrant un script python (en utilisant la bibliothèque Polyglot) qui calcule automatiquement le score des polarités des avis clients et qui nous fournit son appréciation quant aux services proposés par l'hôtel.

Bibliographie

Management

Ben Ammar Mamlouk, Z. 2015. « Des passerelles sémantiques pour un changement de paradigmes », *Ambivalences*, Publications LARIME, ESSECT.

Cristol, D. 2015. « Ambivalence du management numérique », *Ambivalences*, Publications LARIME, ESSECT.

Isaac, H., Campoy, E., Kalika, M. 2007. « Surcharge informationnelle, urgence et TIC. L'effet temporel des technologies de l'information », *Management et avenir,* n°13, vol.3.

Lesca, H. 2015. « Ambivalences, contradictions, signaux faibles et opportunités », *Ambivalences*, Publications LARIME, ESSECT.

Linguistique

Buvet, P-A. 2009. « Quelles procédures d'étiquetage pour la gestion de l'information textuelle électronique ? », *L'information grammaticale*, 122, Peters.

Buvet, P-A. 2015. Sous-presse « Linguistique et intelligence ». In : Journée d'étude *Linguistique et...* (éd. Jan Goes, Salah Mejri et Olivier Soutet), Presses Universitaires d'Artois.

Moscarola, J. 2018. « Chapitre 9. Analyse de données textuelles, lexicales et sémantiques », *Faire parler les données. Méthodologies quantitatives et qualitatives*, sous la direction de Moscarola Jean. EMS Editions, p. 191-217.

Neveu, F. 2004. *Dictionnaire des sciences du langage*. Edition Armand Colin.

Informatique

Cetro, R. 2011. « Outils de traitement des langues et corpus spécialisés : l'exemple d'Unitex ». *Cahiers de recherche de l'École doctorale en linguistique française*, N° 5, 2011, p. 49-63.

Issac, F. 2009. « Place des ressources lexicales dans l'étiquetage morphosyntaxique ». *L'Information Grammaticale*, N° 122, p. 10-18.

Laporte, E. 2009. « Concordanciers et flexion automatique ». *Cahiers de Lexicologie*, Centre National de la Recherche Scientifique, N° 94, Vol. 1, p. 91-106.

Paumier, S. 2016. *Manuel d'utilisation : UNITEX 3.1*, Université Paris-Est Marne-la-Vallée, http://www-igm.univ-mlv.fr/~unitex [consulté le 15 janvier 2022].

Notes

1. https://www.tripadvisor.fr [consulté le 15 janvier 2022].
2. Pour des raisons de confidentialité, nous ne fournissons pas le nom de l'hôtel.