Du présentiel au distanciel : le TAL pour comparer image voulue et image perçue

Pauline Soutrenon

UNIVERSITÉ Grenoble Alpes

Du présentiel au distanciel : le TAL pour comparer image voulue et image perçue

Pauline Soutrenon

Sous la direction de Thomas Lebarbé
Tuteur entreprise : Stéphane Labartino

Laboratoire LITT&ARTS
UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de Master 2 Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2016-2017

Mis en page avec la classe thloria.

UNIVERSITÉ Grenoble Alpes

Du présentiel au distanciel : le TAL pour comparer image voulue et image perçue

Pauline Soutrenon

Sous la direction de Thomas Lebarbé
Tuteur entreprise : Stéphane Labartino

Laboratoire LITT&ARTS
UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de Master 2 Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2016-2017

Mis en page avec la classe thloria.

Remerciements

Je tiens d'abord à remercier Thomas Lebarbé pour avoir pris le temps de me suivre pendant ce stage, pour ses conseils, pour l'aide qu'il m'a apportée et toute sa bienveillance.

J'aimerais également remercier Stéphane Labartino pour m'avoir fait confiance sur son projet et pour m'avoir également suivie tout au long de ce stage.

Un grand merci à mes camarades de classe du Master Industries de la Langue. Tout particulièrement, Ali, Anne-Laure, Doriane, Louise et William pour leur soutien, leur aide qu'ils m'ont apportée.

Merci également à Matteo et Rémi pour leur soutien et leurs encouragements. Pour finir, je tiens à remercier ma mère, pour son soutien dans toutes les épreuves et moments difficiles, pour ses encouragements et pour les relectures.

UNIVERSITÉ Grenoble Alpes

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : **SOUTRENON**

PRENOM : **Pauline**

DATE : **2 juin 2017**

SIGNATURE : 

Table des matières

Table des figures

Liste des tableaux

Introduction

1 Objectifs du mémoire

Le mémoire que nous présentons est un projet de recherche en collaboration entre l'Université Grenoble Alpes et l'entreprise COMONGO. Cette entreprise assure un service de conseil en gestion d'image. Elle souhaite opérer une transition d'un modèle de fonctionnement en présentiel vers un modèle totalement distanciel, assisté par des outils de Traitement Automatique des Langues (TAL) pour identifier les similarités et dissemblances entre l'image voulue d'une personne physique ou morale et l'image perçue par un panel d'usagers.

L'objectif de ce travail est de donner une première approche du projet. Premièrement, nous avons récupéré les données à partir d'une interface de saisie que nous avons développé (prototype) de façon à obtenir un corpus analysable. Cette interface de saisie est également très importante pour l'entreprise puisque c'est un premier outil visible qui lui permet d'entamer sa démarche. Deuxièmement, nous avons effectué les premières analyses linguistiques du corpus en le traitant d'abord manuellement puis en effectuant une simulation par des ressources. Nous avons tenté d'identifier les ressources sémantiques nécessaires afin d'effectuer la simulation par les ressources sur notre corpus. Cela nous permet de voir si celles-ci seront suffisantes ou non pour notre projet. L'analyse consiste à repérer les informations utiles dans les énoncés (en utilisant la méthode de l'entreprise), à comparer les textes les uns par rapport aux autres et à identifier les points communs ou les points de divergence (en utilisant, par exemple, l'intersection lexicale, la proximité sémantique ou encore la polarité des opinions).

Ce mémoire ouvre les perspectives d'un travail de recherche doctorale dans le cadre d'une convention de collaboration.

Nous abordons dans un premier temps et dans la suite de cette introduction, la demande de transition de COMONGO, en nous focalisant sur l'entreprise, sur sa demande et sur la description du projet. Puis, nous nous consacrons dans le chapitre 1 page 21 aux hypothèses que le projet implique. Nous présentons ensuite la méthode mise en place dans le chapitre 2 page 35 et dans le chapitre 3 page 59, nous analysons les résultats attendus et obtenus sur l'étude β. Enfin, nous concluons sur notre travail et présentons les perspectives de ce mémoire et notamment la construction incrémentale des ressources.

2 Demande de transition de COMONGO

2.1 Présentation de l'entreprise et de son fonctionnement

La société COMONGO est une start-up créée il y a plus d'un an et demi (en juillet 2015) dont le président fondateur est Stéphane Labartino. Cette start-up accompagne les personnes physiques ou morales (par exemple, les entreprises) dans leur définition d'image en utilisant une méthode issue de la communication et éprouvée depuis plus de dix ans.

COMONGO peut être consultée par des clients dans le but d'évaluer leur image de marque et de l'améliorer. Par exemple, un client pourrait consulter COMONGO pour savoir si son image (ou par exemple, son domaine d'activité) est bien perçue par le grand public ou par des personnes dont l'avis compte pour lui (leaders d'opinions).

2.2 L'image de personne morale ou physique

L'image d'une personne morale ou physique correspond aux représentations qu'un public cible peut avoir d'une personne morale ou physique.

Selon l'INSEE (l'Institut National de la Statistique et des Etudes Economiques), au sens du droit français, une personne morale est « un groupement doté de la

personnalité juridique qui se compose d'un groupe de personnes physiques réunies pour accomplir quelque chose en commun ». Quant à la personne physique, il s'agit, toujours selon l'INSEE, d' « un être humain doté, en tant que tel, de la personnalité juridique ». Dans notre cas, une personne physique peut être une personnalité (un homme politique, un dirigeant d'entreprise, etc.), tandis qu'une personne morale peut être une grande entreprise ou une start-up qui souhaite savoir, par exemple, comment son domaine d'activité est perçu ou savoir si sa mission environnementale est bien perçue, ou encore avoir une perception de ses compétences.

L'image voulue d'une personne physique ou morale se distingue de l'image perçue. L'image voulue correspond à la façon dont la personne aimerait être perçue. Tandis que l'image perçue est toujours en fonction d'un public cible, la personne physique ou morale n'a pas le contrôle sur cette image. Le public cible a des opinions qui peuvent être positives ou négatives. Il y a une notion de polarité qui correspond à un certain discours et vocabulaire.

L'image a une grande valeur : c'est quelque chose de monnayable qu'il faut savoir estimer. L'image a également une valeur très volatile qui s'entretient et s'estime. Etudier une image et l'évaluer peut servir à remédier à des défauts d'image, ce qui demande des moyens. La création et l'évaluation d'une image d'une personne morale ou physique est un processus long (pouvant s'étaler sur des mois) et coûteux. Il y a donc un marché disponible pour le projet que COMONGO souhaite développer.

2.3 La démarche d'entreprise

La démarche de l'entreprise s'inscrit dans une demande de transition. L'entreprise souhaite que le processus d'aide à la définition d'image réalisé jusque là manuellement soit automatisé. Il s'agit donc de développer un outil sémantique d'aide à la définition d'image qui s'appuierait sur la méthode utilisée par le consultant (Stéphane Labartino).

Lorsqu'un client de COMONGO souhaite évaluer son image, il doit d'abord décrire au consultant son image voulue, la façon dont il aimerait être perçu. Les clients de COMONGO sont généralement des entreprises. Mais cette même méthode

peut s'appliquer à des hommes politiques, des entreprises nationales et internationales, etc.

Puis, un panel (des groupes de personnes, nommés *focus groups*, que nous décrivons plus tard) est constitué par le client et le cas échéant, complété par le consultant. La méthode se base ensuite sur un questionnaire qui est un ensemble de questions ouvertes (ou phrases inductives) : quatre ou cinq questions ouvertes sont posées au panel par le consultant (accompagnement humain). Les réponses à ces questions constituent l'image perçue.

Un bilan est ensuite fait au client avec une synthèse de l'image perçue avec des représentations visuelles.

2.4 Description du projet

Réalisation d'une application sémantique

Le projet global de l'entreprise COMONGO est de créer une application pour évaluer une image de marque d'une personne physique ou morale. Nous développons une application fondée sur la sémantique des sentiments afin d'analyser l'impact d'une image ou de la définir. Le premier but de cet outil est de pouvoir collecter du corpus pour ensuite l'analyser.

Notre démarche est empirique c'est-à-dire qu'elle est guidée par les données. Les données sur lesquelles nous travaillons sont principalement sous forme de questions/réponses (pour les sentiments perçus c'est-à-dire les réponses des publics cibles interrogés). Celles que nous obtenons avec l'outil numérique sont différentes de celles dont nous disposons actuellement puisque le format de saisie des réponses a changé.

Description des données

Il y a trois ensembles de données : l'image voulue de l'entreprise qui constitue l'attente du client, l'image perçue par les *focus groups* (c'est-à-dire les réponses aux phrases inductives) et la synthèse produite.

Plusieurs questions se posent quant à l'image voulue de l'entreprise. Il s'agit

tout d'abord de déterminer comment elle est conçue : c'est le client qui travaille sur son image voulue avec le consultant avant que les *focus groups* aient lieu. Concernant le volume des données, il s'agit de données plutôt petites (la taille des *focus groups* allant de 6 à 10 participants). Ces caractéristiques doivent être prises en compte dans la manière de traiter la donnée.

Les réponses obtenues sont des données recueillies dans des conditions réelles par des utilisateurs réels (les audités). Ces réponses permettent de créer l'image perçue. Les caractéristiques de l'image perçue sont semblables à celles de l'image voulue notamment au niveau de la volumétrie : il s'agit de petites données. Mais cela pose aussi d'autres questions telles que l'imperfection des données.

Suite aux *focus groups*, les données sont analysées dans le but de produire une synthèse (besoin de synthétiser). Cette synthèse de l'image perçue doit ensuite être comparée avec l'image voulue du client. Le fait de croiser ces deux textes permet d'en faire ressortir ce qui émerge le plus.

3 Domaines scientifiques abordés

Notre travail se situe à l'intersection de différents domaines scientifiques du TAL dont : l'*opinion mining*, les ressources sémantiques et les interfaces comme le montre la figure 1 page suivante.

3.1 *Opinion mining*, existant et usages

Avec les nombreux blogs, forums, réseaux sociaux, etc., l'*opinion mining* est un domaine du TAL très étudié : un numéro de la revue TAL [El-Bèze et al., 2010] est d'ailleurs consacré à cette thématique.

Lorsque l'on parle de l'*opinion mining* (qui signifie fouille d'opinions en français), nous nous apercevons qu'il n'y a pas vraiment de consensus sur l'utilisation de ce terme. En effet, plusieurs termes sont utilisés comme : détection d'opinions ou analyse de sentiments (*sentiment analysis*). Ces termes signifient la même chose et dans notre démarche, nous utiliserons le terme d'opinion qui fait référence à

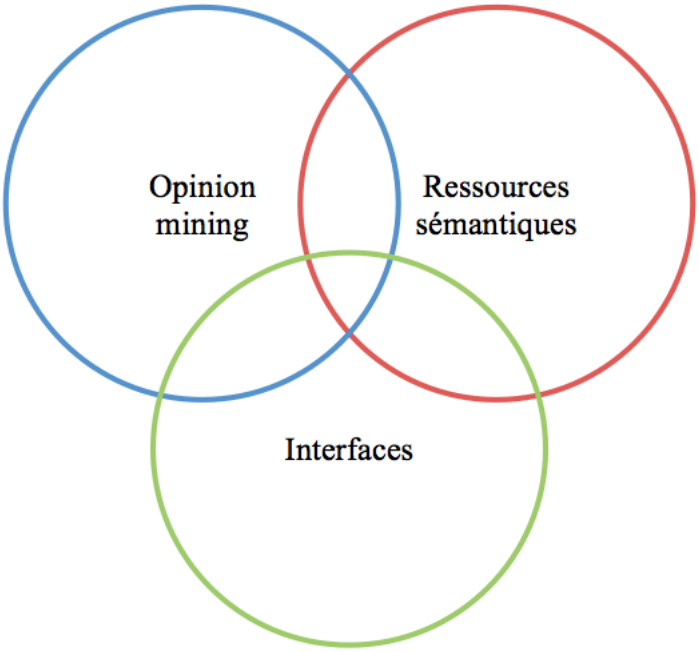


FIGURE 1 – Domaines scientifiques abordés

un jugement. Bien que, dans notre démarche, il s'agisse plutôt de l'expression du ressenti.

[Esuli & Sebastiani, 2006] définissent l'*opinion mining* comme une sous-discipline du TAL permettant de chercher une opinion dans un document :

> « Opinion mining (OM) is a recent subdiscipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses » [Esuli & Sebastiani, 2006].

Quant à [El-Bèze et al., 2010], l'*opinion mining* permet selon ces auteurs :

> « d'extraire et de catégoriser selon leur polarité les opinions ou les sentiments exprimés explicitement dans des textes. »

Nous utilisons le terme d'*opinion mining* comme le processus permettant de détecter des opinions à l'aide d'outils du TAL.

Concernant les usages de l'*opinion mining*, [Brun, 2011] évoque certains systèmes utilisant le *deep learning* pour détecter les opinions. Le *deep learning* est une méthode d'apprentissage automatique des données qui s'effectue sur du *big data* (qui signifie de grosses données). Il s'agit donc de données quantitatives. Par exemple, de très grandes quantités de données comportant une opinion peuvent être collectées sur internet et être utilisées par la suite par un système d'apprentissage automatique.

A la différence des grandes données utilisées pour faire de l'*opinion mining*, les données sur lesquelles nous travaillons dans notre projet de recherche sont petites (données qualitatives de petits groupes de personnes). Pour notre problématique, le *deep learning* n'est pas utile puisque nous travaillons sur de petites données.

De plus, dans le *big data*, nous pouvons détecter le bruit (les solutions non pertinentes trouvées par le système) mais pas le silence (les solutions pertinentes non trouvées). Or, il est intéressant et important, d'étudier le bruit et le silence sur nos données afin d'évaluer l'outil.

Pour notre projet, nous avons pu nous positionner par rapport à l'approche générale de l'*opinion mining* qui consiste à travailler sur de grandes quantités de données avec du *deep learning*. Dans notre projet, nous ferons de l'*opinion mining* mais sur de petites données (des données qualitatives) et c'est ce en quoi nous nous différencions de l'approche générale de l'*opinion mining*.

3.2 Ressources sémantiques, existant et usages

Selon [Gala & Brun, 2012], « les ressources lexicales sont cruciales pour de nombreuses applications de traitement automatique de la langue ». Par conséquent, il nous faut certainement utiliser des ressources sémantiques pour notre projet afin de détecter les opinions.

Nous nous demandons s'il est nécessaire d'utiliser une ressource sémantique existante ou s'il est préférable de constituer notre propre ressource pour l'analyse d'opinions.

Certaines ressources sémantiques existantes permettent de déterminer la polarité des opinions. Pour caractériser la connotation générale d'un texte, ou dans

notre cas, caractériser des mots ou suites de mots, il paraît nécessaire d'utiliser ou de constituer une ressource lexicale de polarité. Cela peut permettre, par exemple, la désambiguïsation.

Nous développons ici les ressources en anglais car ce sont les premières que nous avons trouvées. En effet, de nombreuses ressources sur la polarité des opinions sont en anglais. De plus, les ressources que nous avons trouvées en français sont détaillées dans la section 2.3.2.0 page 55 puisque notre corpus est uniquement en français pour le moment. Cependant, il est intéressant de connaître les ressources sémantiques en anglais pour les perspectives de notre travail. Nous serons amenée à travailler sur des données en anglais pour notre projet et ces ressources peuvent nous être utiles lorsque nous aurons des études à analyser en anglais.

Selon [Abdaoui et al., 2016], plusieurs méthodes existent pour construire ce type de ressources :

— manuellement avec des annotateurs qui assignent la polarité
— automatiquement en utilisant des dictionnaires
— automatiquement en utilisant des corpus (annotés ou non)

La polarité est assignée en donnant un score (valeur numérique) aux mots pour tenter de définir s'ils sont plutôt positifs, neutres ou négatifs. C'est le cas par exemple de **Sentiwordnet**[1] [Esuli & Sebastiani, 2006] qui est une ressource lexicale libre pour l'*opinion mining*. Les synsets (ensemble de synonymes reliés entre eux par des relations sémantiques) de Wordnet (G. Miller & C. Fellbaum) sont utilisés et associés à des scores positifs, neutres ou négatifs. Le tableau 2 page suivante, donne des exemples de synsets issus de la ressource Sentiwordnet associés à leurs scores de positivité (première colonne du tableau) et de négativité (seconde colonne). Les scores vont de 0 à 1. Si les deux scores sont égaux à 0 alors il s'agit d'un synset neutre. Une définition des différents synsets est également donnée et l'information #1 présente après chaque synset indique qu'il s'agit ici, pour tous les termes, de leur sens premier.

Nous pouvons constater que les synsets *paleogeography*, *musicology*, *lexicology* ou *phonology* sont indiqués comme étant des termes neutres (les scores de positivité

1. `http://sentiwordnet.isti.cnr.it` (consulté le 27/04/2017)

POS	NEG	Synset	Définition
0.25	0	ontology#1	(computer science) a rigorous and exhaustive organization of some knowledge domain [...]
0.25	0	cosmology#1	the metaphysical study of the origin and nature of the universe
0.125	0	theology#1	the rational and systematic study of religion [...]
0	0	paleogeography#1	the study of the geography of ancient times [...]
0	0	musicology#1	the scholarly and scientific study of music
0	0	lexicology#1	the branch of linguistics that studies the lexical component of language
0	0	phonology#1	the study of the sound system of a given language [...]
0	0.125	paleoclimatology#1	the study of the climate of past ages

FIGURE 2 – Exemples de Sentiworndet

et négativité sont tous deux égaux à 0). En revanche, les trois synsets *ontology*, *cosmology* et *theology* sont indiqués comme étant des termes positifs avec un score de 0.25 et 0.125 pour *theology*. Le dernier synset, *paleoclimatology*, est lui, indiqué comme négatif.

Parmi tous ces noms de disciplines, nous constatons que les scores varient (avec des synsets neutres ainsi que des synsets plus ou moins positifs et négatifs). Nous pouvons nous demander comment ces scores sont attribués et en quoi, par exemple, le synset *lexicology* serait neutre tandis que le synset *ontology* serait positif, ou en quoi, *cosmology* serait 0.125 plus positif que *theology*.

Cette ressource pose une question essentielle pour l'*opinion mining* : dans quelle mesure un terme peut-il être positif, négatif ou neutre ? Une part de subjectivité peut influencer l'attribution à un score et il semble difficile d'attribuer une polarité.

Nous nous sommes également aperçue que Sentiwordnet dépend d'une interprétation contextuelle. Des termes jugés comme négatifs dans un domaine, pourraient, en fonction d'un autre domaine, s'avérer neutres voire positifs, c'est-à-dire

que certains termes peuvent avoir une polarité différente en fonction du domaine.

D'autres auteurs ont également travaillé sur la polarité [Agarwal & Bhattacharyya, 2006] et [Vegnaduzzo, 2004] mais en se concentrant plus particulièrement sur les adjectifs. [Agarwal & Bhattacharyya, 2006] proposent d'ajouter un nouveau lien dans Wordnet (G. Miller & C. Fellbaum) avec un score de polarité. Tandis que [Vegnaduzzo, 2004] a travaillé sur les adjectifs subjectifs pour créer une ressource lexicale en utilisant le moins possible d'autres ressources :

> « The focus of this work is to investigate the possibility of learning useful resources while at the same time reducing to a minimum the use of knowledge-based resources like annotated data and preprocessing tools. » [Vegnaduzzo, 2004]

Travailler seulement sur la polarité peut ne pas être suffisant, certains auteurs annotent les émotions. C'est le cas de [Mohammad & Turney, 2013] qui ont construit la ressource **EmoLex** en anglais par *crowdsourcing*. En plus de la polarité, des émotions sont associées aux mots parmi les six émotions basiques de [Ekman, 1992] (la colère, la peur, la surprise, la tristesse, la joie et le dégoût) auxquelles ont été ajoutées deux autres émotions : l'anticipation et la confiance.

Au lieu de cela ou en parallèle, s'intéresser à la similarité sémantique peut permettre de traiter les termes proches ayant la même portée. Dans ce cas, les synsets de Wordnet (G. Miller & C. Fellbaum) peuvent être adaptés.

Il existe également d'autres approches que nous devons considérer comme les thesaurus distributionnels. Ce sont des outils utilisés dans le domaine de la recherche d'information et qui peuvent être applicables à notre sujet concernant la recherche d'opinions dans un texte (*opinion mining*) :

> « La sémantique distributionnelle a pour objet de construire des thésaurus (ou lexiques) automatiquement à partir de corpus de textes. Pour une entrée donnée (ie. un mot donné), ces thésaurus recensent des mots sémantiquement proches en s'appuyant sur l'hypothèse qu'ils partagent une distribution similaire au mot d'entrée. » [Claveau & Kijak, 2015].

Pour notre projet, nous avons besoin d'utiliser les deux types de ressources sémantiques que nous avons présentés. Combiner un travail sur la polarité et sur la similarité sémantique s'avère utile.

Bien que la ressource Sentiwordnet soit intéressante pour détecter la polarité d'une opinion dans un texte, il faut être vigilant quant à son utilisation et peut-être réfléchir à un système de listes d'exclusion. Les listes d'exclusion peuvent, par exemple, permettre de ne pas utiliser tel terme avec un score négatif s'il s'avère positif dans un domaine donné.

Cette ressource est en anglais, or, dans notre démarche avec l'entreprise et dans le cadre du mémoire, nous avons d'abord travaillé sur le français. Nous ne pouvons donc pas utiliser tout de suite Sentiwordnet. Une des solutions est de constituer notre propre ressource, spécifique au domaine dans lequel nous travaillons.

3.3 Interfaces, existant et usages

Dans ce projet, la question des usagers et de la littératie numérique est à prendre en compte.

Il y a deux catégories d'usagers : celui qui fournit les données et celui qui les traite. Par conséquent, il y a différents niveaux de compétence. La question des usagers est très importante, il est nécessaire de savoir à qui nous nous adressons afin de fournir un environnement adapté.

Les usagers n'ont pas tous le même rapport aux outils numériques. C'est ici que la littératie numérique intervient. Selon l' [OCDE, 2000], la littératie est : « l'aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités ». La littératie numérique peut donc être définie comme la compétence de chaque usager à utiliser un outil numérique. Lors de la réalisation de l'interface, nous avons essayé de prendre en en compte cela et de voir l'acceptabilité (si l'outil est accepté par les usagers) et l'utilisabilité (comment l'outil est utilisé).

Pour accompagner les différents types d'utilisateurs sur l'interface (consultant, client, audités) il est préférable de mettre en place un petit tutoriel afin de guider

le client dans sa démarche de création d'étude (sujet, image voulue, personnes à auditer), ou de guider les audités dans le questionnaire.

Les théories de l'attention sont également à prendre en compte. [Carr, 2012] évoque un problème de lecture avec l'utilisation d'internet : plus nous utilisons internet et moins nous sommes attentifs sur de longs textes et il devient difficile de se concentrer sur une seule chose.

Pour notre projet, nous devons proposer un outil simple, intuitif et centré sur le sujet puisqu'il est utilisé par différents types d'utilisateurs (le consultant, les clients et les audités) ayant des niveaux de compétences différents.

Le fait de changer de démarche en passant par une interface numérique a des conséquences sur la qualité des données produites. En effet, d'après la théorie de l'attention, il est difficile de rester concentré sur une seule chose avec les outils numériques ou avec internet.

Chapitre 1

Hypothèses

1.1 Hypothèse 1 : d'une pratique présentielle à une pratique distancielle numérique

La démarche de l'entreprise s'inscrit dans une démarche de transition. COMONGO a une pratique actuelle qui est présentielle et accompagnée par un consultant. L'entreprise souhaite garder cette pratique mais changer de modalités par les outils numériques et par l'accompagnement humain.

Deux modalités vont changer ce qui nous pousse à réaliser la transition en deux étapes pour étudier les changements de la première modalité (les outils numériques) puis de la seconde (l'accompagnement humain).

La première étape est de passer par un outil numérique tout en gardant l'accompagnement humain du consultant (pratique présentielle numérique). La seconde étape est de passer en distanciel sans accompagnement humain et toujours avec l'outil numérique de la première étape (pratique distancielle numérique).

La demande de l'entreprise amène à une première hypothèse forte : nous pouvons faire une transition d'une pratique présentielle pour aller vers une pratique distancielle numérique. Passer d'une pratique de *focus groups* à une pratique numérique sous-tend que les productions et leurs analyses vont varier.

1.1.1 Description des trois pratiques

Pratique actuelle : présentiel papier

La pratique actuelle de COMONGO, dans l'accompagnement de la définition d'image d'un client, s'effectue en interrogeant des publics visés. C'est un processus qui passe généralement par le biais de *focus groups* (mais qui peut se passer par entretiens téléphoniques individuels si les audités ne peuvent pas se déplacer). Un *focus group* est un petit groupe de personnes (les audités) constituant un public cible. La figure 1.1 montre la pratique actuelle. Dans un premier temps, le client décrit et construit son image voulue avec le consultant. Dans un deuxième temps, des *focus groups* sont réalisés (production des réponses) afin d'obtenir l'image perçue. Enfin, une analyse est effectuée et le bilan est ensuite restitué au client.

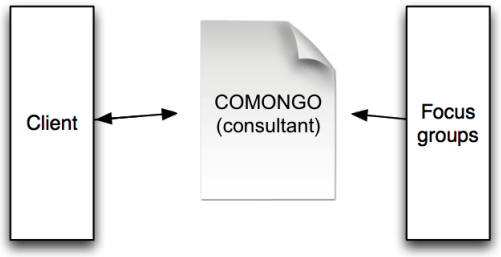


FIGURE 1.1 – Schéma de la pratique actuelle : présentiel papier

Pour une étude de définition d'image, des *focus groups* homogènes de 6 à 10 participants (par exemple, en fonction des tranches d'âge ou du sexe) sont constitués soit par le client, qui choisit des personnes dont l'avis compte pour lui, soit par le consultant, lorsque par exemple, le client n'a pas assez de personnes à auditer. Dans le cas où le sujet s'adresse plutôt au grand public, c'est le consultant qui constitue lui-même les *focus groups*.

Les *focus groups* se déroulent lors de séances de deux heures (ou plus) en présentiel avec un système de questions/réponses. Pendant ces séances, quatre ou cinq questions ouvertes (ou phrases inductives) sont posées. Les types de questions sont toujours les mêmes et ces phrases sont réutilisées pour les différentes études (nous reviendrons sur les types de phrases inductives dans la section 2.1.1.0).

Avant de commencer, le consultant présente le déroulement de la séance au *focus groups*. Il fait également signer un accord de confidentialité précisant que la participation des audités a pour objet d'échanger et de recueillir leur opinion personnelle sur le sujet qui va leur être présenté. Le sujet revêt un caractère confidentiel, les interrogés doivent donc s'engager à ne divulguer à quiconque les informations échangées lors de la séance.

Le consultant précise également que quatre ou cinq questions (cela dépend du sujet) seront posées sous la forme d'une phrase inductive. Les audités devront la terminer avec tout ce qui leur vient à l'esprit. Si rien ne vient à l'esprit, « *rien* » peut constituer une réponse.

Afin que l'étude et les résultats soient précis, la méthode doit empêcher tout type d'influence. Il est donc demandé aux audités de ne pas interagir avec les autres participants pendant la session de questions/réponses. Les participants ont un temps d'échange à la fin de la séance lors du recueil des données (le consultant note les réponses au fur et à mesure et les affiche sur un écran).

Pour finir, le consultant informe les audités que les réponses seront collectées de manière anonyme (l'identité des audités n'est pas inscrite lors de la collecte des réponses) : le client demandeur de l'étude n'aura pas accès aux réponses individuelles. Néanmoins, s'il le souhaite, il pourra voir les réponses brutes, telles qu'elles auront été saisies à l'écran.

Le processus des *focus groups* se déroule globalement de la façon suivante :

- Pour chaque question :
  - Question orale
  - Réponse des audités
  - Une fois que chacun a répondu, la question suivante est posée
  - Répétition de ce processus jusqu'à la dernière question
- Une fois que les audités ont répondu à toutes les questions, pour chaque question :
  - Pour chaque audité :
    - Restitution d'une idée
    - L'idée est notée par le consultant

— Répétition du processus jusqu'à épuisement des idées

Nous décrivons ce processus en détails.

- Pour chaque question :
  - La question est posée oralement par le consultant et elle est visible (projetée sur un écran).
  - Les audités doivent ensuite répondre sur papier libre. Les réponses sont personnelles (le papier sur lequel les réponses sont écrites n'est pas conservé à la fin de la séance) et chacun peut noter ce qu'il souhaite (des mots, des phrases). Il s'agit donc de texte libre dont la forme peut varier d'une personne à une autre (liste de mots, phrases, etc.).
  - Une fois que chacun a terminé de noter ses réponses, le consultant présente à l'oral et en projection la seconde question (phrase inductive).
  - Et ceci, jusqu'à la dernière question.
- Une fois que les audités ont répondu à toutes les questions à l'écrit et individuellement, pour chaque question :
  - Pour chaque audité :
  - Les audités restituent à voix haute et chacun à leur tour une idée notée en réponse à la question.
  - Les idées sont notées par le consultant dans le document qui est projeté et donc visible par tous. Parfois, il est demandé de préciser les réponses pour bien saisir ce que l'audité à voulu exprimer. Si une idée a déjà été évoquée, le consultant note qu'elle a été évoquée plus d'une fois (et autant de fois qu'elle a été évoquée).
  - On répète ce processus jusqu'à épuisement des idées et pour chacune des quatre ou cinq questions.

Pratique transitionnelle : numérique présentiel

Nous avons vu précédemment que la transition s'effectue en deux étapes, nous avons décrit la pratique actuelle (à gauche du schéma 1.2 page suivante) et la pratique voulue (à droite du schéma 1.2 page ci-contre) est décrite dans la partie suivante.

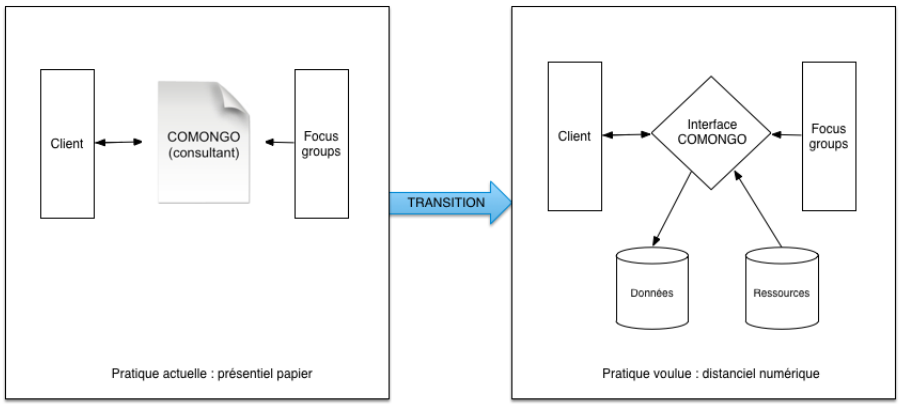


FIGURE 1.2 – Schéma de transition : de la pratique actuelle vers la pratique voulue

La première étape de transition - numérique présentiel - vise à faire utiliser aux audités un outil numérique en séance présentielle, c'est-à-dire lors des *focus groups* avec un accompagnement humain de la part du consultant. Nous avons donc développé un premier outil, une interface numérique de saisie, qui permet aux audités de saisir leurs réponses aux questions (il s'agit d'un questionnaire).

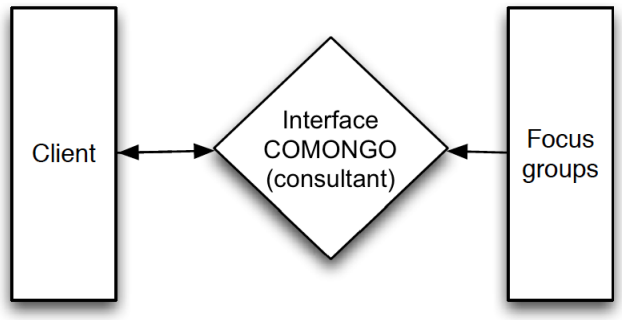


FIGURE 1.3 – Schéma de la pratique de transition : numérique présentiel

Par conséquent, la démarche reste la même, le changement réside dans le fait que les audités utilisent un outil numérique au lieu d'écrire leurs réponses sur papier. Bien que le consultant soit présent (pour donner des explications ou répondre aux éventuelles questions des audités), un tutoriel concernant l'accord de confiden-

tialité, l'anonymat des réponses et le déroulement du questionnaire a été intégré à l'interface de saisie avant de commencer le questionnaire (la première page de ce tutoriel se trouve en Annexe B.1 page 90). En effet, l'outil doit contenir ce que le consultant explique à propos de la garantie de l'anonymat ainsi que de l'accord de confidentialité. Il fallait également donner une explication quant au déroulement du questionnaire. Après plusieurs tests sur des études fictives avec différents audités en modalité individuelle, nous nous sommes rendu compte que la présence du tutoriel était très importante. Cela permet notamment de rassurer les audités sur la garantie de l'anonymat.

Cet outil nous a permis de recueillir des données et de les constituer en corpus. C'est un premier objet d'étude et d'analyse linguistique pour identifier les besoins en ressources linguistiques et les mécanismes numériques nécessaires à la réalisation du projet. Cela nous a également permis de voir comment les audités se servent de l'outil mais aussi de nous rendre compte de la question de la variation dans les productions. Cette variation est d'abord due à l'impact numérique lors de la transition de la pratique présentielle papier à la pratique numérique présentielle. Puis, elle est due à l'impact d'une pratique individuelle lors de la transition de la pratique numérique présentielle à la pratique numérique distancielle en individuel.

La seconde étape est de passer au numérique distanciel. L'outil sera utilisé en distanciel sans accompagnement humain, nous allons à présent décrire cette pratique voulue.

Pratique voulue : numérique distanciel

COMONGO voudrait à présent passer par une démarche utilisant un outil numérique en distanciel. La demande de transition est motivée par un besoin réel de l'entreprise. L'entreprise a besoin, dans sa pratique, d'intégrer une intelligence numérique afin de gagner du temps et de faciliter le processus d'aide à la définition d'une image d'un client. Par conséquent, elle veut pouvoir disposer d'un outil numérique simple, fiable, interactif et disponible en ligne. Cet outil sera mis à disposition du consultant, des clients et des audités. Cela pourra également permettre à l'entreprise d'étendre sa clientèle (en s'adressant à une personne physique, à des

annonceurs ou encore à des agences).

Le processus global de fonctionnement de l'outil est pour l'instant décrit de la façon suivante :

— Définitions d'un sujet et de l'image voulue par le client
— Génération d'un questionnaire
— Envoi du questionnaire
— Récupération des réponses
— Elaboration d'une synthèse de l'image perçue
— Comparaison de l'image perçue et de l'image voulue
— Préconisation au client des axes de communication

Nous décrivons ce processus plus en détails :

— Un client définit un sujet, son image voulue et renseigne les adresses e-mails des personnes dont l'avis compte pour lui (ce sont les audités). Il renseigne le nombre de *focus groups*, le nombre de participants par groupe et des groupes homogènes sont constitués.
— Un questionnaire (sur le sujet précisé par le client) est ensuite généré avec les quatre ou cinq questions ouvertes issues de la méthode du consultant. Ce questionnaire est envoyé aux audités qui doivent le remplir. Leurs réponses aux questions constituent l'image perçue du client.
— Après avoir récupéré les données que constituent les réponses, une synthèse de l'image perçue est créée automatiquement.
— Puis, cette synthèse de l'image perçue est comparée à celle de l'image voulue, rédigée préalablement par le client. Le but de cette comparaison étant d'obtenir les éléments qui se ressemblent et ceux qui se différencient. Cette comparaison des deux synthèses peut être représentée dans le but de la présenter au client et de lui donner un retour : sous forme de diagrammes de Venn (facilement compréhensibles pour comparer deux choses) ou de nuages de mots. Nous pourrions laisser un choix entre plusieurs représentations au client.
— L'outil peut enfin préconiser au client des axes de communication avec les changements à opérer.

Contrairement à la pratique actuelle, la pratique voulue ne se fera plus en présentiel lors de *focus groups* et passera par le biais d'un outil numérique (interface de saisie), comme le montre le schéma 1.4.

Les questionnaires seront directement envoyés aux adresses e-mails des audités et ils pourront ainsi répondre au questionnaire à distance. Au lieu d'écrire leurs réponses sur papier, les audités entreront directement leurs réponses dans l'interface.

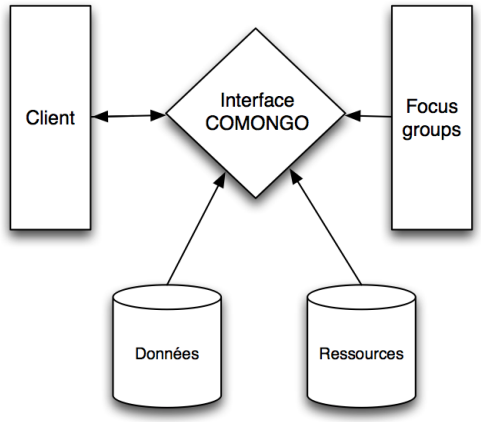


FIGURE 1.4 – Schéma de la pratique voulue : numérique distanciel

1.1.2 Comparaison des trois pratiques

Les trois situations présentent des modalités différentes, nous comparons (observations qualitatives) ces trois situations (états) qui constituent trois ensembles de données (corpus) différents.

Présentiel papier

Comme nous l'avons décrit précédemment, la pratique actuelle est en présentiel et les audités répondent librement sur du papier.

Le cadre de la production des réponses est contraint : le consultant est présent pour donner des explications, en séance de *focus groups* il peut y avoir un effet de

groupe, le temps et le moment des séances sont également contraints. Une séance dure environ deux heures sur un horaire fixé au préalable.

Les réponses sont personnelles et après restitution, le papier sur lequel les réponses ont été notées n'est pas conservé. Les audités peuvent donc répondre sans avoir peur d'être jugé, par exemple, sur leur orthographe ou leur écriture.

Présentiel numérique

Le passage du présentiel papier au présentiel numérique s'effectue par le biais d'une interface numérique sur laquelle les audités saisissent directement leurs réponses. Il s'agit là, de la première modification de la pratique actuelle.

Le cadre reste le même : une séance de deux heures environ avec l'accompagnement du consultant et la possibilité de répondre au questionnaire en *focus groups* (donc toujours un possible effet de groupe). Il s'agit donc toujours du même cadre contraint.

Les réponses sont toujours personnelles, cependant, elles sont conservées (enregistrées dans la Base de Données) mais anonymisées. Il s'agit là, de la seconde modification par rapport à la pratique actuelle.

Les deux changements évoqués ci-dessus auront certainement des conséquences sur la donnée produite qu'il nous faudra prendre en compte. Premièrement, le changement de format (passage au numérique par le biais de l'interface) risque d'influencer l'attention des utilisateurs et cela pourra avoir un impact sur la qualité des réponses. Deuxièmement, le fait de répondre sur une interface numérique et ensuite d'envoyer les réponses peut poser des problèmes à certains audités. De plus, les réponses sont conservées. C'est pourquoi, il nous a été nécessaire de garantir l'anonymat des audités pour qu'ils puissent donner des réponses sans se sentir jugés. Tout comme l'anonymat était garantit dans la pratique actuelle, lorsque les réponses sont notées par le consultant lors de la restitution, l'identité des audites n'apparaît pas.

Nous avons mis en place un tutoriel avant chaque questionnaire permettant de rassurer l'audité sur l'anonymat de ses réponses.

Distanciel numérique

Une fois que nous aurons testé l'outil en présentiel lors des *focus groups*, il s'agira de passer en distanciel. Nous allons donc réutiliser notre interface pour que les audités puissent saisir leurs réponses.

Lors du passage du numérique présentiel au numérique distanciel, d'autres choses vont varier. La majeure modification réside dans le fait que le cadre ne sera plus contraint comme il l'était dans la pratique actuelle et transitionnelle. En effet, concernant la pratique voulue, les utilisateurs n'auront plus à assister à des séances de *focus group* et pourront répondre aux questions à distance. Cela suppose que le consultant ne sera plus présent pour accompagner les audités, il n'y aura plus d'effet de groupe (les audités répondront au questionnaire de manière individuelle). Il n'y aura également plus de contrôle du temps (c'est-à-dire combien de temps les audités vont prendre pour répondre au questionnaire) ni du moment (c'est-à-dire à quel moment les audités répondront au questionnaire : à la maison, au bureau, dans les transports en commun, voiture, etc.).

Les réponses sont toujours personnelles et conservées de façon anonyme dans notre Base de Données. En effet, les réponses sont associées à un identifiant mais ne sont pas associées à l'identité des audités.

D'autres problématiques vont se poser avec le changement de cadre. C'est pourquoi la question de l'ergonomie sera très importante en distanciel. L'audité se retrouvera seul face à l'outil et il faudra donc que cet outil soit simple d'utilisation et très intuitif. Le cadre non contraint pourra avoir une influence sur les données notamment si l'audité n'est pas attentif. Le cadre contraint et le déroulement spécifique des séances de *focus groups* obligent les audités à être attentifs à la question et à prendre le temps de répondre. Une question va alors se poser en distanciel : comment faire pour garder l'attention des audités tout au long du questionnaire et comment faire pour qu'ils prennent suffisamment de temps pour répondre ? Pour essayer de pallier cela et le fait que le consultant ne sera plus présent, nous avons prévu de garder le tutoriel mis en place pour la pratique transitionnelle et de l'améliorer.

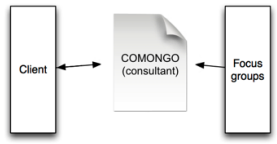
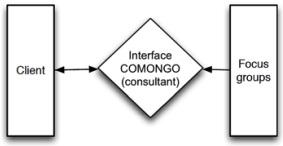
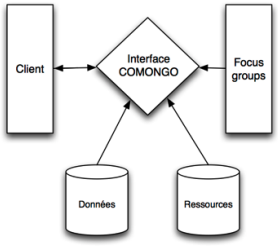
<table>
<tr><th>Pratique actuelle :
présentiel papier</th><th>Pratique transitionnelle :
présentiel numérique</th><th>Pratique voulue :
distanciel numérique</th></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>Etat initial</td><td>Etat transitionnel : numérique</td><td>Etat final : distanciel individuel</td></tr>
<tr><td>Réponses personnelles
Papier non conservé</td><td colspan="2">Réponses personnelles
Données conservées
Données anonymisées
Données analysables</td></tr>
<tr><td>Réponses libres sur papier</td><td colspan="2">Réponses sur l'interface numérique</td></tr>
<tr><td colspan="2">Cadre contraint :
- Pression du consultant
- Effet de groupe en focus group
- Temps et moment contraints</td><td>Cadre non contraint :
- Pas de pression du consultant
- Pas d'effet de groupe (individuel)
- Temps et moment non contraints
(pas de contrôle)</td></tr>
</table>

TABLE 1.1 – Comparaison des trois pratiques

La comparaison des trois pratiques est présentée dans le tableau 1.1. Les changements d'une pratique à une autre ont été mis en couleurs (il s'agit d'une liste non exhaustive, d'autres changements auxquels nous n'avons pas encore pensé peuvent survenir). Pour résumer, les deux changements majeurs de la pratique actuelle à la pratique transitionnelle apparaissent en vert, à savoir : les réponses sur l'interface et le fait que les réponses soient conservées mais anonymisées. En rose apparaissent les changements majeurs de la pratique transitionnelle vers la pratique voulue, à savoir : le cadre de la production qui n'est plus contraint, il n'y aura plus de pression du consultant ni d'effet de groupe et nous n'aurons plus de contrôle sur les conditions dans lesquelles les audités répondront au questionnaire (le temps et le moment).

1.2 Hypothèse 2 : utilisation du TAL dans l'aide à l'analyse des productions

A partir du moment où l'on peut récupérer de façon numérique les productions des audités, c'est-à-dire leurs réponses, les outils du TAL vont nous permettre d'accompagner la démarche de conseils. L'outillage TAL de la démarche est fondé sur un principe où on fait le travail d'un point de vue humain pour des principes que l'on pourrait mécaniser.

Il nous faut comparer les données produites dans le cadre de questionnement de l'entreprise, pour un client donné et pour une question donnée, il y a plusieurs réponses de plusieurs personnes donc il faut les comparer. Les outils du TAL vont nous aider à effectuer les comparaisons entre les réponses mais vont également nous aider à comparer l'image voulue à l'image perçue. Dans quelle mesure et quelle fiabilité ?

Après avoir observé nos données, les outils du TAL peuvent nous aider sur trois niveaux d'analyse que nous avons identifiés comme nécessaires. Nous développons ces différents niveaux, qui sont trois aspects différents, dans les trois sections suivantes et qui amènent la question des ressources sémantiques à utiliser.

1.3 Hypothèse 3 : intersection lexicale

Le premier niveau sur lequel nous avons travaillé avec notre corpus concerne l'intersection lexicale. En effet, nous avons remarqué, en étudiant les données, que des termes identiques apparaissaient entre les réponses des audités aux mêmes questions. Identifier les mêmes termes permettrait de dire si ce sont les mêmes propos ou non et donc d'identifier des points de recouvrement, exprimant des idées similaires et qui sont interessantes pour obtenir l'image perçue.

Nous posons donc notre troisième hypothèse de l'intersection lexicale. Il serait possible de déterminer les points de rencontre dans les productions des audités à l'aide des termes identiques. Nous nous sommes demandée s'il valait mieux utiliser les mots en commun, les lemmes en commun ou les racines en commun. Il faut

peut-être graduer l'intensité TAL en fonction de ce que l'on souhaite obtenir.

1.3.1 Lexique commun

Nous entendons par lexique une suite de caractères située entre deux espaces. Si nous utilisons cette approche, nous entendons par lexique commun les termes strictement identiques au niveau du genre, nombre, personne et mode. Des termes tels que *administratif* et *administrative* ayant une forme différente mais constituant la même entrée lexicale, ne seront pas regroupés dans la même catégorie. Or, ceci nous pose un problème puisque nous souhaitons que ces termes soient regoupés dans la même catégorie. C'est pourquoi nous n'avons pas utilisé cette approche.

1.3.2 Lemmes communs

La seconde approche consiste à utiliser les lemmes communs. Un lemme peut avoir plusieurs formes, ce qui nous permettrait de prendre en compte les termes tels que *administratif* et *administrative* afin de les indiquer comme termes identiques. Cela nous permet de ne pas tenir compte du genre, du nombre, de la personne ou du mode pour un terme donné. Pour un verbe par exemple, nous prenons en compte toutes les formes fléchies de l'infinitif. Nous avons choisi de travailler sur cette approche dans l'annotation de notre étude.

1.3.3 Racines communes

La troisième approche consiste à nous intéresser aux racines communes. Nous entendons par racine la forme commune à différentes variantes telles que les dérivations morphologiques. Cependant, avoir une racine commune ne veut pas dire que la sémantique est dans la même direction. C'est pourquoi nous n'avons pas utilisé cette approche. Par exemple, pour les termes *intéressant*, *inintéressant* et *intérêt* qui ont la même racine, *intéressant* et *inintéressant* seront regroupés alors qu'ils sont antonymes et qu'ils évoquent des idées opposées.

1.4 Hypothèse 4 : proximité sémantique

Le second niveau sur lequel nous avons travaillé est celui de la proximité sémantique. Pour aborder la proximité sémantique, nous utilisons le terme d'éléments sémantiquement proches ou de concepts similaires. Il s'agit de suites contiguës de un ou plusieurs mots (séquences ou segments). Etant donné qu'il n'y a pas de consensus sur le terme « mot », nous considérons comme mot la suite de caractères séparée par des espaces.

Il s'agit de notre quatrième hypothèse, nous pensons que la proximité sémantique, tout comme l'inersection lexicale, peut nous aider à trouver dans notre corpus des concepts qui se rapprochent et qui divergent pour effectuer des comparaisons.

1.5 Hypothèse 5 : polarité des opinions

Le troisième et dernier niveau qui nous intéresse est la polarité des opinions. Nous utilisons le terme d'éléments (ou termes) de polarité positive ou négative. Comme nous l'avons vu dans la section 0.3.2 page 15, repérer la polarité des opinions dans les textes produits paraît essentiel afin de comparer l'image souhaitée et l'image perçue.

Nous posons donc comme cinquième hypothèse, qu'il nous faut reconnaître la polarité des opinions afin de trouver des points de recouvrement et de divergence dans les réponses pour pouvoir effectuer la comparaison entre l'image souhaitée et l'image perçue.

Deux questions se posent concernant l'intersection lexicale, la proximité sémantique et la polarité des opinions : pouvons-nous trouver une ressource suffisante pour percevoir la polarité en français sur notre corpus et est-ce que cela peut nous permettre de comparer, voire de synthétiser ?

Chapitre 2

Méthode

2.1 Constitution du corpus

2.1.1 Construction de l'outil numérique

La construction de l'outil numérique vient de deux besoins : celui de l'entreprise qui souhaite distancialiser une pratique et celui du chercheur de produire des données. Cet outil d'analyse participe au fonctionnement de l'entreprise (besoin de l'entreprise) mais permet également de collecter un corpus (besoin pour le travail de recherche).

Afin de répondre à ces deux besoins, nous avons mis en place un outil composé de deux interfaces : une interface de gestion (permettant de gérer la création et la clôturation des études et permettant également d'accéder aux résultats d'une étude) et une interface de production (questionnaires proposés aux audités permettant de récupérer leurs réponses). La figure 2.1 page suivante présente un diagramme synthétique de l'outil que nous avons mis en place avec les différents usagers (le consultant et le panel d'audités) et les deux interfaces (en haut l'interface de gestion et à droite l'interface de production).

Nous avons développé cet outil dans le but d'avoir un contrôle sur l'outil et d'avoir une donnée proprement structurée. Nous nous sommes donc posée la question de l'encodage des données.

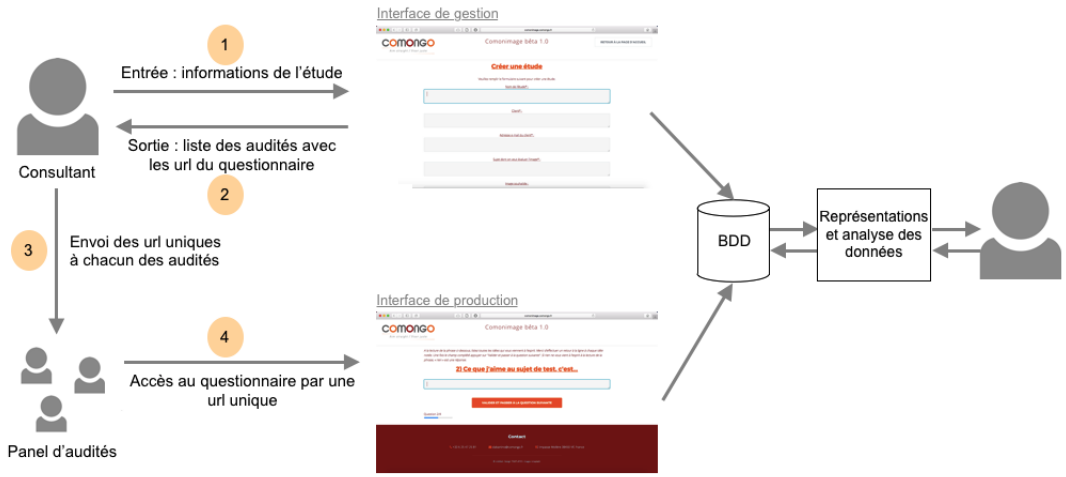


FIGURE 2.1 – Représentation schématique du fonctionnement de l'outil

Globalement, l'outil fonctionne selon les étapes suivantes :

1. Sur l'interface de gestion, le consultant crée une étude en entrant toutes les informations nécessaires dans un formulaire. Ces informations sont enregistrées dans la Base de Données (BDD).

2. L'interface de gestion donne en sortie la liste des audités avec pour chacun, une url unique vers le questionnaire.

3. Le consultant se charge ensuite d'envoyer un mail à chacun des audités avec l'url correspondante.

4. Les audités reçoivent chacun un mail avec une url unique renvoyant vers le questionnaire. Leurs réponses sont enregistrées dans la Base de Données (BDD).

Cet outil nous a permis de tester la pratique de transition (outil et accompagnement humain), de générer les différentes données (structurées dans une Base de Données) et de les stocker (fonctionnalité d'export et d'archive au format XML pour pouvoir analyser/travailler sur ces données et les réutiliser).

Interface de gestion

Une interface de gestion a été mise en place afin de créer les études, d'afficher les études précédentes selon leur statut (brouillon, en cours ou clôturée) et d'ajouter un nouvel utilisateur.

La création d'une nouvelle étude s'effectue en renseignant des informations telles que le nom, le donneur d'ordre (client), le sujet, la liste des audités et les phrases inductives sélectionnées pour l'étude. Un aperçu du processus de création d'une étude est disponible dans l'annexe A page 85 et des explications concernant les phrases inductives sont données dans la section 2.1.1.0 page suivante.

Lorsque nous souhaitons consulter les études précédentes, nous pouvons, soit afficher toutes les études précédentes soit les afficher selon leur statut. Une étude peut comporter un des trois statuts et en fonction du statut, différentes actions sont possibles :

— Brouillon : l'étude a été créée mais la liste des audités et le nombre de groupe n'ont pas été renseignés. Les informations sur l'étude ne sont donc pas suffisamment complètes pour que le questionnaire soit envoyé. Le consultant peut modifier, supprimer ou envoyer l'étude. Il peut également travailler sur l'image souhaitée.
— En cours : l'étude a été créée et toutes les informations obligatoires ont été renseignées. En sortie, le consultant a une liste d'audités avec pour chacun une url unique renvoyant vers le questionnaire. Le consultant peut clôturer l'étude ou consulter son état d'avancement pour savoir si tous les audités ont répondu ou non.
— Clôturée : l'étude est terminée, le consultant a obtenu les réponses des audités. Il peut donc afficher les résultats, cloner l'étude (reprendre certaines informations de l'étude pour en créer une nouvelle) et l'archiver au format XML (deux possibilités sont proposées : archiver l'étude sur le serveur ou exporter l'étude sur la machine).

Interface de production

Une interface de production (interface de saisie) a été conçue pour recueillir les réponses des audités (corpus). Il s'agit de la pratique de transition. Cette interface peut être utilisée selon une des deux modalités suivantes : une modalité individuelle et une modalité *focus group*. Par conséquent, les deux aspects présentiel (modalité *focus group*) et distanciel (modalité individuelle) ont déjà été développés. Cela permet de garder la démarche utilisée par le consultant pendant la phase de transition.

En effet, pour certaines études, le consultant peut effectuer les séances de questionnement en individuel : la séance de questionnement se déroule avec un seul audité à la fois qui répond au questionnaire sur l'interface de production en présence du consultant. Pour d'autres études, le consultant peut choisir d'effectuer les séances de questionnement en *focus group* : plusieurs audités répondent en même temps au questionnaire sur l'interface de production toujours en présence du consultant. La différence avec la première modalité réside dans le fait que lorsque les audités répondent au questionnaire, ils doivent attendre que tout le monde ait répondu pour passer à la question suivante.

Les figures 2.2 page suivante et 2.3 page 40 montrent cette seconde modalité. Deux navigateurs ont été ouverts avec pour chacun l'identité d'un audité différent. La figure 2.2 page suivante montre l'écran de l'audité 1, une page du questionnaire est affichée. L'audité 1 n'a pas encore répondu à la question et n'a pas encore cliqué sur le bouton pour passer à la question suivante. La figure 2.3 page 40 montre l'écran de l'audité 2 qui a déjà répondu à la question et doit patienter tant que l'audité 1 n'a pas cliqué sur le bouton pour passer à la question suivante.

A la fin du questionnaire, lorsque les audités ont répondu, ils peuvent revoir leurs réponses et les modifier (copie d'écran présentée dans l'Annexe B.2 page 91).

Phrases inductives

Les questions sont posées aux audités sous forme de phrases inductives afin de recueillir leurs opinions personnelles sur un sujet donné. Ces phrases inductives suivent un certain schéma, ce qui suppose donc une attente d'un point de vue

FIGURE 2.2 – Copie d'écran de l'interface production avec la modalité *focus group* : audité 1

lexical et sémantique. En effet, les phrases inductives sont construites de façon à obtenir un certain type de réponses sous forme de mots-clés, de groupes de mots ou de phrases courtes. D'une certaine manière, les phrases inductives conditionnent les réponses pour faciliter leur traitement par la suite.

Il existe six types de phrases inductives : évocation, rétention, adhésion, rejet, unicité, attente de communication. Pour chaque étude, au moins quatre questions sont sélectionnées.

Le fait que les productions soient différenciées en fonction du type de la question va nous permettre de voir, si pour les questions du même type, des différences ou des points communs émergent.

La formulation des phrases inductives est très importante, nous avons pu le constater lors de deux tests différents (les études β et δ que nous présentons dans la section 2.2.2 page 43) visant à tester l'interface de production. La formulation est notamment importante pour la pratique numérique et distancielle. En effet, le

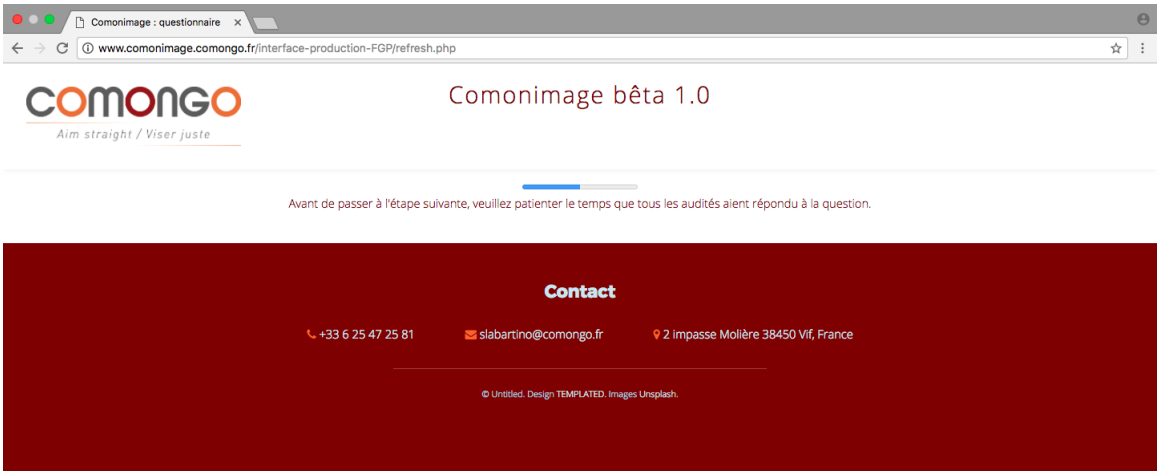


FIGURE 2.3 – Copie d'écran de l'interface production avec la modalité *focus group* : audité 2

consultant ne sera plus présent pour aider les audités s'ils ne comprennent pas la question.

Les phrases inductives font partie du cadre de production en présentiel, le consultant accompagne les audités dans la rédaction de leurs réponses notamment en répétant plusieurs fois la phrase. Avec la transition numérique, cela va changer et le cadre de production ne sera plus le même.

2.1.2 Structure du corpus

Comme nous l'avons décrit précédemment, une fois l'étude clôturée, il est possible de l'archiver sur le serveur ou de l'exporter sur une machine au format XML. Ce mode de gestion nous permet d'annoter de l'information (nous détaillons le processus d'annotation dans la section 2.3 page 49) et de montrer à travers le balisage les différents niveaux d'annotation : l'intersection lexicale, la proximité sémantique et la polarité des opinions. Cela permet également le traitement des données pour les analyses du chercheur.

Ces fichiers nous permettent non seulement de pouvoir travailler sur les données et de les traiter mais aussi, cela aide l'entreprise pour garder une trace des études clôturées dont les résultats sont déjà disponibles.

Nous avons choisi de ne pas utiliser la TEI pour diverses raisons. L'entreprise ne partage pas ses données puisque celles-ci sont confidentielles. Or, l'un des avantages de la TEI est justement le partage de la donnée. De plus, l'utilisation de marqueurs simples suffit à notre annotation, nous n'avons pas besoin de toutes les informations, notamment sur les métadonnées, que la TEI demande ni des outils développés autour de la TEI.

C'est pourquoi nous avons utilisé une DTD (Document Type Definition, voir Annexe C.1 page 94) constituée de manière ad hoc, pour les besoins de l'étude. L'archivage et l'export des études se font au format XML mais deux types de structures de fichiers ont été créés. Pour l'archivage, le fichier XML se compose d'une balise racine « étude » qui contient quatre autres balises : les données sur l'étude, les données sur les audités, les données des phrases inductives ainsi que les données sur les réponses. En revanche, pour l'export, nous avons une balise racine « étude » pouvant contenir des questions. Les questions contiennent ensuite les réponses des audités. Cette différence de format pour l'export nous permet d'avoir toutes les réponses des audités à une même question et de pouvoir les comparer en utilisant la méthode du consultant qui s'effectue par type de questions.

Pour l'instant, nous avons créé un fichier par étude mais il faudrait peut-être prévoir de pouvoir sortir un fichier contenant l'intégralité des études clôturées et donner la possibilité de les trier (par études, par types de phrases inductives, par audités, par donneur d'ordre, par sujet, par ordre chronologique, etc.). Cela nous permettrait d'étudier les données en fonction de ces différents axes de lecture.

Nous avons jusqu'à présent travaillé sur trois études différentes, nous disposons donc de trois fichiers distincts contenant chacun une étude. Le détail des études est effectué dans la section 2.2.2 page 43.

2.2 Traitement des données

2.2.1 Méthode du consultant

La méthode du consultant est une démarche intuitive avec une part de subjectivité[2]. Nous essayons de l'objectiviser et de la formaliser parce que la machine n'a pas cette subjectivité et il faut donc proposer des mécanismes afin de mettre en place cette méthode.

Le consultant traite d'abord les données par phrases inductives avant de les traiter dans leur globalité. Il s'agit donc de traiter chaque phrase inductive de l'étude à part et de mettre en couleurs les idées convergentes.

Nous commençons généralement par la phrase inductive de type évocation qui est toujours la première phrase inductive présentée aux audités. C'est notamment les réponses à cette première phrase inductive qui vont nous permettre de construire l'image perçue sous forme d'un circept, sans regarder l'image voulue construite préalablement.

Un circept est une représentation schématique permettant de mettre en vis-à-vis six à huit idées pour pouvoir les visualiser. La figure 2.4 page ci-contre est un exemple de représentation en circept avec six idées mises en relations. Les rectangles représentent chacun une idée différente. Avec les axes, les idées s'opposent ou s'associent : par exemple, l'idée de *bienveillance* en haut à droite s'associe à l'idée en bas à gauche d'*environnement agréable*, etc.

Les deux questions suivantes, qui sont généralement de type adhésion et rejet, nous donnent des premières indications sur la polarité. Les adhésions vont représenter des choses que les audités aiment et les rejets vont représenter l'inverse. Les réponses de type adhésion sont les points sur lesquels nous pouvons nous appuyer en termes de communication. En revanche, les rejets sont plutôt des choses qu'il va falloir gommer.

Enfin, les phrases inductives de type unicité (ce que l'on retient du sujet) nous permettent de valider les premières conclusions.

2. La méthode du consultant et de l'entreprise étant confidentielle, nous ne la décrivons pas en détails.

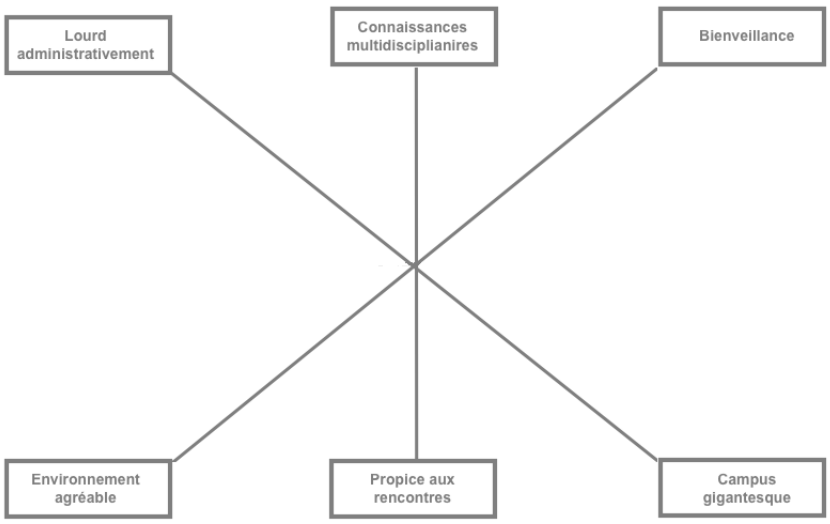


FIGURE 2.4 – Représentation en circept de l'étude β

Après avoir analysé les productions des audités, nous pouvons ensuite regarder l'image voulue et la comparer avec l'image perçue. Nous devons identifier l'écart entre les deux afin de donner un ensemble de conclusions aux clients qui sont des préconisations.

2.2.2 Volumétrie des données

La volumétrie des données est un aspect intéressant à prendre en compte. En effet, cela permet de mieux caractériser la donnée sur laquelle nous travaillons.

Nous avons travaillé sur trois études distinctes que nous nommons α, β et δ pour des raisons de confidentialité. Le tableau 2.1 page suivante présente les trois études.

Pour chacune des études, cinq questions ont été posées, le nombre d'audités et de réponses est très variable. L'étude α est la plus conséquente : elle a été réalisée en présentiel papier avec 6 *focus groups* pour un total de 28 audités et donc de 140 réponses. L'étude β est fictive, elle a été effectuée en présentiel numérique dans le but de tester l'interface de production. Constituée de 5 audités (un seul *focus*

	Etude α	**Etude** β	**Etude** δ
Nombre de questions	5		
Nombre d'audités	28	5	3
Nombre de réponses obtenues	140	25	15
Pratique	présentiel papier	présentiel numérique	présentiel numérique
Modalité	6 *focus groups*	1 *focus group*	individuel

TABLE 2.1 – Répartition des études

group), nous avons pu recueillir 25 réponses au total. La dernière étude, l'étude δ, est également fictive. Elle a été réalisée dans le même but que l'étude β à la différence que δ a été faite en individuel avec 3 audités et donc un total de 15 réponses.

La volumétrie des données que nous détaillons pour chaque étude a été calculée en « nombre d'idées ». La figure 2.5 page ci-contre qui présente les réponses de l'audité 69 à une question donnée de l'étude β permet d'expliciter cette métrique.

Lorsque les audités répondent sur l'interface de production, nous leur demandons d'effectuer un retour à la ligne après chaque idée notée. La volumétrie des réponses a donc été mesurée en comptant le nombre d'idées entrées sur l'interface par chacun des audités et pour chaque question. Comme nous pouvons le voir, l'audité 69 a effectué un retour à la ligne après chaque idée ce qui nous permet de calculer la volumétrie en nombre d'idées. Nous comptons donc au total 7 idées pour l'audité 69.

Il nous a semblé plus pertinent et plus significatif de mesurer la volumétrie des études en termes de nombre d'idées plutôt qu'en comptant le nombre de mots ou de caractères notamment à cause de la variation présente dans les productions. En effet, les audités sont libres de répondre avec des mots-clés, des groupes de mots ou des phrases plus construites.

Etude α

L'étude α avait été réalisée avec la pratique présentielle papier lors d'un *focus group*. Nous avons récupéré ces données que le consultant avait noté lors de la restitution et nous les avons intégrées dans un document XML. Les conditions de

Nouvelle réponse de 69 :
Une bonne université
Un paquebot administratif difficile à manœuvrer
Un espace de rencontre
Une source de stress par moment
Une période de transition dans la vie
Un campus agréable

FIGURE 2.5 – Exemple des idées pour une réponse d'un audité à une question donnée pour l'étude β

collecte de corpus de cette étude sont différentes des deux autres études que nous présentons plus loin.

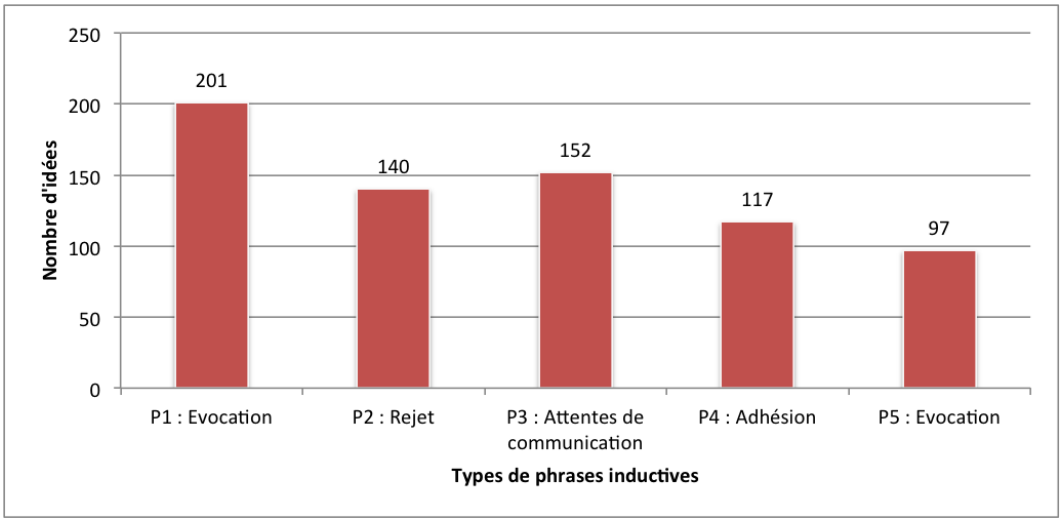


FIGURE 2.6 – Volumétrie des données sur l'étude α

L'étude α comporte le plus de réponses : 707 idées. En effet, plusieurs groupes témoins ont été constitués au lieu d'un seul pour les deux autres études. La figure 2.6 montre le nombre d'idées de la totalité des audités en fonction du type de phrase inductive. Les phrases inductives ont été présentées aux audités en respectant l'ordre suivant : évocation, rejet, attentes de communication, adhésion et évocation.

Nous constatons que nous obtenons le plus haut score du nombre d'idées pour la

première phrase inductive qui est de type évocation. La dernière phrase inductive, qui était également de type évocation (sur le même sujet mais sur un autre aspect) obtient le plus faible score, 97 idées sur la totalité des 707 idées recueillies pour l'étude α.

Pour l'étude α, les rejets (140 idées) sont plus importants que les adhésions (117 idées). Si on obtient plus de réponses sur les phrases inductives de type rejet que celles du type adhésion, cela signifie que les audités ont beaucoup de choses à dire sur le sujet.

Le nombre d'idées a plutôt tendance à diminuer au fur et à mesure où nous avançons dans les questions, excépté pour la phrase inductive 2 concernant les rejets des aduités par rapport au sujet.

Nous pouvons donc nous questionner sur le fait que les audités sont peut-être moins concentrés au fur et à mesure de l'avancement du questionnaire.

Etude β

L'étude β est d'origine factice. Elle a été réalisée avec des étudiants de l'IAE dans le cadre d'une simulation afin de tester le bon fonctionnement de l'interface de production en modalité *focus group* (et le bon enregistrement des données dans la Base de Données). Cela a également permis de faire découvrir et de faire comprendre le principe et la méthode de COMONGO aux étudiants de l'IAE qui ont travaillé sur un autre aspect du projet. Nous utilisons les données tests de cette étude pour donner des exemples dans le cadre de notre mémoire. Cela nous évite de donner des informations confidentielles puisque cette étude n'a pas de valeur d'étude pour l'entreprise COMONGO.

L'étude β comporte 92 idées soit nettement moins que l'étude α. La figure 2.7 page ci-contre montre le nombre d'idées de la totalité des audités en fonction du type de phrase inductive. Les phrases inductives ont été présentées aux audités en respectant l'ordre suivant : évocation, adhésion, rejet, unicité et attentes de communication. Nous avons constaté, en travaillant sur ces données, que nous obtenons plus de réponses (en termes de nombre d'idées) sur les trois premières questions que sur les deux dernières.

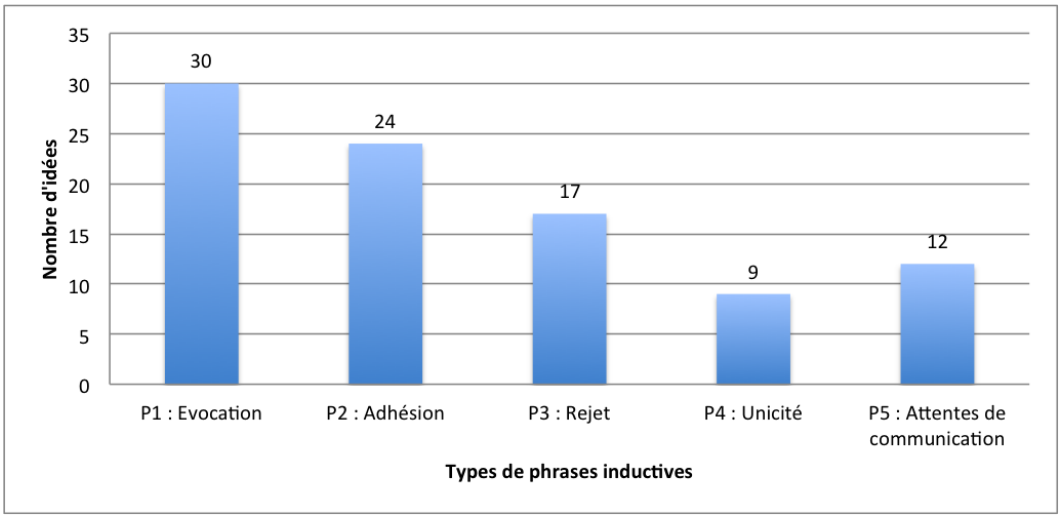


FIGURE 2.7 – Volumétrie des données sur l'étude β

Les trois premières phrases obtiennent les plus forts taux de nombre d'idées notamment pour la phrase inductive de type évocation qui obtient un score de 30 idées. Les scores d'adhésion et de rejet sont plutôt conséquent mais restent plus faibles que la première phrase inductive (respectivement 24 et 17 idées). Cela peut s'expliquer par le fait que pour la phrase inductive de type évocation, les audités donnent toutes les réponses que le sujet exprime (adhésions et rejets mélangés). Ensuite, ils effectuent en quelque sorte un tri dans leurs idées pour noter les adhésions pour la phrase adhésion et de même avec les rejets. Pour l'étude β, les rejets sont moins importants que les adhésions.

Les deux dernières questions qui étaient de type unicité et attentes de communication sont moins importantes (respectivement 9 et 12 idées). En effet, nous avons moins de réponses sur la phrase inductive de l'unicité (ce qui rend le sujet unique) puisque les audités essaient d'être le plus précis. La phrase inductive sur les attentes de communication comporte également moins de réponses puisque c'est une sorte de conclusion au questionnaire et que les audités ont presque tout dit dans les trois premières questions. Un point important à prendre en compte est que la dernière question (attentes de communication) a posé problème aux audités sur le plan de la compréhension. Plusieurs audités ont demandé des explications concer-

nant la question et c'est peut-être la raison pour laquelle nous obtenons moins d'idées. La formulation des phrases inductives est donc très importante comme nous avons pu le voir dans la section 2.1.1.0 page 38.

Cependant, tout comme pour l'étude α, nous pouvons nous questionner sur le fait que les audités sont peut-être moins concentrés au fur et à mesure de l'avancement du questionnaire.

Etude δ

L'étude δ est aussi d'origine factice. Elle a été réalisée dans le cadre d'une simulation afin de tester le bon fonctionnement de l'interface de production en modalité individuelle.

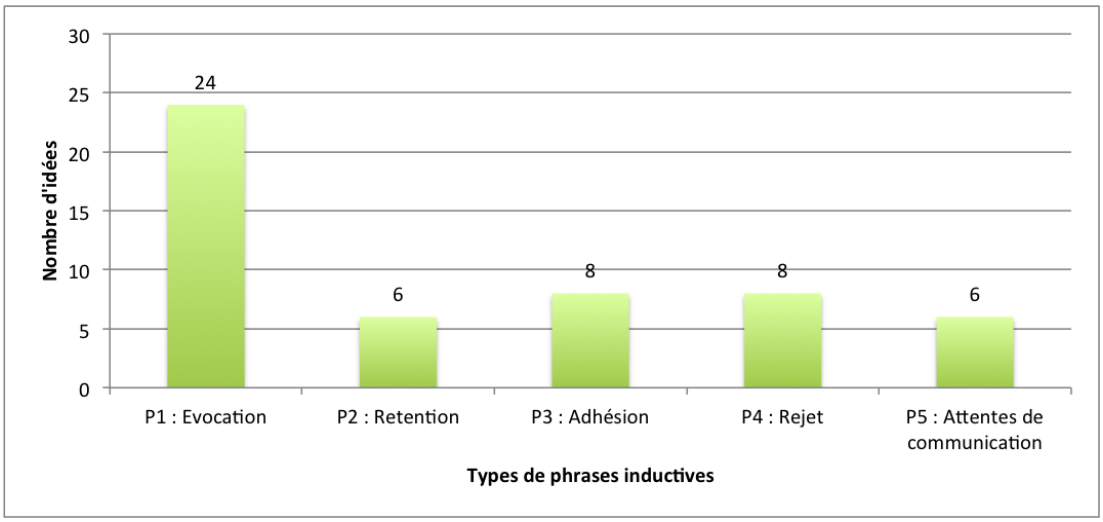


FIGURE 2.8 – Volumétrie des données sur l'étude δ

L'étude δ est celle qui comporte le moins de réponses mais aussi le moins d'audités (seulement trois audités). Cependant elle n'en est pas mois intéressante pour autant. 52 idées au total ont été collectées. La figure 2.8 montre le nombre d'idées de la totalité des audités en fonction du type de phrase inductive. Les phrases inductives ont été présentées aux audités en respectant l'ordre suivant : évocation, rétention, adhésion, rejet et attentes de communication.

Comme pour les deux études précédentes, nous constatons que la première

question qui est toujours de type évocation obtient le nombre d'idées le plus haut (24 idées soit presque la moitié du nombre total d'idées pour cette étude, ce qui constitue donc un nombre conséquent). Un point important à noter est que, tout comme l'étude β, la dernière question (attentes de communication) a posé problème aux audités sur le plan de la compréhension. C'est peut-être la raison pour laquelle nous obtenons, pour cette étude, moins d'idées. Les idées des quatres dernières questions obtiennent sensiblement les mêmes scores avec, également, une légère diminution au fur et à mesure de l'avancement dans le questionnaire.

Pour conclure sur la volumétrie de ces trois études, nous pouvons tout d'abord dire qu'il y a une tendance a obtenir plus de nombre d'idées pour la première phrase (toujours de type évocation). Nous remarquons également que, pour les trois études, nous pouvons observer une diminution générale du nombre d'idées plus les audités avancent dans le questionnaire. Cela peut être dû aux phrases inductives mais aussi au fait que les audités peuvent être moins concentrés au bout de quelques questions.

2.3 Attendu et obtenu

Nous avons construit un attendu dans le but de le comparer et de le contraster avec ce que nous avons obtenu en utilisant la simulation par les ressources sur notre corpus. Cela nous permet de nous rendre compte à quel point les ressources peuvent être suffisantes ou non pour notre projet. C'est pourquoi constituer l'attendu est très important.

Description de la méthode pour constituer l'attendu et l'obtenu :

1. Constitution de l'attendu par analyse linguistique et annotation manuelle du corpus
2. Constitution de l'obtenu par simulation des ressources sur le corpus
3. Comparaison de l'attendu et de l'obtenu

2.3.1 Démarche pour constituer l'attendu

La première étape de la méthode que nous avons mis en place est la constitution de l'attendu. Notre corpus est constitué de trois études clôturées (α, β et δ). Nous nous focaliserons sur l'étude β pour plusieurs raisons. Tout d'abord parce qu'elle est d'origine factice (faite dans un contexte de test de l'interface) et que cela nous permet de nous en servir d'exemple dans notre mémoire sans donner d'information confidentielle. De plus, elle nous semble plus concrète car nous l'avons organisée avec le consultant et nous avons travaillé ensemble sur cette étude avant de l'annoter. L'étude α était déjà existante et les résultats avaient déjà été traités par le consultant. Pour l'analyse, nous n'avons pas pris l'étude δ non plus puisqu'elle contient peu d'informations (en effet, seulement trois audités ont répondu). L'étude β nous semble suffisante pour donner des premières analyses. Elle nous a permis de nous rendre compte rapidement des différents phénomènes.

Afin de constituer l'attendu, nous avons donc travaillé sur l'étude β. Nous avons commencé par effectuer une analyse manuelle des données avec la méthode du consultant afin de mieux comprendre les attentes de l'entreprise. Puis, nous avons effectué une annotation manuelle de cette étude au format XML avec le logiciel Oxygen XML Editor et une DTD (Document Type Definition) constituée de manière ad hoc (Annexe C.1 page 94), pour les besoins de l'étude. L'annotation a consisté à repérer dans les réponses à chaque phrase inductive quelles sont les informations utiles et les points de rencontre du point de vue de l'intersection lexicale, de la proximité sémantique et de la polarité. Les résultats des trois études sont présentés dans le chapitre 3 page 59.

La copie d'écran 2.9 page suivante est un extrait de trois réponses d'audités différents à une question donnée pour l'étude β qui a été présentée plus haut. Chaque nouvelle question est affichée avec le titre de la question et son type entre parenthèses. Chaque nouvelle réponse est précédée de la mention « Nouvelle réponse de ... » ainsi que de l'identifiant de l'audité.

Les termes sans intérêt ont été masqués pour des raisons de lisibilité et afin de mieux se concentrer sur l'annotation.

L'annotation a été faite sur les trois niveaux présentés dans le chapitre 1

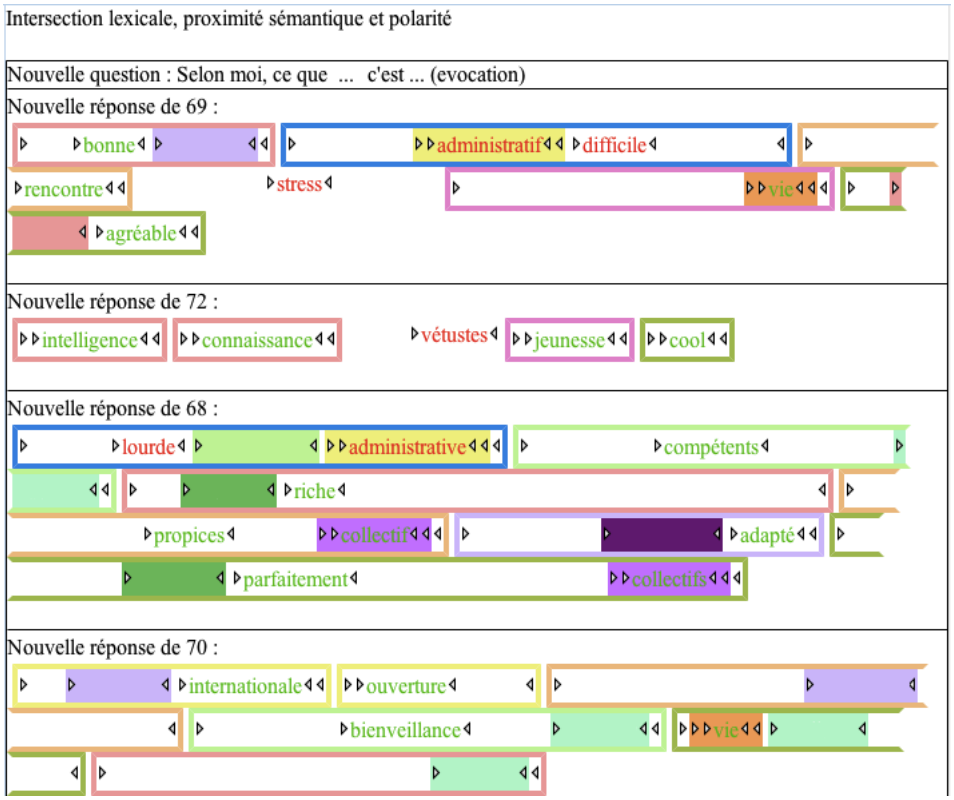


FIGURE 2.9 – Copie d'écran d'un extrait d'annotation de trois réponses d'audités différents à une question donnée pour l'étude β

page 21. Ces trois niveaux d'annotation ont été mis en avant différement :

1. Intersection lexicale : fond coloré
2. Proximité sémantique : bordure colorée (encadrement)
3. Polarité des opinions : texte coloré (en vert, les éléments de polarité positive et en rouge les éléments de polarité négative)

Les résultats de l'annotation manuelle sont présentés dans le chapitre 3 section 3.1 page 59.

Intersection lexicale

Pour parler de l'intersection lexicale, nous utilisons le terme de termes identiques. Nous avons constitué l'attendu tel que perçu dans la démarche. Sont consi-

dérés comme termes identiques, les termes ayant un même lemme (même entrée lexicale). Par exemple, les termes *administratif* et *administrative* sont issus du même lemme : *administratif*. Il s'agit de la même idée et donc de termes identiques. Nous travaillons par question, c'est-à-dire que nous avons choisi d'annoter les termes identiques apparaissant au moins deux fois dans les réponses à une même question. Seuls les mots pleins ont été annotés avec des balises XML puisqu'ils constituent des éléments de sens. Etant donné qu'il n'y a pas de consensus sur le terme « mot », nous considérons comme mot la suite de caractères séparée par des espaces.

L'annotation de l'intersection lexicale est visible sur la figure 2.9 page précédente avec la coloration du fond. Les fonds de la même couleur représentent des termes identiques.

Proximité sémantique

Pour parler de la proximité sémantique, nous utilisons le terme d'éléments sémantiquement proches ou de concepts similaires. Il s'agit de suites contiguës de un ou plusieurs mots. L'annotation sur la proximité sémantique s'est également faite sur les mots pleins.

Les éléments sémantiquement proches ont été créés et annotés artificiellement avec des balises XML. L'annotation de la proximité lexicale est visible sur la figure 2.9 page précédente avec la coloration de la bordure (encadrement). Les encadrements de la même couleur représentent des éléments proches sémantiquement (concept similaire).

Polarité des opinions

Pour parler de la polarité des opinions, nous utilisons le terme de termes ou éléments de polarité positive ou négative. L'annotation de la polarité s'est également faite uniquement sur les mots pleins et sur les mots considérés comme pertinents.

L'annotation de la polarité est visible sur la figure 2.9 page précédente avec la coloration du texte. En vert, les éléments de polarité positive et en rouge, les éléments de polarité négative. Nous avons annoté la polarité des opinions selon ces

deux valeurs parce que nous nous sommes rendu compte qu'annoter un élément avec l'étiquette neutre n'était pas forcément utile pour rechercher l'expression des opinions.

Pour annoter la polarité, nous avons utilisé la méthode du consultant (méthode intuitive), mais nous avons également utilisé la typologie des questions comme un indicateur de la polarité. Par exemple, les réponses à une question de type adhésion sont plutôt positives puisque nous cherchons ce que les audités aiment. Tandis que les réponses à une question de type rejet sont plutôt négatives puisque nous cherchons à savoir ce que les audités n'aiment pas.

Nous avons annoté la polarité des mots mais en travaillant sur le corpus, nous nous sommes rendu compte que la polarité pouvait se trouver à plusieurs niveaux : sur les mots mais aussi sur des groupes de mots. Il serait intéressant de travailler sur ces groupes de mots (suites de caractères de plus d'un mot ou segments). Cependant, pour ce mémoire nous avons seulement considéré la polarité des mots afin de pouvoir faire une comparaison avec la simulation par des ressources, puisque ces ressources considèrent uniquement les mots.

2.3.2 Simulation par des ressources

La seconde étape de notre méthode est la constitution de l'obtenu par simulation des ressources sur le corpus. Cette simulation a été effectuée sur les trois niveaux que nous avons évoqué : l'intersection lexicale, la proximité sémantique et la polarité des opinions. Cela nous permet de voir ce que nous pouvons obtenir en utilisant des ressources existantes. Nous ne sommes pas dans l'idée de faire mieux que l'humain mais plutôt d'imiter la démarche de l'humain. La simulation par les ressources nous permet simplement de voir s'il est suffisant d'utiliser et d'adapter des ressources existantes ou s'il est préférable de constituer nous-mêmes une ressource pour notre projet. Les résultats de ces simulations sont présentés dans le chapitre 3 section 3.2 page 64.

Nous avons gardé la même méthode d'annotation :

1. Intersection lexicale : fond coloré

2. Proximité sémantique : bordure colorée (encadrement)
3. Polarité des opinions : texte coloré (en vert, les éléments de polarité positive et en rouge, les éléments de polarité négative)

Nous avons trois études, un fichier d'annotation a été récupéré pour chacune des études (α, β et δ). Nous avons dupliqué ces fichiers pour la simulation par les ressources. Cela nous a permis de mettre en vis-à-vis les fichiers de l'attendu avec ceux de l'obtenu. Nous présentons dans cette section, pour chaque niveau, la ou les ressource(s) choisie(s).

Intersection lexicale

Pour l'intersection lexicale, nous avons trouvé deux lexiques qui nous donnent les formes lemmatisées puisque nous avons travaillé sur les lemmes afin de constituer l'attendu.

Le **Lexique des Formes Fléchies du Français** (LEFFF[3]) de [Clément et al., 2004] a été construit automatiquement à partir de corpus et par validation manuelle. Il donne de nombreuses informations dont les traits morphologiques, mais aussi le lemme associé à chaque forme. Cette ressource comporte 404 483 formes fléchies et environ 54 524 lemmes.

La seconde ressource trouvée pour les formes lemmatisées est **Lexique 3**[4] de [New et al., 2001] qui comporte 135 000 mots du français avec 55 000 lemmes ainsi que d'autres informations.

Nous avons choisi d'utiliser une seule ressource pour la lemmatisation la ressource LEFFF qui comporte un grand nombre de formes fléchies.

Proximité sémantique

Différentes ressources sont à notre disposition pour simuler la proximité sémantique et ainsi trouver des concepts similaires dans notre corpus.

3. `http://www.labri.fr/perso/clement/lefff/` (consulté le 09/05/2017)
4. `http://www.lexique.org` (consulté le 09/05/2017)

La première ressource que nous avons trouvée est celle du **Dictionnaire Electronique des Synonymes** (DES[5]) du Centre de Recherche Inter-langues sur la Signification en COntexte (CRISCO). Cette ressource, constituée à partir de dictionnaires, contient 49 443 entrées. Les mots sont liés entre eux par des relations de synonymie (plus de 200 000 relations synonymiques) et l'ensemble des mots synonymes est appelé clique.

La seconde ressource, **Diko**[6], est issue du jeu **JeuxDeMots** (JDM[7]) [Lafourcade, 2007]. Diko comprend des relations lexico-sémantiques entre termes. Cette ressource est la version dictionnaire du réseau lexical de JeuxDeMots et a été construite manuellement.

Nous avons choisi le Dictionnaire des Synonymes (DES) en prenant une distance de 1 arc au sein du réseau sémantique.

Polarité des opinions

Dans cette section, nous présentons différentes ressources en français pour détecter la polarité des opinions que nous pourrions utiliser et compléter.

Il existe peu de ressources en français sur la polarité des termes et encore moins de ressources sur les émotions. Une telle ressource permettant de déterminer la polarité est complexe à construire puisqu'elle est subjective et contextuelle.

CASOAR [Asher et al., 2008] est une « ressource construite manuellement à partir de corpus et contient des termes polarisés » [Abdaoui et al., 2016]. Le problème est qu'elle n'est pas disponible. Cela l'exclut puisque nous avons besoin de simuler la ressource sur notre corpus. Nous avons également trouvé la ressource

Polarimots[8] [Gala & Brun, 2012] est une ressource construite semi-automatiquement à partir de la ressource lexicale Polymots [Gala & Rey, 2008]. Polarimots comporte un peu plus de 7 000 termes et permet de donner une polarité positive, neutre ou négative aux différents termes.**Affects Lexicon** [Augustyn et al., 2006] a été « construite automatiquement et recense environ 1 200 termes du français » [Ab-

5. `http://www.crisco.unicaen.fr/des/` (consulté le 27/04/2017)
6. `http://www.jeuxdemots.org/diko.php` (consulté le 27/04/2017)
7. `http://www.jeuxdemots.org/jdm-accueil.php` (consulté le 27/04/2017)
8. `http://polarimots.lif.univ-mrs.fr` (consulté le 09/05/2017)

daoui et al., 2016]. Nous n'avons pas choisi les ressources Polarimots et Affects Lexicon qui contiennent beaucoup moins de termes que les ressources que nous présentons par la suite.

Certaines ressources, également issues du jeu JeuxDeMots [Lafourcade, 2007], peuvent être utilisées afin de déterminer la polarité ou les émotions associées. **Emot**[9] est un jeu issu de JeuxDeMots. Cette ressource a été construite par *crowdsourcing*. Cependant, Emot ne contient que des sentiments ou émotions associés à des mots par les joueurs, or la polarité nous intéresse également. Il est « possible de calculer une polarisation pour un terme auquel des termes de sentiment ont été associée » [Lafourcade et al., 2015] mais nous n'avons pas choisi cette ressource.

LikeIt[10] [Lafourcade et al., 2015] est un jeu de « consensus à vote » également issu de JeuxDeMots pour définir la polarité d'un terme. Un terme est proposé aux joueurs et ils doivent voter pour attribuer une polarité entre positif, neutre ou négatif. Il s'agit donc également d'une ressource constituée par *crowdsourcing* et contenant plus de 360 000 termes. A chaque terme est associé un « triplet de nombre de vote pour chacune des trois polarités possibles » [Lafourcade et al., 2015]. Cette ressource comporte des mots ou des suites de plusieurs mots. Elle comporte également beaucoup d'entités nommées. Cependant, selon les auteurs, cette ressource comporte certains biais. Le premier biais concerne les termes polysémiques. Un terme peut avoir plusieurs sens et ces différents sens peuvent ne pas avoir la même polarité. Comment le joueur va donc choisir entre ces différents sens pour annoter la polarité ? Le second biais concerne la « divergence de perception » : certains termes peuvent avoir une polarité double. Comment choisir ? Le dernier biais de cette ressource est que parfois, les joueurs ont plutôt tendance à favoriser la polarité positive pour un terme qui serait relativement neutre.

La ressource récente **French Expanded Emotion Lexicon**[11] (FEEL) comprend 14 000 termes catégorisés en deux polarités (négative ou positive) et en six émotions (les émotions de [Ekman, 1992] : peur, dégoût, colère, joie, tristesse, sur-

9. `http://www.jeuxdemots.org/emot.php` (consulté le 27/04/2017)
10. `http://www.jeuxdemots.org/likeit.php?what=about` (consulté le 27/04/2017)
11. `http://advanse.lirmm.fr/feel.php` (consulté le 27/04/2017)

prise). Il s'agit d'une traduction de l'English Emotional Lexicon NRC-Canada [Mohammad & Turney, 2013]. Afin de constituer FEEL, les auteurs [Abdaoui et al., 2016] ont d'abord traduit automatiquement les 14 000 termes de l'anglais vers le français en utilisant six traducteurs en ligne. Puis, ces termes ont été vérifiés et validés manuellement par un traducteur humain. Les émotions ont ensuite été annotées et vérifiées par trois annotateurs à cause de la subjectivité de la tâche (mais cela a été fait sur un sous-ensemble).

Nous avons choisi d'utiliser les deux dernières ressources pour la simulation de la polarité par des ressources sur notre corpus : LikeIt qui prend en compte seulement la polarité et FEEL qui comporte également les émotions. Nous avons regardé pour chaque mot plein présent dans l'étude β s'il apparaissait dans les deux ressources et les avons annotés.

3.3 Définition des métriques d'évaluation

Afin d'évaluer la simulation par les ressources sur notre corpus, nous définissons quelques métriques que nous utiliserons dans le chapitre 3 page 59 à savoir : les mesures de rappel et de précision, le silence et le bruit ainsi que la F-mesure. Ces métriques permettent d'évaluer la robustesse d'un système.

Le rappel permet de donner la proportion des solutions pertinentes sur la totalité des solutions. Le rappel se calcule en divisant le nombre de solutions pertinentes proposées (c'est-à dire le nombre de solutions correctes proposées) par le nombre total de solutions pertinentes à trouver :

$$\text{Rappel} = \frac{\text{nombre de solutions pertinentes proposées}}{\text{nombre de solutions pertinentes à trouver}} \ . \tag{1}$$

La précision est la proportion des solutions trouvées par un système qui sont pertinentes. Plus la précision est forte, plus le système parvient à refuser les solutions non pertinentes. Elle se calcule en divisant le nombre de solutions pertinentes proposées par le nombre total de solutions pertinentes proposées :

$$\text{Précision} = \frac{\text{nombre de solutions pertinentes proposées}}{\text{nombre de solutions proposées}} \ . \tag{2}$$

Plus les résultats de ces mesures sont proches de 1, plus la précision et le rappel sont forts. « Un rappel de 100% indique que toutes les frontières de référence ont été

correctement trouvées. Une précision de 100% signifie que l'ensemble des frontières proposées par le système sont correctes » [Bouchekif, 2016].

A partir de cela, nous pouvons calculer les mesures de silence et de bruit.

Le silence correspond aux solutions pertinentes qui auraient dû être trouvées par le système :

$$\text{Silence} = \frac{\text{nombre de solutions pertinentes non proposées}}{\text{nombre de solutions pertinentes à trouver}} \ . \tag{3}$$

Le bruit correspond aux solutions non pertinentes qui n'auraient pas dû être trouvées (les solutions non pertinentes trouvées par un système) :

$$\text{Bruit} = \frac{\text{nombre de solutions non pertinentes}}{\text{nombre de solutions proposées}} \ . \tag{4}$$

La F-mesure peut permettre de mieux évaluer la robustesse puisqu'elle est calculée à partir de la précision et du rappel :

$$\text{F-Mesure} = \frac{2 \times \text{rappel} \times \text{précision}}{\text{rappel} + \text{précision}} \ . \tag{5}$$

Pour évaluer la simulation des ressources sur le corpus en utilisant les mesures du rappel et de la précision, nous avons besoin de définir ce que nous considérons comme documents pertinents (corrects) du point de vue de l'intersection lexicale, de la proximité sémantique et de la polarité. Nous adaptons donc les métriques sur les trois niveaux dans le chapitre 3 page suivante.

Chapitre 3

Analyse des résultats de l'étude β

3.1 Résultats attendus

Nous présentons les résultats attendus ainsi que leurs observations pour l'étude β (étude présentée dans le tableau 3.1) sur chacun des trois niveaux d'analyse par le TAL. Nous avons annoté ce que nous considérons comme intersection lexicale, proximité sémantique et polarité pour chaque ensemble de réponses aux questions (c'est-à-dire pour chaque réponse de tous les audités à la même question).

3.1.1 Intersection lexicale

Au niveau de l'intersection lexicale, nous abordons les résultats du point de vue de la lemmatisation sur l'étude β. La figure 3.1 page suivante présente ces résultats

	Etude β
Nombre de questions	5
Nombre d'audités	5
Nombre de réponses obtenues	25
Pratique	présentiel numérique
Modalité	1 *focus group*

TABLE 3.1 – Présentation de l'étude β

en nous donnant pour chaque terme, le nombre total d'occurences annotées avec en couleurs le nombre total d'occurences du terme par audités. Par exemple, si nous prenons le terme *campus* (apparaissant en premier sur la figure), nous pouvons voir que nous avons repéré 11 occurences au total dont 3 occurences pour l'audité 1 (en bleu foncé), 2 pour l'audité 3 (en vert), 1 pour l'audité 4 (en violet) et 5 pour l'audité 5 (en turquoise). Nous avons un total de 76 occurrences de termes annotées manuellement pour l'intersection lexicale et nous pouvons constater que nous avons 23 termes distincts : campus, université, discipline, étudiant, etc. Cela constitue notre attendu.

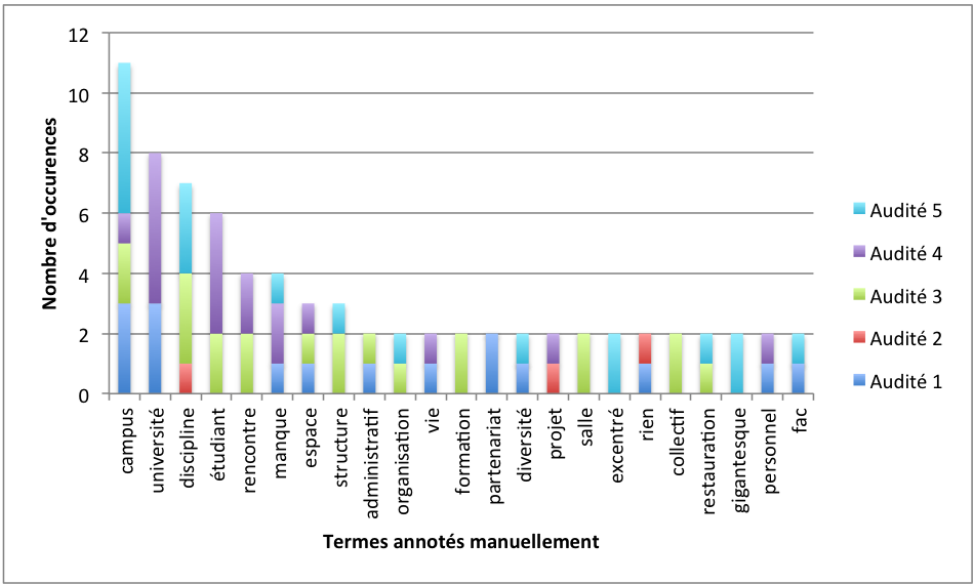


FIGURE 3.1 – Intersection lexicale : résultats attendus sur l'étude β

Le terme *campus* est le terme le plus fréquent (11 occurrences). Si un terme est beaucoup utilisé chez différents audités et sur plusieurs questions, ou si, au contraire, on ne retrouve un terme qu'une seule fois, c'est qu'il a son importance. En revanche, ce n'est peut-être pas pertinent si on retrouve un terme identique partout chez un même audité et c'est notamment le cas du terme 4.*campus* qui apparait 5 fois dans les réponses de l'audité 5.

De la même façon, ce n'est pas pertinent si un terme présent dans la phrase inductive se retrouve dans les réponses. Et plus particulièrement, si ce terme se

retrouve dans toutes les réponses d'un même audité plus d'une fois. C'est le cas du terme *université* qui était présent dans les phrases inductives présentées aux audités pour l'étude β. Nous pouvons constater qu'il s'agit du second terme le plus fréquent que l'on retrouve seulement chez deux audités (les audités 1 et 4). L'intersection lexicale nous permet, dans un premier temps, de voir ce qui émerge le plus mais les termes les plus fréquents ne sont donc pas forcément les plus significatifs et il faut peut-être passer au-delà de ces termes pour nous intéresser aux mots moyennement fréquent comme *discipline* (utilisé ici dans le sens de matière) et *manque*.

Nous remarquons que nous obtenons beaucoup de noms et quelques adjectifs (20 noms et 3 adjectifs). Les termes nominaux et adjectivaux sont les plus intéressants puisquex chaque idée a souvent une forme nominale ou adjectivale plutôt que phrastique.

3.1.2 Proximité sémantique

Pour l'étude β, nous pouvons observer sur la figure 3.2 page suivante que nous avons trouvé manuellement 11 concepts sémantiques différents sur la totalité des questions. Nous avions d'abord fait un premier travail avec le consultant qui nous a aidé à repérer les concepts similaires dans l'étude β. Cela nous a aidé à la réalisation du circept de l'image perçue de l'étude β (voir la figure 2.4 page 43). Notre annotation manuelle est donc fondée sur cette première approche.

Le concept sémantique le plus fréquent de l'étude β est l'idée de *bonne structure* avec 21 occurences dans les réponses des audités. Le second concept *lourd administrativement* a été constitué à partir d'éléments considérés comme sémantiquement proches dont : « une trop lourde organisation administrative » et « un paquebot administratif difficile à manoeuvrer ».

3.1.3 Polarité des opinions

Au niveau de la polarité des opinions, sur les 633 tokens de l'étude β, nous en avons annotés 105. L'annotation est une tâche complexe puisqu'elle est subjective.

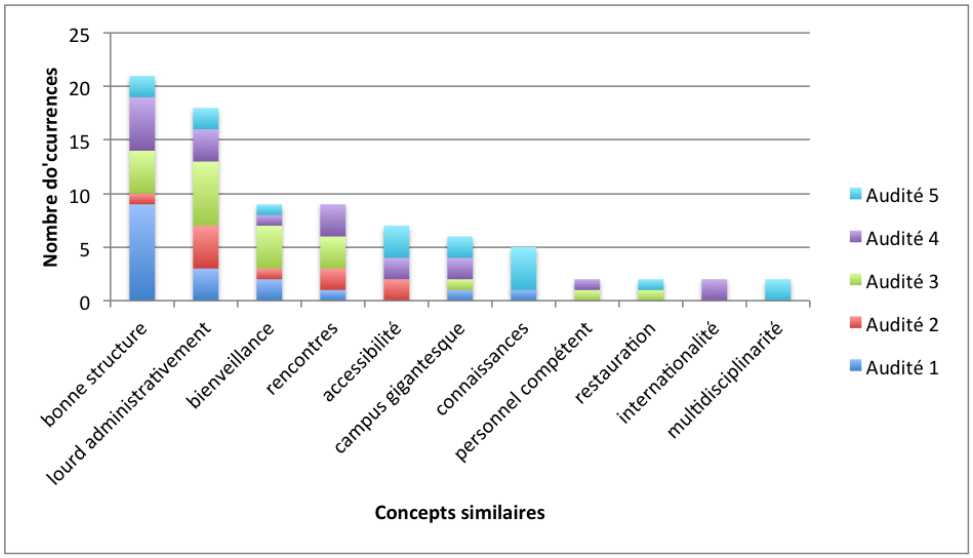


FIGURE 3.2 – Proximité sémantique : résultats attendus sur β

Il est difficile de rester constant dans ses choix. Par exemple, pour cinq termes nous avons identifié une suite contiguë de deux mots au lieu d'annoter un seul mot. Ces cinq termes ont été annotés comme formant un tout : *peu de* (annoté deux fois), *manque de* et *de qualité* (annoté deux fois). C'est pourquoi, nous avons au total 100 termes annotés. La figure 3.3 page ci-contre présente la répartition des termes positifs et négatifs de ces termes. Nous avons annoté plus de mots positifs (71 occurrences) que négatifs (29 occurences). Nous avons essayé, dans notre démarche d'annotation, d'être le plus objective possible mais une part de subjectivité subsiste. C'est pourquoi nous n'avons pas annoté les termes pour lesquels nous avions des doutes.

Comme nous l'évoquions précédemment, la typologie des questions nous permet d'assigner plus facilement la polarité aux différents termes présents dans le corpus. La figure 3.4 page 64 nous donne toutes les réponses de tous les audités à la question adhésion de l'étude β. Comme nous pouvons le constater, nous obtenons plus d'éléments positifs (en vert) que négatifs (en rouge). Seul un élément (« circulation ») est annoté comme étant négatif. Cet élément n'est d'ailleurs pas vraiment d'ordre négatif puisqu'il est précédé du terme « peu de » que nous avons annoté positif et comme formant un tout. « Circulation » ne devrait donc pas être

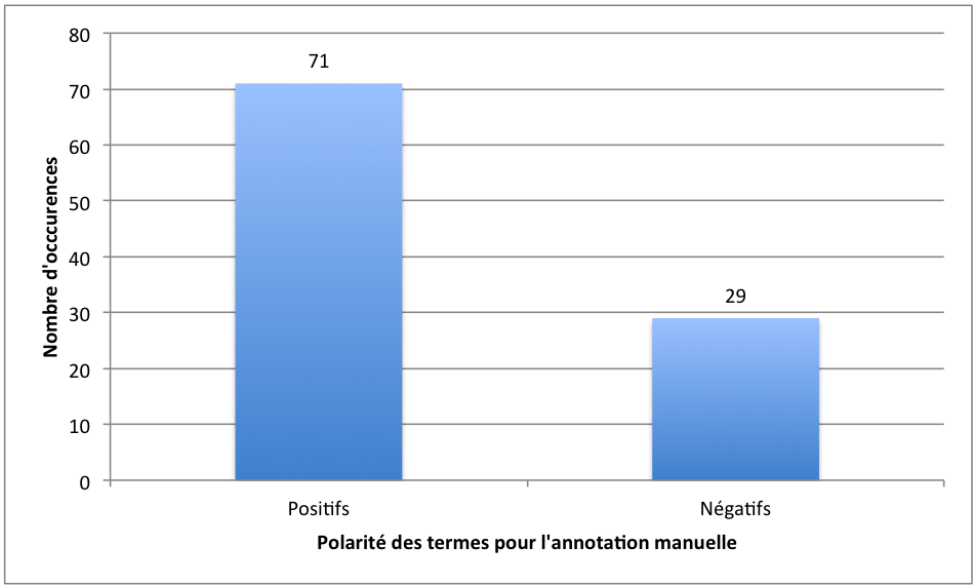


FIGURE 3.3 – Polarité des opinions : annotation manuelle des termes positifs et négatifs

identifié comme un terme négatif mais nous l'avons fait afin de pourvoir effectuer une comparaison avec les ressources qui assignent une polarité aux mots.

La typologie des questions qui nous aide à définir la polarité s'applique également pour les questions de type rejet comme nous le montre la figure 3.5 page 65. Cette figure présente toutes les réponses de tous les audités à la question rejet de l'étude β et nous pouvons constaté que la majorité des termes annotée est de polarité négative (en rouge).

L'annotation de la polarité s'est révélée difficile notamment sur le répérage des frontières de la polarité. Nous nous sommes demandée s'il valait mieux annoter des mots ou des ensembles de mots. En effet, un mot peut avoir une certaine polarité mais dans le contexte d'une phrase ou d'un groupe de mots, il peut avoir une toute autre polarité. Par exemple, le mot *efficace* peut avoir une connotation positive mais si le terme *peu* le précède alors la polarité devient négative.

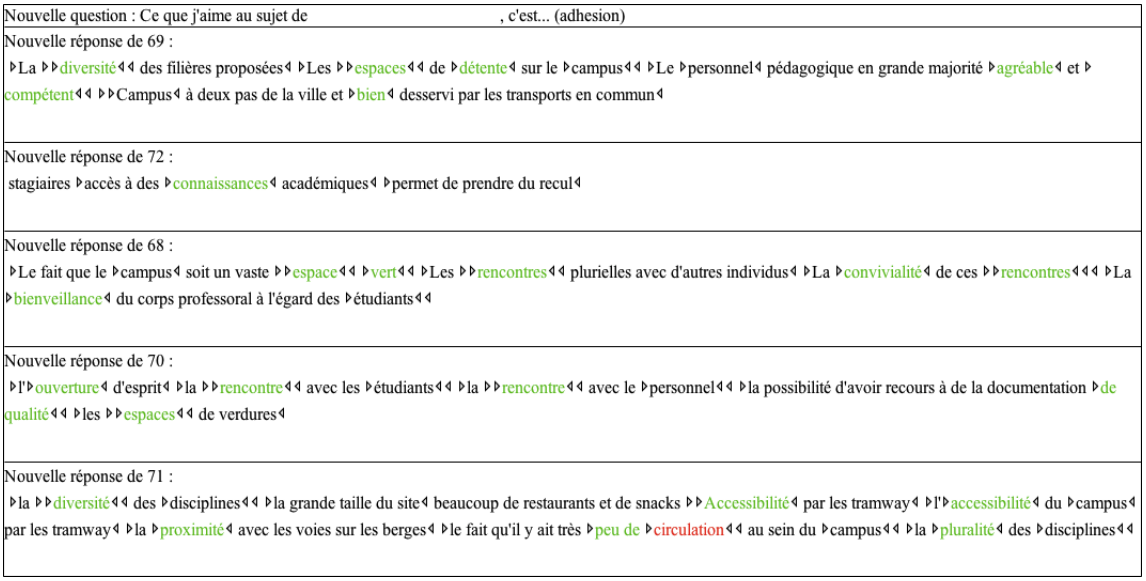


FIGURE 3.4 – Polarité positive : exemple d'annotation manuelle des réponses de cinq audités à une question de type adhésion (étude β)

3.2 Résultats obtenus par simulation des ressources

Nous présentons les résultats obtenus ainsi que leurs observations pour l'étude β sur chacun des trois niveaux d'analyse. Les ressources sont toujours créées dans un contexte particulier pour un domaine donné et un type d'application. La simulation par les ressources sur l'étude β permet de voir ce que ces dernières apportent, si elles sont suffisantes ou s'il est nécessaire d'aller plus loin notamment en utilisant une construction incrémentale des ressources.

3.2.1 Intersection lexicale

Pour l'intersection lexicale, après simulation par la ressource du LEFFF (la description de la ressource se trouve dans la section 2.3.2.0 page 54) sur l'étude β, nous constatons (figure 3.6 page 66) que le nombre total d'occurences (63 occurrences) est légèrement en dessous du nombre attendu (76 occurrences). Le nombre de termes annotés a également diminué. En effet, seulement 19 termes distincts apparaissent sur la figure 3.6, contre 23 dans les résultats attendus. Quatre termes ont

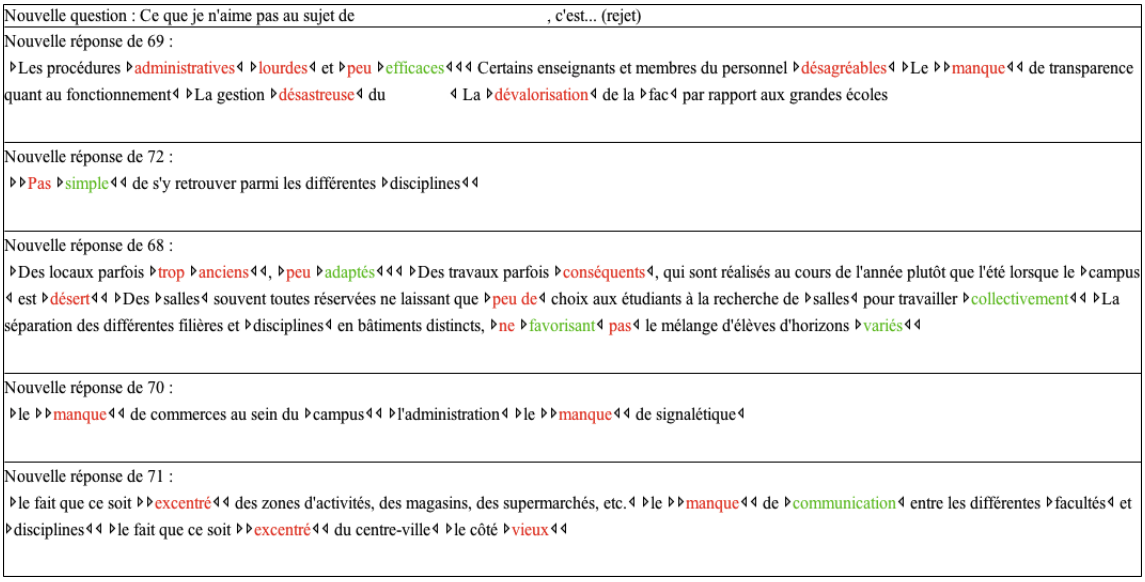


FIGURE 3.5 – Polarité négative : exemple d'annotation manuelle des réponses de cinq audités à une question de type rejet (étude β)

disparu : *étudiant*, *structure*, *collectif* et *fac*. L'absence des trois premiers termes s'explique par le fait que nous ne les avons pas trouvés dans la ressource utilisée, il s'agit donc d'un inconvénient de cette ressource puisque des termes plutôt courants ne sont pas présents. En revanche, pour le dernier terme, *fac*, il s'agit d'un autre problème. En effet, dans notre étude nous avons deux occurrences : *faculté* et *fac*, dans le sens de division d'une université, Unité de Formation et de Recherche (UFR). Nous considérons, dans notre approche, que ce sont des termes identiques puisque *fac* est l'apocope de *faculté*. Or, la ressource du LEFFF considère que ce sont deux lemmes différents, comme nous pouvons le voir dans la troisième colonne du tableau 3.2 : *faculté* et *fac*.

Le tableau 3.2 page suivante montre des exemples de termes annotés manuellement et de termes annotés avec la ressource du LEFFF. Lorque le terme a été trouvé dans le LEFFF et qu'il est correct (c'est le cas du terme *université*), on donne la valeur de 1. En revanche, si nous n'avons pas trouvé d'entrée lexicale correspondante à notre terme, nous donnons la valeur de 0 (*étudiant*).

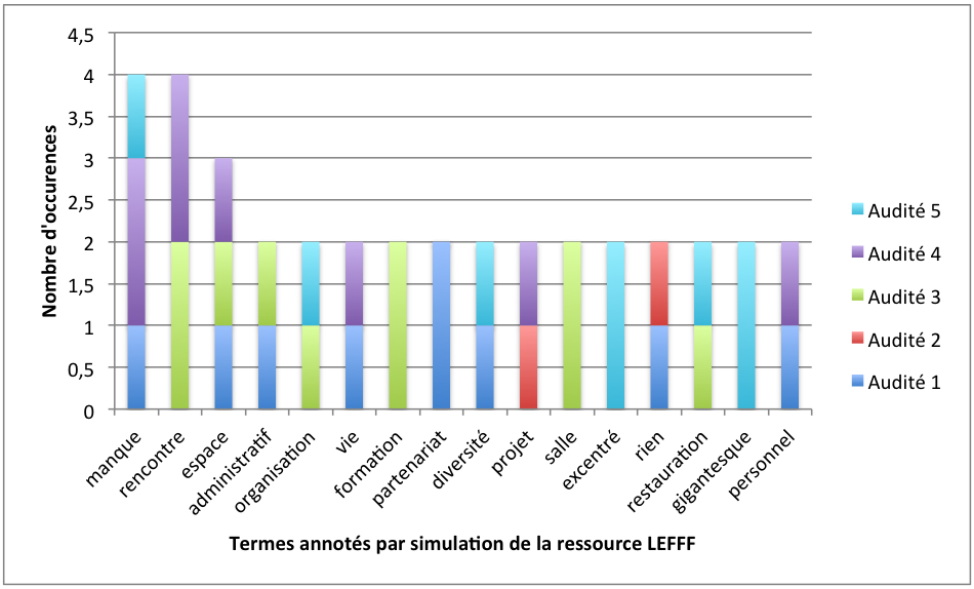


FIGURE 3.6 – Intersection lexicale : résultats obtenus sur l'étude β

Termes	Annotation manuelle	Lemme du LEFFF	Résultat
université	université	université	1
étudiant	étudiant		0
étudiants	étudiant		0
faculté	faculté	faculté	0
fac	faculté	fac	0

TABLE 3.2 – Comparaison de l'intersection lexicale

3.2.2 Proximité sémantique

Nous avons fait la simulation par les ressources avec le DES et DIKO (la description des deux ressources se trouve dans la section 2.3.2.0 page 54)

Pour chaque mot présent dans le DES, plusieurs cliques sont données et nous avons donc dû faire un choix en fonction du terme à identifier. Nous avons pu observer que, bien souvent, il y avait deux cliques pour le même mot alors que nous avons identifié un seul et même concept manuellement. C'est le cas de *faculté* qui est dans la même clique que celle d'*université* et celle de *campus* (*université* et *campus* sont vus comme deux cliques différentes alors que nous avons vu, dans

l'attendu, un seul concept). C'est aussi le cas du terme *intelligence* qui apparait dans deux cliques distinctes, celle de *connaissance* et celle d'*ouverture d'esprit* mais c'est aussi le cas de *pluralité* qui apparaît avec
textitdiversité et *majorité*. Il s'agit peut-être d'un défaut de représentation du point de vue de l'application sur nos données. Peut-être que la ressource du DES est trop précise ou subtile et ne correspond pas vraiment à nos données.

Nous avons trouvé au total 9 cliques dans le DES :

— académie, collège, école, faculté, institut, université
— écolier, élève, étudiant
— agréable, aimable, beau, bon, charmant, doux, exquis, gentil, gracieux, joli
— compréhension, conception, connaissance, entendement, faculté, intelligence
— bigarré, composite, disparate, divers, hétéroclite, hétérogène, varié
— agencement, arrangement, composition, disposition, ordonnance, ordre, organisation, structure
— colossal, démesuré, énorme, extraordinaire, gigantesque, grand, immense, monumental, titanesque
— diversité, multiplicité, multitude, pluralité
— espace, lieu, région, zone
— ancien, antique, archaïque, démodé, dépassé, désuet, fossile, périmé, suranné, vétuste, vieilli, vieillot, vieux

Nous avons également trouvé 9 termes associés dans DIKO qui correspondents aux cliques du DES :

— université, faculté, fac, école
— structure, organisation
— agréable, bon
— étudiant, élève
— connaissance, intelligence
— varié, divers
— espace, lieu, zone
— vétuste, vieux
— gigantesque, grand

Terme	Annotation manuelle	LikeIt	Polarité FEEL	Emotions associées FEEL
Agréable	positif	positif	positif	0 ;0 ;0 ;0 ;0 ;0
Bâtiment		neutre	positif	0 ;0 ;0 ;0 ;0 ;0
Avoir recours		positif	positif	0 ;0 ;0 ;0 ;0 ;0
Manque de	négatif	négatif		
Multidisciplinaire	positif	positif		
Accessibilité	positif	positif		
Parfaitement	positif	positif		

TABLE 3.3 – Comparaison de la polarité

Comme nous pouvons le constater, les deux ressources repérent les mêmes mots, par exemple les termes associés *étudiant, élève* de DIKO correspondent à la clique *écolier, élève, étudiant* du DES.

3.2.3 Polarité des opinions

Pour la polarité, la simulation des ressources s'est effectuée avec LikeIt et French Expanded Emotion Lexicon (FEEL), la description des deux ressources se trouve dans la section 2.3.2.0 page 55. La simulation par ces ressources n'est pas évidente. En effet, presque tous les termes sont ambigus. C'est le cas par exemple du terme *bâtiment* qui peut être utilisé pour désigner une construction ou dans le domaine maritime un grand bateau. Dans l'étude β, il s'agit d'une construction (figure 3.3). LikeIt et FEEL ne donnent qu'une seule entrée lexicale pour ce terme sans distinction de sens, comment pouvons-nous savoir si le sens donné correspond bien à notre terme ?

Comme nous pouvons le voir dans le tableau 3.3, les termes facilement polarisables sont bien reconnus (par exemple, *agréable*). Les deux ressources proposent des mots mais aussi des locutions (telles que *avoir recours*) qui peuvent être très utiles et que nous n'avons pas annoté puisque nous nous étions focalisée sur les mots et non sur les groupes de mots en pensant que cela permettrait une meilleure comparaison. Bien que comprenant 14 000 termes, nous notons que FEEL ne donne pas d'entrée lexicale pour de nombreux mots, certains mots que nous considérons comme plutôt courants manquent : *manque*, *multidisciplinaire*, *accessibilité* et *par-*

faitement. Nous avons également constaté que pour de nombreux mots, la ressource n'associe pas d'émotions.

3.3 Comparaison et évaluation de l'attendu et de l'obtenu

Les comparaisons et évaluations présentées ci-dessous ont été effectuées à l'aide des métriques présentées dans la section 2.3.3 page 57.

3.3.1 Intersection lexicale

Un exemple d'annotation est diponible en Annexe D.1 page 95 sur les réponses de quatre audités à la même question de l'étude β. Nous avons adpaté les métriques à nos besoins, nous considérons comme solutions pertinentes, les termes qui ont été trouvés dans la ressource du LEFFF et qui ont le lemme attendu dans l'annotation manuelle (résultat attendu).

Nous avons de bons résultats avec la simulation par le LEFFF. En effet, comme le montre le tableau 3.5 page 73, le score de précision est à 1 et les scores du rappel et de la F-mesure sont plutôt hauts (respectivement 0.829 et 0.906). Nous avons obtenu un peu de silence (solutions pertinentes qui n'ont pas été trouvées) puisque certains termes (*étudiant*, *structure* et *collectif*) n'ont pas été trouvés ou ne correspondaient pas à nos attentes (*faculté* et *fac*). Cependant, nous n'avons pas obtenu de bruit, c'est-à-dire qu'aucune solution non pertinente n'a été trouvé et que la ressource est capable de les refuser.

Pour conclure, nous pouvons dire que l'utilisation de la ressource du LEFFF et de la lemmatisation s'avèrent utiles pour repérer l'intersection lexicale. Cependant, il y a un inconvénient à utiliser la lemmatisation. Par exemple, le nom *discipline* et l'adjectif *disciplinaire* n'ont pas les mêmes lemmes, or, il s'agit de la même idée et la lemmatisation ne nous permettra pas de le prendre en compte. Ce cas ne s'est pas présenté dans l'étude β mais il pourrait apparaître dans les études à venir et dans ces cas-là, la racinisation semble plus adaptée bien qu'elle présente aussi des

Rappel =	0,829
Précision =	1
Silence =	0,171
Bruit =	0
F-mesure =	0,906

TABLE 3.4 – Intersection lexicale : calculs des métriques

inconvénients.

3.3.2 Proximité sémantique

Comme nous pouvons le constater sur la figure 3.7 page ci-contre, nous obtenons plus de concepts similaires avec l'annotation manuelle. Cependant, ces résultats ne sont pas vraiment comparables puisque dans l'attendu nous avons annoté des suites de mots qui correspondent à des concepts alors que les ressources utilisées sont monolexicales et n'annotent qu'un seul mot. C'est pourquoi nous n'avons pas calculé les métriques pour la proximité sémantique simulée avec le DES et DIKO. Ce n'est pas suffisant et il nous faudra trouver des solutions permettant de repérer la proximité sémantique.

3.3.3 Polarité des opinions

Afin d'évaluer l'écart entre l'attendu et l'obtenu au niveau de la polarité, nous avons adapté les métriques présentées précédemment. Nous considérons comme corrects les termes annotés avec la même polarité que dans l'annotation des résultats attendus. Un exemple mettant en vis-à-vis les trois annotations sur les réponses de quatre audités à la même question de l'étude β est diponible en Annexe D.2 page 96. Nous considérons comme :

— nombre de solutions pertinentes proposées, le nombre de termes annotés correctement par les ressources
— nombre de solutions proposées, tous les termes proposés par les ressources (annotés correctement ou non)

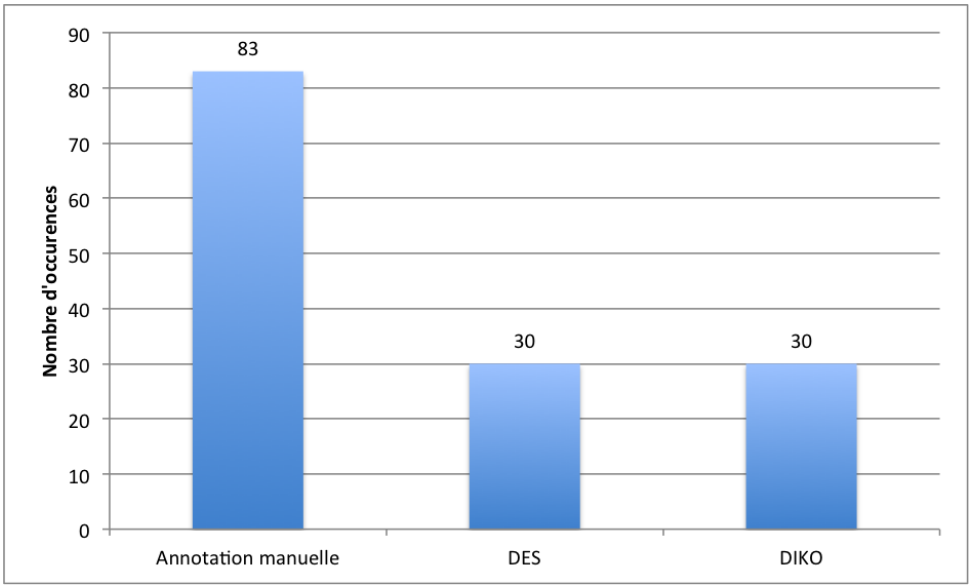


FIGURE 3.7 – Proximité sémantique : nombre d'occurrences obtenues sur l'étude β selon les trois annotations

— nombre de solutions pertinentes non proposées, ce qui n'a pas été annoté mais qui était correct
— nombre de solutions non pertinentes proposées, les termes annotés mais qui ne sont pas corrects

La figure 3.8 page suivante présente le nombre de termes annotés avec leur polarité pour chacune des trois annotations : l'annotation manuelle, la simulation avec la ressource LikeIt et la simulation avec la ressource FEEL. Nous pouvons remarquer pour chacune des trois annotations, la majorité des termes qui sont annotés sont positifs. Nous pouvons aussi observer que les deux annotations faites à partir des ressources LikeIt et FEEL (à droite de la figure) obtiennent un plus grand nombre de termes ayant une polarité positive. Alors que nous avons annoté 71 termes positifs, LikeIt et FEEL en ont trouvé respectivement 180 et 188. Cela vient du fait que nous avons annoté seulement les termes que nous trouvions pertinents. Peu de mots ont été annotés manuellement pour une seconde raison : dans un souci d'objectivité, les termes pour lesquels nous avions des doutes et pour lesquels la polarité n'était pas évidente à déterminer n'ont pas été annotés. L'annotation est

une tâche complexe, elle pourrait être faite par plusieurs annotateurs afin de valider les choix qui ont été faits et ainsi rendre l'attribution de la polarité plus subjective.

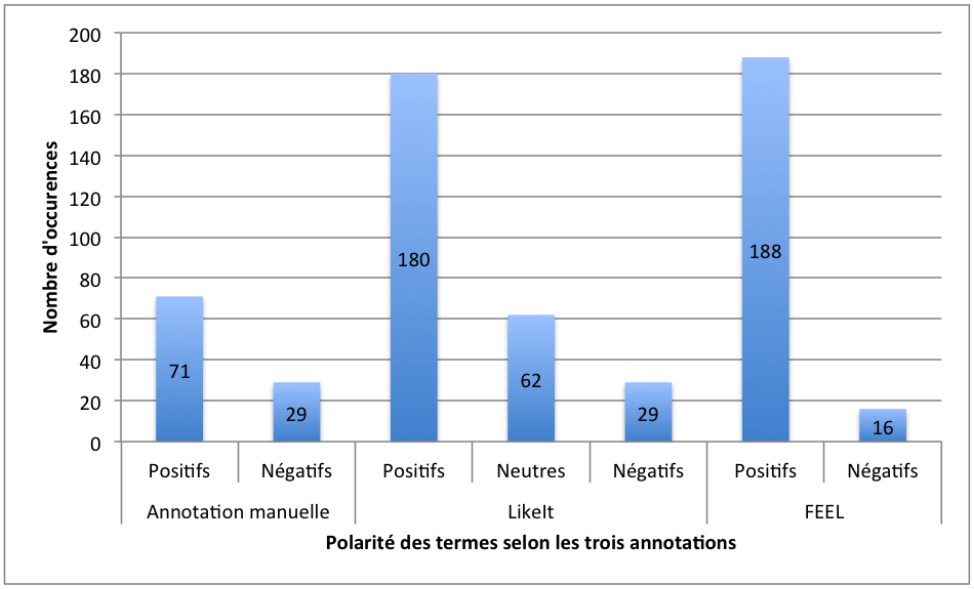


FIGURE 3.8 – Polarité des opinions : résultats des trois annotations

Le tableau 3.5 page suivante donne les résultats des calculs des différentes métriques pour LikeIt et Feel. Pour LikeIt, nous obtenons un fort rappel et par conséquent, le silence est bas. Cette ressource permet donc de donner un grand nombre de solutions pertinentes parmi toutes les solutions trouvées. La raison pour laquelle LikeIt obtient un silence si faible est due au fait qu'un seul terme n'a pas été trouvé sur les 100 termes attendus, il s'agit de *pas*. En revanche, la précision est assez basse et le bruit est élevé, c'est-à-dire que la ressource a donné des solutions non pertinentes. Pour FEEL, le rappel est moyen et le silence bas. Ces deux scores sont moins bons que pour LikeIt. En effet, sur la totalité des solutions pertinentes (au nombre de 100), 29 solutions pertinentes n'ont pas été trouvées, contre 1 solution pertinente non trouvée par LikeIt. La précision est basse et le bruit élevé comme pour LikeIt.

LikeIt obtient un meilleur taux de rappel que FEEL. Quant à la précision, elle est légèrement plus forte pour FEEL que pour LikeIt mais ce n'est pas significatif.

Métrique	LikeIt	FEEL
Rappel	0,840	0,690
Précision	0,310	0,338
Silence	0,010	0,290
Bruit	0,631	0,647
F-mesure	0,453	0,454

TABLE 3.5 – Polarité des opinions : calculs des métriques (LikeIt)

En prenant en compte les quatre mesures du rappel, de la précision, du silence et du bruit, nous appercevons que les résultats sont meilleurs pour LikeIt. Bien que le rappel pour les deux ressources soit plutôt bons, les scores de F-mesure sont très bas et cela nous montre que ces ressources ne sont pas suffisantes pour la tâche de l'annotation de la polarité.

Pour conclure sur les résultats de l'étude β, sur le plan de l'intersection lexicale, les résultats sont très positifs, nous avons obtenu un score de précision de 1. En revanche, les apports du TAL sur la proximité sémantique et la polarité des opinions ne sont pas suffisants en l'état. Ces ressources peuvent nous permettrent d'extraire seulement un peu de vocabulaire pertinent des sentiments perçus. Travailler au niveau des mots ne semble pas être suffisant, il faut prendre en compte le contexte dans lequel le mot apparait et travailler sur les suites contiguës de plus d'un mot. La négation est également très importante à prendre en compte. Nous pouvons améliorer ces résultats en constituant nous-même notre propre ressource. C'est l'aspect que nous développons en tant que travail à venir 3.2 page 76 avec notre sixième hypothèse.

Conclusion et perspectives

1 Conclusion

Pour conclure, nous avons pu aborder au cours de notre stage autant les aspects industriels au sein de COMONGO avec le développement d'une interface de saisie pour collecter du corpus, que les aspects de la recherche avec les problématiques scientifiques posées par la demande de l'entreprise, notamment au niveau de *l'opinion mining*, des ressources sémantiques et des interfaces. Nous avons ainsi pu faire une étude des existants et des usages afin de situer notre démarche.

Dans le chapitre 1, nous avons émis l'hypothèse qu'une transition d'une pratique présentielle papier à une pratique distancielle numérique est possible en passant par une phase de transition. Les outils du TAL sont nécessaires à la réalisation de notre projet comme nous avons pu le démontrer dans ce travail. Enfin, les trois dernières hypothèses nous ont permi d'identifier des ressources pour l'intersection lexicale, la proximité sémantique et la polarité des opinions.

Nous avons exposé, dans le chapitre 2, la méthode mise en place pour la constitution du corpus et le traitement des données. Nous avons ensuite présenté la façon dont ont été constitué l'attendu et l'obtenu (simulation par les ressources) ainsi que les métriques utilisées pour l'évaluation.

Enfin, dans le chapitre 3, nous avons abordé les résultats de l'étude β du point de vue de l'analyse linguistique. Après avoir évalué la simulation par les ressources sur les trois niveaux - intersection lexicale, proximité sémantique et polarité des opinions - nous concluons que ces dernières ne sont pas suffisantes pour notre projet. Il est nécessaire de construire une ressource adaptée à nos besoins en rajoutant

un module incrémental que nous développons dans la section 3.2 suivante.

2 Construction incrémentale des ressources

Le prototype que nous avons développé pour collecter notre corpus fonctionne. Il nous faudrait à présent construire une ressource suffisante pour analyser l'image voulue et l'image perçue. Nous souhaiterions utiliser un mécanisme de construction incrémentale des ressources. Cette hypothèse de travail est très importante pour la continuation du projet, ce serait un moyen de continuer à la fois le travail sur le corpus tout en enrichissant nos ressources, notamment du point de vue de la proximité sémantique et de la polarité des opinions.

La construction incrémentale des ressources constitue notre sixième et dernière hypothèse qui est la suite de notre travail sur ce projet. Permettre à l'utilisateur ou au consultant de compléter l'annotation automatique augmente la base de connaissances linguistiques du système. Chaque nouvelle étude se fonde donc sur une ressource améliorée par la précédente. Le recueil des données va permettre d'enrichir automatiquement nos ressources.

Au niveau de l'intersection lexicale, nous pouvons utiliser la lexicalisation, lemmatisation et racinisation. Au niveau de la proximité sémantique, nous avons simulé le résultast de deux ressources sur notre étude et nous nous sommes rendu compte que cela ne correspondait pas tout à fait à nos attentes. En effet, les ressources que nous avons trouvées ne sont pas adaptées à notre problématique. Au niveau de la polarité des opinions, certaines choses qui devraient être polarisées ne le sont pas et inversement, certaines choses qui ne devraient pas être polarisées le sont. Les ressources concernant la polarité des opinions ou les émotions associées ne correspondent pas à nos attentes et nous paraissent insuffisantes pour notre projet. En effet, l'attribution de la polarité est toujours fondée sur une question de point de vue. Les ressources sur la polarité sont constituées par rapport à quelque chose de particulier, un domaine spécifique.

Puisque les ressources existantes ne sont pas suffisantes pour correspondre à nos besoins, construire nos ressources de façon incrémentale serait un moyen, tout

en travaillant sur le corpus, d'enrichir nos ressources sur les trois niveaux (intersection lexicale, proximité sémantique et polarité des opinions). Ce serait un moyen d'identifier les termes identiques, de construire des ensembles de synonymes et de constituer un lexique polarisé. Cela nous permettrait de disposer d'une ressource spécifique à notre projet et au domaine dans lequel nous travaillons. L'objectif est de construire une ressource suffisante pour analyser la différence entre l'image voulue et l'image perçue.

L'outil que nous avons développé sera réutilisé et nous pourrons ainsi utiliser l'expertise de l'usager qui va enrichir ces ressources. Pour la proximité sémantique, nous pouvons créer une table dans la Base De Données avec un identifiant pour chaque liste de mots et nous devons utiliser la méthode incrémentale pour l'enrichir. Nous pourrions par exemple, proposer une liste de mots ou des « cliques » (comme dans le Dictionnaires Electronique des Synonymes du CRISCO) aux audités de façon interactive et leur demander si ce clique correspond à ce qu'ils ont voulu dire. Il s'agit d'utiliser en quelque sorte du *crowdsourcing*. Pour la polarité, comme nous l'avons vu précédemment, il existe différentes méthodes afin de construire une ressource permettant de définir la polarité d'un terme. Nous pourrions donc nous appuyer sur notre corpus (annoté ou non) afin de créer notre propre ressource (adaptée à nos données). Détecter seulement la polarité peut paraître insuffisant, nous pourrions également ajouter des critères de l'ordre du subjectif/objectif ou nous intéresser aux émotions associées aux opinions, comme le font certaines ressources.

3 Perspectives de travail

Les perspectives de travail sur les trois ans à venir sont multiples. Il faudra continuer à travailler sur les aspects évoqués dans ce mémoire, à savoir : le travail sur l'interface de saisie dans le but de l'améliorer, les analyses linguistiques du corpus et l'identification des ressources nécessaires. De nouvelles tâches vont également se rajouter telles que l'approche de la donnée par représentations et l'évaluation des ressources que nous aurons construites.

3.1 Amélioration de l'outil

L'amélioration de l'interface et plus particulièrement l'amélioration de l'interface de production va poser d'autres questions. Par exemple, en distanciel, il faut pouvoir gérer la possibilité que l'audité arrête le questionnaire en cours. Plusieurs solutions peuvent être proposées : nous pouvons sauvegarder les questions auxquelles il a déjà répondu mais pour cela il faut qu'il ait au moins terminé une question. Nous pouvons également le prévenir qu'il peut sortir et revenir au questionnaire plus tard. Ensuite, il faudrait pouvoir envoyer automatiquement un message à l'audité pour le relancer.

Au niveau de l'export des études, dans l'interface de gestion, il faut pouvoir donner le choix entre différents axes de lectures au consultant. Par exemple, pouvoir afficher les résultats de toutes les études clôturées, les premiers résultats des études en cours, etc.

A partir des informations entrées dans la Base De Données et avec l'expertise des audités, il nous faudra construire un outil permettant d'analyser les données que nous pourrons ensuite ajouter à l'application développée pour COMONGO.

3.2 Analyse des données

Dans le travail à venir, il s'agira de comparer l'étude β, que nous avons analysée dans le chapitre 3 page 59, avec les deux autres études α et δ. La comparaison avec l'étude α sera très importante puisqu'elle a été réalisée avec la pratique présentielle papier et qu'elle nous permettra de nous rendre compte de l'impact du numérique dans la variation des productions.

Lorsque nous aurons de nouvelles données, il s'agira de voir si les analyses sont suffisantes et de voir dans quelle mesure une amélioration est envisageable. Il faudra également pouvoir analyser les études selon différents axes de lecture, en prenant en compte : une seule réponse d'un seul audité à une question, toutes les réponses du même audité à différentes questions, toutes les réponses de tous les audités à la même question ainsi que toutes les réponses de tous les audités aux différentes questions.

3.3 Identification et construction des ressources sémantiques nécessaires

Nous allons devoir constituer notre propre ressource de façon incrémentale. L'objectif est, comme nous l'avons évoqué dans la section 3.2 page 76, de construire une ressource suffisante pour analyser la différence entre l'image voulue et l'image perçue. La technique de construction des données peut-être mise en place sur les trois années à venir.

Du point de vue de l'intersection lexicale, nous pouvons proposer une coloration des termes identiques lors de l'affichage des résultats, en parcourant le premier texte et en regardant pour chaque mot, s'il est présent dans les textes suivants. Comme nous l'avons évoqué dans le chapitre 3, les ressources sont insuffisantes sur le plan de la proximité sémantique puisqu'elles sont monolexicales alors que nous travaillons sur des suites contiguës de plus d'un mot. Concernant la polarité, nous faisons la même remarque, affecter une polarité seulement sur les mots n'est pas suffisante, il faut également s'occupper de la polarité des segments (ou groupes de mots). Nous pourrions effectuer des expérimentations avec le module Python Textblob[12] qui permet de réaliser plusieurs tâches du TAL comme l'étiquetage morphosyntaxique, la traduction, etc., mais il permet également de faire de l'analyse de sentiment mais en anglais. Ce module prend un texte en entrée, et fournit en sortie des scores de polarité (de -1.0 à 1.0) et de subjectivité (de 0.0 à 1.0, 0.0 étant objectif et 1.0 subjectif) pour chaque phrase.

3.4 Approche de la donnée par représentations

Il s'agira également de produire des représentations pour accompagner la démarche analytique. L'approche de la donnée par représentations est un aspect essentiel du travail à venir. En effet, il faudrait pouvoir fournir des représentations synthétiques et analytiques permettant de visualiser le résultat des traitements TAL.

Il existe différents modes de représentation, il nous faudra donc choisir ou

12. `https://textblob.readthedocs.io/en/dev/` (consulté le 01/06/2017)

combiner ce qui nous paraîtra le mieux adapté à notre problématique. Un seul mode de représentation n'est peut-être pas suffisant, c'est pourquoi, il nous faudra probablement effectuer des tests en proposant différentes représentations. Une représentation possible est le diagramme de Venn.

Ces représentations nous permettront de voir s'il y a une adéquation entre l'image voulue et l'image perçue et elles nous permettront aussi d'identifier ce qui ne va pas et d'où vient le problème. Si quelque chose ne va pas, il faudra alors remettre en question les modalités de représentations, modèles et ressources : soit le modèle de représentation n'est pas adéquat, soit les modèles de calculs ne sont pas suffisants ou incomplets, soit les ressources ne sont pas adaptées ou insuffisantes.

3.5 Evaluation des ressources

Pour finir, nous devrons tester nos ressources afin de voir si l'outil fonctionne, si des améliorations sont à prévoir et si le consultant est satisfait. Le consultant a un besoin et une capacité d'interprétation, il est capable de dire si globalement cela convient. C'est pourquoi l'attendu est très important. Si les résultats obtenus sont différents, il faudra revoir les ressources. C'est donc un cycle empirique sur la conception de l'outil : l'outil évoluera en fonction de l'obtenu.

Nous évaluerons donc nos ressources sur le principe attendu/obtenu à l'aide des métriques d'évaluation du bruit (les solutions non pertinentes trouvées par un système) et du silence (les solutions pertinentes trouvées par un système) que nous avons utilisé dans le chapitre 3 page 59.

Bibliographie

[Abdaoui et al., 2016] Abdaoui, A., Azé, J., Bringay, S., & Poncelet, P. (2016). FEEL : a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, (pp. 1–23).

[Agarwal & Bhattacharyya, 2006] Agarwal, A. & Bhattacharyya, P. (2006). Augmenting wordnet with polarity information on adjectives. In *3rd International Wordnet Conference, Jeju Island, Korea, South Jeju (Seogwipo)*.

[Andreani, 2011] Andreani, V. A. (2011). *Immersion in scientific and technical documents : units, theoretical models and processes*. Theses, Université de Grenoble.

[Asher et al., 2008] Asher, N., Benamara, F., & Mathieu, Y. Y. (2008). Distilling Opinion in Discourse : A Preliminary Study. In *22nd International Conference on Computational Linguistics (COLING2008)* Manchester, United Kingdom.

[Augustyn et al., 2006] Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., & Rinck, F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.* Grenoble, France.

[Bouchekif, 2016] Bouchekif, A. (2016). *Structuration automatique de documents audio*. PhD thesis, Université du Maine.

[Brun, 2011] Brun, C. (2011). Un système de détection d'opinions fondé sur l'analyse syntaxique profonde. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles* Montpellier, France : Association pour le Traitement Automatique des Langues.

[Carr, 2012] Carr, N. (2012). Internet rend-il bête ? réapprendre à lire et à penser dans un monde fragmenté (the shallows), paris, robert laffont, 2011 (2010 pour l'édition originale), 324 pages.

[Claveau & Kijak, 2015] Claveau, V. & Kijak, E. (2015). Thésaurus distributionnels pour la recherche d'information et vice-versa. *Document Numérique*, 18(2-3), 101–121.

[Clément et al., 2004] Clément, L., Lang, B., & Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *LREC 04* (pp. 1841–1844). Lisbonne, Portugal.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, (pp. 169–200).

[El-Bèze et al., 2010] El-Bèze, M., Jackiewicz, A., & Hunston, S. (2010). Numéro thématique « opinions, sentiments et jugements d'évaluation ».

[Esuli & Sebastiani, 2006] Esuli, A. & Sebastiani, F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation LREC'06* (pp. 417–422).

[Gala & Brun, 2012] Gala, N. & Brun, C. (2012). Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse de sentiments. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles* (pp. 495–502). Grenoble, France : Association pour le Traitement Automatique des Langues.

[Gala & Rey, 2008] Gala, N. & Rey, V. (2008). Polymots : une base de données de constructions dérivationnelles en francais à partir de radicaux phonologiques. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles* Avignon, France : Association pour le Traitement Automatique des Langues.

[Lafourcade, 2007] Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing* (pp.7̃). Pattaya, Chonburi, Thailand.

[Lafourcade et al., 2015] Lafourcade, M., Le Brun, N., & Joubert, A. (2015). Vous aimez ?...ou pas ? LikeIt, un jeu pour construire une ressource lexicale de polarité. In *TALN : Traitement Automatique des Langues Naturelles* Caen, France.

[Mohammad & Turney, 2013] Mohammad, S. & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

[New et al., 2001] New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du francais contemporain sur internet : Lexique. *L'année psychologique*, 101(3), 447–462.

[OCDE, 2000] OCDE (2000). *La littératie à l'ère du numérique*. Technical report, Organisation de Coopération et de Développement Numérique.

[Vegnaduzzo, 2004] Vegnaduzzo, S. (2004). Acquisition of subjective adjectives with limited resources. In *Actes AAAI spring symposium on exploring attitude and affect in text : Theories and applications* Stanford, USA.

Annexe A

Copies d'écran de l'interface de gestion

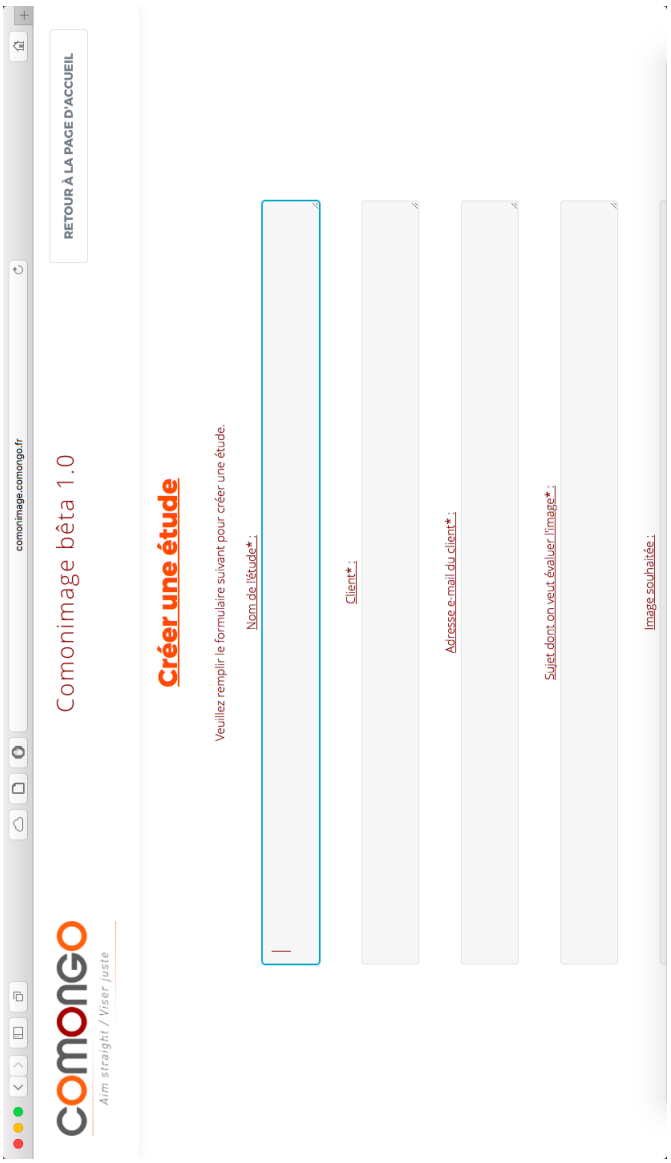


FIGURE A.1 – Copie d'écran de l'interface gestion : création d'une étude (1)

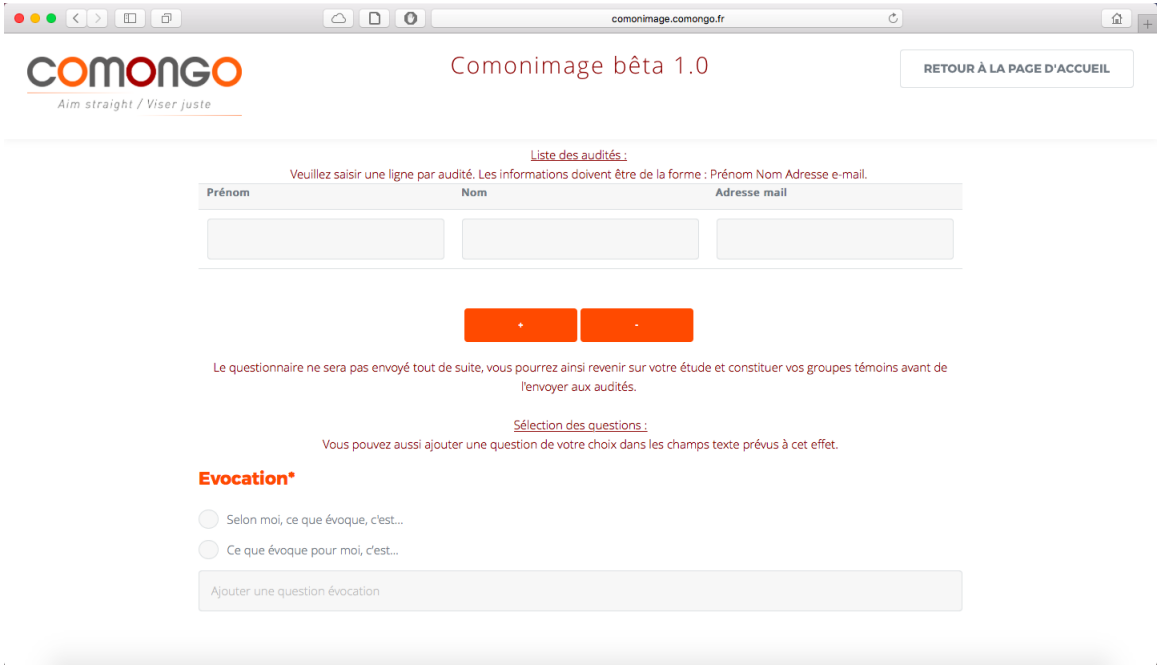


FIGURE A.2 – Copie d'écran de l'interface gestion : création d'une étude (2)

Annexe B

Copies d'écran de l'interface de production

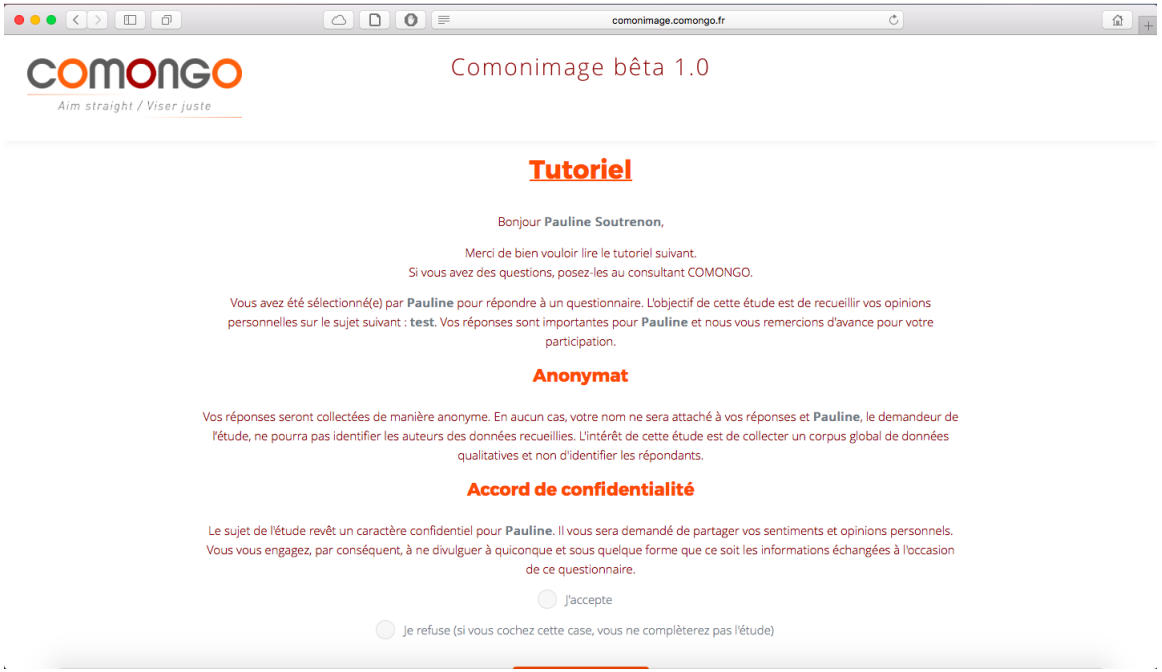


FIGURE B.1 – Copie d'écran de l'interface production : page de tutoriel

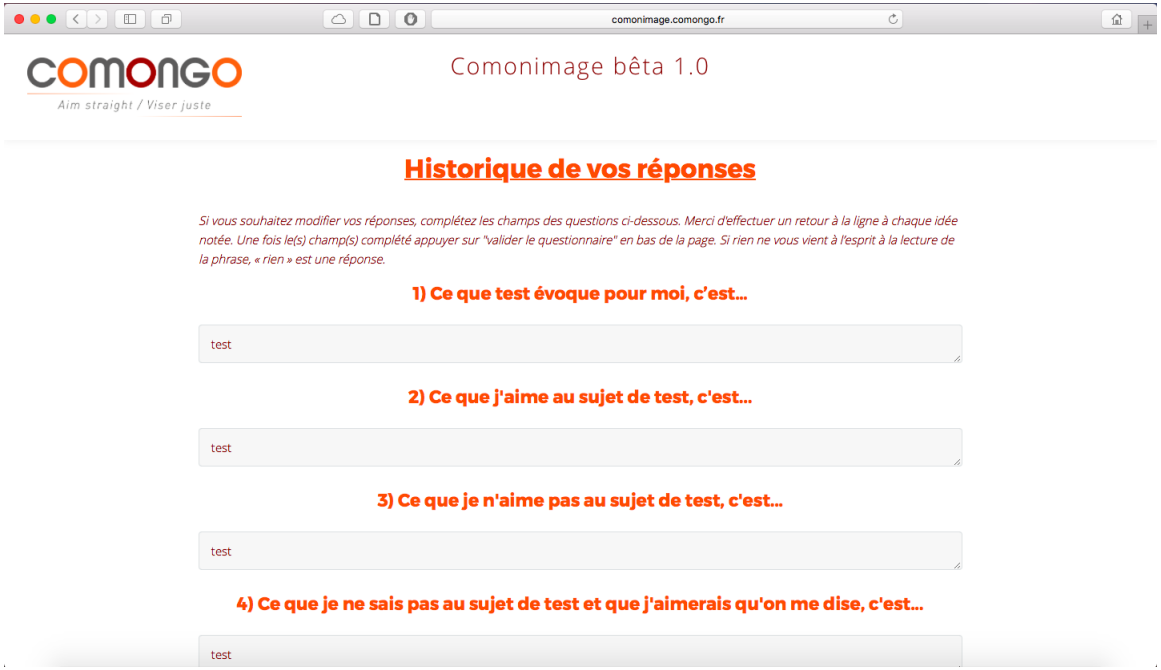


FIGURE B.2 – Copie d'écran de l'interface production : historique des réponses

Annexe C

Document d'annotation : DTD

```
<!ELEMENT corpus (etude+)>
<!ELEMENT etude (question+)>
<!ELEMENT question (reponse+)>
<!ELEMENT reponse (#PCDATA | racine | syno | senti)*>
<!ELEMENT syno (#PCDATA | senti | racine)*>
<!ELEMENT racine (#PCDATA | senti)*>
<!ELEMENT senti (#PCDATA | senti | racine)*>

<!ATTLIST etude
    idEtude CDATA #REQUIRED
    idCreateur CDATA #REQUIRED
    nom CDATA #REQUIRED
    numeroDossier CDATA #REQUIRED
    donneurOrdre CDATA #REQUIRED
    contact CDATA #REQUIRED
    sujet CDATA #REQUIRED
    statut CDATA #REQUIRED
    groupesTemoins CDATA #IMPLIED
    dateCreation CDATA #REQUIRED
    dateEnvoi CDATA #REQUIRED
    dateCloture CDATA #REQUIRED
>

<!ATTLIST question
    idQuestion CDATA #REQUIRED
    texteQuestion CDATA #REQUIRED
    type CDATA #REQUIRED
>

<!ATTLIST reponse
    idReponse CDATA #REQUIRED
    idAudite CDATA #REQUIRED
    numeroGroupe CDATA #IMPLIED
>

<!ATTLIST senti
    polarite (positif|neutre|negatif) #REQUIRED
    likeIt CDATA #IMPLIED
    feel CDATA #IMPLIED
    commentaire CDATA #IMPLIED
>

<!ATTLIST syno
    idSyno CDATA #REQUIRED
    crisco CDATA #IMPLIED
    diko CDATA #IMPLIED
    commentaire CDATA #IMPLIED
>

<!ATTLIST racine
    idRacine CDATA #REQUIRED
    lefff CDATA #IMPLIED
>
```

FIGURE C.1 – DTD (Document Type Definition) utilisée pour l'annotation de l'étude β

Annexe D

Comparaison des résultats attendus et obtenus sur l'étude β

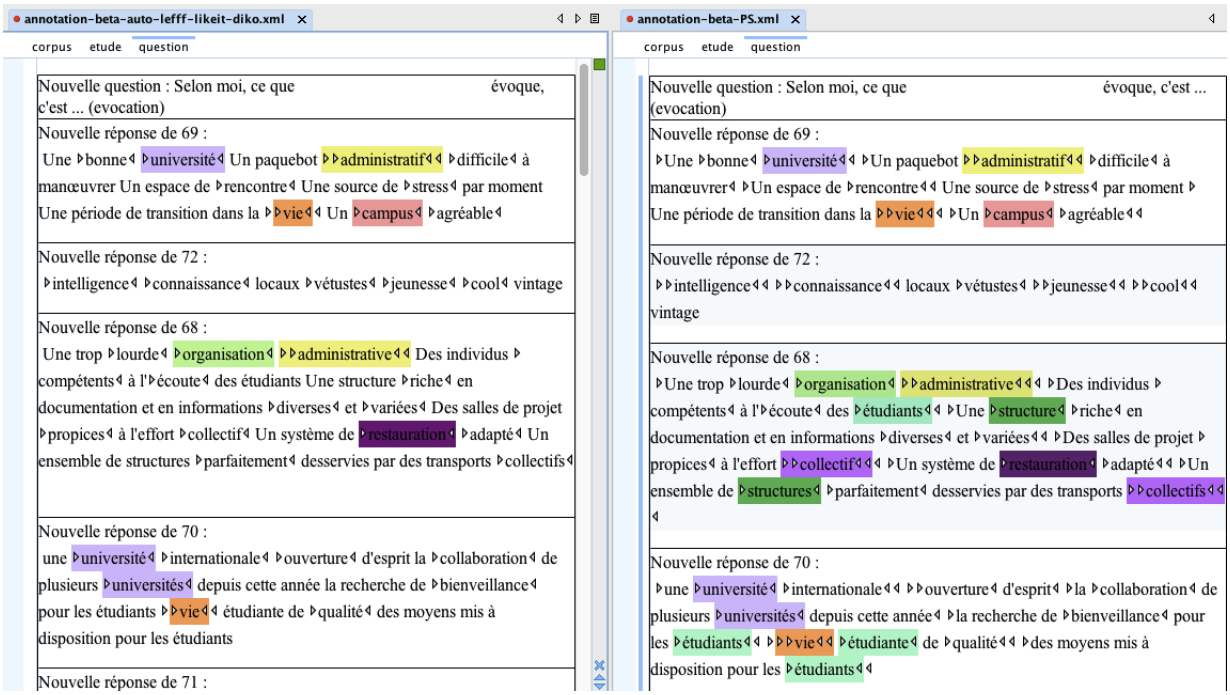


FIGURE D.1 – Intersection lexicale : exemple d'annotation sur les réponses de 4 audités à la même question (étude β)

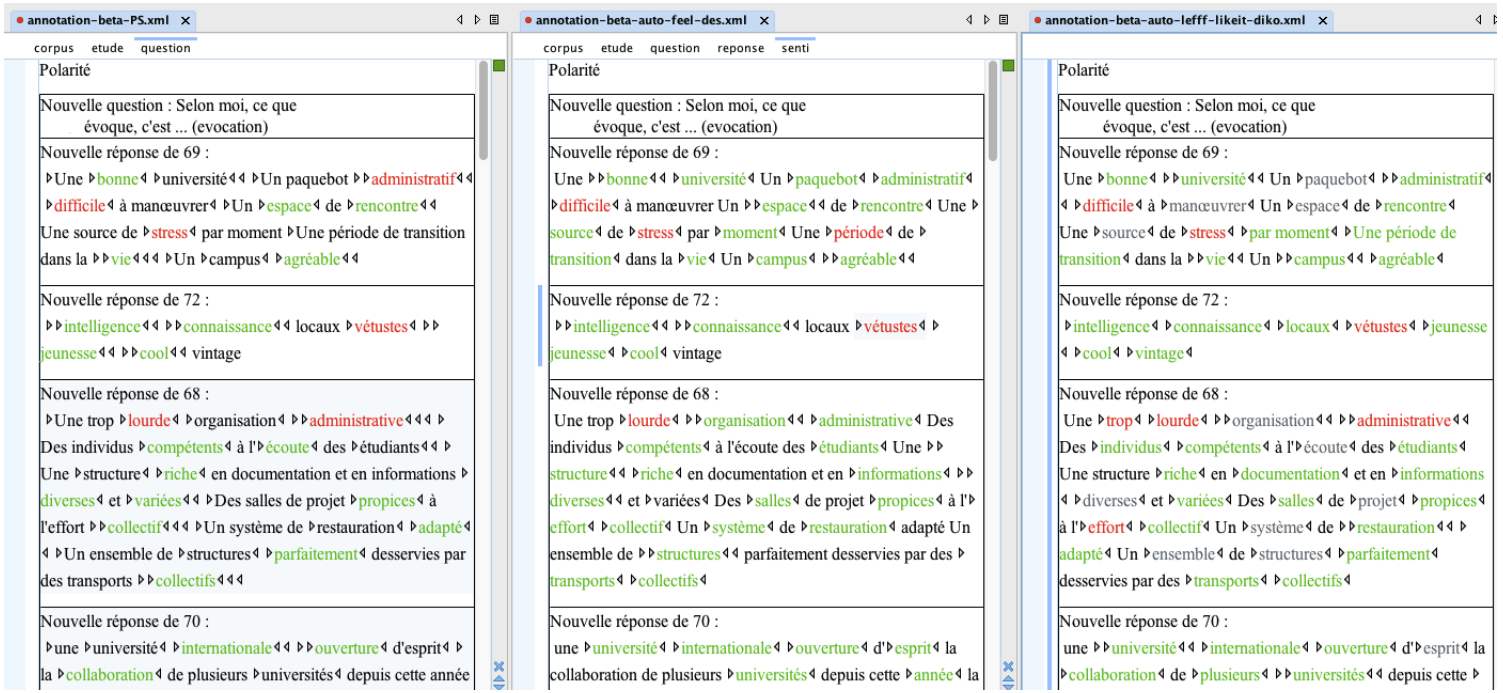


FIGURE D.2 – Polarité des opinions : exemple d'annotation sur les réponses de 4 audités à la même question (étude β)

Résumé

Le travail que nous présentons se situe dans le cadre d'un projet de recherche en collaboration entre l'Université Grenoble Alpes et l'entreprise COMONGO dont le coeur de métier est l'accompagnement et la gestion d'image des personnes morales et physiques. Notre démarche a consisté dans un premier temps à transposer une pratique en *focus group* vers une pratique distancielle numérique. Dans un second temps, il s'est agi d'intégrer les outils du Traitement Automatique des Langues, notamment les ressources sémantiques, à cette démarche professionnelle d'entreprise.

Cette transformation d'une pratique métier nous a menée à poser trois grandes hypothèses : la transition numérique a un impact sur la qualité des données ; les ressources sémantiques permettent une meilleure appréhension des données textuelles traitées mais s'avèrent insuffisantes après simulation ; une démarche incrémentale d'amélioration des ressources doit être envisagée afin d'obtenir des traitements optimaux.

Cette première expérimentation sur données réelles permet de poser les bases d'un projet de recherche et développement à plus long terme au sein de la société COMONGO alliant les domaines de la linguistique de corpus, du Traitement Automatique des Langues et des sciences de l'information et de la communication.

Mots-clés: TAL, opinion mining, ressources sémantiques, interfaces numériques, transition numérique

Abstract

The work that we present is part of a collaborative research project between the University Grenoble Alpes and the company COMONGO which provides image management of legal and physical persons. In a first step, we have transposed a

focus group based approach into a distant digital practice. In a second step, we have investigated the integration of NLP tools, in particular semantic resources, into this professional approach.

This digital transformation of a practice led us to make three main assumptions : the digital transition has an impact on the quality of the data ; the semantic resources allow a better understanding of the textual data processed but are insufficient after simulation ; an incremental approach to resources improvement should be considered in order to obtain optimal results.

This first experimentation with real data allows us to lay the foundations for a longer-term research and development project within the company COMONGO combining the fields of corpus linguistics, Natural Language Processing and Information and Communication Sciences.

Keywords: NLP, opinion mining, semantic resources, digital interfaces, digital transition