
Analyse des données qualitatives avec le Sphinx

Les études « quali »

Les situations de recherche

Les « études quali » se définissent bien souvent par opposition aux « études quanti ». On indique ainsi que les informations étudiées proviennent de sources multiples documents, écrits ou discours et qu'elles sont analysées sans recourir au calcul.

Comme nous le verrons cette simplification est abusive même si elle définit assez bien les situations de recherche dans lesquelles on a recours aux approches qualitatives.

Les enquêtes : recueillir des informations nouvelles.

Le qualitatif apparaît avec la présence de questions ouvertes dans les questionnaires. Celles-ci échappent aux analyses statistiques qui font l'efficacité des dépouillements classiques et sont trop souvent tout simplement oubliées au moment de l'analyse.

Avec les guides d'entretien, les interviews non directives, les entretiens de groupe, la conversation remplace le questionnaire. Le recueil des informations et leur analyse deviennent bien plus difficile (il faut enregistrer, retranscrire...) et implique tout un savoir faire bien spécifique. Il s'agit là de privilégier l'écoute plutôt que le questionnement, la compréhension et l'analyse en profondeur plutôt que le dénombrement.

Les documents : exploiter les informations existantes

Avant de lancer l'enquête on a bien sûr étudié la bibliographie sur le sujet et les documents qui peuvent nous renseigner. Comme le font les historiens la quête porte alors

sur les traces qui renvoient au domaine étudié (documents juridiques, transactionnels, commerciaux, comptes rendus, annonces et communications de toute sorte...).

L'information recueillie peut être très abondante et hétérogène et pose les mêmes difficultés de compréhension et d'analyse.

L'évolution des technologies de l'information et Internet notamment, bouleverse l'accès aux données que nous venons d'évoquer.

Les enquêtes via le web simplifient considérablement la collecte des réponses aux questions ouvertes, les interviews par mail ou par dialogue électronique (chat) offrent de nouvelles opportunités...

Les bases de données bibliographiques, les moteurs de recherches permettent au chercheur de récupérer très facilement références et publications. En généralisant le format numérique les outils de traitement de texte rendent potentiellement accessibles toutes sortes de fichiers et archives électroniques.

Ces évolutions mettent à disposition du chercheur un matériau considérable et posent de nouveaux problèmes : comment traiter des informations dont l'abondance peut très vite décourager et dont la complexité pose des problèmes spécifiques pour lesquelles la statistique n'offre pas de réponse immédiate.

Les méthodes

Ceci nous conduit à une autre caractéristique des études « quali ».

Le terme renvoie à l'approche traditionnelle des textes par lecture et annotation sans autres instruments que le stylo le papier dans la tradition des études littéraires et de la critique. Le travail du chercheur consiste à rendre compte des textes étudiés en citant des extraits et en produisant un nouveau texte compréhensif et démonstratif dont, pour faire bref, la qualité ne tient qu'au talent de lecteur et d'écrivain du chercheur.

Avec l'analyse de contenu ou analyse thématique, le classeur et le stylo de couleur s'ajoutent à la panoplie des instruments de recherche. Le travail de lecture devient plus systématique, le système de compréhension s'explicité en une grille qui guide le classement des citations et le repérage par couleurs ou annotations des passages du texte. Cet effort de méthode débouche naturellement sur le dénombrement des thèmes, si on admet que la répétition ou la fréquence peuvent également faire sens et que l'argument du décompte renforce la démonstration.

L'appellation « quali » évoque ainsi le travail artisanal prudemment démarqué des méthodes scientifiques et de la statistique, mais l'approche des textes a aussi de tout temps été marquée par la fascination pour le chiffre ou le recours au comptage.

Dans la tradition de la kabbale, de l'exégèse et dans le travail des moines copistes les chiffres et les nombres guident vers les significations cachées ou permettent de contrôler l'exactitude des manuscrits. Des études sur la bible aux concordances de Saint Augustin le texte analysé est mis en fragments puis en cartes perforées... Les premières analyses par ordinateurs ont été effectuées en 1945 par IBM qui pour reconstituer les concordances de Saint Augustin.

Devenu donnée textuelle, le texte échappe à la tradition « quali » et se trouve aussi concerné par les problématiques « quanti » comme l'ont montrés des travaux sur l'étude de l'attribution des œuvres littéraires et l'usage déjà très ancien des techniques informatique par les services de renseignements...

En banalisant ces méthodes et en les rendant accessibles, les évolutions technologiques ajoutent l'ordinateur, le moteur de recherche et le logiciel d'analyse de données textuelle à la boîte à outil du chercheur.

La spécificité des données textuelles et les voies d'analyse assistée par ordinateur.

Réponses aux questions ouvertes dans les enquêtes, interviews ou entretiens de groupe, revue bibliographique,

étude de documents, ces corpus ont en commun d'être essentiellement formés de données textuelles.

Elles ont la complexité et l'ambiguïté de la langue. Leur sens n'est pas fixé a priori comme c'est le cas pour les données quanti pour lesquelles les unités de mesures (variables numériques) et les états observés (variables nominales) sont fixés par un accord tacite de compréhension entre le répondant et l'enquêteur. La signification des données textuelles est à découvrir dans le sens des mots des phrases et de l'organisation du discours. Chacun de ces niveaux (lexique, syntaxe, rhétorique) apporte une contribution au sens dégagé par la lecture.

Cette propriété justifie la méfiance de ceux qui pensent qu'on ne peut pas faire l'économie d'une lecture humaine et qu'une machine ne peut pas comprendre un texte. Toutefois sans pour autant souscrire à la prétention des tenants de l'intelligence artificielle nous pouvons accepter que l'ordinateur et sa puissance de calcul apportent une aide pour l'approche des corpus textuels et notamment lorsque ceux-ci sont de très grande taille.

Nous montrerons dans ce qui suit que l'ordinateur peut être utilisé comme un auxiliaire pour rendre plus systématique les approches classiques et augmenter la productivité du chercheur et la qualité de ses résultats.

- En effet, les fonctions de recherche peuvent considérablement accélérer le travail traditionnel nécessité par l'usage de la citation ou la production de verbatim
- Pour l'analyse de contenu, la construction d'une grille thématique et la codification du corpus peuvent se faire en utilisant les ressources de l'informatique. Le travail de conception devient ainsi plus explicite et rigoureux et on gagne en souplesse et productivité pour la lecture et la codification

Mais il peut aussi être mis en œuvre d'une manière beaucoup plus automatique en fournissant, à partir du repérage des

formes graphiques qui composent le texte, des indications sur son contenu.

- L'analyse de contenu peut ainsi être automatisée en construisant les listes de mots ou dictionnaires correspondant aux différentes catégories de la grille thématique. La fréquence avec laquelle ces éléments se trouvent dans le texte sert d'indicateurs pour mesurer la présence et l'intensité de ces catégories. On peut parler ici de sémiométrie puisque l'usage des dictionnaires permet de « mesurer » les significations ainsi repérées dans le texte
- L'analyse lexicale offre une autre voie. Elle consiste, sans aucun a priori sur le texte à utiliser la redondance des données de langue et la statistique pour substituer à la lecture du texte celle d'éléments lexicaux qui en sont extraits : listes des mots les plus fréquents (mots clés), cartes visualisant la manière dont les mots se trouvent associés (zones thématiques) ou l'effet des circonstances ou des contextes (mots spécifiques)

Ces approches sont particulièrement bien adaptées au traitement de très gros corpus. Plus le corpus est volumineux, plus on gagne de temps et plus les régularités et effets statistiques sont significatifs.

Outils et stratégies d'analyse

Tout ceci ne peut bien sûr se faire qu'avec l'aide de logiciels adaptés. L'offre est désormais assez abondante, mais quel logiciel choisir ? Tout dépend du type d'analyse envisagée et de degré de spécialisation des fonctions proposées.

On peut distinguer 3 grandes classes de logiciels

- Les logiciels de recherche. Ils permettent de retrouver dans le texte des passages en fonction du contenu ou du contexte et mettent en œuvre des procédures plus

ou moins sophistiquées pour produire du verbatim (Lexico, WordMapper, Diction, Sphinx Lexica...)

- Les logiciels d'analyse thématique et de contenu ils aident au repérage, à la codification et à l'organisation des idées du texte, ainsi qu'à leur analyse et à leur synthèse (Nud'ist, Atlas tj, Modalisa, Sphinx Lexica...)
- Les logiciels d'analyse de données textuelles, ils abordent le texte par le biais de la statistique (Spad T, Alceste, Hyperbase, Sphinx Lexica, ...)

Ces logiciels sont plus ou moins spécialisés sur leur fonction principale. Alceste par exemple est très spécialisé, Modalisa l'est beaucoup moins. Dans ce qui suit nous nous référerons principalement à Sphinx Lexica qui est un outil très généraliste bien qu'à premier abord il semble s'adresser plus spécifiquement au traitement d'enquête.

Quelque soit l'outil, le chercheur a un rôle essentiel. C'est lui qui pilote le logiciel et c'est lui qui lance sélectionne les citations et le verbatim, élabore la grille thématique lit et code le texte, c'est enfin lui, qui seul est capable d'interpréter et de donner sens aux résultats des statistiques lexicales.

Pour cela il lui faudra maîtriser le passage obligé de l'acquisition des données et de l'intégration du corpus dans le logiciel et selon ces choix produire du verbatim, mener une analyse thématique de contenu, ou faire de l'analyse de données textuelles.

Les techniques avec lesquelles il devra se familiariser reposent également sur une bonne connaissance des propriétés de la langue des textes et discours qu'elle permet de construire.

Quelques connaissances utiles sur les propriétés de la langue, des textes et des idées...

Les grandes étapes

L'acquisition des données textuelles

La première chose à faire consiste à mettre le texte sous une forme utilisable par le logiciel. Cela implique non seulement de l'avoir sous une forme numérique en le saisissant dans un traitement de texte ou en le recopiant depuis Internet par, mais aussi à le découper en distinguant les différents éléments qui le composent, bref le mettre sous la forme d'une 'table de données'.

Tout dépend alors des circonstances.

Les questionnaires

S'il s'agit d'étudier les réponses aux questions ouvertes d'une enquête par questionnaire faite avec le logiciel, le texte est acquis au moment de la saisie des questionnaires pour le 'enquêtes papier crayon' ou directement entré par le répondant lorsqu'il répond à une enquête internet.

Les interviews non directives

Pour les interviews non directives le travail est plus complexe car il faut d'abord « mettre le texte dans Sphinx ».

La méthode la plus simple consiste à le retranscrire dans un questionnaire Sphinx composé des éléments suivants :

- 1- Des questions d'identité pour enregistrer le nom et les caractéristiques de l'interviewé
- 2- Une question pour noter le texte de la question et une autre pour le texte de la réponse

Pour une interview on saisira ainsi autant d'observations que d'échange question réponse auquel il a donné lieu. Il faudra en outre pour chaque nouveau couple répéter les questions d'identité.

Si les interviews ont déjà été saisies dans un traitement de texte on importe directement dans Sphinx le fichier qui les contient. Ce travail nécessite le respect des consignes suivantes :

1/ le fichier à importer doit être enregistré au format texte. Les fichiers de Word ne sont reconnus que s'ils ont été enregistrés sous ce format.

2/ afin de distinguer le texte des questions et des réponses et indiquer quand on passe d'une interview à une autre, il faut ajouter des repères dans le texte. Ces repères doivent respecter des règles qui permettront à l'ordinateur de les reconnaître et de les interpréter.

Par exemple :

```
Interview > Pierre
Q > Que pensez vous de ....
R> Mon opinion sur .....
Q> mais encore...
R> et bien voilà....
....
....
Interview> Jean
Q>.....
R>.....
Q>.....
R>.....
```

Les indications *Interview >*, *Q >*, *R>* signalent le nom de l'interview, une question ou une réponse. On les appelle des balises. Elles sont toujours placées en début de ligne et se terminent par > (ou un autre caractère qui n'est pas utilisé autrement dans le texte)

Le texte consécutif à chaque balise est reporté dans une variable qui lui correspond.

L'exemple ci-dessus sera ainsi converti en un questionnaire de 3 questions ou variables : Interview, Q et R. La table de données correspondante aura 3 colonnes et autant de lignes que de couple questions réponses.

Si seul le texte des réponses a été saisi il suffit d'ajouter en début le nom de la première interview puis celui de la deuxième et ainsi de suite pour repérer le passage d'une interview à l'autre. On appelle ces indications des jalons.

Elles sont en général notées dans le texte comme ci-dessous :

[J]=Pierre]
Mon opinion sur..... Et bien voilà....
.....
[J]=Jean]
Bla bla bla bla.... Truc....
.....

Dans ce cas l'importation du texte conduit à un questionnaire de 2 questions. La première indique de quelle interview il s'agit. La deuxième contient le texte découpé en fragments (paragraphe, phrases ou séquence de mots de longueur égale...). Chaque fragment du texte correspond à une observation.

On peut compléter les annotations qui jalonnent l'ensemble des interviews (*jalons*) par des annotations ponctuelles utilisées par exemple pour ajouter des commentaires (*marques*). Ces annotations, signalées par une indication mise entre crochet (par exemple [M=commentaire]) permettent de distinguer le contenu des commentaires de celui de l'interview.

Les bases de données et l'utilisation des balises

Le texte à analyser peut également provenir d'une application informatique : logiciels de messagerie, banque documentaires... dans ce cas les données sont structurées par des balises : indications placées au début de chaque élément qui définissent la nature du texte consécutif. Par exemple pour une base de données bibliographique :

Titre : les misérables
Auteur : Victor Hugo
Editeur : Hachette
Résumé : Histoire de Jean Valjean et consorts....
Titre : les sequestrés d'Altona
Auteur : Jean Paul Sartre
Editeur : Seuil
.....
ou pour une base de mail
de : andré lucas

à : annie2 ;jean3
objet : rendez vous
message : demain à la première heure.....

Dans ces exemples titre, auteur, éditeur, résumé, de, à, objet, message sont des balises. En les reconnaissant, l'ordinateur pourra la table de données correspondante.

Analyse d'une collection de documents quelconques

C'est le cas par exemple lorsque les données à analyser sont composées d'articles de presse ou d'autre sources documentaires constituant le corpus de l'étude. C'est à l'analyste d'ajouter dans le fichier texte où il a rassemblé tous ces éléments les annotations qui permettront d'indiquer qu'on passe d'un article à un autre ou d'une source à une autre. L'ordinateur pourra alors construire une table dans laquelle sera notée le nom de l'article ou de la source d'une part et le contenu d'autre part. Si les différents textes sont longs on peut en outre décider de les fragmentés paragraphes ou phrases.

Analyse de sites web ou une page de liens

Le logiciel permet d'aspirer directement le contenu de sites ou de pages sélectionnées par un moteur de recherche. Cette possibilité ne permet malheureusement pas d'accéder aux contenus des sites dynamiques. Dans ce cas il faut procéder manuellement en parcourant le site et recopiant le texte dans un questionnaire conçu à cet effet.

Produire des extraits ou faire du verbatim

Faire du verbatim (ou des citations) est la méthode la plus utilisée dans les études qualitatives. Ces citations peuvent être choisies en lisant le texte ou sélectionnées de manière systématique selon le contexte ou selon le contenu.

Verbatim par contexte

Savoir qui dit quoi ou sélectionner les citations selon les circonstances ou toute autre information contenue dans les réponses aux questions fermées. Ce type de sélection peut être plus ou moins complexes : on peut n'utiliser qu'un seul critère, par exemple ce que disent les femmes ou en combiner plusieurs en s'intéressant aux femmes de moins de 30 ans et diplômées Dans les enquêtes de satisfaction le contexte pertinent sera plutôt tous ceux ou celles qui déclarent être insatisfait, l'identité apparaissant comme la signature de chaque citation...

Verbatim selon le contenu

Sélectionner les citations en fonction de ce qui est dit. Cette approche privilégie le contenu qui peut être repéré automatiquement par la présence dans le texte d'un ou plusieurs mots ou par un travail de codification préalable (voir analyse de contenu)

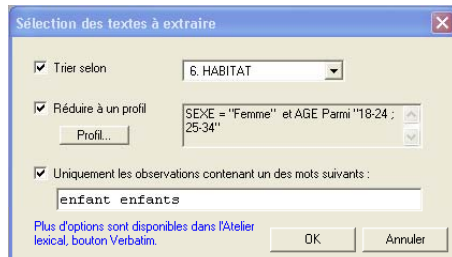
Les 2 approches par le contexte et par le contenu peuvent bien sûr être combinées pour restituer le texte par fragments relatifs aux différents angles de vue ou problématiques de l'étude.

Faire du verbatim avec Sphinx

La sélection de verbatim dans Sphinx peut se faire soit à partir la fonction Etudier les textes de la partie classique du logiciel ou directement dans les tableaux de bord de l'environnement multimédia.

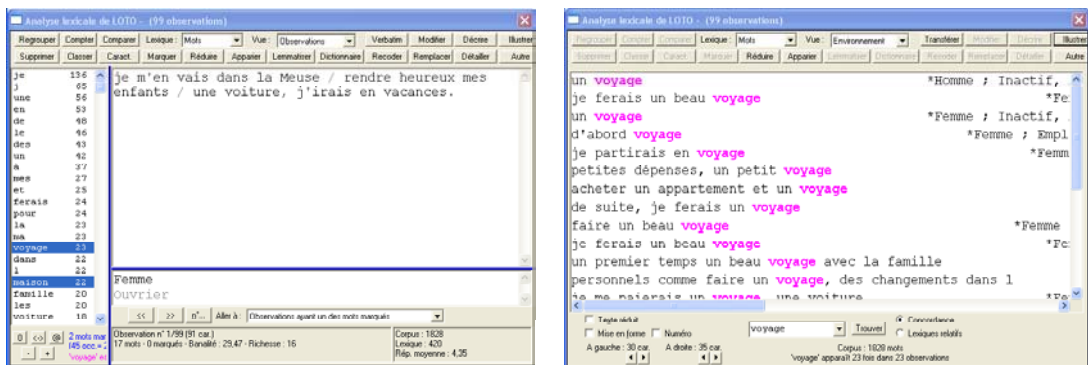
Recoder / Etudier les textes / Verbatim :

L'exemple ci-dessous indique le dialogue correspondant à la sélection des réponses de jeunes femmes citant les mots enfant ou enfants. La liste des réponses correspondantes étant donnée par type d'habitat.



Atelier lexical

Dans l'atelier lexical (Etudier les textes / Atelier lexical) un double clic dans un mot de la fenêtre du lexique permet d'afficher toute les fragments de texte contenant le ou les mots sélectionnés. La longueur du fragment se règle avec le bouton Apparier et les indications de mise en forme figurant en bas de l'écran. Le bouton Illustrer permet d'ajouter une signature (sexe, csp). Utiliser Transférer pour enregistrer ou recopier dans un traitement de texte.



Dans l'environnement multimédia

On peut faire figurer dans un tableau de bord la liste des réponses correspondant à un profil donné. Par exemple dans un tableau de bord on peut faire figurer les commentaires des insatisfait et ou de satisfaits.

Verbatim des remarques

REMARQUES Les insatisfaits (note <5)	REMARQUES Les satisfaits (note >8)
<p>Au sujet de la livraison du poisson frais, il arrive chez moi avec des dates d'emballage datant de six jours. Je trouve cela inadmissible, car le poisson n'est plus frais et il sent. Après de multiples réclamations, rien n'a changé. Par conséquent, dernièrement, j'ai cessé mes commandes de poisson frais chez valibest ou en tout cas, je n'ai cessé de le renvoyer. Aujourd'hui, j'ai tendance à me ravitailler en surgelé. C'est vraiment regrettable.</p> <p>J'ai été très contrarié par une livraison totalement erronée. Le poisson avait été livré à une autre adresse. Lorsqu'il est arrivé chez nous, après réclamation, il était, bien entendu, avarié.</p> <p>Je réclame plus de jours de livraison. En outre les camions doivent être plus spécifiques. Il arrive que les poissons voisinent avec les conditionnements de fruits et légumes. Il en résulte des rejets de poisson qui, fort malencontreusement, se mélangent aux odeurs de fruits et légumes. Les conditionnements doivent être séparés les uns des autres.</p>	<p>(où déposer les colis etc.)</p> <p>Attention à la qualité ! Parfois elle pêche !</p> <p>Au niveau de la livraison, je reprocherai que la marchandise reste sur le qual. Soit le chauffeur livre trop tôt, soit on ne l'attend pas scanner. Des boîtes tagons, il devrait passer à la réception et signaler qu'il a des colis à déposer plutôt que de les laisser traîner sur le qual.</p> <p>Au niveau des commandes, il est regrettable qu'une fois le fax de commande passé, vers 18 H - 18 H30, un rajout par fax ne puisse plus être pris en compte par vos services, car dans la restauration on travaille souvent à la demande. De plus, on ne peut pas se faire livrer le lundi et le mercredi. Or on travaille beaucoup les produits frais et si on a besoin ces jours là de produits en dépannage, on doit s'adresser à un concurrent (Annemasse Primeurs).</p> <p>Depuis notre réouverture, je reçois tout le temps des livraisons qui ne sont pas conformes, alors que le BL est lui bien conforme à ma commande. Je commande, par exemple, des poireaux entiers et je reçois des poireaux émincés. Donc, si je n'ai pas le temps de contrôler à la livraison et que je ne m'en aperçois pas, cela me pose des problèmes lorsque je veux préparer ces produits. Est-ce que c'est dû à des erreurs du préparateur ?</p>

L'exemple ci-dessus est tiré du tableau de bord d'une enquête de satisfaction. On produit pour cela le tableau de la variable texte en choisissant dans l'onglet *Calcul* l'option *Mise en classe des réponses* et en fixant le profil de la sélection. Dans l'onglet *Tableau* on indique que les effectifs et pourcentage ne doivent pas être affichés.

Dans le cas d'une analyse de contenu, la sélection peut être effectuée en fonction des variables de la grille thématique. On peut également utiliser des calculs d'intensité lexicale pour afficher les réponses exprimant le plus les idées correspondant à une liste de mots clés (dictionnaire).

Faire de l'analyse de contenu

Cette méthode consiste à lire l'ensemble du corpus en repérant les thèmes ou idées qu'il contient pour ensuite

produire du verbatim par thèmes et / ou mener une analyse statistique des thèmes.

Les étapes du travail sont les suivantes. Illustrons les sur le cas de l'analyse des questions ouvertes dans les questionnaires.

Définir la grille des thèmes

Elle organise la description des idées susceptibles d'être présentes dans le texte. Par exemple pour analyser les réponses à la question : » Si vous gagniez au loto, que feriez vous ? » on distingue :

les actions : consommer, investir, donner

les personnes concernées : moi, mes proches, les gens...

la tonalité de la réponse : neutre, sceptique, humour....

Ajouter au questionnaire les variables thématiques

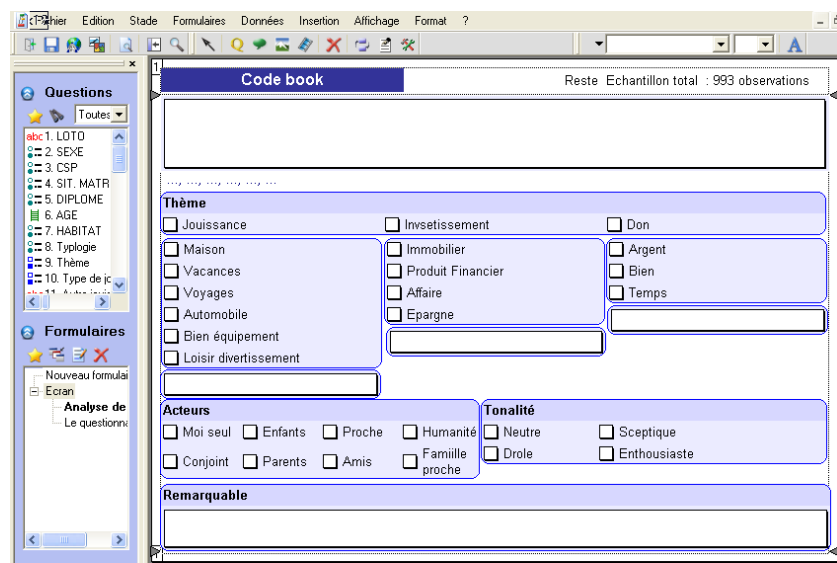
Elles décrivent les thèmes et leurs catégories (modalités)

The screenshot shows a software window titled "LISTE DES QUESTIONS" with a menu bar (Edition, Stade, Rédiger, Organiser, ?) and a toolbar. The main area contains a table with the following data:

Variable	Libellé	Modalités de réponse / Contrôles
Questionnaire		
1. SEXE	Sexe	Homme ; Femme ;
2. CSP	CSP	Agriculteur ; Commerçant, artisan ; Cadre.Prof.In
3. SIT. MATRIMOI	Situation matrimoniale	Célibataire ; Marié(e) ; Vivant maritalement ; Veuf
4. DIPLOME	Diplôme	Aucun ; CEP BEPC ; CAP BEP ; Bac ; Bac+2 B
5. AGE	Classe d'âge	18-24 ; 25-34 ; 35-49 ; 50-64 ; 65+ ;
6. HABITAT	Taille de la commune	- de 2 000 ; 2 000-20 000 ; 20 000 - 100 000 ; + d
7. LOTO	Si vous gagniez le gros lot au Loto, que feriez-vo	
Grille thématique		
8. Action	Type d'action évoquée dans la réponse	Consommation ; Investissement ; Don ;
9. Type de jouissz	Biens ou services	Equipements ; Services ; Conso ;
10. Type d'investit	Quel type d'investissement choisiriez-vous?	Biens ; Service financier ; Activité éco ;
11. Tonalité	Tonalité	Neutre ; Enthousiasme ; Sceptique ; Humoristiqu
12. Type de don	Type de don	Bien ; Argent ; Soi ;
13. Acteur	Acteur	Moi ; Conjoint ; Enfants ; Parents ; Proche ; Amis
14. Remarques	Recopier les réponses intéressantes	
Variables lexicales		


Au stade du formulaire, faire modification du questionnaire et ajouter les questions correspondant à la grille
Créer le code book

C'est l'interface dans laquelle les idées du texte seront codées. Elle se présente comme un formulaire qui présente la réponse texte à analyser et les « questions » de la grille thématique. On peut rajouter une nouvelle variable texte pour recopier les expressions savoureuses et les retrouver plus facilement.



Au stade questionnaire, faire formulaire multimédia et créer un nouveau formulaire, puis disposer les variables correspondant à la réponse et à la grille. On peut ajouter une légende pour rappeler l'identité du répondant.

Lire interpréter et coder le contenu

Le code book défini à l'étape précédente est « lancé » à partir du module opérateur. On peut ainsi parcourir l'ensemble des réponses (flèches du haut de l'écran), ou seulement celles qui contiennent tel mot et répondent à telle identité (sélection d'un profil ) ou plus simplement celles qui ne sont pas encore codées (non réponse pour la variable thème).

Strates Observation 964

A coder pour LOTO Avec "enfant" : 208 observations

Code book Reste A coder pour LOTO Avec "enfant" :

donner aux enfants qu'ils en profitent / placer.

Homme, Retraités, Veuf(ve), Bac+4 DESS, 65+, 20 000 - 100 000

Thème

<input type="checkbox"/> Jouissance	<input checked="" type="checkbox"/> Investissement	<input checked="" type="checkbox"/> Don
	<input type="checkbox"/> Immobilier	<input type="checkbox"/> Argent
	<input type="checkbox"/> Produit Financier	<input type="checkbox"/> Bien
	<input type="checkbox"/> Affaire	<input type="checkbox"/> Temps
	<input checked="" type="checkbox"/> Epargne	

Acteurs

<input type="checkbox"/> Moi seul	<input checked="" type="checkbox"/> Enfants	<input type="checkbox"/> Proche	<input type="checkbox"/> Humanité
<input type="checkbox"/> Conjoint	<input type="checkbox"/> Parents	<input type="checkbox"/> Amis	<input type="checkbox"/> Famille proche

Tonalité

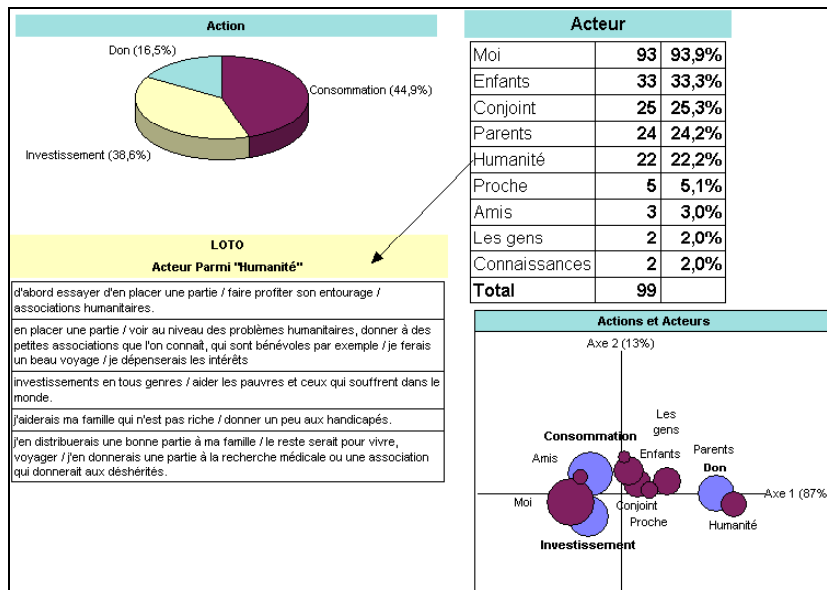
<input checked="" type="checkbox"/> Neutre	<input type="checkbox"/> Sceptique
<input type="checkbox"/> Drole	<input type="checkbox"/> Enthousiaste

Remarquable

Analyser la fréquence des thèmes et faire du verbatim

Une fois toutes les réponses lues et codées, l'analyse s'achève par l'étude statistique des thèmes et leur illustration par le verbatim qui leur correspond. Toutes les ressources statistiques sont alors disponibles pour confronter les thèmes entre eux ou les croiser avec les autres variables de l'enquête.

Si la grille thématique est pertinente les résultats seront à coup sûr intéressants.



Le travail de préparation pour les interview non directives

Pour les interviews non directives le travail est plus complexe car il faut d'abord « mettre le texte dans Sphinx ».

Mettre le texte dans Sphinx.

La méthode la plus simple consiste à directement retranscrire le texte dans un questionnaire Sphinx composé des éléments suivants :

- 3- Des questions d'identité pour enregistrer le nom et les caractéristiques de l'interviewé
- 4- Une question pour noter le texte de la question et une autre pour le texte de la réponse
- 5- Les questions relatives à la grille thématique

Lorsque ce ` questionnaire ` est construit il n'y a plus qu'à le mettre en forme pour disposer d'un `code book` bien présenté. On peut alors écouter l'enregistrement de

l'interview et entrer le texte. La codification peut se faire en même temps ou après que tout a été retranscrit.

Si les interviews ont déjà été saisies, on importe directement dans Sphinx le fichier qui les contient. Ce travail nécessite le respect des consignes suivantes :

1/ le fichier à importer doit être enregistré au format texte. Les fichiers de Word ne sont reconnus que s'ils ont été enregistré sous ce format.

2/ afin de distinguer le texte des questions et des réponses et indiquer quand on passe d'une interview à une autre, il faut ajouter des repères dans le texte. Ces repères doivent respecter des règles qui permettront à l'ordinateur de les reconnaître et de les interpréter.

Par exemple :

```
Interview > Pierre
Q > Que pensez vous de ....
R> Mon opinion sur .....
Q> mais encore...
R> et bien voilà....
....
....
Interview> Jean
Q>.....
R>.....
Q>.....
R>.....
```

Les indications *Interview >*, *Q >*, *R>* signalent le nom de l'interview, une question ou une réponse. On les appelle des balises. Elles sont toujours placées en début de ligne et se terminent par > (ou un autre caractère qui n'est pas utilisé autrement dans le texte)

Le texte consécutif à chaque balise est reporté dans une variable qui lui correspond.

L'exemple ci-dessus sera ainsi converti en un questionnaire de 3 questions ou variables : Interview, Q

et R. La table de données correspondante aura 3 colonnes et autant de lignes que de couple questions réponses.

Si seul le texte des réponses a été saisi il suffit d'ajouter en début le nom de la première interview puis celui de la deuxième et ainsi de suite pour repérer le passage d'une interview à l'autre. On appelle ces indications des jalons. Elles sont en général notées dans le texte comme ci-dessous :

[J]=Pierre]
Mon opinion sur..... Et bien voilà....
.....
[J]=Jean]
Bla bla bla bla.... Truc....
.....

Dans ce cas l'importation du texte conduit à un questionnaire de 2 questions. La première indique de quelle interview il s'agit. La deuxième contient le texte découpé en fragments (paragraphe, phrases ou séquence de mots de longueur égale...). Chaque fragment du texte correspond à une observation.

On peut compléter les annotations qui jalonnent l'ensemble des interviews (*jalons*) par des annotations ponctuelles utilisées par exemple pour ajouter des commentaires (*marques*). Ces annotations, signalées par une indication mise entre crochet (par exemple [M=commentaire]) permettent de distinguer le contenu des commentaires de celui de l'interview.

Analyse de données textuelles et approximation lexicale

Cette approche s'impose lorsque le corpus est très volumineux. Elle permet de gagner beaucoup de temps mais requiert un savoir faire spécifique autant pour bien maîtriser les méthodes mises en œuvre que pour en interpréter les résultats.

Approcher le texte par le lexique

L'idée est simple : prendre connaissance du texte à partir de des mots les plus fréquemment utilisés. L'informatique et la statistique permettent de faire cela très vite quelque soit la taille du corpus. Plus il est volumineux, meilleurs sont les résultats. Il faut ensuite pouvoir se faire une idée du texte à partir d'une simple liste de mots classés par fréquences décroissantes et bien choisir cette liste.

L'exemple ci-dessous donne les différents lexiques tirés de l'étude sur le rêve des français.

Lexique brut		Sans mots outils		Lemmatisé		Les substantifs		Les verbes	
je	1182	maison	355	maison	354	maison	354	acheter	350
j	653	enfants	286	acheter	350	voyage	310	faire	296
de	613	voiture	198	voyage	310	enfant	259	placer	195
une	566	voyage	192	faire	297	voiture	194	donner	187
en	464	ferais	182	enfant	264	argent	118	voyager	148
à	412	voyages	158	placer	230	famille	117	aider	122
le	385	achèterais	150	voiture	194	monde	81	partir	91
des	362	famille	122	donner	187	don	75	travailler	91
un	362	argent	118	voyager	148	placement	71	profiter	81
maison	355	acheter	115	aider	122	appartement	65	investir	76
enfants	286	donnerais	113	argent	118	reste	65	mettre	71
la	286	placerais	111	famille	117	association	57	changer	63
pour	280	faire	107	partir	91	immobilier	55	aller	59
l	257	reste	97	travailler	91	partie	54	arrêter	58
mes	244	monde	95	monde	81	oeuvre	51	améliorer	48
les	240	placer	84	profiter	81	vie	49	prendre	38
et	208	travailler	79	investir	76	tour	45	vivre	37
voiture	198	vacances	79	don	75	vacance	43	partager	36
dans	194	achète	69	immobilier	74	achat	41	payer	33
voyage	192	voyagerais	68	beau	72	loisir	39	rester	33
ferais	182	immobilier	67	mettre	71	besoin	36	épargner	32
voyages	158	partie	64	placement	71	placer	35	avoir_besoin	30

Les mots outils sont révélateurs de l'énonciation, ici la fréquence des *je* et *j* est vraiment remarquable. Il faut descendre plus bas pour trouver les premiers mots pleins, ou les supprimer (Lexique sans mots outils).

La lemmatisation ramène chaque mot à sa forme racine : l'infinitif des verbes, le masculin singulier des noms et adjectifs. Enfin, présenter le lexique par catégories grammaticales permet de focaliser l'attention sur les objets (substantifs), les actions (verbes) et évaluations (adjectifs).

On peut enfin chercher à grouper les termes du lexique en utilisant des dictionnaires de termes équivalents pour encore réduire la variété lexicale et mieux approcher les différentes idées du texte et leur importance.

DON	INVESTISSEMENT	JOUISSANCE	➔	Présence des thèmes		
enfant	placer	maison		Nombre de réponses évoquant un des mots du thème		
donner	investir	acheter		#JOUISSANCE	770	78%
famille	de_côté	voyage		#DON	494	50%
don	épargner	voiture		#INVESTISSEMENT	411	41%
association	construire	voyager		Total	993	
ami	crédit	partir				
partager	épargne	monde				
distribuer	rembourser	améliorer				
malheureux	fructifier	tour				
enfants	intérêt	vacance				
femme	rapporter	plaisir				
handicapé	rente	faire_plaisir				
mari		faire_le_tour				
parent		pays				
		bateau				
		étranger				
		robe				

Segments répétés et cartes d'association lexicales

Les lexiques donnent très rapidement un aperçu du texte analysé mais ils peuvent aussi conduire à des interprétations erronées. Il faut donc vérifier et resituer chaque mots dans son contexte en revenant au texte (verbatim) ou d'une manière plus synthétique en cherchant les segments répétés et en produisant des cartes d'associations lexicales.

Les segments répétés (séquences de mots répétés à l'identique) renvoient les rigidités du texte, les formules toutes faites ou la langue de bois. Il permettent aussi de soulever bien des ambiguïtés (arrêter de travailler) et révèlent les leitmotifs du corpus.

Segments répétés

acheter maison	140	14%
placer argent	58	6%
faire voyage	55	6%
arrêter travailler	52	5%
tour monde	44	4%
faire profiter	43	4%
donner enfant	40	4%
acheter voiture	37	4%
aider enfant	33	3%
investir immobilier	33	3%

Les cartes d'association lexicales

D'une manière moins rigide que les segments répétés la statistique des associations lexicales (via l'analyse factorielle des correspondances multiples) donne une idée de la propension à associer les mots les uns aux autres ou au contraire à ne pas les faire coexister dans une même expression.

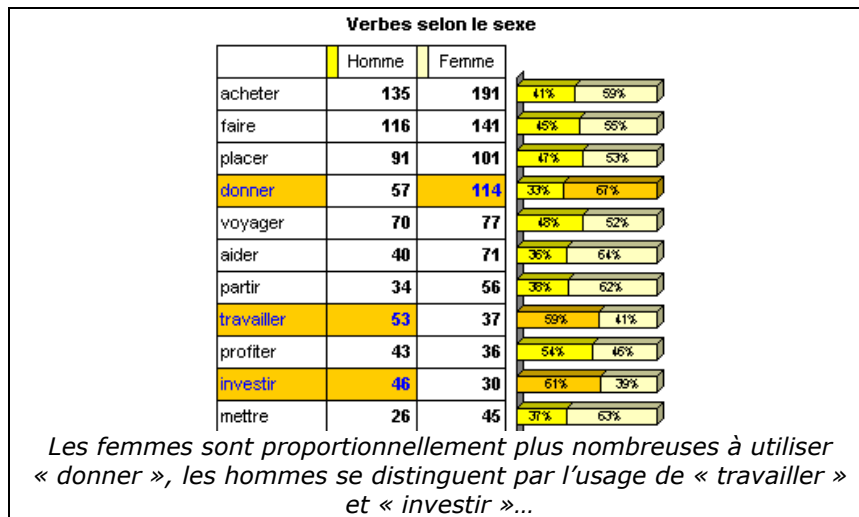
Ainsi, la carte ci dessous laisse apparaître dans les constellations proches ou distinctes les réseaux sémantiques, modèles cognitifs ou configurations mentales qui dans leur répétition structurent le discours. Ici on peut opposer les évocations généreuses à droite, à celles de l'utilitarisme à gauche....

La lecture de ces cartes conduit à identifier les thématiques du corpus. Elles ont l'avantage d'être produites sans biais cognitifs et de réduire très significativement la masse d'information qui devient ainsi partageable.... et discutable.

9 857 mots sur un total de 17 743. Leurs réponses sont en moyennes plus longues que celle des hommes, mais la catégorie la plus prolixe est les commerçants artisan.

Les mots utilisés sont ils les mêmes suivant l'identité de celui qui parle, les circonstances ou toute autre information qui situe le texte analysé ?

On peut le savoir en croisant par exemple les réponses à une question fermée avec les mots les plus couramment utilisés.



Plus directement on peut sélectionner la liste des mots sur représentés dans telle ou telle catégorie et obtenir ainsi les mots spécifiques qui les caractérisent. Ces listes peuvent être cartographiées pour mettre en évidence des zones de langages.

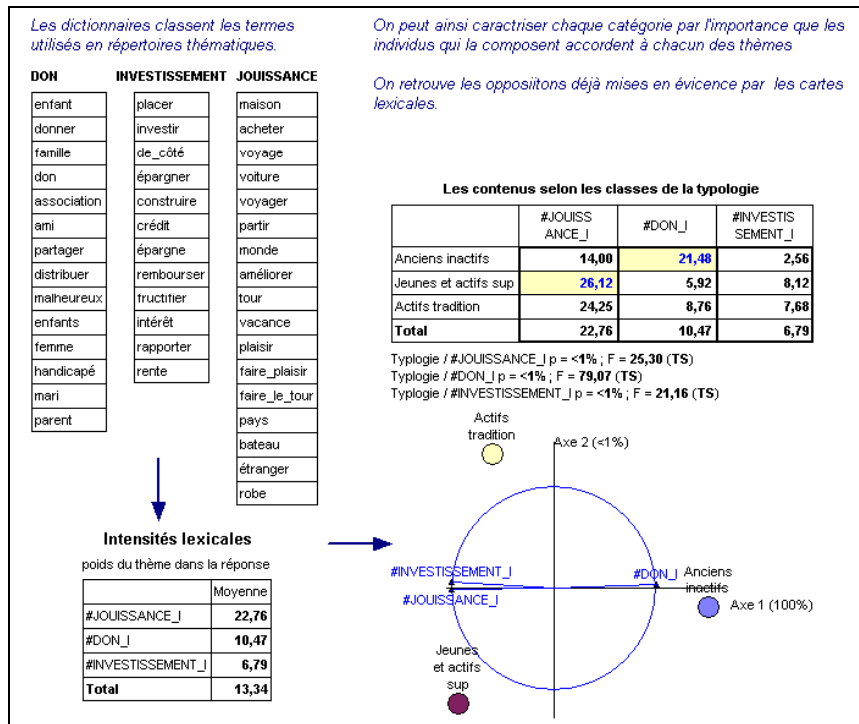
18-24	Nb.	25-34	Nb.	35-49	Nb.	50-64	Nb.	65+	Nb.
voiture	50 (1,54)	immobilier	24 (1,57)	gens	16 (1,64)	enfant	89 (1,95)	enfant	83 (2,74)
appartement	23 (2,12)	achat	19 (1,67)	pays	12 (1,90)	monde	20 (1,53)	oeuvre	16 (2,93)
loisir	10 (1,53)	crédit	10 (2,00)	recherche	11 (1,92)	besoin	10 (1,72)	don	13 (1,62)
parent	5 (2,49)	personne	8 (1,80)	mari	8 (2,33)	cancer	9 (2,07)	besoin	10 (2,59)
quelque chose	5 (1,76)	travail	8 (1,69)	aventur	7 (2,04)	gens	9 (1,64)	part	7 (4,67)
meuble	5 (1,57)	commerce	6 (1,80)	entourage	7 (1,74)	pauvre	8 (2,16)	malheureux	6 (4,00)
action	4 (2,18)	soleil	5 (1,63)	épargne	7 (1,63)	jour	5 (2,59)	filles	6 (3,50)
entreprise	3 (1,63)			dette	6 (1,90)	intérêt	3 (1,86)	cancer	5 (1,73)
				action	5 (1,59)	retraite	3 (1,69)	pauvre	4 (1,62)
				sida	5 (1,59)	Femme	3 (1,55)	bien	3 (2,80)
				temps	5 (1,59)			retraite	2 (1,70)
								mer	2 (1,56)

Les jeunes se distinguent par un vocabulaire utilitariste (voiture, appartement, meuble) à la différence des plus âgés qui privilégient l'aide et le don....

Intensités lexicales et lexicométrie

On peut aussi orienter la recherche par rapport à ce qu'on pense trouver. Comme on le ferait en lisant tout le texte pour repérer les idées qui s'y trouvent on confie cette tâche à l'ordinateur. Pour cela on dresse des listes de termes exprimant les idées que l'on cherche à repérer. Ces listes appelées dictionnaire permettent de calculer le nombre de fois où l'un des termes se trouve dans la réponse analysée. L'intensité lexicale est calculée comme le rapport entre ce nombre et le nombre total de mots de la réponse. Le poids de l'idée correspondante est « mesuré » par se rapport ou intensité lexicale.

Cet usage de données lexicales pour mesurer l'importance d'une idée permet de parler de lexicométrie. Ainsi, à partir d'une expression libre on mesure l'importance des éléments du modèle pour ensuite poursuivre les analyses comme si ces mesures étaient établies en recueillant des opinions sur des échelles.



Le calcul des intensités lexicales permet de mettre en évidence les différences de réaction selon la catégorie de français.

Comment procéder

Dans la parties classique de Sphinx

Lemmatiser : Panneau de commande : Créer les variables lexicale ou dans l'Atelier Bouton Lemmatiser.

Calculer les lexiques : Panneau de commande ou dans l'atelier bouton Réduire Regrouper Supprimer pour agir sur les mots marqués dans le lexique.

Segments répétés : Panneau de commande, ou dans l'atelier article Expression du menu Lexique puis bouton Segments

Cartes d'associations lexicales. Dans l'atelier :

1/ Sélectionner les mots à considérés dans la fenêtre du lexique

2/ Créer une variable codant la présence de ces mots : Recoder, Fermé sur les mots marqués du lexique. Nommer la nouvelle variable

3/ Lancer l'analyse factorielle des correspondances multiples à partir de la variable qui vient d'être créée : Menu Approfondir (menus déroulants du haut de l'écran), Analyse factorielle des correspondances multiples.

4/ Pour construire une typologie des thèmes : bouton *Typologie* et définir interactivement les classes

Bilan lexicaux : Panneau de commande, Bilan par catégories ou tableau de bord en croisant une variable texte avec l'option de calcul *Nombre de mots* avec une autre variable fermée. Choisir les indicateurs (moyenne, somme, part...) dans l'onglet *Tableau* .

Tableaux croisés lexicaux : Panneau de commande ou atelier lexical menu Vues, article Tableau lexical. Bouton Croiser pour sélectionner la variable à croiser avec le texte.

Mots spécifiques : Panneau de commande ou dans l'atelier
1/ calculer la liste article *Mots spécifiques* du menu Vue puis bouton *Croiser* pour choisir la variable et *Select*. Pour fixer les seuils de spécificités et de fréquence.

2/ produire la liste des réponses caractéristiques de chaque catégorie (verbatim des réponses contenant le plus de mots spécifiques) : bouton *Réponses caractéristiques*.

Intensités lexicales. La définition des dictionnaires peut se faire dans un traitement de texte ou en sélectionnant les mots dans l'atelier lexical. La procédure la plus rapide consiste à utiliser une thématique (ensemble de dictionnaires correspondant chacun à un thème) : atelier lexical, bouton *Recoder* et *Analyse thématique*. Sélectionner les fichiers des dictionnaires ou ouvrir une thématique existante puis *Recoder*. Pour chaque dictionnaire de la thématique une variable numérique contenant l'intensité lexicale de la réponse ou du fragment est créée.


Ces variables peuvent être utilisées pour créer une typologie : Menu *Approfondir*, *Classification automatique*.

Dans les tableau de bord multimédia.

La lemmatisation et la définition des variables d'origine lexicales dont le calcul des intensités lexicale et le calcul des doivent se faire dans la partie classique et l'atelier lexical. Il en également ainsi pour le calcul de mots spécifiques.

Calcul des lexiques : Choisir l'option *Mise en classe des mots* dans l'onglet *Calcul*. Le bouton *Dictionnaire* permet de supprimer les mots outils et, ou de restreindre le lexique à une liste ou au contenu d'un dictionnaire. Dans le même onglet les articles du menu *Simplifier* permettent de contrôler la longueur de la liste en fixant le nombre de lignes ou la fréquence minimum.

Tableau croisés et cartes lexicales.

Une variable texte présentée comme un lexique (mise en classes des mots) peut être croisée avec n'importe quelle autre variable avec la possibilité d'utiliser des représentations cartographiées des tableaux croisés ainsi obtenus (*Carte* dans onglet *Graphique* ou raccourcis *montrer la carte* ).

Ceci permet notamment de caractériser les réponses selon les différentes catégories de répondant.

Dans le tableau de groupe ci-dessus caractérisant les réponses par sexe et CSP, la variable texte et calculée selon

l'option *Nombre de mots* de l'onglet *Calcul* et dans l'onglet *Tableau*, moyenne somme et effectif ont été sélectionnés.

Privilégier une approche ou les combiner ?

Les approches que nous venons d'évoquer sont très différentes.

Le verbatim est de loin la pratique la plus courante et la plus simple à mettre en œuvre.

L'analyse de contenu par recodification présente l'avantage de reposer sur une thématique qui en elle-même est déjà une contribution à la compréhension du texte. La codification permet de préciser avec la rigueur de la statistique comment cette thématique s'applique et de mettre en évidence les interdépendances entre thèmes ainsi que rechercher des explications contextuelles.

D'autre par le verbatim associé à chaque catégorie de codification permet d'illustrer de manière concrète les idées générales et peut être d'en affecter la compréhension en leur donnant plus de vigueur ou de pertinence.

Ces 2 méthodes souffrent également de la subjectivité qui fatalement accompagne leur mise en œuvre dans le choix final de telle ou telle citation et dans la décision de coder de telle ou telle manière. L'usage des outils informatiques et statistiques permet de mieux contrôler ces biais subjectifs grâce à la formalisation et aux possibilités de recoupement. Enfin malgré ses nombreux avantages l'analyse de contenu est très consommatrice de temps et de ce fait très coûteuse pour l'analyse des gros corpus.

L'analyse des données textuelle offre une voie bien différente et beaucoup plus technique. Elle présente l'avantage d'une réelle objectivité dans la production de substituts lexicaux (listes, cartes...). Bien que la subjectivité intervienne à nouveau au moment de la lecture et de l'interprétation elle bénéficie ainsi d'un crédit de scientificité que les autres approches n'ont pas.

D'autre part, en détournant le regard du sens de surface vers l'interprétation des actes de langage (tout ce que le choix des mots révèle dans les répétitions ou les absences) ce procédé crée les conditions d'une distance critique et créatrice. Au risque bien sûr d'importants contre sens dont il faut se prémunir par des précautions élémentaires.

- ses méthodes ne sont applicables que sur de très gros corpus, là où les grands nombre et la statistique peuvent légitimement faire sens

- un patient retour au texte reste absolument indispensable pour contrôler les interprétations rapidement acquises par ces techniques. A nouveau le verbatim s'impose

- enfin certains aspects des contenus sont très difficiles à appréhender ainsi des évaluations ou jugement. Il est facile de voir qu'il est question de travail ou de prix mais beaucoup plus difficile d'établir si le travail ou les prix sont évoqués positivement ou négativement. Les formes lexicales de l'évaluation, du jugement sont en effet très variées et complexes à appréhender... (pas trop, trop, trop peu).

L'analyse lexicale doit être déconseillée pour traiter les réponses ouvertes dans des enquêtes sur un petit nombre de répondants (moins de 200). L'analyse de contenu prendra moins de temps et les résultats en seront à coup sûr probants.

Il est en revanche des situations où seule, compte tenu du temps ou des budgets disponible, l'analyse des données textuelles est envisageable. C'était par le exemple le cas pour l'analyse des 45000 pages du débat national sur l'avenir de l'école.

Mais le plus souvent ces méthodes gagent à être mises en œuvre de manière complémentaire. Dans un premier l'analyse lexicale utilisée de manière exploratoire permet de rapidement prendre connaissance du corpus et dans les cas heureux de faire jaillir des pistes d'interprétation qu'on n'aurait peut être jamais empruntées autrement. C'est ainsi que la thématique émergente du texte peut être complétée par les problématiques, modèles et systèmes d'interprétation

généraux qui permettront de construire une grille thématique plus pertinente. Enfin le travail méthodique d'analyse de contenu et de codage permettra si le corpus n'est pas trop volumineux de rigoureusement catégoriser les idées présentes dans le texte. Sinon il faudra lire le lexique pour le ventiler dans les dictionnaires qui permettront par le calcul des intensités lexicales d'automatiser le codage des idées présentes dans le texte.

Enfin après l'exposé des résultats statistiques le texte reprendra ses droits grâce aux citations et verbatim qui ajouteront aux chiffres le pouvoir du sens singulier des phrases ou des paroles.