# Méthodologie de la recherche

**Marie-Laure Gavard-Perret**, **David Gotteland**,
**Christophe Haon**, **Alain Jolibert**

ISBN : 978-2-7440-7241-3

## Chapitre 7 – Le changement et l'innovation

### Complément 1
### Exemple d'analyses de matériels iconiques (p. 251)

**1 - Analyse d'évocations graphiques**

À partir de Demory et Lancestre, 1983

Les évocations graphiques analysées par les auteurs ont été produites à l'aide du terme « assurances ». Les auteurs s'attachent à interpréter et donner du sens aux dessins réalisés par les sujets. Ainsi, selon eux, « l'assurance apparaît comme une puissance dédaigneuse et redoutable face à laquelle le client se sent angoissé et accablé ». À titre d'illustration, ils donnent quelques exemples de dessins qui expriment ces idées : « un aigle emportant dans ses serres un mouton du troupeau, des immeubles géants et inhumains, un nuage pesant sur les "pauvres humains", etc. ».

Dans ce même travail, les auteurs analysent les évocations graphiques portant cette fois sur le « courtier d'assurances ». Il ressort des dessins du groupe des perceptions majoritairement négatives du courtier qui s'expriment graphiquement par des représentations telles que : « des doigts aux ongles crochus ».

*[B. Demory et A. Lancestre (1983),* Le marketing qualitatif – Des produits nommés désirs*, Paris, Chotard et associés.]*

**2 - Analyse sémiotique de publicités**

À partir de Zhao et Belk, 2008

Les auteurs de cette recherche utilisent une approche sémiotique afin d'examiner les rapports entre globalisation et localisme dans la publicité chinoise des années 1930, période qui, pour ces chercheurs, correspond aux premières confrontations du local avec le global à Shanghai. Les supports étudiés sont les calendriers publicitaires, les Yuefenpai, très populaires dans ce pays.

Ils ont recours aux principes de la rhétorique visuelle et de la sémiologie Saussurienne. Ils s'attachent donc à comprendre le processus de construction du sens à partir des signes contenus dans ces compositions publicitaires. Comme dans le cas d'une analyse de contenu classique, les deux auteurs ont procédé à une analyse individuelle avant de comparer leurs résultats.

Le premier niveau d'analyse considéré est le niveau de la dénotation. Il s'agit alors de rechercher la représentation de la globalisation d'une part et du localisme d'autre part.

Le second niveau est celui de la connotation, au travers des styles, idées, valeurs symboliques exprimés par les signes visuels et linguistiques. À ce niveau, l'attention est également orientée sur la composition et la rhétorique des images.

*[X. Zhao et R.W. Belk, « Advertising consumer culture in 1930s Shanghai – Globalization and localization in Yuefenpai »,* Journal of Advertising*, 37, 2, 2008, p. 45-56.]*

## Complément 2
## Conditions préalables à une bonne analyse quali (p. 255)

Préalablement à l'analyse à proprement parler, le chercheur doit :

- faire le choix des matériels qui formeront le corpus à analyser ;
- établir des hypothèses/objectifs ;
- formuler des indicateurs qui rendront possible l'interprétation terminale et qui serviront de base à la définition d'un ensemble de règles de découpage du corpus tout d'abord, puis de catégorisation et codage.

À ce stade, certaines conditions doivent être respectées afin d'améliorer la qualité de l'analyse qualitative et, plus largement, celle de la recherche concernée. Ainsi, il est nécessaire de prendre en compte de façon exhaustive l'ensemble des composants du corpus retenu, et ce sans a priori aucun. Le respect de cette condition permettra de garantir l'exhaustivité du matériel analysé, et donc une rigueur scientifique supérieure. Imaginons un chercheur en stratégie des entreprises, désireux d'analyser les visions stratégiques de ces dernières, telles qu'elles apparaissent au travers de leur discours contenu dans le rapport annuel d'activités. Il doit donc commencer par définir, en fonction de son objectif de recherche, l'envergure à donner à son corpus et le type de matériaux à réunir de façon pertinente. En l'occurrence, il peut vouloir focaliser son attention uniquement sur les entreprises qui affichent les meilleures performances sur un territoire national donné. Cela le conduira alors à délimiter son corpus aux rapports d'activités des 100 entreprises classées comme les plus performantes en France par un journal économique reconnu.

Dans certains cas, il est possible de déroger à cette condition d'exhaustivité en ne considérant qu'un échantillon du champ d'origine, mais en veillant cependant à ne pas enfreindre la règle de la non-sélectivité qui consiste à s'assurer de réunir tous les éléments qui correspondent aux critères de constitution définis par le chercheur de façon rigoureuse. Ainsi, notre chercheur pourrait, soit de manière aléatoire, soit sur la base de quotas déterminés au vu des qualités possédées par les 100 entreprises considérées (effectifs, C.A., secteur d'activité, etc.), réduire son analyse à un échantillon représentatif de la population concernée.

Dans la situation de recherche décrite, le chercheur respectera aussi la condition d'homogénéité, puisqu'il ne regroupera au sein de son corpus que des documents de même nature et donc comparables. Il serait problématique par exemple que certaines entreprises soient décrites par le biais de leurs rapports d'activités alors que d'autres le seraient au travers d'un entretien avec leur dirigeant*.

* *Ce qui ne signifie pas qu'un corpus ne peut pas compter différentes sources d'informations. Par exemple, le corpus pourrait tout à fait comprendre (de façon systématique) un entretien du dirigeant et le rapport d'activités par entreprise concernée*.

## Complément 3
## Débat sur la terminologie relative à la catégorisation des données (p. 256)

Pour Bardin (2003), il s'agit d'une part des « unités d'enregistrement » ou unités élémentaires de signification qu'il faudra coder, catégoriser et, si besoin, décompter et, d'autre part, des « unités de contexte ». L'unité de contexte renvoie, pour cet auteur, à « l'unité de compréhension pour coder l'unité d'enregistrement ». Autrement dit, il faut choisir cette unité de contexte de telle manière que l'unité d'enregistrement garde du sens relativement à son contexte d'utilisation, ce qui suppose une unité de contexte suffisamment englobante, mais, dans le même temps, il ne faut pas non plus trop élargir l'unité de contexte car des unités trop larges sont plus chronophages en termes de traitement car elles supposent une prise de connaissance plus longue du contexte pour chaque unité d'enregistrement considérée. Ainsi, l'analyste peut faire le choix d'unités d'enregistrement plutôt fondées sur des critères de signification, comme le thème par exemple, ou plutôt fondées sur des critères linguistiques, comme le mot par exemple. Il est cependant souvent difficile de distinguer précisément ce qui relève uniquement de la sémantique de ce qui relève uniquement de la linguistique, les deux étant étroitement liés au sein d'une communication quelle qu'elle soit. La phrase peut aussi bien être considérée comme composant linguistique que comme structure de sens. Pareillement, la phrase peut, selon les objectifs du chercheur, les spécificités du corpus, le genre d'analyse développée, aussi bien être traitée comme l'unité d'enregistrement au sein d'unités de contexte plus vastes que sont les paragraphes, que comme unité de contexte pour l'unité d'enregistrement « mot ». Ces choix ne sont pas sans conséquences dès lors qu'on procède à des comptages sur les unités considérées.

Ainsi, Blanchet et Gotman (2001) évoquent, eux, la notion « d'unité de découpage » définie comme « le fragment de discours portant une signification » et qui, dans le cas d'une analyse thématique (voir partie suivante), correspond au fragment de discours porteur d'un thème. Nous verrons d'ailleurs dans la partie suivante que certaines formes particulières d'analyses recourent à des modes de découpage spécifiques (par exemple, l'analyse propositionnelle du discours ou APD). Ladwein (1996), lui, distingue les « unités de production » (par exemple les entretiens retranscrits) des « unités de traitement » ou « items » (mots, thèmes). Quant à Saunders (2003), il mentionne des « unités d'information »* dont il précise qu'elles correspondent à des bribes (bits) ou à des morceaux plus gros (chunks) de données, pertinents et qui seront attachés aux catégories définies.

De même, les notions de codes, mais plus encore de catégories, ainsi que d'une manière générale tous les types d'annotations/classements qui peuvent être opérés sur un corpus, font l'objet d'un vocabulaire mal stabilisé entre les auteurs, ce qui conduit Paillé et Mucchielli (2003) à déclarer que : *« Il est bien difficile, à l'heure actuelle, de s'y démêler, particulièrement en ce qui concerne le type et le niveau des annotations utilisées par les analystes, les chercheurs n'ayant pas de langage commun sur cette question. »* Ainsi, alors que Bardin (2003) dit des catégories qu'elles sont *« des rubriques ou classes qui rassemblent un groupe d'éléments (unités d'enregistrement dans le cas de l'analyse de contenu) sous un titre générique, rassemblement effectué en raison des caractères communs de ces éléments »* et ne semble donc aucunement distinguer la rubrique de la catégorie, Paillé et Mucchielli (2003), eux, instaurent une nette différence entre ces deux notions. En effet, pour ces derniers, *« la rubrique renvoie à ce dont il est question dans l'extrait du corpus faisant l'objet de l'analyse mais ne renseigne en aucune façon sur ce qui a été dit à ce propos »*, alors que la catégorie peut prendre deux sens distincts. Elle peut être entendue dans un sens très général (« générique »), celui d'une classe regroupant des objets de même nature et ils considèrent que, dans ce cas, sa définition se rapproche de celle de la rubrique ou du thème, ou dans un sens beaucoup plus étroit, inspiré notamment par le courant de recherche de la théorie enracinée (grounded theory), qui correspond alors à la « désignation substantive d'un phénomène ». Ils font le choix de ne retenir que ce sens spécifique et utilisent autrement les termes « rubrique » ou « thème », considérant que « la catégorie se situe, dans son essence, bien au-delà de la simple annotation descriptive ou de la rubrique dénominative. Elle est l'analyse, la conceptualisation mise en forme, la théorisation en progression ». Ils parlent alors plus précisément de « catégories conceptualisantes ». Weber (1990), bien que n'instaurant pas cette différenciation entre catégorie générique/ catégorie conceptualisante, précise qu'une catégorie peut être composée d'une ou plusieurs unités ayant un sens similaire et que cette similarité peut aller d'une totale synonymie à une similarité plus basée sur des connotations communes.)

* *Traduction libre de « units of data »*

*[L. Bardin,* L'Analyse de contenu*, PUF, Paris, 2003.*

*A. Blanchet et A. Gotman,* L'Enquête et ses méthodes : l'entretien*, Nathan Université, Paris, 2001.*

*Ladwein R.,* Les Études marketing*, Économica, Paris, 1996.*

*M. Saunders, P. Lewis et A. Thornhill,* Research methods for business students*, Essex, Pearson Education, 2003.*

*P. Paillé et A. Mucchielli,* L'Analyse qualitative en sciences humaines et sociales*, Armand Colin, Paris, 2003.*

*R.-P. Weber,* Basic content analysis*, Sage Publications, Newbury Park, 1990.]*

---

## Complément 4
## Les différentes formes de catégorisation (p. 258)

Un premier cas de catégorisation correspond à la simple affectation des unités d'enregistrement dans des catégories préalablement identifiées et qui renvoient à des classifications théoriques déjà constituées. Bardin (2003) intitule cette manière de faire : « procédure par boîtes ». Le chercheur a alors surtout à déterminer quels sont les indices, représentatifs d'une catégorie particulière, sur lesquels il devra s'appuyer pour opérer l'attribution des unités d'enregistrement dans les catégories adéquates. Il doit aussi veiller à ce que chaque catégorie soit précisément définie et que soit parfaitement clarifié ce qu'elle comprend et ce qu'elle ne comprend pas. Cette situation peut être rapprochée de ce que certains auteurs nomment « codage a priori » (Stemler, 2001). Stemler considère qu'il y a codage a priori dès lors que *« les catégories ont été établies préalablement à l'analyse sur la base d'une quelconque théorie »*[*]. Il indique qu'il est nécessaire qu'un accord soit obtenu quant aux catégories avant que les données ne soient codées à partir de la liste des catégories définies. Il souligne le fait qu'un travail de révision des catégories peut être effectué au fur et à mesure du codage, de telle manière que ces dernières soient améliorées et tendent vers l'idéal d'exclusion (ou exclusivité) mutuelle et d'exhaustivité prescrit par divers auteurs, dont Weber notamment. En effet, si les catégories se chevauchent, les codeurs risquent de faire des erreurs d'affectation. À ce sujet, Bardin (2003) parle d'exclusion mutuelle, alors que Ladwein (1996) parle, lui,

d'exclusivité mutuelle. Concernant l'exhaustivité, préconisée particulièrement par Berelson (1952), de nombreux auteurs ont depuis relâché la contrainte. Ladwein (1996) par exemple explique que *« comme cette condition est difficile à réaliser […] on accepte souvent une catégorie par défaut (la catégorie « divers ») à condition que ce soit la catégorie dans laquelle on classe le moins d'items »*, quand Blanchet et Gotman (2001) considèrent que : *« Une analyse de contenu doit pouvoir rendre compte de la quasi-totalité du corpus (principe d'extension). »*

Un deuxième cas de catégorisation consiste, à partir des données qualitatives et par une description minutieuse, proche du contenu explicite, à s'attacher à nommer ce dernier. Cette approche correspond à ce que Paillé et Mucchielli (2003) appellent « description analytique » : *« L'appellation de la catégorie ne contient aucun ajout de nature conceptuelle par rapport à l'expérience rapportée ou observée. Le niveau d'inférence de la catégorie est donc peu élevé. »* Le codage ne s'appuie donc pas sur des catégories pré-existantes, il émerge des données, d'où l'appellation « codage émergent » ou « procédure par tas » pour Bardin (2003). À un second stade cependant, l'activité peut devenir plus interprétative, dans la mesure où le chercheur peut vouloir déterminer, – par des rapprochements/regroupements entre les catégories, la création de « méta-catégories », la mise en perspective avec des connaissances scientifiques existantes, etc. –, les explications théoriques du contenu qu'il a catégorisé pour pouvoir lui donner du sens. On se situe alors dans le cadre d'une « déduction interprétative » pour reprendre les termes de Paillé et Mucchielli (2003).

Un troisième cas de catégorisation consiste, toujours en partant des données collectées, à chercher à en induire une interprétation pour dégager une théorisation possible. Cette configuration d'analyse est qualifiée « d'induction théorisante » par Paillé et Mucchielli (2003). La détermination des catégories se fait en restant proche du contenu du corpus et de sa structuration, mais en tentant de faire apparaître une conceptualisation des propos tenus, des expériences décrites, des phénomènes dévoilés, sans s'appuyer sur des constructions théoriques établies. Paillé et Mucchielli (2003) parlent alors de « construction discursive originale ». C'est ce type de conceptualisation, induite par les données, qui est à l'œuvre dans le cas des approches de « théorie enracinée » (grounded theory).

** Traduction libre de : the categories are established prior to the analysis based upon some theory*

*[L. Bardin,* L'Analyse de contenu*, PUF, Paris, 2003.*

*S. Stemler,* An overview of content analysis, Practical Assessment, Research & Evaluation*, 7, 17, 2001, disponible sur http://PAREonline.net/getvn.asp?v=7&n=17*

*R.-P. Weber,* Basic content analysis*, Sage Publications, Newbury Park, 1990.]*

*B. Berelson, Content analysis in communication research, Free Press, New York, 1952.*

*Ladwein R.,* Les Études marketing*, Économica, Paris, 1996.*

*A. Blanchet et A. Gotman,* L'Enquête et ses méthodes : l'entretien*, Nathan Université, Paris, 2001.*

*P. Paillé et A. Mucchielli,* L'Analyse qualitative en sciences humaines et sociales*, Armand Colin, Paris, 2003.]*

## Complément 5
## Codage dans le cas d'une approche de théorie enracinée (grounded theory) (p. 258)

Lors d'un codage qui s'inscrit dans une perspective de théorie enracinée (grounded theory), trois formes de codage apparaissent (Strauss et Corbin, 1990 ; Lincoln et Guba, 1985) :

- le codage ouvert (open coding) ;
- le codage axial (axial coding) ;
- le codage sélectif (selective coding).

Le codage ouvert permet, ainsi que nous l'avons mentionné précédemment, de catégoriser les phénomènes conformément à ce qu'indiquent les données brutes du corpus. À ce niveau, les catégories correspondent à des concepts regroupés et catégorisés, c'est-à-dire au premier stade de construction d'une théorie enracinée.

Ensuite, le chercheur procède à un codage axial qui cherche à établir des connexions entre une catégorie particulière et des sous-catégories qui peuvent lui être attachées et qui renvoient à des dimensions, caractéristiques, propriétés de la catégorie.

Enfin, l'analyste réalise un codage sélectif qui a pour but de mettre en relation les différentes catégories obtenues de façon à pouvoir les articuler autour d'une ou deux catégories centrales et formuler des propositions théoriques. Il s'agit d'une phase d'intégration destinée à structurer un cadre théorique.

*[A. Strauss et J. Corbin,* Basics of qualitative research : Grounded theory procedures and techniques*, Sage, London, 1990.*

*Y.-S. Lincoln et E.-G. Guba,* Naturalistic Inquiry*, Sage, Beverly Hills, 1985.]*

## Complément 6

## Extrait d'une grille de codification relative à des rapports d'activité (p. 259)

| Extrait | Énoncé | Rubrique | Code | Catégorie | Code |
|---|---|---|---|---|---|
| « Financièrement, nous avons renforcé nos marges de manœuvre. Le chiffre d'affaires consolidé, de 45 milliards d'euros, augmente de 7,4 %. Cette croissance est surtout organique : en France, où le marché a pourtant connu une nouvelle étape d'ouverture à la concurrence, les recettes croissent de 5 %. Hors de France, la croissance à périmètre et change constants atteint 16,4 % en Europe et 11 % dans le reste du monde, témoignant d'investissements passés pertinents et de synergies effectives. La rentabilité du Groupe s'améliore fortement, puisque le résultat net courant passe de 0,2 à 1 milliard d'euros. EDF enregistre ainsi un bénéfice net de 857 millions d'euros, après avoir intégré le surcoût de plus de 300 millions d'euros dû aux achats d'électricité pendant la canicule, ainsi que les intérêts attachés au paiement à l'État d'un impôt relatif au réseau d'alimentation générale, demandé par la Commission européenne. » | EDF enregistre des résultats satisfaisants en France comme à l'international. | Résultats | RESUL | Le succès des choix stratégiques | SUCC_STRAT |
| La stratégie du Groupe a pour objectifs :<br>- le développement de ses activités d'exploration et de production ;<br>- le renforcement de sa position parmi les leaders sur les marchés du gaz naturel et du GNL de par le monde ;<br>- la consolidation de ses parts de marché dans le marketing en Europe, tout en se développant sur les marchés en croissance rapide du Bassin méditerranéen, d'Afrique et d'Asie ;<br>- la rationalisation de son portefeuille Chimie en donnant la priorité à l'amélioration de la rentabilité, au développement des activités pétrochimiques et de spécialités, et à la constitution en octobre 2004, d'une nouvelle entité décentralisée comprenant les Produits Vinyliques, la Chimie Industrielle et les Produits de Performance. | Total s'appuie sur différentes stratégies de croissance pour consolider et développer ses positions dans le monde et sur la rationalisation de ses activités pour améliorer sa rentabilité. | Stratégie | STRAT | Leviers stratégiques utilisés | LEV_STRAT_UTIL |

Le 1er extrait est tiré du rapport d'EDF et le second du rapport de Total (données collectées par les auteurs du présent chapitre mais n'ayant pas encore donné lieu à publication).

---

## Complément 7
## Exemple d'utilisation des *verbatim* (p. 260)

Extrait de l'article de **Charles-Pauvers *et alii* (2007) à propos de la gestion des ressources humaines dans les centres d'appels** : « La première mission avancée spontanément par les salariés est l'apport d'une réponse à un client qui appelle : ils considèrent réaliser une mission de prise d'appel (apporter une réponse au client et le conseiller) et, pour certains, un diagnostic.

"… Concrètement, c'est réceptionner des appels des clients et voir les demandes des clients en fonction de leur dossier, en fonction de leur forfait ou factures plus souvent ; et répondre et rechercher la réponse dans leur dossier par rapport au problème qu'ils posent."

La seconde mission, vente de services et développement de l'usage, plus récente, est vue comme une mission commerciale. Ce double aspect est ressenti comme une contradiction et le conseiller éprouve une difficulté à "se positionner" face à son client et face à sa mission.

"…c'est un métier pénible psychiquement, je veux dire, on peut s'user au bout de trois quatre ans quoi.

[Vous la sentiez, vous cette usure-là, au téléphone ?] Ouais ça se sent, c'est-à-dire que, il faut être disponible vraiment à fond tout le temps et on peut pas prendre un appel comme ça, quand on appelle un service comme ça ; ou un CAP, ou une administration, on a l'impression d'être le premier mais en réalité, la personne en est peut-être à son centième appel, faut toujours faire comme si on ne… voilà et ça ouais, ça prend, faut prendre sur soi et qu'à la fin, je crois qu'on s'use quasiment… on peut pas faire ça toute une carrière, c'est certain". »

*[B. Charles-Pauvers, C. Urbain et E. Le Quentrec, « Pratiques de gestion des ressources humaines et performance commerciale - Le cas d'un centre d'appels »,* Revue Française de Gestion*, 33, 176, 2007, p. 21.]*

---

## Page 263 – Voir complément 1, exemple 2

---

## Page 264 - Voir compléments 10 et suivants

---

## Complément 8
## Différences majeures entre les logiciels Spad-T, Lexica, Alceste (p. 266)

À partir de Helme-Guizon et Gavard-Perret, 2004

- Sphinx lexica et Spad-T reposent sur une classification ascendante hiérarchique.

Ces logiciels partent des mots => la construction des catégories est fonction des cooccurrences.

- Alceste fonctionne sur la base d'une double classification descendante hiérarchique.

Ce logiciel part du texte global => il procède par partitions successives du corpus et met ainsi au jour des classes de mots.

- Alceste, du fait de la méthode de classification adoptée (descendante), tend à maximiser les différences à l'intérieur d'un corpus, mais ne garantit pas une totale homogénéité des classes obtenues.
- Spad-T et Lexica, du fait de la méthode de classification adoptée (ascendante) maximisent l'homogénéité des catégories obtenues, mais au détriment des différences inter-catégorielles.
- Alceste cherche avant tout à *« rendre compte de l'organisation interne d'un discours plutôt que rendre compte de différences statistiques entre les divers textes d'un corpus »* (Reinert).

Ce logiciel est donc plus pensé dans une logique d'analyse du discours que dans une logique de statistique lexicale, mais il nécessite un corpus homogène. Ce critère d'homogénéité conduit à délaisser une partie du corpus et l'on peut perdre ainsi le vocabulaire rare.

- Spad-T ne comporte ni dictionnaire par défaut, ni lemmatisation, ni découpage automatisé du texte.

Il est donc difficilement utilisable pour des corpus complexes comme des corpus d'entretiens car cela suppose de lourdes opérations manuelles (de regroupement de mots d'une même famille, de construction de dictionnaires de mots outils, etc. et de pré-découpage du texte.

- Alceste peut convenir pour une recherche sans a priori dans laquelle le chercheur n'a pas d'objectif précis d'analyse ni de stratégie particulière d'analyse du contenu, puisque, dans tous les cas et sans paramétrages ou choix particuliers de l'analyste, il proposera des classes. Charge ensuite à l'analyste d'essayer de les interpréter et de leur donner du sens !
- Sphinx Lexica et surtout Spad-T supposent une stratégie de recherche préalable car les possibilités d'action, les choix à faire et les manipulations sont nombreuses (création de variables ; regroupements ; suppression ; etc.). Sans intention particulière et sans procédure d'analyse en tête, le chercheur peut se perdre dans la masse des données et la palette des fonctionnalités offertes ou, ne pas réussir à « faire parler » le corpus sur la seule base des statistiques lexicales (qui, elles, requièrent relativement peu d'initiatives).

Complémentarités possibles :

- Alceste permet de faire émerger des classes rapidement et Lexica de les préciser au moyen de la statistique lexicale et des comptages sur les individus. Avec Alceste, on peut ainsi poser rapidement des hypothèses interprétatives à partir des structures textuelles dévoilées, tandis que Spad-T ou Lexica apporteront une vérification de ces hypothèses à la lumière de la statistique lexicale.
- Mais Lexica peut aussi être utilisé en premier afin de « dégrossir » l'analyse, alors qu'Alceste, plus précis en matière de classification, affinera ce premier travail. Lexica s'avère utile également pour intervenir sur ce qu' Alceste a ignoré et parfaire ainsi l'analyse.

*[A. Helme-Guizon et M.-L. Gavard-Perret , « L'analyse automatisée de données textuelles en marketing : comparaison de trois logiciels »,* Décisions Marketing*, 36, n° spécial « Études Qualitatives », 2004, p. 75-90.]*

---

## Complément 9

## Choix de la structuration du corpus, conséquences du formatage retenu et extraits de corpus correspondant à différents formatages (p. 268)

### *Choix de formatage et conséquences*

Le fichier contenant les données textuelles à analyser à l'aide du logiciel Sphinx doit être un fichier « texte » (.txt). Il est possible également de travailler à partir de fichiers Excel ou Access. En revanche, on ne peut pas importer directement un ficher de type Word (traitement de texte). Il est conseillé de passer son fichier au correcteur orthographique avant de l'importer dans Sphinx Lexica, car un mot mal écrit sera traité comme une forme graphique spécifique par le logiciel et ne sera donc pas compté avec la forme graphique correspondant à ce même mot correctement écrit (sauf si on tient à conserver les mots exacts utilisés par les répondants, même s'ils comptent des fautes, parce que c'est important pour le type d'analyse envisagé). Il faut aussi veiller à ne pas utiliser certains paramètres et fonctions des traitements de texte qui introduisent des caractères spéciaux dans le corpus, caractères que le logiciel Sphinx Lexica ne sait pas reconnaître ou remplace automatiquement par d'autres de manière pas toujours satisfaisante.

### *Formatage par annotations*

Les textes annotés conviennent très bien au traitement de collections d'articles, de bases de données bibliographiques, de chapitres d'un livre, etc.

Dans ce cas, au début de chaque texte considéré, une annotation (le jalon) va préciser le nom du texte et ses caractéristiques, comme l'indiquent les exemples ci-après. Ainsi, dans le premier exemple qui concerne un corpus constitué de rapports d'activité, le jalon (J) mentionne le nom de l'entreprise, son secteur industriel et l'année du rapport. Dans le deuxième exemple, on pourrait procéder de cette manière pour une collection d'articles, si aucune information particulière relative à l'article n'était intéressante pour l'analyste. Dans le troisième exemple, il pourrait s'agir d'articles, de références bibliographiques, d'entretiens, etc. identifiés par le nom auquel chacun d'entre eux se rattache.

[J = Rhodia, chimie, 2004]

ou

[J = article 1]

ou

[J = JOLIBERT]

L'analyste peut aussi ajouter une marque (M) pour distinguer différents niveaux de fragments à l'intérieur d'un même document délimité par les jalons (par exemple les questions des réponses). Ainsi, dans le cas d'une utilisation pour des entretiens, on pourrait avoir la structuration suivante :

[J= DUPONT, F, 32, employée]

[M = q] et [M = r]

La table qui serait constituée par le logiciel sur la base de ces indications serait la suivante :

| **Entretien** | **Texte** | **Statut** |
|---|---|---|
| DUPONT, F, 32, employée | AZERTYUIOPQSDFGHJKLMWXCVBN ?.<br><br>NBVCXWMLKJHGFDSQPOIUYTREZA. | Q<br><br>R |

L'analyste dispose par conséquent d'une variable texte et de deux variables signalétiques. Il ne pourra pas (sans manipulation spécifique et création de nouvelles variables) examiner l'effet du genre de la personne, par exemple sur les propos tenus, puisque le genre n'est pas mentionné comme un repère particulier susceptible de donner lieu à la constitution d'une variable. Il en va de même dans cet exemple pour l'âge ou la catégorie socioprofessionnelle du répondant.

A contrario, si l'analyste fait le choix de baliser le corpus, il pourra agir de la manière suivante.

*Formatage par balises*

Le balisage convient mieux à des documents ou entretiens pour lesquels des données de contextualisation sont importantes ou dont le découpage en rubriques, thèmes, etc. est indispensable. Ainsi, si l'on reprend l'exemple indiqué ci-dessus, l'analyste introduira les balises suivantes :

<Entretien> DUPONT

<Sexe> F

<Age> 32

<CSP> EMPLOYEE

<Q> AZERTYUIOPQSDFGHJKLMWXCVBN ?

<R> nbvcxwmlkjhgfdsqpoiuytreza.

Grâce à cet ensemble de balises, la table suivante sera constituée.

| **Entretien** | **Sexe** | **Âge** | **CSP** | **Q** | **R** |
|---|---|---|---|---|---|
| DUPONT | F | 32 | employée | AZERTYUIOPQSDFGHJKLMWXCVBN ? | nbvcxwmlkjhgfdsqpoiuytreza. |

L'analyste a alors à sa disposition deux variables texte et quatre variables signalétiques. Il pourra ainsi aisément confronter les discours des hommes à ceux des femmes (idem en fonction de l'âge ou de la CSP).

## Extraits de corpus de rapports d'activité formaté soit avec annotations, soit avec balises sous Sphinx Lexica

***Formatage par annotations***

[JT= A1, Total, Industrie_pétrochimique]

« Stratégies - En 2003, le Groupe, qui a pris le nom de Total, a été l'un des acteurs les plus dynamiques et les plus performants de l'industrie pétrolière mondiale. En l'espace de quatre ans, la production d'hydrocarbures du Groupe a crû de 23 %. Cette progression s'appuie sur une stratégie clairement définie et des perspectives visant à concilier l'amélioration des performances et le respect des engagements pris par le Groupe en faveur d'un développement responsable et durable.

La stratégie du Groupe a pour objectifs : […]

[…] L'objectif est de former un nouveau groupe chimique ayant vocation à devenir indépendant et qui, doté d'une structure financière solide, puisse s'assurer un développement pérenne. »

[JT= A2, Vivendi_universal, Communication]

[…]

***Formatage par balises***

<extrait> A1

<entreprise> Total

<secteur> Industrie_pétrochimique

<T1> « Stratégies. En 2003, le Groupe, qui a pris le nom de Total, a été l'un des acteurs les plus dynamiques et les plus performants de l'industrie pétrolière mondiale. En l'espace de quatre ans, la production d'hydrocarbures du Groupe a crû de 23 %. Cette progression s'appuie sur une stratégie clairement définie et des perspectives visant à concilier l'amélioration des performances et le respect des engagements pris par le Groupe en faveur d'un développement responsable et durable.

La stratégie du Groupe a pour objectifs : […]

[…] L'objectif est de former un nouveau groupe chimique ayant vocation à devenir indépendant et qui, doté d'une structure financière solide, puisse s'assurer un développement pérenne. »

<extrait> A2

<entreprise> Vivendi_universal

<secteur> Communication

<T2> Vivendi Universal, ayant maintenant assaini sa situation de trésorerie, peut mettre en œuvre une stratégie de (…)

## Exemple de formatage de corpus pour le logiciel Alceste

L'exemple est extrait de la recherche menée par les auteurs relativement à la personnalisation sur Internet.

0001 *age_35 *sexe_masc *CSP_cadre *étude_bac *exp_+5a *duree_-2h *typ_ADSL *lie_dom *nbrha_1a2 *type_prdtcult *typ_voy *Q_0

Personnalisation... C'est-a-dire ? Est-ce que ca peut être euh...

Personnalisation c'est, en fonction du but recherche, euh... lors de consultations sur Internet, c'est d'avoir le service que l'on recherche dans le meilleur délai.

0001 *age_35 *sexe_masc *CSP_cadre *étude_bac *exp_+5a *duree_-2h *typ_ADSL *lie_dom *nbrha_1a2 *type_prdtcult *typ_voy *Q_1

Oui, tout à fait. Sur des sites..., des sites par exemple basés sur le euh... sur le commerce en ligne, donc une fonctionnalité avec une convivialité d'utilisation, avec euh... services complémentaires de mémorisation des commandes passées ; donc c'est vraiment adapté pour que l'internaute revienne sur ce site.

0001 *age_35 *sexe_masc *CSP_cadre *étude_bac *exp_+5a *duree_-2h *typ_ADSL *lie_dom *nbrha_1a2 *type_prdtcult *typ_voy *Q_2

Les deux, navigation à titre personnel et à titre professionnel.

0001 *age_35 *sexe_masc *CSP_cadre *étude_bac *exp_+5a *duree_-2h *typ_ADSL *lie_dom *nbrha_1a2 *type_prdtcult *typ_voy *Q_3

Non, j'ai jamais constaté des efforts de personnalisation faits par des sites sans que je les aies demandés ; non parce qu'en général, les sites que je consulte, je les consulte dans un objectif bien déterminé, donc je prends pas le temps vraiment de... de regarder l'architecture du site, mais simplement d'aller à l'essentiel, là où je souhaite aller. Et donc, lors de euh... Comment dire... lors de recherches sur ces critères là, si le site ne répond pas dans un délai bref à ce que je recherche et bien, je le quitte et je vais prendre le suivant.

La ligne, dite étoilée, permet de caractériser les différents fragments composant le corpus. Plus précisément :

0001 signifie qu'il s'agit du premier entretien ; pour le 2$^{e}$ entretien, on indiquera 0002

*age_35 indique que le répondant a entre entre 35 et 49 ans ; les autres modalités sont *age_18 (âge compris entre 18 et 24 ans), *age_25 (âge compris entre 25 et 34 ans), * age_50 (âge au delà de 50 ans)

*sexe_masc indique que le répondant est de sexe masculin (vs féminin : *sexe :fem)

Et ainsi de suite pour les autres *CSP_cadre (CSP) *étude_bac (niveau d'études) *exp_+5a (nombre d'années depuis lesquelles le répondant navigue sur Internet) *duree_-2h (durée hebdomadaire de connexion) *typ_ADSL (type de connexion) *lieu_dom (lieu de connexion) *nbrha_1a2 (nombre de produits achetés en ligne au cours des 3 derniers mois *type_prdtcult (type de produits achetés- culturels) *typ_voy (type de produits achetés- voyages) *Q_2 (question)

Le nombre de variables illustratives n'est pas limité.

Il a été fait le choix d'identifier les questions auxquelles les réponses se rapportent Q_0, Q_1, Q_2 etc. Ce parti pris de recherche s'est révélé pertinent car il nous a permis de déceler que les individus parlaient de personnalisation passive (e-mails sollicités, pubs, spams, etc.) alors qu'il leur était posé une question sur la personnalisation active (produit, page web, etc.). Et vice-versa. Nous en avons déduit une certaine confusion quant à l'objet de la recherche. Ceci nous a suggéré d'adopter une autre approche d'analyse que celle initialement envisagée selon la comparaison personnalisation active/passive.

---

## Complément 10
## Présentation et comparaison de différentes catégories de lexiques (p. 269)

***Comparaison des lexiques brut, réduit et lemmatisé pour le corpus des réponses à la question ouverte : « Si vous gagniez au loto ? » (990 réponses)***

[Sphinx Développement, **http://www.lesphinx-developpement.fr/**]

| **Lexique** | | **Lexique réduit** | | | | **Lexique lemmatisé** | | | |
|---|---|---|---|---|---|---|---|---|---|
| je | 1182 | maison | enfants | voiture | voyage | 355 | maison | faire | acheter | 363 |
| j | 653 | ferais | | | | 286 | voyage | | | 357 |
| de | 613 | voyages | | | | 196 | enfant | | | 350 |
| une | 566 | achèterais | | | | 192 | placer | | | 311 |
| en | 464 | famille | | | | 182 | voiture | donner | voyager | 288 |
| à | 412 | argent | | | | 158 | aider | | | 237 |
| le | 385 | acheter | | | | 150 | | | | 197 |
| des | 362 | | | | | 122 | | | | 187 |
| un | 362 | | | | | 118 | | | | 147 |
| maison | 355 | | | | | 115 | | | | 122 |

Le lexique réduit aux seuls mots pleins permet très rapidement de repérer les idées majeures exprimées par les sujets et leurs intentions principales (maison, voyage, voiture) mais aussi les bénéficiaires de ces projets (enfants, famille). Le lexique lemmatisé précise encore le sens de certaines formes graphiques et permet de faire des comptages plus justes. Ainsi, le mot « maison » passe de 355 occurrences à 363, soulignant le fait que le mot était utilisé au singulier mais aussi au pluriel. Pareillement, le verbe « acheter » révèle toute sa puissance une fois toutes ses formes conjuguées regroupées. De même, d'autres formes graphiques élémentaires qui ne faisaient pas partie des 10 premières dans le lexique réduit atteignent les premières places dans le lexique lemmatisé (placer, donner, aider). Il est important de souligner qu'avec ces 10 mots, on arrive à refléter pratiquement le tiers du corpus.

***Taille de corpus et de lexique pour deux corpus de nature différente***

| Corpus de slogans publicitaires (Gavard-Perret *et alii,* 1995) et entretiens non directifs sur les transports (Durrande-Moreau, 1994) | Taille du corpus total | Taille du lexique total | Taille du corpus réduit | Taille du lexique réduit |
|---|---|---|---|---|
| Corpus slogans (23 78 observations) | 15 610 | 2 440 | 7 422 | 2 314 |
| Corpus entretiens non-directifs sur les transports (10 observations) | 59 282 | 3 626 | 22 893 | 3 417 |

*[M.-L. Gavard-Perret, J. Moscarola et P. Domenjoz, « Langage publicitaire et contexte d'émission et de diffusion. Analyse d'annonces publicitaires imprimées de langue anglaise »,* Actes du XI[e] congrès de l'Association Française du Marketing*, 11, Reims, 11-12 mai, 1995, p. 1245-1274.*

*A. Durrande-Moreau ,* Qualité de service et perception du temps : l'attente, propositions théoriques et étude empirique*, Thèse, Université de Grenoble II, 1994.]*

***Les lexiques catégorisés issus du corpus des réponses à la question ouverte : « Si vous gagniez au loto ? » (990 réponses)***

[Sphinx Développement, **http://www.lesphinx-developpement.fr/**]

| **Noms** | | **Verbes** | | **Adjectifs** | |
|---|---|---|---|---|---|
| maison | 354 | acheter | 350 | beau | 72 |
| voyage | 310 | placer | 195 | petit | 41 |
| enfant | 259 | donner | 187 | humanitaire | 32 |
| voiture | 194 | voyager | 148 | grand | 29 |
| argent | 118 | aider | 122 | bon | 26 |
| famille | 117 | partir | 91 | immobilier | 19 |
| monde | 81 | travailler | 91 | tout | 13 |
| vacances | 79 | profiter investir changer | 81 | nouvel | 12 |
| don | 75 | | 76 | personnel | 11 |
| placement | 71 | | 63 | gros | 8 |

Ainsi, le lexique des 10 verbes les plus occurrents fait nettement ressortir différents groupes d'actions envisagées : celles qui se rattachent à l'acte de consommer de façon hédoniste (acheter, voyager), celles qui ont trait à l'acte de tirer des profits des sommes gagnées (placer, investir) ou encore celles qui se réfèrent plutôt à un acte altruiste (donner, aider). Un autre groupe renvoie à l'idée de modifier son existant (changer, partir). D'autres verbes, en revanche, nécessitent d'être précisés (notamment par leur contexte d'utilisation dans la phrase ; nous verrons ce point ultérieurement) pour qu'on puisse leur conférer un sens sans ambiguïté ni contresens. C'est le cas des verbes « profiter » (s'agit-il de profiter de la vie, de faire profiter son argent ou encore de faire profiter ses enfants des gains acquis ?). Il en est de même pour « travailler » dont le sens reste encore incertain. Enfin, si le verbe « changer » exprime clairement l'idée générale de modification, il mérite néanmoins d'être précisé afin de connaître l'objet ou le sujet de cette envie de changer (la maison, la voiture, le travail, son style de vie, son époux ou épouse, etc. !).

Le lexique des 10 noms communs les plus fréquents fait apparaître différents groupes « d'objets » : ceux sur quoi s'appliqueront les actions projetées tout d'abord (vacances/voyage ; voiture ; maison ; placement ; don) et ceux qui bénéficieront des actions (enfant ; famille). Quelques mots ne sont pas suffisamment explicites pour pouvoir être cernés immédiatement (« argent » : sera-t-il donné, investi, dépensé, etc. ?; « monde » : est-ce l'intention de donner de l'argent à tout le monde, d'aider son petit monde personnel, de faire le tour du monde, etc. ?).

Quant au lexique des 10 adjectifs les plus cités, c'est indéniablement le plus équivoque car, par sa nature même, l'adjectif se rapporte à un nom commun et, sans ce nom, son sens devient ambigu. Seul un petit nombre d'adjectifs garde une signification suffisante pour pouvoir être interprété en première lecture : « humanitaire », « immobilier » et dans une certaine mesure « nouvel » qui évoque sans nul doute l'idée de changement, de renouvellement (il restera à préciser à quoi il s'applique cependant : vie, maison, voiture, travail, etc.). Les autres supposent forcément une mise en contexte pour reprendre tout leur sens, d'autant que plusieurs d'entre eux pourraient voir leur signification changer radicalement s'ils étaient associés à une négation.

---

## Complément 11

## Lexiques de segments répétés et de formes graphiques les plus occurrentes (p. 269)

***Lexique des 8 segments répétés les plus occurrents du corpus des réponses à la question ouverte : « Si vous gagniez au loto ? » (990 réponses)***

[Sphinx Développement, **http://www.lesphinx-developpement.fr/**]

| **Les 8 segments répétés les plus occurrents** | |
|---|---|
| acheter maison | 140 |
| placer argent | 59 |
| faire voyage | 55 |
| arrêter travailler | 52 |
| tour monde | 50 |
| faire profiter | 44 |
| donner enfant | 40 |
| acheter voiture | 37 |

L'exemple précédent montre bien qu'avec les huit segments répétés les plus occurrents, il est possible de cerner assez précisément les idées clés du corpus et de lever certaines incertitudes mentionnées précédemment. Il devient clair par exemple que « monde » prend surtout son sens au travers du « tour du monde » rêvé par les répondants. Un sens principal du mot « travailler » s'éclaire : il s'agit « d'arrêter de travailler » ! Pareillement, « profiter » se précise : pas seulement profiter égoïstement ou faire profiter son argent, mais aussi et surtout « faire profiter » d'autres personnes de son gain et notamment « donner à ses enfants ».

***Lexique des 15 formes graphiques les plus fréquentes et lexique des 15 segments répétés les plus fréquents (corpus « Personnalisation sur Internet », Helme-Guizon et Gavard-Perret)***

Cet exemple est développé plus précisément dans le chapitre de A. Helme-Guizon et M.-L. Gavard-Perret, « L'analyse de données textuelles avec Sphinx - Une application à la personnalisation sur Internet », *in Analyse Statistique de Données Textuelles en Sciences de Gestion – Concepts, Méthodes, Applications*, dirigé par C. Gauzente et D. Peyrat-Guillard, Éditions EMS, 2007.

| Les 15 mots les plus fréquents | | Les 15 segments répétés les plus fréquents | |
|---|---|---|---|
| euh | 969 | Peut-être | 131 |
| site | 756 | ça peut | 39 |
| ça | 753 | ça va | 27 |
| d | 557 | centres d'intérêt | 22 |
| j | 533 | titre personnel | 22 |
| l | 432 | aujourd'hui | 22 |
| personnalisation | 390 | techniques de personnalisation | 19 |
| m | 217 | efforts de personnalisation | 19 |
| information | 195 | e-mail | 19 |
| exemple | 188 | Ça peut être | 18 |
| chose | 162 | personnalisation non | 18 |
| s | 137 | veux dire | 18 |
| n | 118 | personnalisation sur Internet | 18 |
| page | 111 | peut être | 18 |
| temps | 104 | partir du moment | 17 |

** avant lemmatisation et réduction*

Le corpus concerné comprend 24 entretiens réalisés sur le thème de la personnalisation sur Internet. Alors que le lexique des 15 formes graphiques les plus occurrentes ne nous apprend rien ou fort peu de choses sur le corpus, le passage au lexique des 15 segments répétés les plus fréquents s'avère beaucoup plus intéressant. Le mot « personnalisation » renvoie à différentes catégories d'idées : les techniques de personnalisation ; les efforts de personnalisation, la personnalisation sur Internet (peut-être par différenciation/opposition avec la personnalisation dans le monde réel ?), de même qu'une autre distinction se fait jour avec « à titre personnel » par rapport vraisemblablement à « à titre professionnel ». De plus, le segment « personnalisation non » mérite une attention particulière puisqu'il souligne le fait que les répondants ont insisté sur la « personnalisation non … sollicitée, demandée, voulue, etc. ». Enfin, l'incertitude et les doutes des répondants sur ce qu'est réellement la personnalisation sur Internet transparaissent nettement de ces quelques segments répétés : « peut-être », « ça peut », « ça va », « peut être », « ça peut être », « veux dire ».

## Complément 12

## Les concordances et lexiques relatifs pour le mot « politique » (exemples tirés du site Sphinx Développement ) (p. 270)

| Concordances pour l'adjectif « politique » |
|---|
| plus que ne l'imaginent les acteurs politiques, les responsables économiques, les intellectuels, |
| je constate que les rouages politiques, économiques et sociaux de notre pays sont atteints |
| le milieu politique donne aux français le spectacle d'un interminable |
| volontiers tous les responsables politiques dans le même panier |
| l'existence d'une alternative politique |
| ce peu de marge à la décision politique |

***Extrait du lexique relatif au mot-pivot « politique » (à partir du lexique lemmatisé)***

| pivot –2 à –1<br>261 mots, 355 occurrences | | pivot +1 à +2<br>150 mots, 351 occurrences | |
|---|---|---|---|
| être-V | 17 | être-V | 13 |
| volonté-N | 7 | économique-A | 11 |
| Europe-P | 5 | avoir-V | 7 |
| pouvoir-N | 5 | commun-A | 6 |
| action-N | 4 | social-A | 6 |
| décision-N | 4 | contractuel-A | 5 |
| responsable-N | 4 | étranger-A | 4 |
| véritable-A | 4 | européen-A | 4 |
| | | pouvoir-V | 4 |

## Complément 13
## Caractérisation statistique de corpus de natures différentes (p. 270)

Adapté de M.-L. Gavard-Perret et J. Moscarola, « Enoncé ou énonciation? Deux objets différents de l'analyse lexicale en marketing », *Recherche et Applications en Marketing*, 13, 2, 1998 p. 31-47.

| | **Nombre observations** | **Taille corpus total** | **Taille lexique brut** | **Répétition** | **Taille corpus réduit** | **Taille lexique réduit** | **Répétition** |
|---|---|---|---|---|---|---|---|
| Corpus de type « discours » | | | | | | | |
| Résumés scientifiques | 90 | 10 352 | 2 378 | 4,35 | 5 208 | 2 220 | 2,35 |
| Tracts de communication aux élections européennes 1994 | 6 | 3 479 | 1 065 | 3,26 | 2 884 | 1 017 | 2,83 |
| Entretiens non directifs sur les transports | 10 | 59 282 | 3 626 | 16,35 | 22 893 | 3 417 | 6,70 |
| Corpus de type « phrases » | | | | | | | |
| Slogans publicitaires de Wellhoff | 2 378 | 15 610 | 2 440 | 6,4 | 7 422 | 2 314 | 3,21 |
| Corpus de type « mots » | | | | | | | |
| Le vendeur en quelques mots | 370 | 1 454 | 460 | 3,16 | 1 171 | 401 | 2,92 |
| Le vendeur idéal en quelques mots | 370 | 2 121 | 485 | 4,37 | 1 487 | 420 | 3,54 |
| Adjectifs associés à des sportifs | 43 | 195 | 98 | 1,99 | 191 | 95 | 2,01 |
| Adjectifs associés à des marques sponsors | 43 | 158 | 114 | 1,39 | 150 | 107 | 1,4 |

Il apparaît nettement que la répétition des formes graphiques diminue fortement dès lors que l'émetteur fait un choix réfléchi (et/ou contraint) de ses mots. Si le répondant à un entretien s'exprime de façon fort répétitive du point de vue des mots utilisés (et donc peut-être des idées émises), ce n'est pas le cas pour les hommes politiques ni pour les publicitaires qui choisissent avec soin les mots, respectivement, de leur profession de foi et de leur slogan, ni pour le chercheur qui doit résumer son article en un nombre limité de mots. De même, les protocoles d'association de mots conduisent à des choix de mots peu redondants.

Si l'on considère la lexicalité, (c'est-à-dire le rapport entre les mots pleins, appelés aussi mots lexicaux ou mots signifiants, et le nombre total de mots d'un corpus) et que l'on ne considère que les corpus de discours et de phrases, on peut alors opposer la faible lexicalité (38 %) des entretiens par rapport à la forte lexicalité des slogans publicitaires (47 %) et des résumés scientifiques (50 %). Alors que dans ces deux derniers cas, l'expression est particulièrement riche en mots pleins avec un mot plein sur deux mots (ou quasiment pour les slogans), celle des entretiens est moins « lexicale ».

## Complément 14

## Spécificités lexicales dans le cas du corpus « Personnalisation sur Internet » pour la variable « Nombre d'achats sur internet au cours des 3 mois ayant précédé l'entretien » (p. 270)

À partir de A. Helme-Guizon et M.-L. Gavard-Perret, « L'analyse de données textuelles avec Sphinx - Une application à la personnalisation sur Internet », *in Analyse Statistique de Données Textuelles en Sciences de Gestion – Concepts, Méthodes, Applications*, dirigé par C. Gauzente et D. Peyrat-Guillard, Éditions EMS, 2007.

| +6 achats | 3_5 achats | 1_2 achats | 0 achat |
|---|---|---|---|
| +énervant | +gagner_du_temps | – énervant | +pratique |
| +servir_rien | +pratique | +agréable | +agréable |
| | +ennuyeux | +sympa | +gagner_du_temps |
| | +perdre_temps | +perdre_temps | +appréciable |
| | +avoir_n_besoin* | +utile | +servir_rien |
| | | | |
| – perdre_temps | -énervant | – avoir_n_besoin | – énervant |
| – sympa | -appréciable | – pratique | – sympa |
| – gagner_du_temps | | | – avoir_n_besoin* |

NB : seules les formes graphiques spécifiques au sens du Chi Deux ont été conservées dans ce tableau. Le signe positif ou négatif indique s'il s'agit d'une spécificité positive ou négative (pour plus de précisions sur la notion de Chi Deux, voir le chapitre 8).

* *la mention _n_ indique la forme négative du verbe dans le lexique lemmatisé*

Par ce calcul de spécificités lexicales, le logiciel dévoile une différence essentielle entre les internautes les plus expérimentés (+ 6 achats) et ceux n'ayant jamais fait d'achat sur Internet. Alors que les premiers semblent avoir perdu toute illusion sur les bénéfices possibles des actions de personnalisation sur Internet, les seconds, a contrario, semblent encore assez séduits par ces dernières.

## Complément 15

## Les intensités lexicales des thèmes de campagne (p. 271)

À partir de M.-L. Gavard-Perret et J. Moscarola, « Enoncé ou énonciation? Deux objets différents de l'analyse lexicale en marketing », *Recherche et Applications en Marketing*, 13, 2, 1998 p. 31-47.

**Les thèmes de la campagne**

| | Europe/ens/ennes | France/ais/aises | Emploi chômage travail | Société social | Vote élection | Économie entreprise |
|---|---|---|---|---|---|---|
| Baudis | **9,91** | **5,66** | 0,94 | 0,47 | 1,42 | 0 |
| Rocard | 5,48 | 1,44 | 1,73 | **4,9** | 0,86 | 0,29 |
| Tapie | 8,03 | 2,92 | 0,73 | 0,73 | 1,46 | 0 |
| Villiers | 7,27 | 2,73 | 1,21 | 0,61 | 0,91 | **0,91** |
| Le Pen | 4,14 | **6,51** | 1,78 | 0,3 | **2,07** | 0,59 |
| Wurtz | **3,19** | **1,14** | **3,86** | 1,14 | 0,91 | 0,23 |
| Ensemble | 6,34 | 3,4 | 1,71 | 1,36 | 1,27 | 0,34 |

NB : les valeurs en caractères gras indiquent une intensité lexicale qui présente un écart statistiquement significatif par rapport à la moyenne.

## Complément 16

### L'utilisation des pronoms personnels, du nom du parti et du nom de la tête de liste selon les candidats caractérisée par le calcul d'intensités lexicales (p. 271)

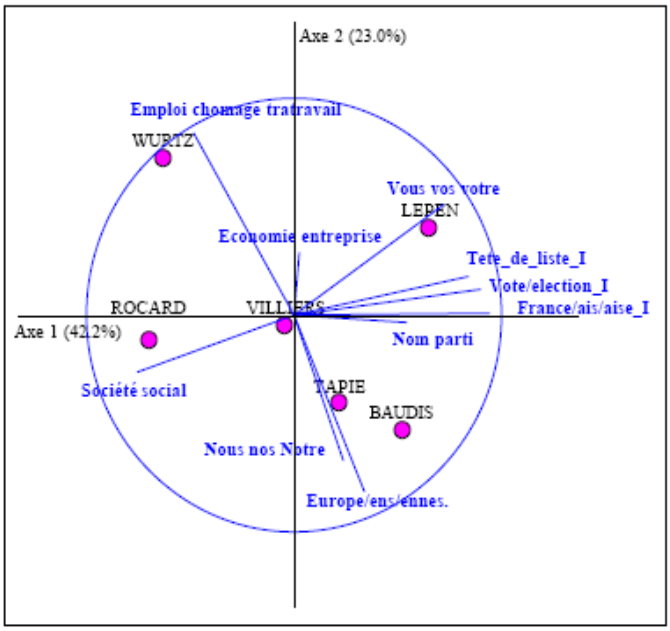
| | Vote | Vote | | Liste |
|---|---|---|---|---|
| Baudis | 2,83 | 2,36 | **2,18** | 0,73 |
| Rocard | 4,32 | **0,29** | 1,19 | 0,3 |
| Tapie | **8,76** | 2,19 | 1,02 | 1,02 |
| Villiers | 6,06 | 2,12 | 0,9 | 0,9 |
| Le Pen | 4,44 | **2,96** | 1,57 | **2,19** |
| Wurtz | **0,91** | 2,51 | 1,13 | 0,25 |
| Ensemble | 4,55 | 2,07 | 1,33 | 0,9 |

Les intensités lexicales sont exprimées en %. Les valeurs en gras signalent un écart significatif à la moyenne.

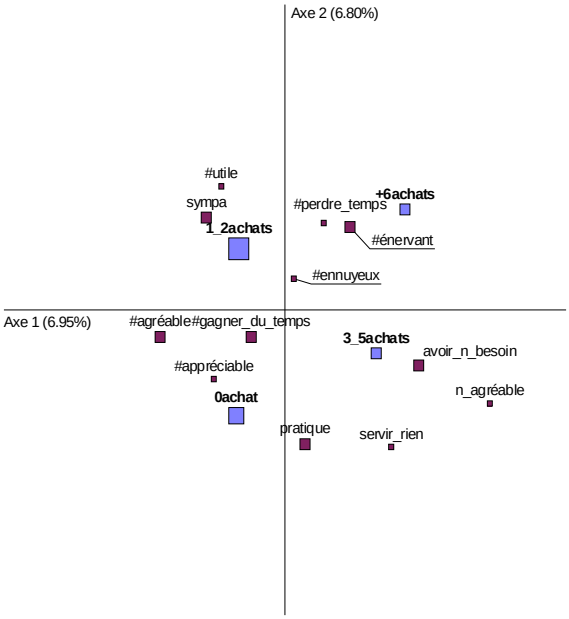
## Complément 17

## Mise en évidence d'associations au travers d'une ACP et d'une AFCM (p. 271)

***Associations thématiques et stylistiques au travers d'une ACP sur des intensités lexicales (corpus élections européennes)***

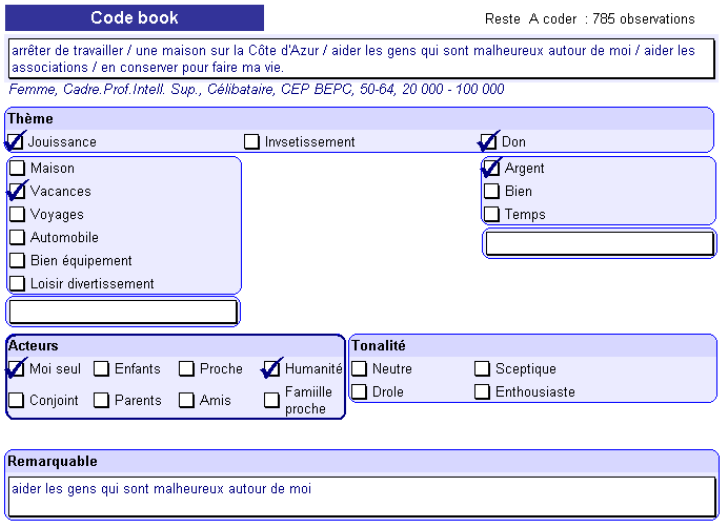



***Associations entre le nombre d'achats sur Internet et les évaluations portées sur les techniques de personnalisation sur Internet à l'aide d'une AFCM***




NB : le symbole # indique que plusieurs formes graphiques ont été regroupées car synonymes.

## Complément 18
## Extrait du code book réalisé pour le corpus Loto (p. 272)




## Complément 19
## Le Kappa de Cohen (p. 273)

La formule de calcul est la suivante :

K = PA – PC / 1 – PC

Où : PA est la proportion d'unités sur laquelle les codeurs sont d'accord ;

et PC est la proportion d'unités pour laquelle l'accord est attendu par effet du hasard uniquement.

Plus on obtient un K proche de 1, et plus la fiabilité entre les codeurs est élevée. Toutefois, des discussions ont eu lieu entre les chercheurs afin de déterminer à partir de quel niveau le coefficient Kappa de Cohen pouvait être considéré comme satisfaisant ou non. Par exemple, Landis et Koch (1977) proposent les repères suivants pour interpréter le coefficient de Kappa (tableau suivant) :

**Repères pour interpréter le coefficient Kappa de Cohen**

| Kappa de Cohen | Fiabilité de l'accord |
|---|---|
| <0.00 | mauvais |
| 0.00-0.20 | faible |
| 0.21-0.40 | modéré |
| 0.41-0.60 | moyen |
| 0.61-0.80 | important |
| 0.81-1.00 | quasiment parfait |

*[J.-R. Landis, G.-G. Koch :* The Measurement of Observer Agreement for Categorical Data*, Biometrics, 33, 1977, p. 159-174.]*