Saadi Lahlou

L'analyse lexicale

Article (Accepted version) (Unrefereed)

(1994) - LAHLOU, Saadi. L'analyse lexicale. *Variances*, n°3, Octobre 1994. pp. 13-24.

L'analyse lexicale

Une nouvelle famille de techniques se développe actuellement, qui permet d'appliquer les méthodes statistiques à du texte. L'analyse lexicale permet des approches exploratoires extrêmement puissantes, notamment pour la constitution de classifications "naturelles", et ouvre à la statistique descriptive le champ des données qualitatives.
Cette famille de méthodes va bouleverser la séparation classique entre approches qualitative et quantitative.
L'évolution des techniques dans le domaine, extrêmement rapide depuis quelques années, aura certainement des implications majeures dans le domaine des études et du marketing. Elle va modifier la conception classique des questionnaires au profit des questions ouvertes. Elle va surtout ouvrir à l'analyse statistique la masse considérable des bases de données en texte intégral.

Statisticiens contre qualiticiens

On sait qu'il y a dans les études, deux écoles: les quantitativistes et les qualitativistes, populations réciproquement hostiles qui se livrent maints combats homériques.
Du côté quantitatif, les statisticiens font un choix a priori des variables. De l'autre, les partisans des méthodes qualitatives renoncent à faire ce choix et explorent de façon ouverte (entretiens, descriptions...) le matériel à analyser. Les premiers savent bien mesurer l'importance relative des paramètres qu'ils ont choisis, et les seconds savent découvrir les bons paramètres. Les premiers reprochent aux seconds le manque de comparabilité de leurs observations, et mettent en avant la puissance des analyses statistiques que leur permet leur choix initial. Ils ont beau jeu de souligner qu'il serait difficile, par exemple, de réaliser le recensement de la population par entretiens semi-directifs, et que la subjectivité de l'enquêteur amène des biais considérables.

De leur côté, les partisans des méthodes qualitatives répondent que les quantitativistes sont bornés dans leur analyse par leur choix préalable des variables, qui risque souvent de ne pas être pertinent. Leur critique peut être illustré par la classification que Borges[1]

[1] BORGES, Jorge Luis (1957).- "La langue analytique de John Wilkins". *Enquêtes*, 1937-1942. N.R.F., Gallimard. p. 44.

fait attribuer par le docteur Franz Kuhn à "certaine encyclopédie chinoise" intitulée "le marché céleste des connaissances bénévoles" où

> "les animaux se divisent en a) appartenant à l'Empereur, b) embaumés, c) apprivoisés, d) cochons de lait, e) sirènes, f) fabuleux, g) chiens en liberté, h) inclus dans la présente classification, i) qui s'agitent comme des fous, j) innombrables, k) dessinés avec un très fin pinceau de poils de chameau, l) et coetra, m) qui viennent de casser la cruche, n) qui de loin semblent des mouches".

Gardons nous de sourire : combien de questions fermées proposent des modalités à choix forcé qui ne comprennent pas une petite part d'arbitraire ? Combien de fichiers utilisés pour nos analyses ne sont pas restreints et biaisés par nos méthodes de recueil ?

Depuis longtemps, on espérait trouver des méthodes qui combinent les avantages des deux approches sans cumuler leurs limites. Il semble que de telles méthodes soient enfin en train de prendre corps, avec une nouvelle famille de techniques issues de l'analyse lexicale. L'analyse lexicale résoud avec élégance la question du choix des traits pertinents, et en décharge le statisticien. Pour caricaturer, celui-ci n'a plus qu'à rassembler le matériau à analyser sous forme de texte décrivant ce qui l'intéresse.

Le problème de la description

La statistique c’est l’art d’étudier les objets nombreux. Cette notion “d’objets nombreux” mérite qu’on s’y arrête un instant. Pour pouvoir qualifier des objets de “nombreux”, il faut qu’on puisse les compter comme une population. Là intervient la première opération de classement, qui considère tous les individus statistiques comme identiques en ce qu’ils sont dans une même classe, la population.
Mais ensuite, le travail du statisticien consiste à considérer cs mêmes individus comme différents, en examinant comment varient leurs traits ("variables", "modalités") dans la population. Tout l'art du statisticien descriptif réside dans cette dialectique subtile entre le même et le différent.
On met dans une même classe des objets analogues, on sépare les objets différents. Réciproquement ce sont précisément les ressemblances et les différences qui définissent les classes. Tout le problème vient de ce que c'est un procédé de type "poule et oeuf" : les classes sont définies comme agrégats d'individus ayant les mêmes traits, et les indvidus sont définis comme appartenant à une classe parce qu'ils ont les traits typiques de la classe. Bref, un individu est dans une classe parce qu'il a les traits typqiues d'un individu de la classe.Alors, comment définir les bons traits quand on ne connaît pas les classes ?

Ces ces traits ne sont pas toujours connus a priori, et si les traits typiques ou les variables sont mal choisis, la classification n'a pas de sens.

Le plus souvent, les traits sont des modalités de *variables*, que le statisticien choisit a priori. Mais rien ne garantit que ces variables sont choisies de la manière la plus pertinente, car un choix optimal suppose que l'on saurait *avant l'analyse* ce qu'il faut observer. On peut se demander, par exemple, ce qui est important à suivre pour comprendre l'évolution de l'indice Nikkei, ou les facteurs d'achat d'un nouveau modèle automobile, voire les raisons qu'ont les entrepreneurs français d'espérer une reprise... Si on veut faire un peu d'économétrie, encore faut-il disposer des facteurs qui expliquent effectivement la variance. Comment trouver les bons paramètres, pour ensuite mesurer leur importance relative ?
La solution que propose l'analyse lexicale n'est pas la meilleure, mais elle est efficace et "bestiale" : elle se propose tout simplement de décrire tout ce qu'on peut, et de faire le tri ensuite ("prenez les tous, Chi deux reconnaîtra les bien"). Comme c'est techniquement impossible avec des variables fermées, elle utilise une description ouverte.
Présentons rapidement cette famille de méthodes, en montrant d'abord comment elles permettent de répondre à la question du classement et du repérage des traits pertinents.

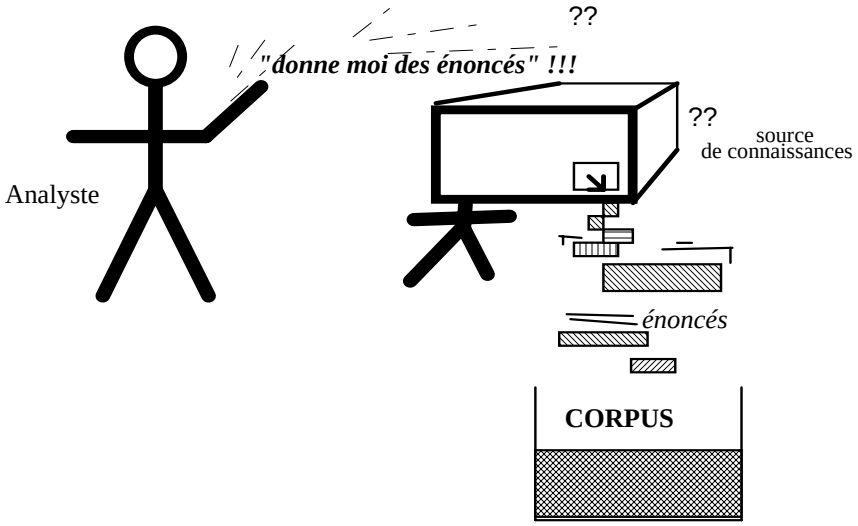
La description en langue naturelle

Nous avons vu les limites d'une description des objets par des variables a priori. Il existe par contre un système de description très large et polyvalent, la langue naturelle. La description d'un objet dans une langue se fait fatalement sous la forme d'une combinaison de mots. L'idée qui préside au traitement par des méthodes d'analyse lexicale est simple : c'est de comparer les objets entre eux sur la base des traits de leur description sous forme de combinaison de mots. Cela revient à considérer que chaque objet est décrit par un nombre non limité de variables, les mots. On pourra alors classer ensemble les objets qui se ressemblent en comparant les combinaisons. Ensuite, on pourra, dans un deuxième temps, regarder les traits typiques de chaque classe obtenue, et c'est ainsi que nous pourrons la caractériser avec les variables les plus pertinentes.

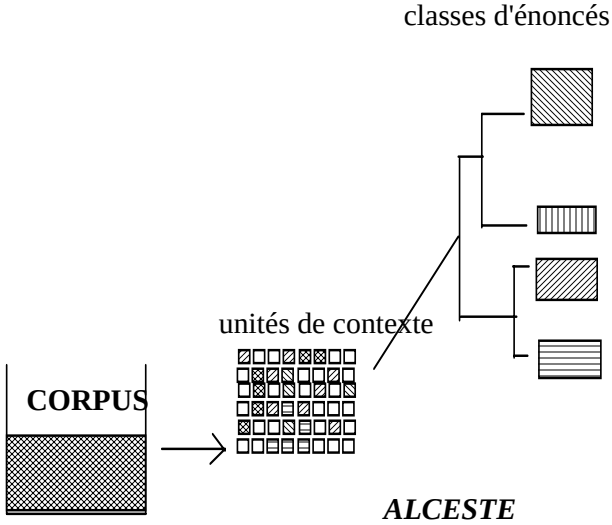
Ceci nécessite une technique qui puisse recueillir les descriptions d'objets (ou de représentations, d'opinions, de concepts...). Dans la pratique, nous aurons recours à une source de connaissance, sorte de boîte noire dont on sait qu'elle est capable de produire des descriptions en langue naturelle (sujet humain, dictionnaire...) et nous allons lui faire *énoncer* les termes associés à l'objet à analyser, en la stimulant pour qu'elle

produise des énoncés tout en gardant son attention sur ce terme central. Et l'on recueillera ces énoncés pour en constituer un corpus (c'est-à-dire l'ensemble de ces énoncés, stockés sous forme écrite).

Procédure de recueil des instanciations en langue naturelle à partir d'une source de connaissances



Ensuite, nous allons découper ce corpus en énoncés de taille à peu près égale, et nous allons classer ces énoncés sur la base de leur contenu lexical, par des méthodes issues de l'analyse des données appliquées à des matrices hypercreuses (c'est-à-dire des tableaux croisés où il y a surtout des cases vides). On obtient ainsi des classes d'énoncés analogues, qui repèrent en quelque sorte les grandes dimensions de l'objet étudié.



Cette méthode permet de repérer les grandes dimensions de l'objet sans faire d'hypothèses préalables. Elle ressemble fortement à l'analyse des données, dont elle est d'ailleurs techniquement dérivée. En fait, la statistique textuelle revient pratiquement à faire de l'analyse multivariée sur un nombre a priori indéterminé de variables.

Un exemple : le traitement de questions ouvertes

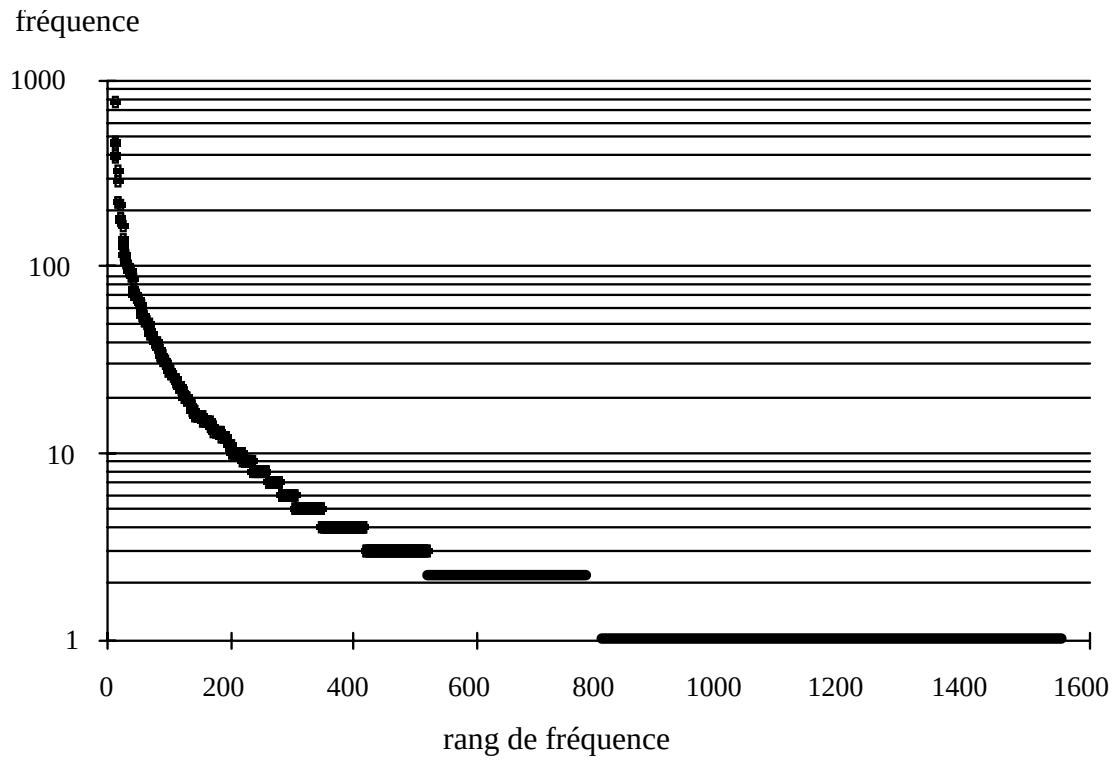
L'analyse lexicale a d'abord été développée à des fins stylistiques. C'est une école vivace mais qui nous concerne peu ici. Les applications qui intéressent plus le statisticien économiste sont celles qui utilisent les techniques d'analyse des donnée développées par Benzecri, et dont la branche la plus fertile fut le traitement des questions ouvertes dans les enquêtes d'opinion, sous l'égide de Ludovic Lebart, au Crédoc. Cette voie s'est largement développée depuis.

Prenons un exemple concret. Nous nous intéressions aux représentations de l'alimentation. Diverses questions fermées posées dans les enquêtes donnaient des résultats peu probants. Nous avons donc posé aux Français la question ouverte *"Si je vous dis "bien manger" à quoi pensez-vous ?"*. La question a été posée dans la vague de printemps (1990) de l'enquête périodique du Crédoc : Aspirations et Conditions de vie des Français. L'échantillon (2000 personnes) est représentatif de la population française adulte résidant sur le territoire métropolitain. La passation des questions se fait en face-à-face, au domicile de l'enquêté, par des enquêteurs professionnels. Les réponses des enquêtés ont été retranscrites directement par l'enquêteur lors de l'interview, puis saisies telles quelles pour un traitement informatique par analyse lexicale. Elles sont très variées, par exemple *: un repas de famille dans la joie ; apéritif, fruits de mer, gigot avec flageolets, plateau de fromages fraisiers, glace vins, champagne, café, pousse café ; regarder le nombre de calories qu'il faut et aussi les vitamines.*

La lecture de ces réponses ne donne pas facilement les grandes dimensions du problème. De fait, des post-codages effectués par des codeurs différents ne donnent pas de résultats concordants. Voyons maintenant l'analyse lexicale.
Les fichiers ont été retranscrits par des opératrices de saisie et reproduisent, aux erreurs de saisie près, ce qu'ont noté les enquêteurs sur les questionnaires. Le corpus a été traité à l'aide du logiciel ALCESTE développé par Max Reinert (voir encadré).

Un première analyse fréquentielle montre que certains mots sont très fréquents, tandis que la plupart sont rares. La courbe des fréquences "en baignoire" est assez classique de ce que l'on obtient dans des enquêtes de ce type. On notera que M*anger* et *bien* apparaissent très fréquemment, par un effet d'écho, les sujets ayant tendance à s'ancrer sur la question pour commencer leur réponse.

Courbe des fréquences des mots dans le corpus Bien_manger



Clé de lecture : chaque point représente un mot. En ordonnée figure la fréquence d'apparition du mot dans le corpus. Les mots ont été rangés de gauche à droite par ordre décroissant d'apparition. On voit ainsi qu'un seul mot apparaît plus de 500 fois, tandis que de très nombreux mots apparaissent avec une faible fréquence.

Près de 150 mots sont des hapax (ils n'apparaissent qu'une seule fois dans le corpus); environ 200 mots apparaissent plus de 10 fois. Voici le début de la table des fréquences : *manger (769) ; bon (716) repas (453) ; ; équilibré (407); et (380) ; pas (400) ; je (290) ; c'est (219) ; ne (218) ; bien (212) ; en (181) ; on (176) ; sain (166); faire (165); que (138) ; viande (134) ; restaurant (132) ; avec (130) ; légume (119) ; chose (118) ; pour (115) ; trop (113) ; se (104) ; faim (101).* Ce premier survol permet déjà d'avoir une première vue du contenu du corpus. C'est l'équivalent des "tris à plat" de la statistique descriptive classique.

On procède ensuite à une classification hiérarchique descendante des énoncés. On fait celle-ci sur la base des racines lexicales (lemmatisation) pour avoir des tableaux moins "creux". La lemmatisation consiste à réduire les verbes à l'infinitif, le pluriel au singulier etc. Pratiquement, on découpe le corpus en petits morceaux, les énoncés (en fait, des "U.C.E." ou Unités de Contexte Elémentaires). On compare ces UCE sur la base de leur contenu lexical, en opérant une classification descendante (voir encadré, mais il existe d'autres méthodes). Les classes obtenues peuvent alors être caractérisées

par les traits qui leurs sont typiques, c'est-à-dire qui, par construction, y sont plus fréquents que dans les autres classes.
La classification descendante construit d'un même mouvement les traits caractéristiques et les classes, car la probabilité que le lexème "a" puisse être un trait provient précisément de ce que "a" est observé sur un certain nombre d'énoncés (uce), mais pas sur tous. En d'autres termes, émergent comme traits classificatoires les traits *discriminants,* ceux qui sont effectivement susceptibles d'être productifs de classifications intéressantes. Ceux-ci permettent de caractériser un nombre suffisant d'expressions à classer, ils sont donc efficaces. Les traits trop répandus (non discriminants) ou les traits trop rares (hapax, traits aberrants) ne sont pas intéressants. C'est là une résolution empirique du problème du choix des traits pertinents. Les traits non discriminants sont rendus inopérants par leur répartition homogène dans les uce, ils constituent le "fond". Les traits qui pourraient être discriminants mais sont présents en trop petit nombre n'ont pas une fréquence assez élevée pour influencer l'analyse (contrairement à ce qui se passe dans les méthodes factorielles classiques où ils provoquent des artefacts et des instabilités) : ils sont également rendus inopérants par la classification descendante.

Voici un exemple simplifié. Il est clair intuitivement que les deux réponses :

`fruit légume viande ou poisson céréales laitages`

`viandes fruits légumes pain et un coup de rouge ! et du fromage`

seront proches l'une de l'autre, mais éloignées de :

`à la bonne cuisine à la maison avec des produits naturels`

qui sera elle-même plus proche de :

`une bonne bouffe à la française plats cuisinés  maison`

Le logiciel retrouve ces proximités en calculant une disance lexicale. Nous calculons les distances sur les réponses considérées comme "sacs de racines de mots". Les deux derniers énoncés deviennent :

`bon+ cuisin+ maison avec produit naturel+`

`bon+ bouffe+ français+ plat+ cuisin+ maison`

Dans ce cas particulier, les deux énoncés ont en commun 3 lexèmes (*bon+, cuisin+, maison*) sur les 6 que chacun contient. Compte tenu de la fréquence d'apparition des mots dans le corpus, la probabilité d'une telle cooccurrence est très faible, et la distance entre les noncés est donc très petite. On comprend bien que ces deux énoncés vont se retrouver dans la même classe, mécaniquement.
Mais examinons maintenant le cas suivant :

`repas copieux avec entrée plat_résistance et dessert`

se retrouve dans la même classe que :

des crudités de la viande et des légumes de la salade et un fruit

ce qui est sémantiquement satisfaisant, mais à première vue surprenant, puisque ces deux énoncés n'ont aucun mot plein (actif) en commun. Cela s'explique par la présence dans la même classe d'autres énoncés qui sont des formulations intermédiaires du même paradigme, ayant des mots communs avec ces deux formes extrêmes, comme :

à un bon *repas entrée plat_résistance* fromage **salade** *dessert* café

la **viande** les **légumes** *dessert* les **fruits** le pain le fromage

qui sont proches des deux formes extrêmes, et également proches entre elles par cooccurrence de ***fromage*** et ***dessert***.

La classification qui combine analogie et contraste, permet donc de rassembler des énoncés de même "sens". Voici les sept classes obtenues :

Un repas complet (15 à 19%)

Il s'agit du modèle traditionnel du repas structuré : *entrée - plat chaud - fromage - dessert*. Voici quelques extraits de réponses typiques : *... manger des légumes des fruits, poisson, fromages laitages et un peu de viande /... un bon repas : entrée plat de résistance fromage salade dessert café /... viandes fruits légumes, pain et un coup de rouge ! et du fromage...* Bien manger, c'est ici ingérer certaines catégories de produits en respectant un ordre formel particulier, celui du repas complet à la française.

Respecter la diététique

Les principes diététiques, largement diffusés par les médias et les spécialistes, apparaissent dans le discours des Français sous deux formes : l'une, positive, met l'accent sur la variété, l'équilibre, le naturel ; l'autre, négative, interdit les excès, surtout de graisse et de sucre.

varié et équilibré (16 à 20%)

Dans ce thème, bien manger c'est avant tout bien s'alimenter. On retrouve à la fois le discours diététique contemporain et la recherche du naturel. Il faut manger *sain, équilibré, varié, naturel*. Bien manger signifie ici incorporer les vertus de la nature en ingérant l'aliment aussi proche que possible de son état naturel. C'est ainsi que l'on refuse les *produits chimiques*, les *conserves*, les *surgelés*, les *colorants*. Le terme *frais*, qui apparaît souvent dans les réponses, a également le sens de "naturel". Quelques extraits de réponses de cette classe très stéréotypée : *... manger équilibré, sans produit chimique ni conservateur /... manger sainement, une nourriture variée et produits frais /... manger des produits sains non pasteurisés, naturels, des produits sains /... équilibre de l'alimentation, nourriture variée et modérée.../... l'équilibre alimentaire protéines lipides glucides, vitamines sels minéraux ...*

pas trop de graisse ni de sucre (11 à 13%)

Ce thème est construit sur l'interdiction et la restriction. *Bien manger* c'est ici proscrire la gourmandise en ce qui concerne le gras et le sucré, mais aussi les *graisses* et les *calories*. Quelques phrases type : *... pas trop gras pas trop de sucre et pas trop de calories /... pas manger trop lourd ... pas sucré car j'ai du diabète /... ne pas abuser des graisses entre autres / régime, cholestérol, trop de sucre /... manger en qualité et non pas en quantité ...*

Écouter son corps

Cette troisième dimension met en scène le mangeur face à la quantité et à la qualité de l'aliment. Là, c'est autant le corps du mangeur que son esprit qui nous répond : il s'agit de se remplir le ventre sans se goinfrer, ce qui rejoint un peu le thème précédent, mais aussi de manger ce qu'on aime : en somme satisfaire les besoins et les désirs, de son corps.

manger juste à sa faim (10 à 14%)

Ce thème contient principalement des prescriptions en termes de quantité. Bien manger, c'est ici : manger en quantité raisonnable, suffisamment pour satisfaire sa faim, mais pas plus. Il s'agit d'une application au domaine alimentaire de la philosophie du juste milieu. Celui-ci est perçu comme une sorte de normalité raisonnable, comme l'indique la fréquence élevée des racines : raisonnable(ment), normal(ement), correct(ement), et des expressions comme *manger à sa faim, ne pas se goinfrer.*

Quelques extraits typiques : *... manger à sa faim, ne pas manger trop riche /... se nourrir normalement pour vivre je parle pas des grandes bouffes /... me nourrir normalement, me nourrir en restant avec une petite faim /... il faut manger pour vivre et non pas vivre pour manger /... manger correctement à ma faim sans faire d'excès ...*

manger ce qu'on aime (14 à 17%)

Avec ce thème apparaissent des notions de désir, de plaisir, d'envie. Le discours n'est pas normatif comme dans les deux thèmes précédents, mais subjectif : les mots *je, moi, me, soi* apparaissent très fréquemment. Il s'agit là encore de réponses en termes d'aliments, mais ce qui compte c'est le plaisir qu'ils apportent au mangeur, comme l'indique la fréquence des mots : *envie, apprécier, plaire, plaisir, palais.* Quelques exemples : *... manger ce qui me plaît, ce que j'ai le plaisir à manger /... se régaler une fois que ça me plaît/...manger quelque-chose qu'on aime/...des spécialités des trucs qu'on ne mange pas souvent /... tout ce qui me fait envie : comme le pinard, le fromage...*

Le repas, lien social

La dernière grande dimension du bien-manger est sociale : familiale (maternelle surtout), amicale et festive, enfin formelle et gastronomique.

bons petits plats (4 à 6%)

Ce petit thème rassemble des réponses orientées autour de la cuisine. Les chaînes les plus fréquentes sont simples, sans équivoque : *bons petits plats, bon petit repas, bon plat bien préparé, bon repas avec bonne cuisine, plat cuisiné maison.* On retrouve dans les réponses des tournures caractéristiques de l'univers affectif et familial dans le français familier, en particulier le qualificatif affectueux *bon petit*. *Cuisine, plat, petit, français, traditionnel, consistant* et *familial* apparaissent avec des fréquences élevées et très significatives. Bien manger signifie ici consommer les produits culinaires domestiques. Quelques exemples : ... *aux petits plats à maman /... une table familiale gaie présence d'harmonie et de bonne cuisine /... une bonne bouffe à la française, plats cuisinés maison /... des bons petits plats, pâtisserie /... faire un bon repas avec maman, la bonne cuisine familiale de mon enfance /... les bons petits plats en famille avec des produits sains ...*

convivial (12 à 14%)

Ce thème est assez proche du précédent. Il s'agit de faire des bons repas en famille ou avec des amis, avec de bons petits plats, du bon vin, bref, *une bonne bouffe*. Les réponses se caractérisent par l'accent mis sur le repas non pas comme occasion de manger, mais de se réunir autour d'une table et de *faire la fête*. Le commensalisme et l'ambiance décontractée apparaissent explicitement ou en filigrane dans les réponses. *Repas, famille, table, amis, gueuleton* et *fête* apparaissent pour décrire un cadre général de convivialité bon enfant. La présence de mots comme *barbecue, vin, super, copains, ambiance* renforcent cet accent mis sur la socialité amicale. Quelques extraits typiques : *faire un super bon repas avec des amis /... repas de famille, bonne chère, bon vin, faire la fête /... bon gueuleton, bon vin /... un bon repas en famille, un bon repas d'anniversaire ou de fête de fin d'année ...* Ce thème est caractéristique des jeunes, habitants de villes moyennes, des hommes, des employés et des ouvriers.

le décorum (6 à 8%)

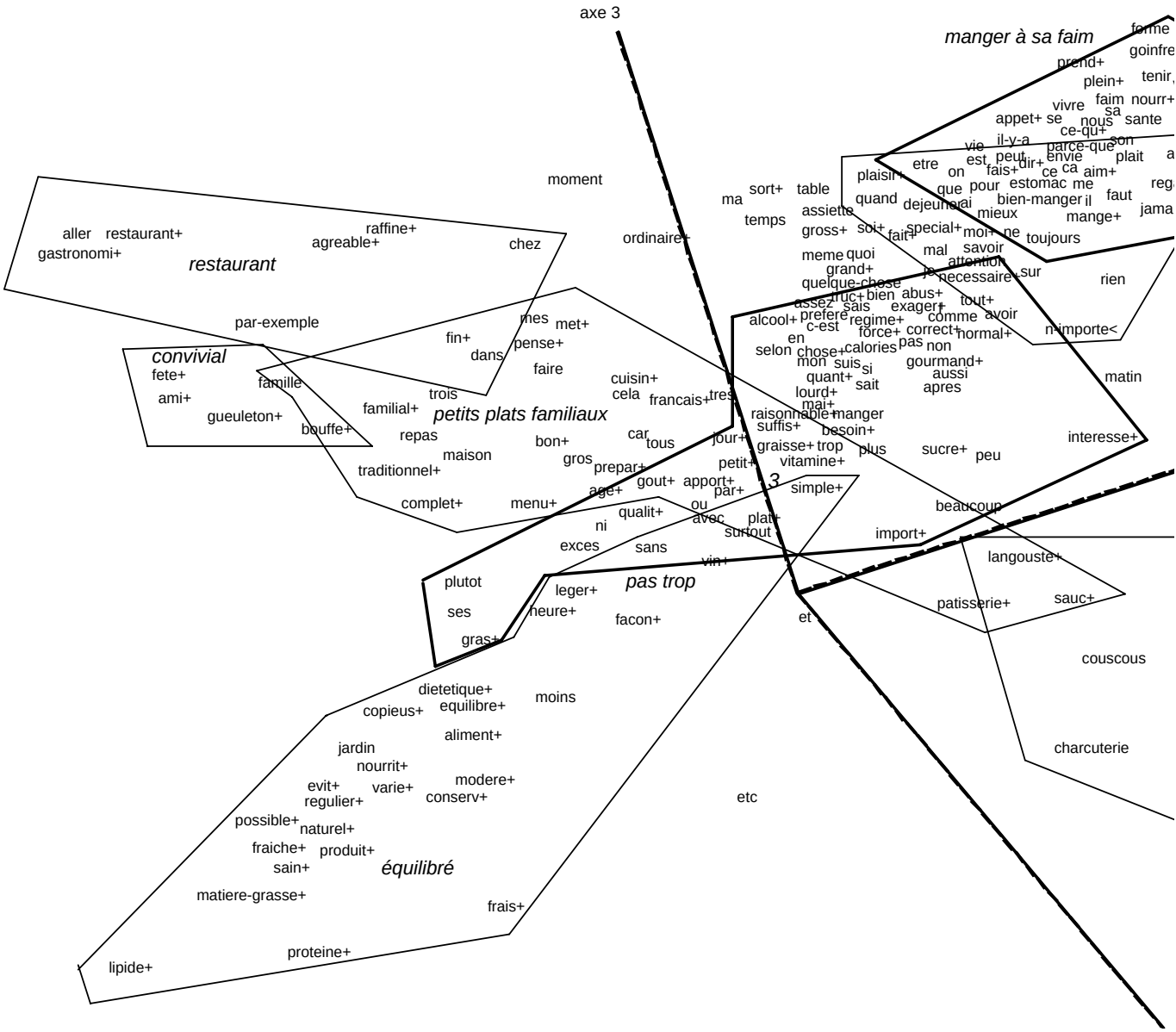
Au rôle social du repas qui apparaît dans les thèmes *petits plats* et *convivial*, ce thème ajoute la dimension gastronomique. On peut le résumer par "aller dans un bon restaurant". C'est principalement le côté gastronomique du restaurant, avec ce que cela implique de rituel et de décorum, qui ressort dans cette classe. C'est pourquoi les repas gastronomiques à domicile (banquet de fête, soupers fins...) se retrouvent dans la même classe. Quelques phrases typiques : ... *restaurant gastronomique/... c'est me payer le*

restau ou un banquet de famille le week-end/... aller dans un beau restaurant avec des menus raffinés, du foie gras/... faire une sortie au restaurant avec ma femme ... Dans plusieurs réponses apparaît le caractère exceptionnel de ce type de repas. Ce thème est fréquent chez les jeunes, les habitants de grandes villes, les hommes, les concubins, des diplômés de l'université ou des grandes écoles, les artisans, commerçants et chefs d'entreprise et les cadres supérieurs ou professions intellectuelles.

Représentation graphique

L'idée est de construire un espace des connotations reproduisant les proximités entre mots et dans lequel deux mots seront d'autant plus proches qu'ils sont associés dans les réponses des individus. On l'obtient par une analyse factorielle sur la base du tableau classes x mots. Une vue en perspective où nous avons entouré les classes permet de donner une vision d'ensemble.

L'espace de connotations du Bien_manger *: projection des traits lexicaux*



L'intérêt d'une telle représentation est qu'elle porte ici sur une vison réellement subjective du problème alimentaire. On peut imaginer qu'en situation de manger, c'est par rapport à une telle structure que le sujet doit prendre position. En particulier, on peut comprendre qu'avant de passer au pôle des objets (***Entrée plat chaud fromage dessert***) qui est un peu le pôle final obligé qui guide le déroulement d'un repas structuré, le sujet va se trouver pris dans des dilemmes entre les pôles diététique (***Pas trop...***, ***Equilibre***), social (***Petits plats***, ***Convivial***, ***Restaurant***), et son désir (***Manger à sa faim***, ***Manger ce qu'on aime***). Par ailleurs, chacun de ces pôles contient des schèmes qui sont parfois alternatifs (***Restaurant*** ou ***Convivial*** ? ; ***Manger à sa faim*** ou ***ce qu'on aime*** ?).

Les avantages de la méthode

On voit que la méthode permet d'abord d'obtenir l'équivalent d'une post codification automatique de la question. Les classes obtenues sont bien meilleures que celles obtenues par post codage manuel, comme l'ont montré plusieurs comparaisons systématiques. Un autre intérêt est que la méthode n'est pas sensible à l'arbitraire du codeur : deux analyses par des opérateurs différents fourniront les mêmes classes.

Ensuite, la méthode permet d'obtenir une visualisation des "champs de connotation" lexicaux. On imagine l'intérêt de ce type de visualisation, notamment en recherche exploratoire à des fins marketing ou publicitaires.
Par ailleurs, la variable ouverte peut évidemment être croisée avec les variables fermées. Il est possible de caractériser les différents types de répondants ; par exemple, les mots *bon, foie gras, vin, français, frites, steak, restaurant, charcuterie, qualité*, ou encore *banquet, pomme de terre, choucroute, sauce, charcuterie, copains, canard* et *boeuf* sont significativement caractéristiques des réponses masculines. Les réponses des femmes sont caractérisées par l'apparition plus fréquente de *équilibré*, *légumes, sain, laitages, vitamines*, et encore de *kilos, lait, varié*... Les mots *bouffe, copains, restaurant, couscous, dessert, gâteau* apparaissent plus fréquemment dans le discours des plus jeunes, tandis que chez les plus âgés les formes *peu, sans, excès, modérément, raisonnable,* ou *cholestérol* sont plus typiques.

Les perspectives ouvertes par l'analyse lexicale pour le traitement des questions ouvertes sont excitantes : d'abord, en raison de la meilleure qualité du traitement obtenu. Ensuite, parce que la problématique du traitement n'a pas forcément à être définie ex ante. Ces techniques restent coûteuses en raison du coût de transcription des questions ouvertes. Elles ne donneront sans doute leur pleine mesure pour les enquêtes que lorsque l'on disposera de méthodes efficaces de transcription automatique (reconnaissance de la parole), dans quatre ou cinq ans.

L'analyse des bases de données en texte intégral

Si l'utilisation de l'analyse lexicale pour le traitement des questions ouvertes représente un progrès majeur, ce n'est sans doute pas là que leur utilisation va le plus bouleverser le métier du statisticien, mais par le fait que l'analyse lexicale va ouvrir aux méthodes de traitement statistique la masse considérable de données stockées sous forme de texte.
En effet, la diffusion de l'informatique, et en particulier du traitement de texte, a commencé à produire une masse impressionnante de données disponibles sous forme de fichiers texte. Même les textes anciens commencent à se transformer en fichiers, par océrisation : par exemple, la nouvelle Bibliothèque Nationale (TGB) a commencé à scanner son fonds documentaire. Cette situation est une malédiction pour le commun

des mortels, car il est bien connu que trop d'information tue l'information. Nous sommes tous submergés par l'information. Mais c'est une bénédiction pour le statisticien, dont le problème principal jusqu'ici avait été l'obtention des données. De cette masse monstrueuse, les techniques d'analyse lexicale vont permettre d'extraire la substantifique moelle, et cela d'autant mieux que les données sont nombreuses. Et, de fait, les techniques d'analyse lexicale, parce qu'elles sont statistiques, donnent de meilleurs résultats sur de larges corpus.
L'idée générale est de faire de l'analyse de contenu en masse, afin de repérer les grands problèmes, les points saillants, les évolutions.

Un exemple : l'analyse du dictionnaire

Nous allons illustrer la puissance de la technique avec l'analyse d'une source simple, le dictionnaire. On restera sur l'exemple du "manger", qui a le mérite d'être simple. Supposons que nous ne sachions pas ce que "manger" veut dire, ce qui est d'ailleurs plus ou moins le cas. Plus exactement, "manger" a trop de significations. Le seul article de définition du Grand Robert fait cinq pages, et recense nombre de sous-sens. La bibliographie sur le sujet recèle des dizaines de milliers d'ouvrages et d'articles.
Prenons une source structurée, sorte de base de données linguistique, et consultons la sur le sujet. Nous avons constitué un corpus de 150 pages environ (300 koctets), composé de la mise bout à bout des définitions de "manger" et des définitions complètes des 144 synonymes, analogues et dérivés que propose le dictionnaire (le Robert fournit une liste de ces synonymes en fin de définition).

Le corpus a été découpé en quelques milliers de fragments correspondant à peu près à des phrases (les phrases longues sont coupées en morceaux). Environ 1000 racines ont été retenus comme variables d'analyse.

Les mots les plus fréquents (plus de cent occurrences) sont :

```
prendre. (662 occurrences) ; se (646) ; en (450) ; manger (443) ;
faire. (258) ; qqn (243) ; table (203) ; alimentation (190) ;
bouche (188) ; repas (166) ; nourrir (158) ; toucher (153) ;
mettre. (120) ; attaquer (114) ;nourriture (107) ; qqch (107) ;
gueule (104) ; avaler (102) ; goût (99).
```

On retrouve sans surprise un effet d'écho à la question avec "*manger*" qui apparaît avec une fréquence très élevée. Les autres mots pleins (hors articles, prépositions etc.) les plus fréquents sont *prendre, , quelqu'un, table, aliment, bouche, repas, nourrir, toucher, attaquer, nourriture, quelque chose, gueule, avaler* et *goût*.

L'analyse livre quatre grandes classes, dominées par une grosse classe qui regroupe près de la moitié du corpus.

Classe 1

Cette classe est de loin la plus grosse. Ses traits caractéristiques sont essentiellement verbaux.

> prendre ; toucher; attaquer ; qqn ; qqch ; entamer ; main ; sujet ; consumer ; ronger ; contact ; arme ; forte ; fondre (sur). ; croquer ; atteindre ; adversaire ; ennemi ; attraper ; feu ; couper ; détruire ; tirer ; pince ; agir ; combattre.

C'est d'autant plus remarquable que le logiciel est "sourd" aux catégories syntaxiques lorsqu'il compose les classes. Cette classe d'appropriation est chargée de connotations violentes, agressives, qui rappellent notre état primitif de chasseur cueilleur. On peut l'appeler la classe PRENDRE, d'après son trait le plus saillant.

Classe 2

Les traits typiques de cette classe sont essentiellement des substances alimentaires, ou des catégories de telles substances.

> aliment ; nourrir ; pain ; nourriture ; régime ; vivre ; jeûne ; subsistance ; végétal ; nécessaire ; fournir ; privation ; lait ; diététique ; diète ; élément ; viande ; nutritif ; sein ; sustenter ; amer, sucre ; enfant ; gâteau ; sève ; affamer ; légume ; liquide ; manquer ; produit ; eau ; boisson ; frais ; animal ; fruit ; maigre ; chair ; idée ; dieu, oeuf.

On peut appeler cette classe NOURRITURES. Il s'agit clairement d'une classe de substances alimentaires.

Classe 3

Les traits typiques sont surtout des substantifs.

> repas ; table ; dîner ; service ; vaisselle ; soupe ; buffet ; invit(é, er) ; servir ; déjeuner ; plat ; restaurant ; dessert ; hôte ; convive ; cantine ; collation ; couvert ; hôtel ; ensemble ; ustensiles ; assiette ; laver ; petit-déjeuner ; régaler ; région ; soir ; heure ; ordonner ; salle.

Cette classe contient des éléments du repas autres que les aliments : la table, les couverts, les convives, et des précisions situationnelles (horaire, occasion). Elle est centrée sur les aspects sociaux, rituels, et instrumentaux de la prise. Elle fournit des *compléments circonstanciels* de la prise alimentaire. On appellera cette classe REPAS.

Classe 4

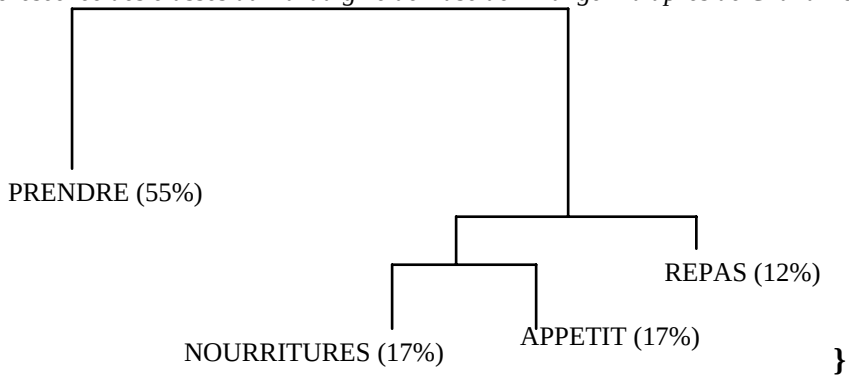
Voici quelques traits typiques.

```
glouton ;  appétit ;  manger ;  gourmand ;  goinfre ;  goulu ;
rassasier ;  empiffrer ;  vorace ;  excès ;  ogre ;  avide ;
assouvir ;  bâfrer ;  faim ;  appétit ;  aspirer ;  carnassier ;
yeux ; avidité ; contenter ; loup ; désir.
```

Il s'agit à l'évidence de mots caractérisant la faim. Celle-ci est cependant présente avec une connotation d'avidité, de voracité, qui caractérise un désir intense. Cette violence, un peu surprenante compte tenu de l'origine du corpus, n'est pas un artefact. On appellera cette classe APPETIT.

Voici comment s'organise l'agrégation de ces classes.

Arborescence des classes du Paradigme de Base de "Manger" d'après de Grand Robert



}

Le noyau définitoire nous livre donc une structure proche du modèle linguistique sujet/verbe/objet/complément.

Le Paradigme de Base du "Manger" d'après le Grand Robert

(verbe)
PRENDRE

(sujet)
APPETIT

(objet)
NOURRITURES

REPAS
(complément circonstanciel)

Ces éléments forment le coeur de la représentation du "manger". Les noyaux sémantiques qui le composent sont remarquablement stables sur le plan statistique. La représentation du manger s'organise donc autour de quatre noyaux de base : le *désir* (faim, appétit, envie...), la *prise* (prendre, attaquer, attraper...), la *substance nutritive* (aliment, pain, viande,..), et le contexte social et instrumental (repas, convives, ustensiles).

Le résultat le plus intéressant et novateur sur le plan technique est que l'on a obtenu la construction d'un sens. En effet, le sens n'est, finalement, que l'association entre des objets. Et l'analyse nous livre, "brut de décoffrage", que *manger*, c'est l'articulation entre les noyaux APPETIT/ PRENDRE/ NOURRITURE/ REPAS. Celle-ci s'articule dans un cadre pragmatique reliant sujet (ou, plus exactement, pulsion), objet, opération, et modalités.

Ceci peut paraître trivial, si nous le savions déjà. Tout l'intérêt réside dans le fait que cette articulation n'est pas une interprétation, ni une définition, mais le *résultat empirique* d'une méthode aveugle et sourde au sens, qui est rigoureuse et reproductible. Que l'on retombe ici sur une évidence est encourageant, et nous incite à penser que, appliquée à des corpus dont le contenu n'est pas connu au préalable, la méthode nous livrera les grandes dimensions du sens sous forme d'une articulation des noyaux de base.

Notons par ailleurs que les résultats ne sont absolument pas triviaux pour les spécialistes du domaine. On retrouve notamment des éléments très intéressants sur le plan de la recherche, comme la présence d'agressivité dans le noyau "prendre", qui confirme des intuitions anciennes de Freud sur le processus "d'incorporation" de la phase orale qui n'avaient pas jusqu'ici reçu beaucoup de confirmations expérimentales :

> "En fait trois significations sont bien présentes dans l'incorporation : se donner un plaisir en faisant pénétrer un objet en soi ; détruire cet objet ; s'assimiler les qualités de cet objet en le conservant en dedans de soi. C'est ce dernier aspect qui fait de l'incorporation la matrice de l'introjection et de l'identification. "(Laplanche **Erreur ! Signet non défini.**et Pontalis**Erreur ! Signet non défini.**, dans l'article "incorporation" du Vocabulaire de Psychanalyse, 1967, 1990 p. 200).

Ce que nous avons vu sur l'exemple du manger peut être réalisé à partir de n'importe quelle base de données en texte intégral. Nous avons par exemple aussi travaillé sur des revues de presse scannérisées (quelques centaine d'articles), sur des bases de données décrivant des objets... Quelques exemples d'applications (classification des petits-déjeuners des français, analyse du discours des firmes agro-alimentaires en matière de diététique, perception de l'effet de serre par les Français, analyse des jugements rendus par les différentes juridictions, classification de produits décrits dans des bases de données relationnelles, évolution sur 70 ans du référentiel de valeurs d'un grand établissement financier...), sont décrits dans l'annexe au cahier de recherche du Crédoc n°48 (voir bibliographie).

Conclusion

L'analyse lexicale, très prometteuse, n'en est qu'à ses premières applications au niveau industriel, même si de grandes entreprises, notamment EDF, commencent à l'utiliser pour leurs études internes. Chacun imaginera les applications qu'il peut en faire dans son domaine particulier.

Mais il serait dommage que la technique, d'usage apparemment facile, soit dévoyée par des analystes incompétents, et que l'on en arrive, comme cela s'est passé pour l'analyse des données, à une situation où, à la suite d'abus, beaucoup sont tentés de jeter le bébé avec l'eau du bain. Raison de plus pour que les statisticiens compétents s'y mettent, avant que d'autres aient commis des dégâts irréparables.
J'encourage donc ceux qui sont intéressés par ces techniques à prendre contact avec moi, dans le cadre du groupe "Analyse lexicale" mis en place par l'ASTEC qui se réunira périodiquement en invitant les utilisateurs en pointe et les développeurs de logiciels. Les participants à ce groupe seront mis directement en contact avec les développeurs de ces techniques. Envoyez votre carte de visite à *Variances*, qui transmettra.

ENCADRE la méthodologie ALCESTE[2].

Les analyses lexicales effectuées par ALCESTE se déroulent en trois parties.

Étape A : préparation et codification du texte initial.

Le texte est découpé en énoncés (unités de contexte initiales "UCI"). Ces unités de contexte sont de l'ordre de quelques lignes. Ces UCI sont ensuite découpées en "UCE" (unités de contexte élémentaire), de deux ou trois lignes. Pour effectuer cette découpe, ALCESTE tient compte, autant que faire se peut, de l'importance relative des marqueurs syntaxiques.
Le texte est débarrassé de tous les accents et de tous les traits d'union éventuels. ALCESTE possède un dictionnaire des locutions qui lui permet de reconnaître les "mots composés" : *à peu près, au contraire, de temps en temps...* Ceux-ci sont ensuite considérés comme étant un seul mot. Ce dictionnaire est nécessairement incomplet mais l'utilisateur a la possibilité de le modifier et d'y introduire des formes nouvelles. On peut choisir également de signaler les locutions en introduisant directement dans le texte des "_", par exemple, pain_au_chocolat.
Il est nécessaire ensuite de choisir les mots qui seront utilisés dans l'analyse ("mots pleins"). Pour ce faire, ALCESTE procède en deux étapes : premièrement, une lemmatisation, deuxièmement une reconnaissance de certains mots-outils.
Pour effectuer la lemmatisation un algorithme reconnaît, à l'aide d'un dictionnaire des racines, les mots-outils et les racines des principaux verbes irréguliers pour les réduire à leur forme infinitive. Les formes non reconnues sont ensuite traitées grâce à un

[2] Analyse Lexicale par Contexte Etabli à partir d'une Segmentation du Texte en Enoncés. Cette méthodologie a été mise au point par Max Reinert (CNRS Toulouse).

algorithme particulier en se rapportant aux mots déjà lemmatisés dans le corpus et à un dictionnaire des suffixes. On a déjà vu les avantages et les inconvénients théoriques de la lemmatisation. Le choix de la lemmatisation est ici nécessaire puisque ALCESTE effectue ses analyses essentiellement à partir des fréquences et cooccurrences des mots.
Les mots pleins sont définis par opposition aux mots outils qu'ALCESTE reconnaît grâce à un dictionnaire modifiable par l'utilisateur. Sont considérés comme mots outils les articles, les pronoms, les auxiliaires *être* et *avoir*, les chiffres et certaines locutions. Les mots pleins sont regroupés dans un fichier DICO, que l'on peut modifier pour réparer d'éventuelles erreurs de lemmatisation.
Seuls les mots pleins dont la fréquence dans le texte (ou "corpus") est supérieure à trois, sont actifs. Toutes les autres formes sont des variables supplémentaires. L'utilisateur peut imposer que certaines formes de fréquence supérieure à trois soient considérées comme variables supplémentaires, en modifiant le fichier DICB qui regroupe toutes les formes lemmatisées ainsi que leur nature. Ceci peut être particulièrement utile dans le cas d'analyses de réponses à des questions ouvertes, afin d'éviter les phénomènes d'écholalie (répétition des termes de la question). On peut inversement rendre actif un mot plein de fréquence inférieure à trois, en le rattachant à un mot lemmatisé actif.

Étape B : classification descendante hiérarchique

On obtient à la fin de l'étape A un dictionnaire des formes contenues dans le texte qui permet de construire un tableau disjonctif à double entrée croisant en ligne les unités de contexte, et en colonne les formes réduites. Les cases de ce tableau valent 1 ou 0, selon que la forme apparaît dans l'UCE ou non. En règle générale, ce tableau est grand et très creux. ALCESTE utilise une méthode de classification descendante hiérarchique qui est particulièrement adaptée pour traiter des tableaux creux.
La première classe analysée contient toutes les unités de contexte. A chaque étape on cherche la meilleure bipartition de la plus grande des classes restantes. L'algorithme permet de conserver 12 classes. Le logiciel procède en trois étapes :
- il cherche le premier facteur de l'AFC du tableau creux,
- il cherche l'hyperplan perpendiculaire au premier axe maximisant l'inertie interclasses des deux nuages,
- il choisit la plus grande des classes restantes, puis réitère.

ALCESTE ne classe pas toutes les UCE : il crée une classe résiduelle d'UCE qui n'appartiennent à aucune classe.
Le choix de la classification descendante entraîne certains problèmes, en particulier la création de classes de petit effectif. Ces classes sont souvent créées par la présence de mots de faible fréquence (cependant supérieure à trois). Elles sont séparées très tôt du reste du corpus de sorte que l'on ne peut pas les éliminer en jouant sur le nombre de classes retenues (inférieur à 12). Mettre en variables supplémentaires les mots à

l'origine de la création de ces classes parasites n'est pas très satisfaisant. En effet, souvent d'autres mots émergent alors et renouvellent le phénomène.

Étape C : aide à l'interprétation des classes

Cette étape crée un certain nombre de fichiers parmi lesquels les plus utiles sont :

- la liste du vocabulaire spécifique de chaque classe (profil). Sont retenus les mots lemmatisés satisfaisant un critère d'appartenance à la classe (chi^2 à un degré de liberté supérieur à 2,7). Pour chaque forme, sont fournis : son ordre d'apparition dans le dictionnaire d'ALCESTE, le nombre d'UCI de la classe et de l'ensemble du corpus qui la contiennent, son chi^2 d'appartenance à la classe, son code d'identification.
- la liste du vocabulaire non spécifique pour chaque classe (antiprofil).
- la liste des réponses les plus caractéristiques pour chaque classe. Elles sont classées par ordre de chi^2 d'association à la classe décroissant.
- l'arbre d'agrégation qui montre dans quel ordre les classes se séparent.

ENCADRE : qui fait de l'analyse lexicale en France ?

Les utilisateurs sont de plus en plus nombreux. D'abord les laboratoires de recherche, qui utilisent les programmes existants, et sont quelques dizaines. On en trouvera notamment dans les actes des 2èmes journées internationales d'analyse statistique des données textuelles, et dans ceux du colloque international Consensus ex Machina, qui ont réuni la plupart des chercheurs du domaine (voir bibliographie). Ensuite, les instituts d'étude : ils sont plus discrets car ces technique font de plus de plus partie de leur fonds de commerce. Rares sont les documents publics disponibles, en dehors de ceux du Crédoc (voir bibliographie). Enfin, certaines grandes entreprises qui disposent de structures de recherche et d'études interne, dont la plus avancée est actuellement EDF, avec son laboratoire de recherches en sciences humaines, le GRETS, qui utilise ces techniques pour des études internes (et confidentielles) en marketing, en communication, et en ressources humaines.

Bibliographie :

Actes des secondes journées internationales d'analyse statistique des données textuelles. Montpellier, 21-22 octobre 1993. Paris, ENST.

BEAUDOUIN, Valérie, LAHLOU, Saadi (1993). L'analyse lexicale, outil d'exploration des représentations. Réflexions illustrées par une quinzaine d'analyses de corpus d'origines très

diverses. Cahiers de recherche du Crédoc, n°48, septembre 1993. (demander aussi les annexes avec plein d'exemples !)

BENZECRI, Jean-Pierre, LEBART, Ludovic, REINERT, Max (1981).- *Pratique de l'analyse des données, Linguistique et lexicologie.* - Paris, Dunod.

Colloque international Consensus ex Machina. Paris, Sorbonne, 20-23 avril 1994. ALLC-ACH-Inalf/CNRS-ENS Saint Cloud.

LAHLOU, Saadi (1992 a).- *SI/ALORS : "BIEN MANGER" ? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus*.- Paris, CREDOC, Cahier de recherche du CREDOC, n°34.

LEBART, Ludovic, SALEM, André (1988).- *Analyse statistique des données textuelles*. Préface de Christian BAUDELOT.- Paris, Dunod.

REINERT, Max (1990).- "ALCESTE, une méthode d'analyse des données textuelles. Application au texte "Aurélia de Gérard de Nerval" *Bulletin de Méthodologie Sociologique*, 26.- pp. 25-54.

YVON, François (1990).- *L'analyse lexicale appliquée à des données d'enquête : état des lieux*. - Paris, CREDOC, Cahier de recherche, n°5.