JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

MATHIEU BRUGIDOU

La combinaison des inférences statistiques, linguistiques et sociologiques dans l'analyse d'une question ouverte

Journal de la société française de statistique, tome 142, n° 4 (2001), p. 105-120

<http://www.numdam.org/item?id=JSFS_2001__142_4_105_0>

LA COMBINAISON DES INFÉRENCES STATISTIQUES, LINGUISTIQUES ET SOCIOLOGIQUES DANS L'ANALYSE D'UNE QUESTION OUVERTE

Mathieu BRUGIDOU *

RÉSUMÉ

L'analyse textuelle présente l'avantage de dissocier assez clairement deux moments de l'analyse des questions ouvertes : d'abord la formalisation des données et leur « traitement » puis l'analyse « sociologique » des résultats. L'interprétation linguistique constitue un moment intermédiaire et est laissée dans l'ombre soit parce qu'elle paraît intuitivement naturelle, soit encore parce que les concepts linguistiques disponibles pour argumenter ces inférences linguistiques sont peu ou mal connus. Cet article propose un protocole permettant d'objectiver en partie ce processus interprétatif en tenant compte des trois moments qui le constituent. Pour ce faire, on s'appuiera sur l'analyse partielle d'une question ouverte posée à un échantillon d'agents d'EDF.

Mots clés : Analyse textuelle, question ouverte, sémantique interprétative.

ABSTRACT

The main strength of computerised textual analysis is to clearly dissociate two moments of the analysis of open-ended questions : first, the formalisation of data, secondly the sociological analysis of results. Most frequently, the users do not take into account the linguistic interpretation because this interpretation seems natural or because linguistic concepts available are little or badly known. The purpose of this paper is to highlight and to formalise the process of interpretation in taking into account these three moments. In that aim, we consider the preliminary analysis of one open-ended question taken from a survey addressing a sample of EDF collaborators.

Keywords : Textual analysis, open-ended question, interpretative semantics.

1. Introduction

L'analyse d'une question ouverte mobilise des compétences multiples. Pour interpréter le « sens » des réponses à une question ouverte, l'analyste, ainsi que

* GRETS/EDF R&D et CIDSP, 1 avenue du Général de Gaulle, 92141 Clamart cedex. E-mail : mathieu.brugidou@edf.fr

Maurice Tournier (Tournier, 1980) l'a naguère montré à propos de l'analyse textuelle, réalise de manière plus ou moins consciente une série d'inférences de nature différente. Grossièrement, trois types de compétences sont mobilisés : statistique, linguistique et sociologique. Classiquement, on définit l'attitude scientifique par la capacité à contrôler les effets de ces différents types d'inférences : il est en effet important d'élucider les « intuitions » sur lesquelles reposent notre saisie ordinaire du sens et d'identifier le type de compétences qui les fondent. Après avoir reconnu la nature statistique, linguistique ou sociologique de notre jugement (qui peut éventuellement mêler ces différentes instances), il devient possible de s'assurer de sa qualité et de contrôler sa validité. L'analyse textuelle constitue une avancée importante dans l'étude des questions ouvertes dans la mesure où elle propose une objectivation partielle de l'analyse. Elle formalise en effet des « données textuelles » grâce aux apports de la linguistique et propose des « classements [1] » de ces données par des traitements statistiques. La statistique définit des règles d'interprétations de ces résultats mais évidemment ne dit rien des inférences qui seront tirées sur le « sens ».

L'analyse textuelle présente ainsi l'avantage de dissocier assez clairement deux moments de l'analyse : d'abord la formalisation des données et leur « traitement » puis l'analyse « sociologique » de ces résultats. Toutefois, on sait que ces deux moments sont dans les faits largement récursifs et ne sont qu'idéalement dissociés (Lahlou, 1995) – en particulier les hypothèses sociologiques sont centrales dans le recueil des données. Par ailleurs, ce deuxième moment n'est que partiellement contrôlé par des hypothèses sociologiques : l'interprétation linguistique constitue un moment intermédiaire et est laissée dans l'ombre soit parce qu'elle paraît intuitivement naturelle, soit encore parce que les concepts linguistiques disponibles pour argumenter ces inférences linguistiques sont peu ou mal connus.

Cet article se propose, d'une part, de montrer la complexité des jugements portés dans ce type d'analyse, qui imbrique des sources de connaissances hétérogènes, et d'autre part, d'esquisser un protocole permettant d'objectiver en partie ce processus interprétatif en tenant compte des trois moments qui le constituent. Pour ce faire, on s'appuiera sur l'analyse partielle d'une question ouverte posée à un échantillon d'agents d'EDF [2].

1. Ces classements dépendent des hypothèses de l'analyste : par exemple classification exploratoire ou tris croisés en fonction de variables sociologiques.

2. Il s'agit d'une question sur l'avenir du nucléaire qui a été posée dans le cadre d'une enquête portant sur la perception des thèmes environnementaux. L'enquête a été réalisée par téléphone au domicile des agents du 12 février 2001 au 3 mars 2001, auprès d'un échantillon représentatif de 1653 agents d'EDF, habitant en France métropolitaine.

2. Inférences statistiques

Je ne discuterai pas ici de manière approfondie de l'analyse statistique qui a été abordée à plusieurs reprises dans ce volume. Je me contenterai de rappeler le contexte de cette question ouverte et certaines caractéristiques propres au protocole Alceste [3].

2.1. De la méthode d'enquête aux données

2.1.1. Les sondages d'opinion

Evoquer les statistiques, c'est nécessairement, ainsi que Xavier Marc (2001) et Dominique Labbé (2001) l'ont chacun montré, s'assurer des protocoles de recueil des données et de leurs formalisations.

Mais, avant même ces considérations, il faut dire quelques mots sur le recueil des données dans les enquêtes d'opinion. On est ici de plain-pied dans des considérations sociologiques : on sait que les réponses à des questions d'opinion, et a fortiori des questions ouvertes, sont d'autant plus fiables que les thèmes abordés ont un « sens » pour les personnes interrogées. Par là, on veut signifier que l'opinion recueillie n'est pas formée de manière artefactuelle par effet d'imposition mais qu'elle préexiste à l'enquête : le thème abordé doit être connu et intéresser les personnes interrogées. La question traitée satisfait à ce double réquisit puisqu'elle aborde l'avenir du nucléaire dans le cadre d'une enquête d'opinion sur les perceptions de l'environnement [4]. Ce thème est familier aux agents d'EDF qui sont sensibilisés par leur métier et l'information interne sur ce sujet.

Le libellé de la question traitée s'insère dans la séquence suivante :

– *Pensez-vous que, d'ici vingt ans, on aura pris la décision d'arrêter les centrales nucléaires en France ? (oui, non, je ne sais pas)*

– *Pourquoi ? (question ouverte, réponse saisie in extenso)*

On recense moins de 2 % de non réponse à cette question, ce qui pour une question ouverte est particulièrement faible. Ce taux confirme l'hypothèse – à vrai dire assez peu risquée – d'un fort intérêt des personnes interrogées pour la question. Par ailleurs, on remarquera qu'un premier chaînage entraîne une préstructuration des réponses : les personnes interrogées doivent produire une réponse argumentée qui justifie (« *Pourquoi ?* ») une première réponse positive, négative ou dubitative.

Ces deux caractéristiques, forte structuration sociologique de l'opinion interne à EDF sur la question du nucléaire et préstructuration par le questionnaire, expliquent sans doute pour une large part la qualité de la classification obtenue par le logiciel Alceste.

---

3. Pour un exposé plus complet sur le logiciel Alceste voir (Reinert, 1994) et (Blot et Le Roux 1992).

4. Le cadre interne de l'enquête ainsi que le contexte des autres questions sont bien sûr non négligeables pour apprécier la portée et le sens des réponses.

2.1.2. Le choix du logiciel et du protocole d'analyse

Le choix du logiciel n'est pas sans conséquence sur le protocole d'analyse, il implique des options méthodologiques fortes notamment dans le cas d'Alceste [5].

Cette méthode privilégie en effet une approche statistique qui identifie dans un corpus de textes donné des sous-ensembles homogènes de verbatims sur la base de leur profil lexical. Elle permet notamment de dégager la structure des textes à partir de proximités lexicales sans identifier *a priori* un point de vue (partition du corpus par locuteur etc.) qu'il s'agirait de caractériser. En revanche, il est possible de rendre compte des regroupements obtenus en projetant les « points de vue » (locuteurs, années etc.) comme des variables illustratives.

Pour cela, les textes sont notamment découpés en « unités de contexte », qui correspondent grossièrement à la phrase considérée comme l'unité sémantique de base. Ce sont des tableaux croisant ces propositions et les mots [6] qui sont l'objet d'un traitement statistique [7] pour constituer des classes au contenu lexical le plus homogène possible. Cette méthode présente certaines limites :

– le regroupement massif et parfois approximatif [8] des formes sous leur racine implique que les variations sémantiques ne sont saisies qu'à très gros traits,

– le découpage des unités de contexte, bien qu'il respecte les ponctuations fortes, est discutable,

– une double classification permet de ne retenir que les verbatims qui constituent le coeur stable des classes et abandonnent une partie parfois importante de l'information,

– les seuils retenus (fréquence minimum d'un mot, longueur de l'UCE [9], etc.) font varier parfois de manière importante l'extension de classes et quelquefois leur nombre [10]. Ce phénomène est en partie lié à la structuration plus ou moins forte du texte autour d'isotopies.

---

5. On trouvera un exemple d'analyse d'une question ouverte avec le logiciel SPAD-T dans (Brugidou, 1998).

6. Il s'agit des termes réduits à leur racine. Le choix linguistique de la « stemmatisation » plutôt que de la lemmatisation est en partie justifié par des considérations statistiques – il s'agit de « remplir » le plus possible des tableaux comportant de nombreuses cases vides en accroissant les fréquences des items. On notera par ailleurs qu'Alceste enlève les accents – ce qui ne va sans inconvénient –, pour des raisons de commodité, ils ont été rétablis dans le graphique factoriel et dans les extraits de listings présentés (où l'on a par ailleurs supprimé des tirets rajoutés par le logiciel).

7. Il s'agit d'une classification descendante hiérarchique, *cf.* (Reinert, 1987).

8. Il est fait avec un succès inégal par le logiciel.

9. Les UCE sont les segments de texte sur lesquels se base l'analyse et sont définis automatiquement par Alceste en fonction du nombre de mots et de la ponctuation.

10. En bref, la complexité des paramètres ne permet pas de restituer au pas à pas la construction des classes. Toutefois, d'une part, la stemmatisation et le découpage en UCE constituent le prix à payer (dans la mesure où il s'agit de « remplir » des tableaux qui seraient autrement trop clairsemés) pour pouvoir analyser des tableaux lexicaux qui ne soient par agrégés sur la base d'information métadiscursive (par exemple des tableaux croisant les mots et les locuteurs). D'autre part, la variation de ces multiples paramètres (fréquence, longueur...) permet de faire apparaître des noyaux sémantiques stables. Le mérite du procédé tient alors dans son caractère frustre : ce qui reste après ces multiples avanies est robuste.

De cela, il découle une série de principes de prudence : nous ne devons accorder qu'un statut d'information indicative à la taille des classes et tenter surtout d'interpréter le contenu sémantique commun de ces classes. On s'attachera ainsi peu au destin singulier des mots, les regroupements des formes sous une lexie sont en effet trop massifs et la méthode ne permet pas de comparer des séries statistiques obtenues par d'autres comptages lexicométriques. En revanche, il paraît nécessaire de privilégier une approche par champs lexicaux pour tenter d'éclairer les cooccurrences mises en évidence par la classification automatique [11]. L'interprétation des résultats d'Alceste s'appuie sur des indicateurs de nature différente que nous allons passer en revue : certains indicateurs sont statistiques, d'autres plutôt linguistiques ou sociologiques. On commentera tout d'abord l'arbre de la classification et l'analyse factorielle. On s'apercevra toutefois que le commentaire de l'analyse factorielle implique des inférences linguistiques ou sociologiques notamment dans l'opération qui consiste à « étiqueter » les axes et les classes. Ensuite, on cherchera à analyser une classe en s'appuyant sur les termes spécifiques et les UCE caractéristiques. Le raisonnement par champ lexical est plutôt de nature linguistique, il s'appuie aussi sur des inférences statistiques (termes spécifiques, UCE caractéristiques, variables illustratives linguistiques : mots outils). Enfin, on s'intéressera aux variables illustratives sociologiques décrivant le statut des locuteurs caractéristiques des différentes classes.

2.2. Des classes aux thèmes

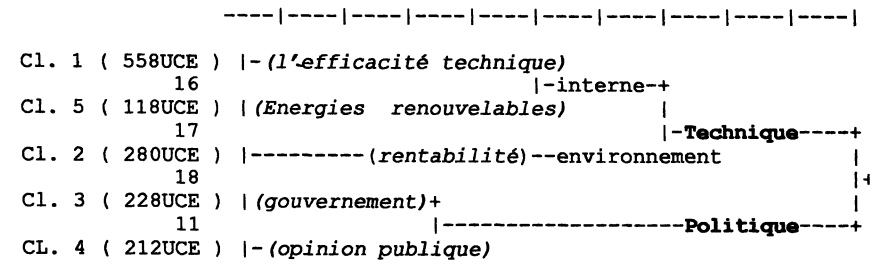
La classification descendante hiérarchique a été effectuée sur les unités de contexte [12]. 73 % des UCE sont classés. Cinq classes de réponses (UCE) sont identifiées à l'issue de cette classification. Le graphique 1 décrit la partition du corpus. La première partition distingue deux sous-ensembles, qui chacun se scinde en deux nouvelles sous-partitions. On constate que les classes 3 et 4 sont en fait très proches et ne sont isolées qu'à la fin de la classification. De fait, différentes classifications faisant varier la longueur des unités de contexte ou retenant l'ensemble de la réponse comme item classé montrent la très forte stabilité des trois premières classes, puisque les classes 1 et 5 s'avèrent, elles aussi, assez proches. On verra que ces proximités ne sont pas des artefacts mais sont bien le fruit de la proximité sémantique des énoncés.

Sur cet arbre, on a inscrit sur les différentes branches des « étiquettes » destinées à faciliter la lecture. Les noms de ces étiquettes sont en fait le résultat d'une série d'actes interprétatifs que nous allons essayer d'identifier et de décrire. Pour l'instant, on se contentera de constater que la première partition distingue des raisons de maintien ou de non maintien du nucléaire que l'on a qualifié de « techniques » en les opposant à des raisons « politiques ». Cette première partition correspond ainsi qu'on le verra au premier axe de

11. Ces points ont notamment été développés dans (Brugidou et Labbé, 2000).

12. Et non sur l'ensemble de la réponse des individus, mais ici les deux options sont pratiquement équivalentes du fait de la taille des réponses à une question ouverte. Le tableau sur lequel porte la classification descendante hiérarchique croise les UCE (soit ici les réponses des individus) et les mots retenus, soit les mots pleins ayant une fréquence supérieure à 4.

l'analyse factorielle construite à partir des classes de réponses (*cf.* graphique 2). Les deux nouvelles partitions de l'arbre correspondent au deuxième axe de l'analyse factorielle : dans les deux cas en effet, on peut opposer des réponses qui font référence à des raisons « internes » (à EDF, à la France) à des réponses faisant référence à des raisons « externes », liées à l'environnement (nouvelles donnes internationales tant au niveau économique que de l'opinion).



GRAPHIQUE 1. – Classification Descendante Hiérarchique - Dendrogramme des classes stables.

Cette première « image » de la classification, bien qu'elle nous donne une assez bonne idée de la taille des classes et de leur proximité doit donc être complétée par une représentation plus qualitative.

3. Inférences linguistiques

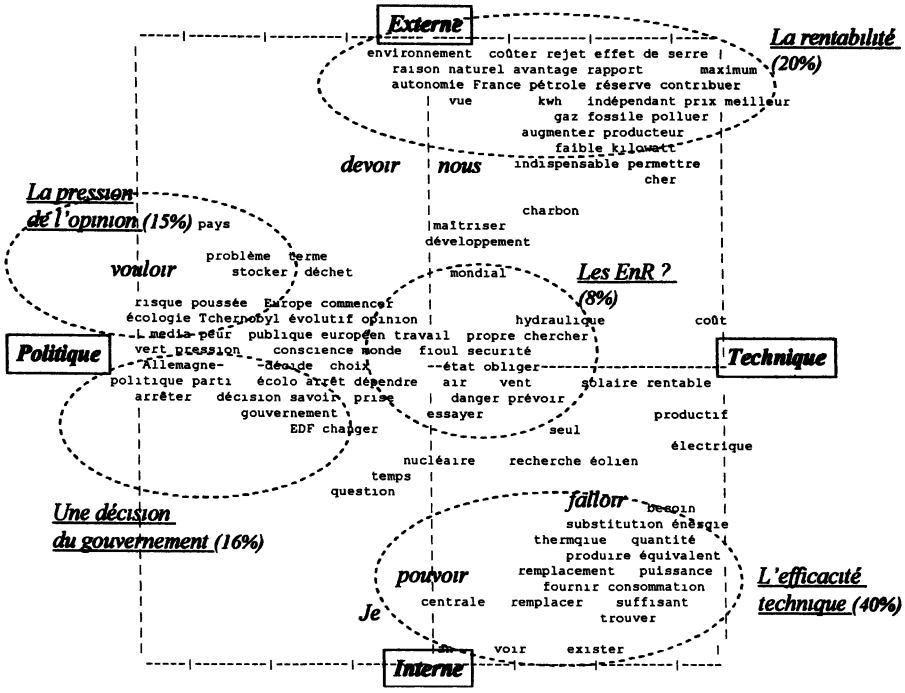
Le graphique 2 construit à partir d'une analyse de correspondance binaire [13] croisant les 5 classes et les mots nous permet de mieux comprendre comment l'analyse combine des inférences statistiques et linguistiques.

L'axe un (37 % de l'inertie) oppose trois classes caractérisées par un vocabulaire « technique », qu'il s'agisse d'un lexique économique (*coût, cher, rentable, prix...*) ou de la production (*énergie, puissance, électrique, produire, solaire* etc.), à deux classes caractérisées par un vocabulaire « politique » (*politique, gouvernement, écologie, opinion publique, Europe...*).

L'axe deux (26 % de l'inertie) différencie des réponses qui font références à des raisons « internes » et celles qui évoquent l'environnement au sens large. Cet axe oppose essentiellement les raisons liées à l'efficacité technique de la production et celles liées aux contraintes économiques (notamment du fait de l'environnement international).

En bref, on recense trois types de raisons « techniques », justifiant la poursuite du nucléaire :

13. L'Analyse de Correspondance Binaire est construite à partir du tableau croisant les 5 classes et les mots retenus par le paramétrage (soit les mots pleins ayant une fréquence supérieure à 4.)



GRAPHIQUE 2. – Visualisation des classes d'arguments (ACB).

• une argumentation technique liée aux contraintes de production : aucune autre source d'énergie n'est assez puissante pour pouvoir actuellement remplacer le nucléaire, (intitulée « *l'efficacité technique* ») [14],

• une argumentation technique à forte dominante économique reliée aux contraintes internationales : le nucléaire est l'énergie la plus rentable (*cf.* le marché) et il permet de lutter contre l'effet de serre (intitulée « *la rentabilité* »),

• une troisième très liée à la première (contrainte technique de production) discute de l'alternative possible représentée par les énergies renouvelables (intitulée « *les Energies renouvelables* » (EnR)),

et deux types de raisons « politiques » :

• une argumentation « politique » reposant sur la pression de l'opinion publique notamment internationale pouvant expliquer l'arrêt d'ici 20 ans du programme nucléaire (intitulée « *La pression de l'opinion* »),

• une argumentation « politique » faisant référence à la situation politique française et à une éventuelle décision du gouvernement (intitulée « *Une décision du gouvernement* »).

14. Cette classe regroupe 558 UCE sur les 1396 réponses classées , soit 40 % des UCE classées

On comprend la très forte proximité entre ces deux classes puisque la décision gouvernementale s'expliquerait par une pression de l'opinion publique nationale et internationale.

On a ici livré en quelques mots un résumé de l'interprétation de ces classes. Celle-ci s'appuie sur des inférences linguistiques réalisées à deux niveaux :

• d'une part, au niveau de l'ensemble du corpus, ce sont des inférences qui construisent la signification d'une classe par opposition ou rapprochement avec les autres classes (exemple de la lecture de l'analyse factorielle réalisée à partir de la classification),

• et d'autre part, au niveau d'une classe particulière, ce sont des inférences qui s'appuient sur les UCE et les termes caractéristiques de la classe et qui reconstituent l'économie interne du thème.

Par ailleurs, on constate que l'interprétation est le fruit d'un va et vient entre ces deux niveaux de lecture : les hypothèses interprétatives formées au palier du texte [15] « cadrent » les hypothèses « locales », lesquelles permettent à leur tour d'affiner les premières hypothèses. Ainsi, on qualifie bien « le local par le global [16] », comme le montre la lecture du graphique factoriel : le système d'opposition et d'équivalence construit par l'analyse des Correspondances Binaires permet de « nommer » les axes. C'est cette première saisie du sens qui contextualise l'interprétation des verbatims caractéristiques.

3.1. De la statistique à la linguistique : le carré sémiotique

Pour interpréter ces données, le « carré sémiotique », structure élémentaire de la signification pour Greimas, fournit une métaphore commode. En interprétant ce graphique, nous avons en effet « modélisé » un système d'identités et de différences que l'on peut décrire par le tableau tétrachorique [17] ci-dessous.

TABLEAU 1. – Carré sémiotique à partir de l'analyse factorielle.

	Interne	**Externe**
Politique	Une décision du gouvernement	La pression de l'opinion publique
Technique	L'efficacité technique	La rentabilité économique

Dans ce carré sémiotique, les différentes classes sont reliées par les relations logiques suivantes :

« Une décision gouvernementale » **et** « La pression de l'opinion publique » (*relation de contradiction*)

« Une décision gouvernementale » **ou** « La rentabilité économique » (*relation de contrariété*)

15. Lecture globale qui caractérise une classe par opposition ou rapprochement avec les autres classes.

16. Sur ces questions voir notamment (Rastier, 1996).

17. Sur ces notions et leur utilisation pour l'interprétation sociologique, voir (Demazière et Dubar, 1997, p. 311 et suivantes).

« La pression de l'opinion publique » **présuppose** « La rentabilité économique » (*relation de présupposition*) [18].

On aurait ici l'exemple d'une de ces typologies qualitatives décrites par D. Demazière et C. Dubar (1997, p. 308) citant Culioli à propos des parcours d'insertion : « les catégories « naturelles », bricolées et instables, qui servent à mettre en récit de nouveaux types de parcours, dessinent des « configurations sémantiques » qui constituent autant de « domaines notionnels » plus ou moins « clairement structurés autour d'attracteurs assurant une certaine fermeture et ouverts vers des complémentaires (inverses ou opposés) par rapport auxquels ils se définissent par négation ».

La stabilité relative de la classification et de la représentation factorielle reflète la stabilité des catégories naturelles analysées (où l'on voit que nous parlons aussi de sociologie). Le graphique factoriel représente de manière géométrique des relations logiques : la conjonction se lit comme un angle à 0°, cependant que la disjonction exclusive se lit comme un angle à 180°. Ainsi, d'un point de vue statistique, nous sommes fondés à conclure que les termes de la classe « efficacité technique » ne sont jamais ou presque associés à ceux de la classe « La pression de l'opinion », il s'agit bien d'une relation disjonctive exclusive. Le passage de la statistique à la linguistique est ici assuré.

3.2. De la linguistique à la sociologie

Chaque classe est décrite par des termes spécifiques [19], des mots-outils et des UCE caractéristiques.

On propose ici de formaliser ces différentes étapes, en ayant recours à des catégories linguistiques simples.

Étape A : vocabulaire spécifique de la classe « efficacité technique »

Le listing Alceste présente les termes de la classe par ordre de spécificité. Pour faciliter l'interprétation, on peut tenter de construire des « champs lexicaux », c'est-à-dire de ranger ces termes en fonction de leur proximité sémantique. Pour cela, on propose d'emprunter quelques notions simples à la sémantique interprétative et d'en faire un usage, sans doute approximatif, mais qui pourra nous aider à argumenter notre interprétation [20]. Ainsi, on regroupe les termes en fonction d'un sème générique qui permet de les ranger en « taxème [21] » ou

18. Ou encore : « L'efficacité technique » **et** « La rentabilité économique »/« L'efficacité technique » **ou** « La pression de l'opinion publique »/« La rentabilité économique » **présuppose** « La pression de l'opinion publique ».

19. « Caractéristiques » au sens statistique. Il s'agit en effet de termes sur-représentés au sens du chi-2 dans les uce de la classe. Le chiffre noté entre parenthèses est l'occurrence du « stemme » dans la classe.

20. L'application des concepts de la sémantique interprétative pour l'interprétation des résultats d'Alceste a été proposée et très finement discutée par Valérie Beaudoin dans son travail de thèse (Beaudouin, 2002).

21. Le « sème générique » (par exemple /puissance/, notée entre simple barre oblique) est le sème récurrent à tous les sémèmes du champ lexical.

TABLEAU 2. – Listing Alceste des termes caractéristiques de la classe.

moyen+(96), remplacement+(74), remplac+er(84), trouv+er(177), energ+16(332), besoin+(76), source+(71), suffis+ant(31), producti+f(64), produire.(46), electri<(64), fourn+ir(22), voir.(48), consommat+ion(18), substitution+(39), thermique+(20), an+(64), centrale+(94), puissance+(21), quantite+(19), demand+er(22), exist+er(11), equival+ent(13), capable+(8), court+(9), grand+(11), forme+(7), instant+(21), jour+(15), solution+(55), systeme+(9), renouvel+er(29), supprim+er(6), techn+16(15), difficile+(13), capacite+(6), face+(7), continu+er(17), faire.(53), fonctionn+er(10), represent+er(7), impossi+ble(4), nouveau+(3), enormement(5), entretien+(3), ouverture+(3), adapt+er(5), class+16(5), industri<(6), perform+ant(6), puiss+ant(9), vala+ble(5), rechange+(5), surgenerateur+(3), necessaire+(5), partie+(5), plan+(4), aspect+(5), avis(14), contrainte+(4), manque+(4), reseau+(2), attendre.(2), diversifi+er(4), prefer+er(2), subvenir.(4), utilis+er(9), esthet+16(2);

champ lexical [22]. Les différents champs lexicaux peuvent être à leur tour rangés sous des ensembles d'un niveau de généralité supérieur appelés « domaines [23] ».

Classement par champs lexicaux :

moyen de remplacement, remplacer, substitution, équivalent, rechange, renouveler, diversifier : taxème /échange/, domaine // équivalence//

puissance, puissant, quantité, capable, capacité : taxème/puissance/
court, grand : taxème /quantité/
besoin, nécessaire, suffisant, énormément : taxème/quantité-évaluatif/
] //La quantité//

énergie (source), électricité, thermique : taxème /énergie/
centrale, surgénérateur : taxème /moyen de production/
production, produire, industrie, fournir, subvenir : taxème/produire/
consommer, demander : taxème /consommer/
] //La production//

fonctionner, entretien, plan, solution, système, technique : taxème/système/
aspect, partie : taxème /partie/
] //Le système//

valable, performant : taxème /performance/
difficile, contrainte, impossible : taxème /difficulté/
] //L'efficacité//

adapter, nouveau, supprimer, continuer : taxème / continuité/
an, instant, jour : taxème /temps/
] //Le temps//

22. Pour F. Rastier, « Le taxème est l'ensemble de rang inférieur (...) on peut lui appliquer cette définition de Coseriu : « structure paradigmatique constituée par des unités lexicales ('lexèmes') se partageant une zone commune de signification et se trouvant en opposition immédiates les unes avec les autres »... ». (Rastier, 1994) p. 49. Pour F. Rastier, l'opposition taxème/domaine reste insuffisante, il a proposé en 1994 d'ajouter un intermédiaire qui est le champ comme classe de travail (et non « de langue »), intermédiaire. F. Rastier (avec la collaboration de Marc Cavazza et Anne Abeillé), (Rastier, 1994).

23. Domaine noté entre double barre oblique : //domaine//.

On constate que la plupart des unités lexicales peuvent être classées dans ces ensembles [24].

Étape B : mots outils spécifiques de la classe « efficacité technique»

Le logiciel propose par ailleurs une liste des mots-outils ou mots grammaticaux « caractéristiques [25] » de la classe.

TABLEAU 3. – Listing Alceste des mots outils caractéristiques de la classe.

entre(5), falloir.(88), paraitre.(4), pouvoir.(73), ne(308), pas(326), rien(19), ici(41), arriere(6), aujourd'-hui(18), derriere(5), lendemain(3), longtemps(3), assez(17), autant(14), plus-d<(15), alors(5), aussi(29), a- la-place(4), encore(25), pour(168), pourquoi(1), il(165), je(97), mon(14), aucun+(13), autre+(181), celles(4), en(124), on(315), quel(3), quelle(9), quelque-chose(6), toutes(10), aura(122), ayant(2), ben(3), vingt(7), a(257), n(6), y(81);

Classement par catégorie grammaticale [26] :

falloir, pouvoir, paraître : [modalisations]

ne, pas, rien, aucun : [négations]

assez, autant, aussi, plus : [mots outils marquant une comparaison]

aujourd'hui, lendemain, longtemps : [adverbes de temps]

ici, arrière, derrière : [adverbes de lieu]

je, mon : [pronoms personnels ou adjectifs possessifs à la première personne]

il, on : [pronom personnel troisième personne]

aura, ayant, a : [auxiliaire avoir]

alors, pourquoi : [adverbes exprimant un rapport logique de causalité]

Étape C : les unités de contexte les plus caractéristiques de la classe « efficacité technique»

La troisième étape de ce parcours interprétatif consiste à reconstituer à partir des verbatims les plus caractéristiques de la classe, le patron linguistique sous-jacent.

On constate que la plupart des verbatims sont formés sur le modèle suivant :

Modélisation du schème logique à partir des catégories lexicales (étape A) et grammaticales (étape B) :

Cette modélisation est approximative, car nombre de verbatims présentent des variations à partir de cette structure. Mais *intuitivement*, la lecture du

24. Non classés : trouver, voir, exister, forme, faire face, représenter, ouverture, avis, manque, réseau, préférer, utiliser, esthétique.

25. On rappellera que les mots-outils constituent des variables illustratives et qu'ils ne concourent pas à la construction de la classe.

26. Notés entre crochet : [mots outils]. Non classés : à la place, encore, pour, autre, celles, en, quel, quelle, quelque chose, toutes, vingt, y.

TABLEAU 4. – Listing Alceste[28] des verbatims les plus caractéristiques de la classe.

1839 17 on a pas d'autres #solutions pour l'#instant/[27] je ne sais pas si d'ici la on aura #trouver une #énergie de #remplacement/ parce que c'est des #puissances de #productions #énormes et je ne #vois pas ce qui pourrait #remplacer/

671 15 à moins que l'on #trouve des nouvelles #sources d'#énergie des #nouveaux #moyens de #production mais d'ici 20 #ans ca sera dur/ les autres #moyens il faudra qu'ils soient aussi #puissants/

249 14 pour #faire #face à la #production. on n'a pas de #moyens de #remplacement. je ne #vois pas ce que l'on pourrait #faire d'autre à la place.

766 14 on aura #besoin et on n'aura pas #trouve d'#énergie de #substitution en #quantité #suffisante.

1187 13 ben, je ne sais pas parce que il faut avoir des #énergies de #remplacement avant de #supprimer les #centrales/ on en #trouve mais de faible #puissance/ on est trop #industrialisé maintenant/

TABLEAU 5. – Modélisation du schème logique.

Modélisation	//temps// ou [adverbe de temps]	[négation + modalisation]	//équivalence//	//production//	[pronom personnel + modalisation]	//production//	[adverbe marquant une égalité]	//quantité//
verbatims	*Actuellement,*	*rien ne peut*	*remplacer*	*le nucléaire,*	*il faudrait*	*une source d'énergie*	*aussi*	*puissante.*

listing montre que ces réponses obéissent bien au même modèle. L'hypothèse d'une telle interprétation est la suivante : au-delà du patron linguistique (ligne « modélisation » du tableau), cette opération nous permettrait de reconstituer un « schème logique » et d'accéder ainsi au niveau cognitif. On peut souligner par ailleurs le caractère extrêmement simple de l'exemple analysé. La plupart du temps, les regroupements ne sont pas aussi facilement réduits à un schème logique sous-jacent et l'on se contente de plusieurs « patrons » plus ou moins complets. Cette analyse atteste ainsi de la qualité de la classification déjà soulignée qui reflète en grande partie la forte structuration des opinions sur le thème traité.

Sur la base de cette analyse, on peut inférer que l'équivalence entre items de la catégorie //production// est réglée par la catégorie //quantité//.

Ces différentes étapes sont très largement imbriquées. En réalité, les étapes A et B sont déterminées par précompréhension intuitive de l'étape C. Comme pour l'interprétation du graphique factoriel, le « global détermine le local ». En effet, déterminer des champs lexicaux ne repose pas seulement sur une connaissance du système de la langue et du sens collecté dans des dictionnaires. Ce classement implique d'abord une saisie de « l'entour pragmatique [29] », c'est-

27. Le caractère / indique ici une relance de la part de l'enquêteur (« *Et encore ?* »).

28. Le dièse signale les mots appartenant à un « stemme » spécifique de la classe.

29. On reprend ici la terminologie proposée par F. Rastier.

à-dire de la situation d'énonciation. Cette opération permet d'activer des « topoï [30] », des lieux communs culturellement sédimentés qui sont pertinents pour l'interprétation de ce type de textes. On lève ainsi certaines ambiguïtés : par exemple, « centrale » ne désigne ni une prison, ni une position géométrique mais une unité de production (ce qu'attestent par ailleurs certains contextes : « centrale nucléaire »). Mais classer les unités lexicales implique aussi une précompréhension du schème logique à travers une lecture intuitive des verbatims et donc une connaissance du « contexte ». Par là, on va associer certains termes : *énergie, électricité, thermique* ont en commun d'être des énergies et de s'opposer à *centrale* et *surgénérateur* qui sont [31] des moyens de production. On peut toutefois considérer qu'au regard des autres champs lexicaux identifiés (/quantité/, /système/, /efficacité/ ...), ces termes, par ailleurs opposés, ont en commun une même dimension sémantique, l'univers de la production.

On voit que la linguistique que nous mettons en œuvre pour interpréter le sens de ces textes n'est pas une linguistique réduite au système de la langue, mais qu'elle inclut, suivant F. Rastier, trois instances immanentes de codification : le système de la langue (« dialecte »), la norme sociale (« sociolecte ») et les usages (« idiolectes »).

4. Inférences sociologiques

Analyser les réponses à une question ouverte, c'est, en dernière instance, réaliser des inférences sociologiques : en quoi l'analyse de ces réponses nous permet-elle en effet une meilleure connaissance de la population étudiée ?

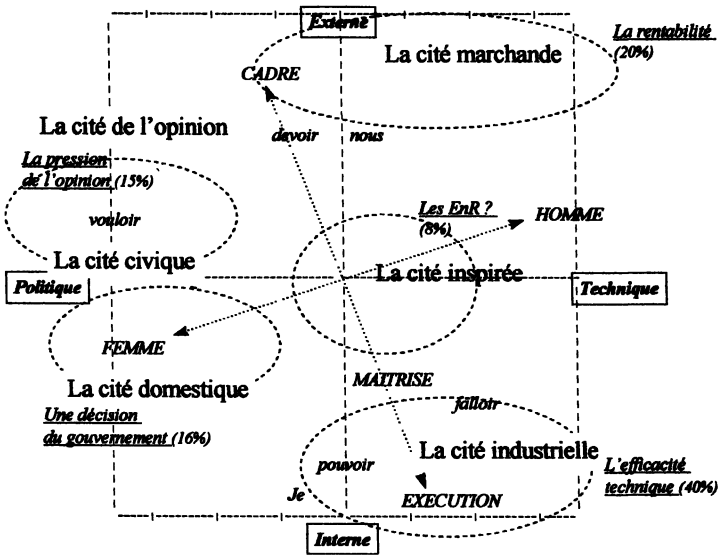
Alceste permet de rapprocher les variables textuelles (mots, verbatims...) et les variables sociologiques, qu'il s'agisse de variables morphologiques (par exemple les caractéristiques socio-démographiques, le statut dans l'entreprise) ou de variables attitudinales (par exemple les réponses aux questions fermées de l'enquête). On distinguera donc ici deux dimensions sociologiques, celle qui décrit la position sociale « objective » (et donc la position d'énonciation) et celle concernant la dimension culturelle.

4.1. Logiques sociales

Le graphique 3, construit à partir du graphique 2 (ACB), permet de restituer la place des variables illustratives et de proposer une interprétation sociologique de ces résultats. Cette dernière représentation met en évidence des variables « objectives » : ainsi, on constate que la position hiérarchique recoupe à peu près l'axe deux du graphe factoriel : les agents d'exécution et dans une moindre mesure les agents de maîtrise valorisent les arguments

30. « L'interprétation de la relation $s2 \rightarrow s4$ est un axiome normatif dépendant de normes socialisées, qui peut s'énoncer : la femme est un être faible. On nommera topos ce genre d'axiome largement attesté (...) une relation d'afférence dont l'interprétant est un topos sera dite socialement normée » (Rastier, 1994, p. 47).

31. *i.e.* « ont en commun d'être » des moyens de production.



GRAPHIQUE 3. – Construction d'un espace théorique à partir des inférences statistique, linguistique et sociologique

«internes», cependant que les cadres mobilisent davantage les arguments «externes». Par ailleurs, on constate une légère dérive selon l'axe un : les agents d'exécution sont plus sensibles aux arguments «techniques», les cadres citant plus des arguments «politiques». Autrement dit, les cadres mettent davantage en avant les contraintes économiques du marché et surtout la pression de l'opinion publique en soulignant les contraintes et les opportunités liées à l'environnement international. Les agents d'exécution argumentent plus sur les contraintes techniques de production et dans une moindre mesure sur les contraintes politiques nationales. Ce même graphique oppose les hommes et les femmes sur l'axe un : ces dernières font davantage état que les hommes d'arguments «politiques» (opinion publique, décision du gouvernement). Les hommes argumentent plus volontiers à partir d'arguments techniques (ce point n'est évidemment pas sans rapport avec le métier des personnes interrogées.)

4.2. Logiques symboliques

L'analyse sociologique ne se limite toutefois pas à la caractérisation objective des répondants, elle porte aussi sur l'identification des logiques argumentatives mises en œuvre dans les réponses. On a vu quels types d'inférences linguistiques nous permettaient de reconstituer les schèmes logiques sous-jacents aux groupes de réponses. Ces analyses peuvent être approfondies par une approche proprement sociologique des topoï mobilisés dans les argumentations. L. Boltanski et L. Thevenot (Boltanski et Thevenot, 1991) ont mis en évidence l'existence de systèmes discursifs (qui sont aussi des systèmes de valeurs) permettant de développer des ressources argumentatives pour fonder le juste ou

l'injuste. L'analyse des logiques argumentatives des groupes de réponses montre à partir de quel système de valeurs ces réponses sont produites. On vérifie dans ces réponses les effets de l'ensemble des « cités » décrites par ces auteurs. La classe de réponses de « l'efficacité technique » se déploie selon la logique de justification de la cité industrielle chère à Saint Simon. En effet, ce qui compte dans ce monde c'est l'efficacité technique, les arguments liés à la rentabilité du produit n'ont pas ici de portée à la différence de l'argumentation construite à partir de la « cité marchande ». On pourrait de la même manière montrer que les arguments « politiques » sont construits à partir de logiques de justification différentes : la cité de l'opinion qui légitime la grandeur sur la force de la réputation et de l'honneur et la cité civique qui fonde le juste sur le dévouement à la cité et à l'intérêt général permettent de rendre compte de ces types de réponses.

5. Conclusion : Un parcours interprétatif

À l'issue de ce « parcours interprétatif » et à partir de la classification et du graphique factoriel, on a pu construire une représentation combinant les trois dimensions de l'analyse :

i) une représentation statistique (groupes de réponses pondérées),

ii) une représentation sémantique structurée par deux axes (technique *versus* politique et interne *vs* externe) avec des types d'arguments (l'efficacité technique, la rentabilité etc.),

iii) une représentation sociologique (logiques sociales, logiques symboliques).

Ce parcours est marqué, d'une part, par l'imbrication de trois dimensions d'analyse (statistique, linguistique et sociologique) et, d'autre part, par une récursivité forte entre ces différentes étapes. Non seulement chaque étape de l'analyse mêle plus ou moins étroitement des inférences de nature différente, mais chaque moment de l'analyse anticipe les étapes qui suivent. Si comprendre les réponses consiste toujours à « interpréter » globalement le sens à partir de ces indices hétérogènes, argumenter ses hypothèses interprétatives revient à distinguer la nature de ces indices et à dissocier les différents moments de l'analyse.

Bibliographie

BEAUDOUIN V. (2002), *Mètre et rythmes de l'alexandrin classique : Corneille et Racine*, Champion, coll. Lettres numériques, Paris.

BOLTANSKI L. et THÉVENOT L. (1991), *De la justification*, Gallimard, Paris.

BLOT I. et LE ROUX D. (1992), L'utilisation du logiciel Alceste au département GRETS. *Note interne EDF-R&D*, HN-52/92/067.

BLOT I., HAMMER B. et LE ROUX D. (1994), Traitement des questions d'opinion « ouvertes » : utilisation d'Alceste, outil d'assistance à l'analyse. *Revue ICO Québec*; 6. (1 & 2).

BRUGIDOU M. (1998), Épitaphes, l'image de François Mitterrand à travers l'analyse d'une question ouverte posée à sa mort. *Revue Française de Science Politique*, vol. 48, n° 1, pp. 97-120.

BRUGIDOU M. et LABBÉ D. (2000), *Le discours syndical français contemporain.* CERAT, Grenoble.

DEMAZIÈRE D. et DUBAR C. (1997), *Analyser les entretiens biographiques, l'exemple des récits d'insertion*, Nathan, Paris.

ESCOFFIER C. (2001), Perception des actions et des services d'EDF à l'intention des clients démunis. Analyse des questions ouvertes de l'enquête SME 2001. *Note interne EDF-R&D*, H71/01/020/A.

LABBÉ D. (2001), Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial. *Journal de la Société Française de Statistique*, 142, 4.

LAHLOU S. (1995), La construction du sens dans l'analyse statistique de données textuelles : théorie et méthodologie illustrées par deux analyses. *Note interne EDF-R&D*, HN- 51/95/012.

LEBART L. et SALEM A. (1993), *Statistique textuelle*, Dunod, Paris.

MARC X. (2001), Les modalités de recueil des réponses libres en institut de sondage. Le rôle de l'enquêteur, les consignes et les procédures de contrôle, les perspectives d'amélioration. *Journal de la Société Française de Statistique*, 142, 4.

RASTIER F. (1996), *Sémantique interprétative*, P.U.F., Paris.

RASTIER F. (1994), *Sémantique pour l'analyse*, avec la collaboration de Marc Cavazza et Anne Abeillé, Masson, Paris.

REINERT M. (1987). « Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud », *Bulletin de Méthodologie Sociologique*, n° 13.

TOURNIER M. (1980), « D'où viennent les fréquences de vocabulaire? La lexicométrie et ses modèles », *Mots*, 1, pp. 189-212.