Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique.

Cyril Labbé, Dominique Labbé

Journées d'études du CIDSP
9 mars 2001

Les nouvelles méthodes d'analyse des entretiens

Discrimination et classement au sein d'un groupe d'entretiens

Le cas du confort électrique

Cyril LABBE
Laboratoire d'Informatique de Grenoble - Université Joseph Fourier

Dominique LABBE
CERAT-IEP, BP 48 – 38040 GRENOBLE Cedex 9
dominique.labbe@iep-grenoble.fr

Résumé :
Présentation du calcul de la distance intertextuelle et de deux méthodes de classification (classification hiérarchique ascendante, analyse arborée). Caractérisation du vocabulaire spécifique des différentes classes. Application à un groupe d'entretiens sur le confort électrique.

Cette recherche a été réalisée avec l'aide du Groupe de Recherche Energie, Technologie et Société, département de la Division Recherche et Développement d'Electricité de France.

Lorsque l'on traite un grand nombre d'entretiens, il est indispensable d'opérer des regroupements. On utilise généralement pour cela des informations extérieures aux textes proprement dits : sexe, âge, profession, proximités partisanes, etc. Le texte des entretiens vient alors en illustration de ces classements. Nous présentons ici la démarche inverse : classer les individus en fonction de ce qu'ils disent et ne faire intervenir qu'ensuite les autres informations.

Nous illustrerons cette démarche à l'aide d'un corpus de 64 entretiens remis par le département GRETS d'EDF[1]. Ces entretiens portent sur le thème du "confort électrique" (voir présentation de ce corpus en annexe). Ils ont été réalisés entre 1994 et 1997 selon une même méthode semi-directive et retranscrits sur fichier électronique de façon très minutieuse. Nous avons corrigé ces retranscriptions (orthographe, graphies aberrantes, ponctuation) et ajouté une série de balises permettant d'isoler les réponses des enquêtés et de ne traiter qu'elles tout en conservant les autres informations (notamment les relances de l'enquêteur).

Le logiciel **normalise** les graphies — par exemple, il réduit les majuscules initiales de phrase, traduit les chiffres, les abréviations… — puis il associe à chaque mot ainsi normalisé un **vocable** (lemme et catégorie grammaticale). En quelque sorte, la **lemmatisation** rattache chaque mot du texte à sa place dans le lexique de la langue française (par exemple, les verbes sont rattachés à leurs infinitifs, les adjectifs à leur masculin singulier, etc). La lemmatisation est effectuée dans le texte, elle est **exhaustive** (tous les mots sont analysés), **sans ambiguïté** (un seul lemme par mot) et **réversible**, c'est-à-dire qu'on peut retrouver le texte original à partir du fichier des lemmes[2].

Le corpus total de ces réponses compte 394.000 mots, attestés sous 11.602 formes normalisées et 6.606 vocables. Pour délimiter quelques grands groupes dans ce vaste ensemble, nous avons mesuré la distance séparant chacun des textes et opéré deux classifications (classification hiérarchique ascendante et analyse arborée). La caractérisation de ces groupes se fera à l'aide de leurs vocabulaires spécifiques.

La distance intertextuelle

Depuis l'indice de Jaccard, de nombreuses formules ont été proposées pour opérer des discriminations et des classements au sein de populations statistiques (Sokal et Sneath, 1973). Toutefois, la quasi-totalité de ces indices se fondent sur la présence ou l'absence d'un caractère, au sein des populations considérées, et ne tiennent pas compte de la fréquence de ce caractère (pour un recensement de ces indices et une synthèse de leurs propriétés : Hubalek, 1982). Cette limitation s'explique par l'impossibilité de recenser exhaustivement la plupart des individus étudiés par les spécialistes du vivant (biologistes, entomologistes, botanistes, etc.).

Pour l'analyse des textes, le problème a été étudié sous le nom de « **connexion** lexicale » définie comme « l'intersection du vocabulaire de deux textes » (Muller 1977). La connexion est donc le complémentaire de la **distance**, terme plus familier en statistique et que nous retenons pour cette raison. Mais en statistique lexicale également, l'attention s'est fixée sur le

[1] Dans le cadre d'une convention de recherche entre le CERAT et le Groupe de Recherche Energie, Technologie et Société, département de la Division Recherche et Développement d'Electricité de France.

[2] La lemmatisation est réalisée par un analyseur syntaxique qui résout les difficultés dues à l'homographie (deux mots différents ayant la même orthographe : par exemple « est », verbe être et point cardinal). Voir Charles Muller, Principes et méthodes de la statistique lexicale, Paris, Hachette, 1977; Dominique Labbé, Normes de saisie et de dépouillement des textes politiques, Grenoble, Cahiers du CERAT, avril 1990.

vocabulaire (ou **lexique)** : ensemble des mots différents (ou **vocables**) utilisés dans un texte, sans tenir compte de leur fréquence (voir par exemple Brunet, 1988, Hubert et Labbé, 1998).

Cette approche présente des inconvénients lorsque les caractères observés sont très inégalement répartis comme c'est le cas des mots dans tout texte écrit en langue naturelle. Dans ce cas, les indices fondés sur la présence ou l'absence donnent un très grand poids aux mots rares et négligent les fluctuations de densité parfois importantes chez les individus présents partout (comme les mots outils ou les verbes usuels).

La statistique lexicale a la chance de pouvoir recenser exhaustivement les individus qu'elle étudie : l'association à chacun de ces vocables de sa fréquence d'emploi donne en effet le nombre total des mots d'un texte (son étendue ou sa **dimension**, notée N). Dans le terme « distance intertextuelle », l'adjectif **textuel** indique que les calculs portent sur l'ensemble des textes (N) et non sur leur seul vocabulaire (V).

Nous avons recherché un indice doté des propriétés suivantes (pour une discussion : Baulieu, 1989) :

— insensible aux différences de taille entre les textes comparés ;

— applicable à plusieurs textes et, potentiellement, à tous les textes d'une même langue ;

— variant « uniformément » — entre 0 (même vocabulaire et fréquence semblable de chacun des mots dans les deux textes) et 1 (aucun vocable en commun) — sans saut ni effet de seuil autour de certaines valeurs ;

— symétrique (soit deux textes A et B alors $\delta_{(A,B)} = \delta_{(B,A)}$) ;

— aussi « transitif » que possible : quand on agrège le vocabulaire de deux textes, les distances de ce nouvel ensemble vis-à-vis des autres textes doit refléter l'ordre des distances antérieures (si $\delta_{(A,B)} < \delta_{(A,C)} < \delta_{(B,C)}$ alors $\delta_{(A,B)} < \delta_{\{A,(B+C)]}$)

— aussi "robuste" que possible (ie une modification marginale dans le vocabulaire d'un des deux textes doit se traduire par une variation marginale de leur distance)…

Soit deux textes A et B et,

V_a et V_b : nombre de vocables dans A et B (vocabulaire) ;

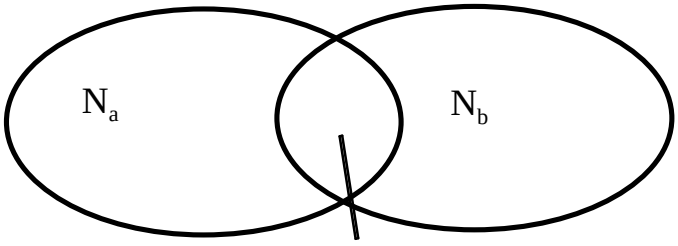
F_{ia} : fréquence du vocable i dans A ;

F_{ib} : fréquence du vocable i dans B.

N_a et N_b : nombre de mots dans A et B (taille) ;

avec $N_a = \sum F_{ia}$ et $N_b = \sum F_{ib}$

La distance absolue entre A et B sera l'union des deux textes moins leur intersection, c'est-à-dire la somme des différences entre les fréquences absolues de chacun des mots des deux textes.



Mots communs à N_a N_b

La distance relative pourra être calculée de deux manières :

$$(1)\quad \delta_{(a,b)} = \frac{\sum_{Va} | F_{ia} - F_{ib} | + \sum_{Vb} | F_{ib} - F_{ia} |}{N_a + N_b}$$

ou :

$$(2)\quad \delta_{(a,b)} = \frac{1}{2}\left(\frac{\sum_{Va}|F_{ia} - F_{ib}|}{N_a} + \frac{\sum_{Vb}|F_{ib} - F_{ia}|}{N_b}\right)$$

La formule (2) est, à la notation près, celle que donne E. Brunet dans Brunet (1988). La distance maximale absolue est égale à $N_a + N_b$.

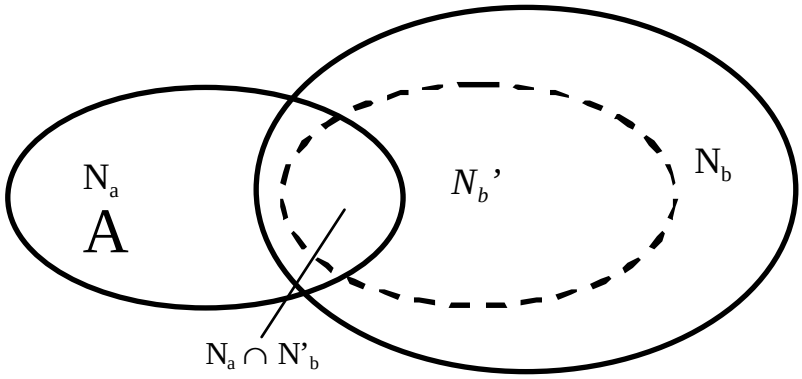
Ces formules classiques ont rencontré deux objections :

— (1) et (2) ne sont équivalentes que quand les textes ont des tailles égales ($N_a = N_b$). Si les deux textes comparés ne partagent aucun vocable, les formules (1) et (2) donnent bien un indice de 1 quelle que soit la taille des textes (ce qui est une des conditions formulées ci-dessus). En revanche, le minimum théorique ne peut atteindre zéro que dans le cas particulier de tailles égales. A partir d'un certain seuil, le vocabulaire du plus grand des deux textes ne peut plus "physiquement" entrer dans le plus petit. Ainsi les 2.152 vocables différents du plus long entretien (PCR08) ne peuvent tous entrer dans le petit entretien (DOM03) qui est long de seulement 1144 mots.

— dans (1) comme dans (2), l'intersection des deux textes est comptée deux fois. On donne donc plus d'importance aux vocables communs qu'aux vocables propres à chacun.

Pour surmonter ces deux objections, nous proposons de simuler la réduction du plus grand des deux textes à la taille du plus petit.

Soit B' cette réduction de B en fonction de la taille de A :



Soit U(a,b), le coefficient de proportionnalité entre A et B :

$$U_{(a,b)} = \frac{N_a}{N_b}$$

Tout vocable i de fréquence F_i dans B aura une fréquence attendue dans A égale à :

$$E_{ia(u)} = F_{ib} * U_{(a,b)}$$

D'où l'on tire que :

$$N'_b = \sum_{Vb} E_{ia(u)}$$

Dans les formules (1) ou (2), on remplace les termes F_{ib} par $E_{ia(u)}$ et N_b par N'_b. Le minimum théorique (zéro) sera atteint quand le petit texte sera une sorte de modèle réduit du grand. Dans ce cas, tous les vocables de A se retrouvent dans B avec une fréquence telle que :

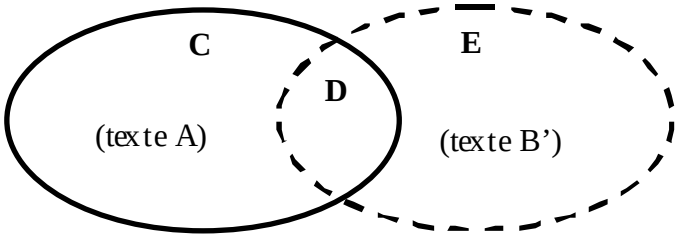
$$F_{ia} = E_{ia(u)}$$

Le numérateur de la formule (2) sera égal à zéro et le dénominateur à : $N_a + N'_b \approx 2\ N_a$. C'est en effet l'effectif maximum que les deux textes peuvent partager s'ils ont même dimension, même vocabulaire et la même fréquence pour chacun des vocables. Le maximum théorique (l'unité) devrait être atteint quand les deux textes n'ont aucun mot en commun. Au numérateur, comme au dénominateur, figureront N_a et N'_b.

Toutefois, cette nouvelle formulation ne répond pas à l'objection concernant le double compte de l'intersection des deux textes et ne résout pas totalement le problème physique mentionné ci-dessus : tous les vocables de B ne peuvent pas théoriquement figurer dans A. Pour tenir compte de ces deux objections, il est proposé de :

- ne considérer qu'une seule fois l'intersection des deux textes ;
- limiter le calcul à l'ensemble des vocables de A mais seulement à V'_b, ensemble des vocables de B dont la fréquence est telle que l'on en attend au moins 1 dans A ($E_{ia(u)} \geq 1$). La somme de ces espérances donne N'_b.

Ce calcul se déroulera en deux temps (voir figure ci-dessous) :



Afin de ne pas compter deux fois les vocables compris dans l'intersection (ensemble D), on considère d'abord la totalité des V_a vocables de A (pour l'ensemble C, on a : $E_{ia(u)} = 0$), puis les seuls $V'_{b(E)}$ vocables de l'ensemble E (c'est-à-dire ceux dont la fréquence dans B est suffisante pour qu'on en attende au moins un dans A alors qu'ils en sont absents).

La distance absolue entre A et B' est alors égale à :

$$D_{V_a,\ b(u)} = \sum_{V_a,\ V'_{b(E)}} \left| F_{ia} - E_{ia(u)} \right|$$

La distance maximale sera atteinte quand A et B' seront totalement disjoints (aucun vocable en commun) et elle sera égale à : $N_a + N'_b$. Pour que, dans ce cas, l'indice soit égal à 1, il faut que ce maximum figure également au dénominateur. La distance relative sera donc :

$$(3)\quad D_{(a,b)} = \frac{\sum_{V_a,\ V'_{b(e)}} \left| F_{ia} - E_{ia(u)} \right|}{\sum_{V_a} F_{ia} + \sum_{V'_b} E_{ia(u)}} = \frac{\sum_{V_a,\ V'_{b(e)}} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

On remarquera que le même résultat, aux arrondis près, peut être obtenu en soustrayant les fréquences relatives de chacun des vocables dans les deux textes comparés, à condition de limiter le calcul à tout le vocabulaire du plus petit des deux textes et à ceux des vocables qui, dans le plus grand, ont une fréquence suffisante pour qu'on en attende au moins 1 s'il avait la taille du plus petit ;

Ainsi calculée, la distance devrait être théoriquement insensible à la taille des textes comparés. Ce n'est pas tout à fait exact lorsque N_a et N_b sont très différents : bien que la corrélation soit faible, il existe alors un lien entre taille et distance. Ce lien provient de deux choses :

— les arrondis. Alors que les fréquences observées sont toujours des entiers, les fréquences théoriques auront presque toujours des décimales qui entreront dans la distance. Ce défaut sera d'autant plus sensible que les mots de basses fréquences occuperont une étendue

importante du plus petit texte. Le calcul est entaché d'une incertitude de .5 pour les fréquences 1, de 0.25 pour les fréquences 2, etc.

Pour le comprendre considérons trois textes A, B, C :

N_a = 1 000 mots ; N_b = 1 500 mots ; N_c = 9 500 mots.

Si tous les vocables de A ont une fréquence de 1, et que ces mêmes vocables se retrouvent tous dans B — la moitié avec une fréquence 1 et l'autre moitié avec une fréquence 2 — et dans C (fréquence 9 ou 10) : la distance A-B calculée, selon (3), sera d'environ 0,3 et celle de A-C : 0,05. Ces deux distances ne devraient-elles pas être nulles puisque les textes comparés sont aussi proches que possible ?

A structure semblable, les distances calculées sur des petits textes risquent donc d'être supérieures à celles calculées sur des textes beaucoup plus grands. Pour limiter partiellement ce premier inconvénient, on élimine du numérateur toutes les distances inférieures à 0.5.

— la déformation de la distribution des fréquences avec l'augmentation de la taille des textes. Dans un texte bref, les basses fréquences occupent toujours une étendue importante. Plus la taille augmente, plus les fréquences élevées prennent de l'importance. Autrement dit, l'évolution de la fréquence des mots en fonction de la dimension n'est pas tout à fait une fonction arithmétique simple comme le postule notre calcul.

Pour limiter ce second inconvénient, nous proposons d'inscrire ce calcul dans un certain intervalle : seuil inférieur de 1.000 mots et différence de taille inférieure à 1/12 entre le plus petit et le plus grand des textes soumis à comparaison. Dans cet intervalle, aucune corrélation entre taille et distance n'a pu être observée jusqu'à maintenant.

En revanche, les tests déjà effectués conduisent à repousser la solution consistant à arrondir à l'unité les valeurs théoriques car l'arrondi introduit des « effets de seuils » qui peuvent être significatifs quand les textes sont petits et de tailles pas trop différentes.

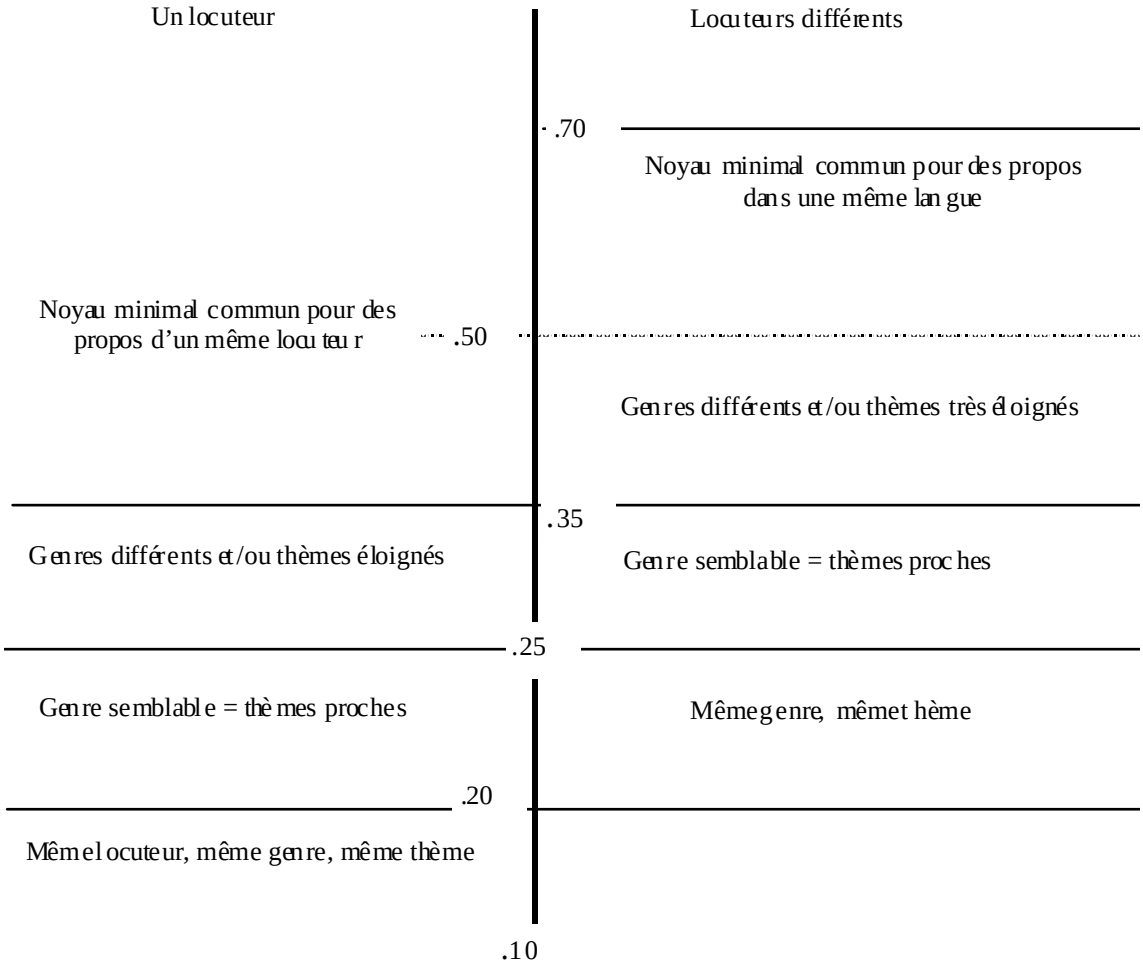
Applications

L'interprétation des résultats est d'une grande simplicité. Par exemple, un indice de .50 signifie que l'intersection des deux textes comparés peut être estimée à la moitié de leur surface totale ; un indice de .25 qu'ils partagent les trois quarts de leurs mots (ou qu'ils en ont un quart de différents), etc. Il devient possible d'établir une échelle des distances qui sera notamment utile pour la classification des entretiens.

Le calcul a été appliqué à divers corpus comptant au total près d'une dizaine de millions de mots (tous dépouillés selon la même norme) : les allocutions du général de Gaulle, de F. Mitterrand, les discours des Premiers ministres canadiens, québécois et français depuis 1945 (Labbé-Monière, 2000), les entretiens radio-télévisés de plusieurs hommes politiques, divers romans des trois derniers siècles (en collaboration avec E. Brunet), des éditoriaux de la presse syndicale (Brugidou-Labbé, 1999), des articles de la presse économique, des transcriptions d'entretiens d'enquêtes (Bergeron-Labbé, 2000)…

Toutes ces expériences ont permis d'établir empiriquement une échelle des distances. Nous présentons ici celle qui a été étalonnée sur les entretiens (Tableau I ci-dessous). Deux dimensions produisent de la distance : le genre (discours familier, soutenu, technique...) et le thème. Enfin, pour un même locuteur, les distances sont toujours inférieures à celles qui peuvent exister entre deux locuteurs différents et contemporains (quand ils traitent d'un même thème), d'où la séparation du tableau en deux parties.

Tableau I. Echelle normalisée des distances entre entretiens



- Les distances inférieures à .20 ne se constatent pas chez deux locuteurs différents (pour des propos appartenant à un même genre avec des thèmes semblables). En dessous de ce seuil, on peut être certain que l'un des deux s'est « inspiré » de l'autre…
- Entre 0,20 et 0,25 la parenté entre les propos est très forte. Si le locuteur est unique, les thèmes changent légèrement. En revanche, pour des locuteurs différents, on peut être certain que le genre et le thème sont semblables ;
- Entre 0,25 et 0,35 un même locuteur a probablement changé de registres et de thèmes. En revanche des locuteurs différents demeurent relativement proches quant au thème et au genre. Ce n'est qu'au-dessus de ce seuil que l'absence de parenté entre les propos devient certaine.

Naturellement, ces seuils ne doivent pas faire oublier que la distance varie uniformément et non par sauts : plus elle augmente, plus les propos sont différents dans leur forme et sur le fond.

Appliqué au corpus "confort électrique", le calcul donne une série d'indications intéressantes. Par exemple, il permet de détecter les entretiens les plus proches et les éloignés (tableau 2). On voit que les paroles les plus semblables ont pratiquement toutes été émises par des locataires de HLM. Par exemple, les propos les plus proches sont tenus par deux couples en situation économique précaire, maîtrisant mal leur chauffage et attendant leurs factures avec une certaine appréhension… (HLM 09 et 12). A l'inverse les propos de ces mêmes locataires semblent nettement séparés de ceux tenus par certains propriétaires de maison

individuelles équipés de planchers chauffants ou "domotisés"). Cela suggère déjà une piste d'analyse.

Tableau 2. Les couples d'entretiens proches et éloignés

Les dix couples d'entretiens les plus proches			Les dix couples d'entretiens les plus éloignés		
Chau HLM 09	Chau HLM 12	0,218	Chau HLM 08	PCR 02	0,470
Clim HLM 02	Chau HLM 12	0,229	Chau HLM 09	PCR 02	0,471
Chau HLM 05	Chau HLM 08	0,230	Fidel 02	Dom 03	0,480
Chau HLM 07	Chau HLM 12	0,230	Clim HLM 04	PCR 02	0,480
Chau HLM 08	Chau HLM 12	0,231	Clim HLM 03	PCR 02	0,484
Fidel 16	Fidel 05	0,232	Clim Pav 08	PCR 02	0,487
Chau HLM 08	Chau HLM 09	0,234	Fidel 02	PCR 02	0,494
Clim HLM 02	Chau HLM 08	0,237	Clim Pav 01	PCR 02	0,496
Chau HLM 05	Chau HLM 12	0,239	Clim Pav 01	Chau HLM 04	0,497
Clim HLM 05	Chau HLM 14	0,239	Chau HLM 10	PCR 02	0,499

Tableau 3. Les entretiens centraux et décalés

Les plus centraux			Les plus décalés		
N°	Titre	Distance	N°	Titre	Distance
1	Chau HLM 13	0,297	55	PCR05	0,348
2	PCR 08	0,299	56	Fidel 13	0,355
3	Fidel 06	0,300	57	Clim Pav 01	0,356
4	Clim HLM 02	0,307	58	Fidel 10	0,360
5	Chau HLM 12	0,310	59	Fidel 15	0,366
6	Chau HLM 14	0,311	60	PCR 01	0,369
7	Clim HLM 05	0,311	61	Fidel 02	0,373
8	Dom 01	0,312	62	Chau HLM 04	0,396
9	Clim Pav 06	0,312	63	PCR 02	0,420
10	Fidel 16	0,312	64	Dom 03	0,423

En cas de corpus important, cette démarche devient vite impraticable. Par exemple, pour les entretiens sur le "confort électrique", il y a 2016 couples possibles à envisager et ce nombre croit exponentiellement avec la taille des corpus. Une autre démarche possible consiste à calculer, pour chaque entretien, la moyenne des distances les séparant de tous les autres. Ceux dont les scores sont les plus faibles peuvent être considérés comme les textes "centraux" et, à l'inverse, plus la distance moyenne augmente, plus le texte est décalé (tableau 3).

Naturellement, il faut ensuite retourner aux textes pour comprendre les raisons qui placent chacun de ces entretiens plus ou moins loin du centre de gravité. Par exemple, deux entretiens se situent au-dessus de 0.40 et sont donc très décalés par rapport aux autres. Dom03 et PCR02 sont des textes brefs (1 144 et 2 089 mots), tout comme Fidel 02 qui figure également en bas du tableau : en présentant le calcul de la distance, nous avons indiqué que la distance pouvait être légèrement plus élevée pour les plus petits textes à cause de l'importance des basses fréquences. Cependant, la petite taille n'explique qu'une faible part de l'écart total (d'autres textes brefs ne sont pas aussi décalés). L'examen de ces textes permet de déceler les véritables causes du décalage. Pour Dom03, le texte le plus décalé, il s'agit d'un couple de cadres supérieurs d'une institution internationale dont les réponses sont presque "diplomatiques" : style écrit, phrases sobres, sans tournures familières ni interjections… Une autre enquêtée se trouve un peu dans ce cas : la responsable de la société d'économie mixte maître d'ouvrage des HLM chauffés à l'électricité (Chau HLM 04) qui a manifestement l'habitude de la "communication". PCR01 et 02 n'ont pas été retranscrits comme les autres entretiens : les questions ont été effacées et les réponses ont été expurgées des hésitations, répétitions et

reprises propres à l'oral. Sans doute ces deux textes sont-ils fidèles dans l'esprit aux propos des enquêtés mais probablement pas à leur style… Enfin, Fidel 02 est un entretien avec une assistante sociale d'Edf qui déclare d'emblée n'avoir rien à dire sur le sujet de l'enquête et qui se contente ensuite de quelques réponses convenues… Autrement dit, le calcul a détecté les individus "aberrants" (qui s'écartent des autres à la fois par le contenu et par le style).

Connaissant maintenant les proximités relatives, on peut se demander s'il existe certains groupes d'entretiens proches ou s'opposant à d'autres. Diverses techniques de classification permettent de constituer ces groupes sans faire intervenir la subjectivité de l'observateur.

Classifications

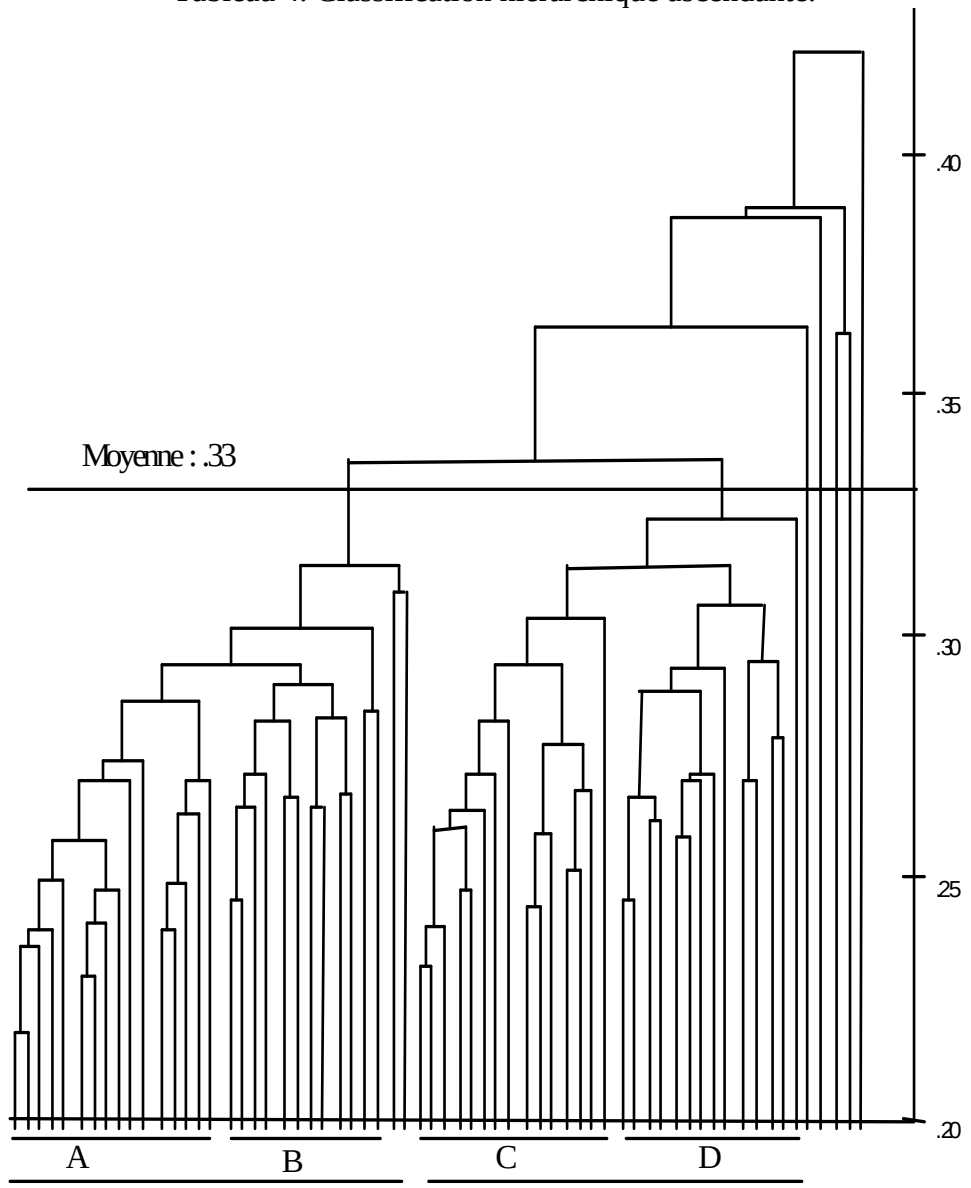
En principe, classer les individus appartenant à une même population consiste à rechercher la meilleure partition possible, c'est-à-dire celle qui minimisera les distances à l'intérieur des groupes constitués et qui maximisera la distance séparant chacun de ces groupes (Sneath et Sokal, 1973 ; Benzecri, 1980 ; Tomassone, 1988). Plusieurs méthodes sont possibles. Nous commencerons par la plus classique (la classification automatique ascendante) avant d'évoquer l'analyse arborée.

La **classification automatique ascendante** est opérée sur la matrice des distances. L'algorithme procède à la construction d'une classe en regroupant les deux textes séparés par la distance la plus faible (ici les entretiens des deux locataires mentionnés ci-dessus), puis il recalcule les distances des autres textes par rapport à ce nouvel ensemble, etc. Et ceci jusqu'à la constitution d'un ensemble unique. Ces regroupements successifs sont représentés dans le dendrogramme ci-dessous. En absisces, figurent, de gauche à droite, les agrégations successives et, en ordonnées, les distances correspondantes à ces regroupements. Le couple formé par les deux entretiens les plus proches figure donc à gauche et la barre horizontale qui les relie est la plus basse du graphe. A l'opposé, l'entretien le plus décalé figure à droite et ne rejoint les autres qu'en dernier et à la hauteur maximale. On ne sera pas surpris de trouver, à part sur la droite du graphe, les cinq entretiens mentionnés ci-dessus qui ne rejoignent les autres qu'au-dessus du seuil de .35.

En dehors de ces cinq individus aberrants, tous les autres sont "mariés" de la "meilleure façon" possible (en utilisant un procédé dichotomique simple). En coupant le graphe, horizontalement et au plus près de l'un des barreaux de l'échelle standardisée, on pourra isoler les groupes de textes très proches, relativement proches, etc. Naturellement, plus l'on s'élève dans le graphe, plus les classes constituées sont hétérogènes et plus l'interprétation des différences devient complexe.

En coupant le graphe ci-dessous autour de la moyenne, on obtient deux grands groupes (notés I et II). En coupant un peu plus bas, on peut en distinguer trois ou quatre (notés A, B, C, D). On remarque simplement que A et B sont plus proches entre eux que C et D…

Tableau 4. Classification hiérarchique ascendante.



De gauche à droite :

I. Groupe A
CHAU_HLM09
CHAU_HLM12
CLIM_HLM02
CHAU_HLM07
CHAU_HLM10
CHAU_HLM05
CHAU_HLM08
CLIM_HLM01
CLIM_HLM06
CLIM_PAV08
CLIM_HLM03
CLIM_HLM05
CHAU_HLM14
CHAU_HLM13
DOM02
DOM04

I. Groupe B
CLIM_PAV02
CLIM_PAV06
DOM01
CLIM_PAV05
CLIM_PAV07
CLIM_PAV10
CLIM_PAV09
CLIM_HLM07
CLIM_PAV01
CLIM_PAV03
CHAU_HLM06
CHAU_HLM11
CLIM_PAV04
CLIM_HLM04

II. Groupe C
FIDEL16
FIDEL06
FIDEL07
FIDEL05
FIDEL04
FIDEL09
FIDEL03
FIDEL08
FIDEL17
FIDEL12
FIDEL01
FIDEL10
FIDEL14
FIDEL15
FIDEL13

II. Groupe D
PCR08
PCR11
PCR04
PCR05
PCR03
PCR10
PCR06
PCR07
PCR09
CHAU_HLM01
CHAU_HLM03
CHAU_HLM02
CHAU_HLM15
FIDEL11

FIDEL02
CHAU_HLM04
PCR01
PCR02

DOM03

La moitié gauche du graphe regroupe tous les "usagers" : les locataires du HLM "haut de gamme" de la région parisienne chauffés à l'électricité (8 se trouvent dans le groupe A et 2 dans le groupe B) ; tous les locataires du HLM du Midi — également haut de gamme mais climatisés (et chauffés) à l'électricité — avec les propriétaires de pavillons (ou d'appartements) équipés d'une climatisation. On trouve, également dans ce groupe, trois entretiens de l'enquête "domotique" : ces personnes utilisent la programmation pour piloter leur chauffage ou leur climatisation…

La moitié droite du graphe rassemble, d'un côté, tous les agents EDF interrogés dans le cadre de l'enquête "fidélisation" — on remarque l'homogénéité relative de ce groupe, avec un peu à l'écart l'entretien n°11… — et de l'autre tous les propriétaires d'un lotissement équipé de "planchers chauffants et rafraîchissants" (sauf les deux dont les transcriptions n'ont pas suivi le modèle général).

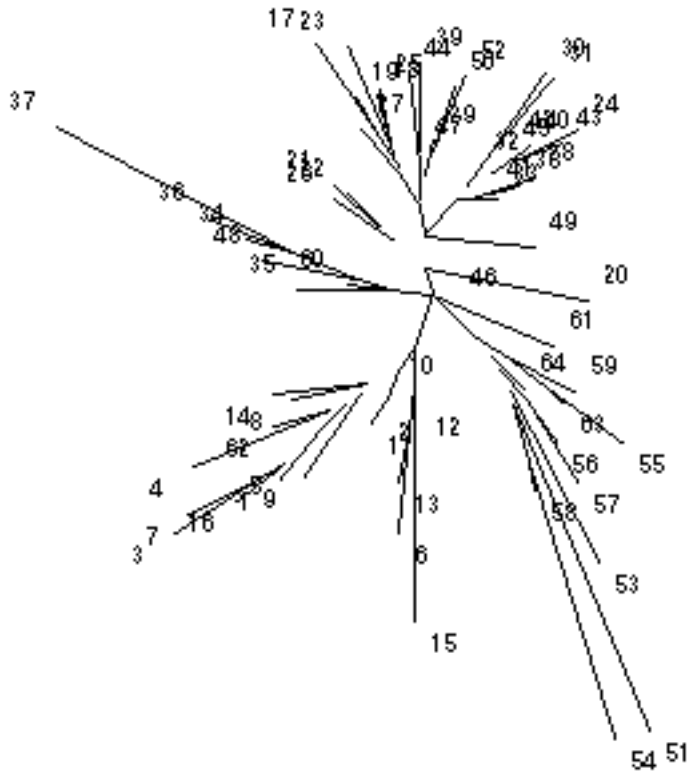
Enfin, dans le groupe D, on remarque 4 entretiens à propos du HLM de la banlieue parisienne (Chau HLM 1, 2, 3, 15) : le fournisseur du procédé de régulation du chauffage, un employé de l'office HLM, l'entreprise qui a posé le chauffage et le responsable EDF qui suit l'expérience . Ces 4 là se retrouvent assez loin des gens qu'ils gèrent mais ils sont bien groupés ensemble et se rattachent finalement aux propriétaires équipés de PCR. L'examen du vocabulaire permet de comprendre la raison de ce regroupement (les thèmes développés sont les mêmes : conception du logement, maîtrise financière et technique des procédés)…

La classification automatique traditionnelle est donc un outil puissant mais elle présente certains inconvénients : elle produit parfois des "effets de chaîne" (conduisant alors à des graphes "en escalier") ; elle efface certaines proximités entre textes en supprimant les sommets qui les relient à cause d'agrégations à un niveau inférieur. Ainsi, certaines questions demeurent sans réponse : n'y-a-t-il aucun lien entre agents EDF et "usagers" ? les propriétaires de l'enquête Clim sont-ils tous plus proches des locataires de HLM ou certains n'auraient-ils pas des affinités avec des propriétaires de l'enquête PCR ? etc. Pour tenter d'éclairer ces questions, sans avoir à explorer en détail la matrice des distances, on utilise des procédés de classification plus complexes, notamment l'analyse topologique et la représentation arborée (Barthélémy-Guénoche, 1988 ; Barthélémy-Luong, 1998 ; Luong, 1994).

Il s'agit d'obtenir, dans un plan, la meilleure représentation possible des distances de chacun des textes à tous les autres. Chaque texte est représenté par une feuille terminale de l'arbre. La distance qui le sépare d'un autre est matérialisée par la longueur du chemin à parcourir sur l'arbre pour unir ces deux textes : dans la figure ci-dessous, la plus courte distance sépare les entretiens n° 43 et 46 (Chau HLM09 et 12) qui sont quasiment confondus en haut à droite du graphe, la plus longue sépare les trois textes les plus décalés du corpus : n° 54 51 (PCR 2, Dom 03) en bas à droite et 37 (chau HLM 04), à l'opposé. Les branches convergent vers les principaux nœuds et forment ainsi des groupes relativement homogènes (Tableau 5 ci-dessous).

X. Luong a aimablement tracé ce graphe à partir de la matrice des distances intertextuelles. Attention : ce graphe a été obtenu en grossissant les faibles écarts relatifs séparant les groupes de textes. Pour cela, on utilise le « théorème » selon lequel la topologie de l'arbre est inchangée lorsque l'on retranche de toutes les distances une quantité légèrement inférieure à la plus petite d'entre elles (ici la distance entre les entretiens 43 et 46 soit .20), de même que, dans le dendrogramme ci-dessus, l'origine a été placée à .20 et non à zéro. Cette valeur n'est d'ailleurs pas choisie au hasard : à l'oral elle semble être la plus petite distance concevable entre deux propos émis sur le même thème par des personnes différentes ne s'étant pas concertées avant l'entretien. Pour les arbres, cette opération ne change pas la disposition relative des textes dans le plan, mais elle réduit la longueur des branches reliant les « feuilles » terminales et grossit les « troncs », c'est-à-dire les sections unissant les principaux nœuds.

Tableau 5. Classification arborée des entretiens sur le confort électrique
(d'après les distances intertextuelles diminuées de .20)



Le graphe isole bien les 3 groupes de textes déjà repérés sur le dendrogramme mais il apporte quelques informations supplémentaires. En premier lieu, on constate une quasi-équidistance de ces groupes indiqués par la position de leurs nœuds respectifs par rapport au tronc central. Deuxièmement, le graphe confirme la grande proximité des groupes A et B, qui peuvent donc être rassemblés en un seul bloc (les "usagers"), et l'homogénéïté du groupe C (les agents EDF). Dans le détail, on trouve confirmation du décalage très grand, par rapport aux autres textes, du troisième entretien de l'enquête Domotique et des deux premiers de l'enquête "PCR" (qui ont été partiellement réécrits) : on trouve ainsi confirmation de l"étrangeté" de ces trois textes mais aussi de leur parenté (ils sont assez proches du style écrit). Enfin, à gauche, dans une feuille très étirée, on retrouve tous les gestionnaires du HLM de la banlieue parisienne, à l'écart du reste.

Autrement dit, l'analyse isole nettement quatre familles de discours sur le "confort électrique". Deux sont assez homogènes : les usagers et les agents d'EDF. Deux sont un peu plus hétérogènes : les propriétaires de planchers chauffants et les gestionnaires de HLM.

Ces résultats appellent un certain nombre de remarques :

— il ne faut pas considérer ces résultats comme triviaux : la lecture "flottante" ne permet pas de dire que le discours des usagers de la climatisation est semblable à celui des utilisateurs du chauffage électrique, ou encore qu'il n'y a pas de différence nette entre les propriétaires et les locataires, lorsqu'ils utilisent un appareillage classique, ou entre les "commerciaux" et les "techniciens" d'EDF ou entre les locuteurs en fonction de leurs catégories professionnelles ;

— l'influence du questionnaire ? Le tableau 3 montre qu'il n'y a pas de corrélation entre l'enquête et le caractère plus ou moins décalé des propos, ce sont donc très probablement des caractéristiques individuelles (thématiques et stylistiques) qui placent les individus plus ou moins loin les uns des autres.

— l'influence des enquêteurs ? En fait, deux enquêteurs sont intervenus, l'un dans trois enquêtes et l'autre dans deux, et les textes de ces entretiens entrent dans des groupes différents (PavClim, ClimHLM et PCR pour l'un et ClimHLM et Dom pour l'autre).

Autrement dit, lorsque les entretiens ont été conduits en respectant les règles de l'art — ce qui est le cas en l'espèce — la classification restitue bien le contenu de la conversation et, secondairement, le style de l'enquêté. Pour mieux appréhender ces contenus, il faut examiner les vocabulaires des différents groupes.

Spécificités du vocabulaire

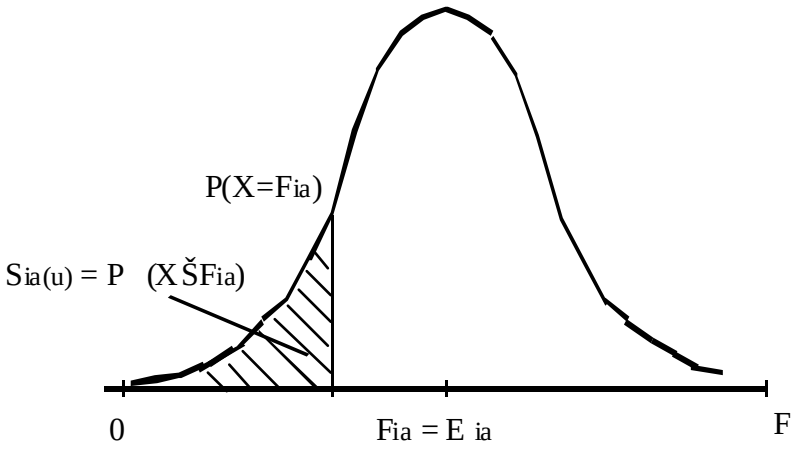
L'étude du vocabulaire des parties d'un corpus suit le raisonnement présenté par P. Lafon (Lafon, 1984 et pour une discussion : Labbé et Labbé, 1994). Un vocable appartient au **vocabulaire spécifique** d'une partie d'un corpus lorsque sa fréquence, dans cette partie, s'écarte *significativement* de celle observée dans l'ensemble du corpus. S'il y en a de "trop", on dit que le vocable est une *spécificité positive* ; dans le sens contraire, la spécificité est dite *négative*. En revanche, si l'écart n'est pas significatif, dans aucune des parties du corpus, on dit que le vocable est "non-spécifique" (on peut le supposer commun à tous les sous-corpus). Autrement dit, la norme de calcul est le corpus entier et la spécificité du vocabulaire de chaque partie est jugée par rapport à ce tout.

En reprenant la notation ci-dessus : N taille du corpus, N_a : taille du groupe A. Un vocable i, ayant une fréquence F_i dans le corpus entier, aura une espérance mathématique dans A égale à $E_{ia(u)}$ calculée selon la même procédure que pour la distance. Si la fréquence constatée (F_{ia}) est différente de la fréquence attendue ($E_{ia(u)}$), quand peut-on dire que le vocable considéré appartient au vocabulaire spécifique du groupe A ?

La réponse réside dans la probabilité de l'événement observé par rapport à l'événement attendu. La variable aléatoire X suit une loi hypergéométrique de paramètres N, N_a, F_i et F_{ia} :

$$(4)\quad P(X = F_{ia}) = \frac{C_{Fia}^{Fi}\, C_{Na - Fia}^{N - Fi}}{C_{Na}^{N}}$$

F_{ia} peut varier entre 0 — aucune occurrence du vocable dans A — et F_i : toutes les occurrences du vocable sont observées dans A ($0 \leq F_{ia} \leq F_i$). A condition que N, F et N_a soient suffisamment grands, les valeurs de X se distribueront selon la fameuse "courbe en cloche", avec un mode pour $F_{ia} = E_{ia(u)}$ (graphique ci-dessous). La spécificité du vocable considéré sera la probabilité pour que la fréquence observée soit **au moins** égale à F_{ia}. C'est la surface comprise sous la portion de courbe allant de 0 à F_{ia} (graphe ci-dessous).



A condition que E_{ia} soit assez grand, S tendra vers zéro quand F_{ia} sera très petite par rapport à E_{ia} et vers 1 dans le cas inverse, avec comme valeur remarquable : S = 0.5 quand $F_{ia} = E_{ia(u)}$.

On peut donc suivre le raisonnement probabiliste classique : un vocable i suit des lois de distribution différentes dans le corpus entier et dans le sous-corpus lorsque S s'écarte du mode de la courbe au-delà, ou en deça, de deux écarts-types, si l'on choisit une marge d'incertitude de moins de 5%, ou de trois écart-types avec une marge inférieure à 1%. Autrement dit, un vocable i est significativement suremployé dans A lorsque $S_{ia(u)}$ est supérieure à .95 ou à .99 et significativement sousemployé lorsque $S_{ia(u)}$ est inférieure à .05 ou .01. Et l'on pourra considérer que sa spécificité est d'autant plus significative que l'indice sera plus proche de zéro ou de 1 (pour les limites du raisonnement, voir Labbé et Labbé, 1994).

Appliquée aux entretiens du groupe A, la méthode permet de caractériser le vocabulaire du noyau central des "usagers" (la liste complète des spécificités est donnée en annexe II). On constate, en premier lieu, que le poids des différentes catégories grammaticales s'écarte assez significativement des autres entretiens (tableau 5 ci-dessous)

Tableau 5. Densité comparée des catégories grammaticales employées par les usagers comparés aux autres entretiens.

Catégories	A Corpus-Sous corpus	B Sous corpus	A/B	B-A%
Verbes	18.6	19.5	104.4	+ 4.4
Formes fléchies	*13.2*	*14.3*	*108.1*	*+ 8.1*
Participes passés	*2.2*	*2.5*	*114.8*	*+ 14.8*
Participes présents	*0.1*	*0.1*	*85.5*	*- 14.5*
Infinitifs	*3.1*	*2.5*	*81.8*	*- 18.2*
Substantif	14.8	12.5	84.6	- 15.4
Noms propres	*0.6*	*0.3*	*52.2*	*- 47.8*
Adjectif	3.5	3.4	95.4	- 4.6
Adj participe passé	*0.3*	*0.2*	*86.3*	*- 13.7*
Pronom	18.4	20.8	112.5	+ 12.5
Pronoms personnels	*10.0*	*11,7*	*117.0*	*+17.0*
Determinant	13.3	11.8	88.6	- 11.4
Nombres	*1.8*	*1.8*	*99.4*	*- 0.6*
Adverbes	11.6	13.4	115.4	+ 15.4
Prépositions	11.3	9.4	83.8	- 16.2
Conjonction	7.1	8.1	114.1	+ 14.1

En fait le discours des usagers est celui qui "colle" le plus à la forme "question-réponse" : les adverbes "oui", "non" et "ne… pas" sont en effet les spécificités les plus remarquables. De même, le discours des "usagers" se caractérise par un excédent significatif du groupe verbal (verbes, pronoms, adverbes). Par rapport aux autres, c'est un discours orienté vers l'action (en premier lieu : *mettre le radiateur, la clim.… être* (ou *avoir*) *chaud* ou *froid…*), et fortement personnalisé (excédent de pronoms personnels et d'abord de la première personne) mais aussi orienté vers le passé (15% de participes passés en plus). En définitive, ces usagers livrent un récit de leur vie quotidienne et (parfois) de leurs difficultés à utiliser les appareils.

En revanche, les formes "dégradées" du verbe — notamment l'infinitif qui est à la frontière entre le verbal et le nominal ou qui signale un discours modalisé — sont nettement sous-employées. Logiquement, on constate un déficit significatif du groupe nominal (substantifs, adjectifs et déterminants) et tout spécialement des noms propres. On dit plus volontiers "ici" ou "là" et l'on ne nomme pas beaucoup les lieux en dehors de la ville d'habitation ; on répugne à employer les marques ; les appareils, comme les entreprises, ou les institutions sont désignés par leur nom commun générique : le *radiateur*, le *clim(atiseur)*, la *machine* ou l'*horloge*, le

gérant, l'*électricité* (pour *EDF*), alors que dans l'autre groupe, on nomme plus volontiers les marques, les institutions et l'on utilise la terminologie technique qui se retrouve toute entière dans les spécificités négatives du vocabulaire des usagers.

Une fois établi le vocabulaire spécifique, le programme relit les entretiens considérés en recherchant les phrases qui contiennent le plus de spécificités positives et le moins de spécificités négatives. Deux tris sont opérés : sur les scores absolus — généralement ce sont des phrases longues qui sont repérées — puis les scores relatifs (scores absolus rapportés au nombre de mots de la phrase) : on repère ainsi les petites phrases "slogans". En quelque sorte, ces phrases sont des "archétypes" qui concentrent en quelques mots plusieurs des thèmes caractéristiques du groupe. Elles figurent à la fin de l'annexe. Ces phrases doivent être interprétées avec prudence :le logiciel repère les vocables spécifiques sans préjuger du sens.

En se limitant aux frontières du corpus, ce raisonnement laisse une question en suspens : le discours tenu par les enquêtés est-il singulier par rapport au reste de la société française et, si oui, en quoi se singularise-t-il ? Pour répondre à cette question, il faudrait disposer d'un échantillon aussi large que possible du français tel qu'il est parlé par nos contemporains. Un tel échantillon n'existe pas et, à notre connaissance, la question n'a jamais été abordée par les linguistes et les sociologues français, depuis la fameuse expérience du "français fondamental" (Gougenheim et Al, 1964).

A défaut de disposer d'un tel étalon, nous proposons une expérience plus modeste : la confrontation d'une partie du corpus — ici la quinzaine d'entretiens des employés d'EDF (groupe C) — avec une base d'entretiens constituée grâce à des travaux antérieurs. Il s'agit de 110 transcriptions, réalisées en suivant les principes exposés ci-dessus, et comportant au total 885.000 mots : huit entretiens radio-télévisés — C. de Gaulle, F. Mitterrand, J. Chirac (Labbé, 1990) — 35 entretiens sur les Français et la politique (réalisés par S. Pionchon en 1994), 15 entretiens sur la vie scolaire et les conduites à risque chez les adolescents (remis par N. Leselbaum et C. de Peretti), 36 entretiens sur les relations professionnelles et la négociation collective au Québec (Bergeron-Labbé, 2000), etc. Même s'il s'agit du plus grand corpus étiqueté existant sur le français oral, il n'atteint pas la taille ni la diversité qu'exigerait un véritable échantillon représentatif. Aussi faut-il prendre les résultats de la confrontation avec prudence (annexe III et tableau 6 ci-dessous).

Tableau 6. Densité comparée des catégories grammaticales employées par les agents EDF comparées au "français oral contemporain".

Catégories	A Corpus de référence	B Corpus étudié	B/A (%)	B - A (%)
Noms propres	0.6	0.5	88.1	- 10.9
Verbes	19.5	18.7	95.7	- 4.3
Formes fléchies	*13.4*	*13.0*	*97.6*	*- 2.4*
Participes passés	*2.7*	*1.8*	*65.8*	*- 34.2*
Participes présents	*0.1*	*0.1*	*97.6*	*- 2.4*
Infinitifs	*3.3*	*3.7*	*112.8*	*+ 12.8*
Substantifs	14.1	14.9	106.0	+ 6.0
Adjectifs	3.5	3.6	100.4	+ 0.4
Adj participe passé	*0.4*	*0.2*	*67.4*	*- 32.6*
Pronoms	19.2	18.4	96.0	- 4.0
Pronoms personnels	*10.1*	*9.7*	*96.5*	*- 3.5*
Determinants	13.3	12.9	96.8	- 3.2
Nombres	*1.3*	*1.1*	*92.5*	*- 7.5*
Adverbes	10.5	11.2	106.8	+ 6.8
Prépositions	11.9	11.8	99.2	- 0.8
Conjonctions	6.6	7.4	112.2	+ 12.2

En premier lieu, les propos des employés d'EDF, interrogés sur la fidélisation au chauffage électrique, ne s'écartent pas considérablement de "l'usage moyen". Ils présentent malgré tout un net déficit en noms propres et en chiffres. Or les noms propres, comme les dates et les chiffres fournissent une bonne partie de l'ancrage spatial et temporel du discours. Le déficit suggère donc une légère mais réelle propension à se réfugier dans les généralités. Cette tendance est confirmée par le déficit du verbe — excepté sous sa forme infinitive qui est la plus proche du groupe nominal, ou du discours de principe ("il va falloir faire" semble être l'une des expressions les plus caractéristiques) — et par la faiblesse relative de la personnalisation du propos et de l'engagement du locuteur dans ce qu'il dit : le pronom "je" figure presque en tête des spécificités négatives alors que "on" et "nous" figurent dans les pronoms les plus employés. Enfin, on notera le fort excédent des adverbes et, en premier lieu, de la négation : le propos est souvent construit en réaction contre des choses déjà dites. Les phrases les plus caractéristiques de l'annexe III montrent que cette réaction vise à la fois l'image négative du chauffage électrique dans l'opinion mais aussi peut-être certaines réformes… Mais, en définitive, une telle attitude ne saurait surprendre : interrogé sur leur lieu de travail par un enquêteur venant au nom du "centre", il est logique que l'enquêté soit relativement prudent dans ses propos même si son anonymat est garanti.

Naturellement, il peut sembler logique de trouver, en tête des spécificités, des mots comme EDF, *électricité* ou *électrique*… Il est peut-être moins banal, étant donné la culture de l'entreprise, de trouver également des vocables comme *client*, *clientèle* ou *commercial*, etc. Au fond, l'outil rend exactement les services qu'on peut attendre de lui. Il est intéressant de disposer d'une synthèse des principaux thèmes caractéristiques développés dans une collection de textes avant de les dépouiller en détail (les 130.000 mots de ce groupe d'enquêtés représentent près de 500 pages dactylographiées en double interligne, soit une vingtaine d'heures d'entretiens). On peut même considérer que l'instrument n'est pas assez discriminant et souhaiter des listes plus courtes, ou plus clairement hiérarchisées. C'est une piste à explorer.

Remarques conclusives.

Tous ces calculs exigent au préalable que les graphies des textes comparés aient été normalisées mais aussi, à notre avis, que les mots aient été « lemmatisés » (rattachés à leurs entrées de dictionnaire ou « vocables »). En effet tout calcul de distance exige au préalable que l’on se mette d’accord sur des unités de mesure un peu comparables au mètre étalon…

Notre présentation est trop brève pour pouvoir évoquer la richesse des pistes ouvertes par ces méthodes. Il faudrait encore examiner les associations entre les mots, les principaux univers lexicaux, les indices stylistiques comme la richesse ou la diversité du vocabulaire…

De même, la « contribution à la distance » des différentes catégories grammaticales (substantifs, verbes, pronoms…) apporte un éclairage intéressant sur les mécanismes de la langue. Par exemple, la distance intertextuelle peut apporter une réponse scientifique à la fameuse question de savoir si le langage diffère selon les classes sociales et le "capital" culturel des individus. Par exemple, dans les 64 entretiens sur le confort électrique, on observe aucune distance significative en fonction de ce critère hormis pour deux cadres supérieurs d'une organisation internationale et d'un office de HLM (dont les propos sont d'ailleurs très brefs). Naturellement, ces différences seraient probablement apparues si l'on avait prolongé les entretiens en passant à des sujets plus abstraits ou si l'on avait demandé aux enquêtés de consigner par écrit certaines de leurs observations. Mais, pour la langue de la vie quotidienne, cette fameuse "misère" (lexicale) des classes populaires est peut-être une pré-notion de la sociologie contemporaine.

Annexe 1

Le corpus des entretiens

	N°	Enquête	Taille (N)	Formes normalisées	Vocables
1	1174	FIDEL17	5 911	1 072	818
2	1175	FIDEL16	9 017	1 190	808
3	1176	FIDEL15	10 377	1 505	1 076
4	1177	FIDEL14	9 925	1 204	802
5	1178	FIDEL13	5 254	1 023	752
6	1179	FIDEL12	7 740	1 338	1 004
7	1180	FIDEL11	4 776	707	514
8	1181	FIDEL10	8 506	1 008	693
9	1182	FIDEL09	7 880	1 261	919
10	1183	FIDEL08	7 371	1 186	867
11	1184	FIDEL06	11 026	1 661	1 188
12	1185	FIDEL07	9 263	1 125	796
13	1186	FIDEL05	5 698	932	688
14	1187	FIDEL04	5 252	838	629
15	1188	FIDEL03	8 519	1 243	863
16	1189	FIDEL02	2 220	467	344
17	1190	FIDEL01	8 711	1 505	1 100
18	4102	CLIM_PAV01	2 409	472	357
19	4103	CLIM_PAV02	5 473	931	699
20	4104	CLIM_PAV03	5 597	862	640
21	4105	CLIM_PAV04	4 225	840	647
22	4106	CLIM_PAV05	4 749	856	625
23	4107	CLIM_PAV06	8 567	1 333	961
24	4108	CLIM_PAV07	4 242	760	558
25	4109	CLIM_PAV08	7 236	1 031	704
26	4110	CLIM_PAV09	3 765	751	556
27	4111	CLIM_PAV10	5 464	935	687
28	5127	CLIM_HLM07	3 590	628	458
29	5128	CLIM_HLM06	5 651	849	635
30	5129	CLIM_HLM05	6 605	997	717
31	5130	CLIM_HLM04	3 712	713	547
32	5131	CLIM_HLM03	5 015	799	590
33	5132	CLIM_HLM02	8 204	1 190	841
34	5133	CLIM_HLM01	7 428	957	692
35	5184	CHAU_HLM01	7 168	1 176	862
36	5185	CHAU_HLM02	6 063	1 099	828
37	5186	CHAU_HLM03	5 487	1 027	778
38	5187	CHAU_HLM04	2 034	636	499
39	5188	CHAU_HLM05	5 895	983	741
40	5189	CHAU_HLM06	3 066	616	451
41	5190	CHAU_HLM07	6 197	961	704

42	5191	CHAU_HLM08	8 312	1 008	700
43	5192	CHAU_HLM09	6 380	876	597
44	5193	CHAU_HLM10	4 986	809	537
45	5194	CHAU_HLM11	3 352	662	480
46	5195	CHAU_HLM12	8 347	1 073	766
47	5196	CHAU_HLM13	10 538	1 500	1 053
48	5197	CHAU_HLM14	6 336	842	604
49	5198	CHAU_HLM15	6 719	1 137	848
50	6112	DOM01	8 849	1 455	1 057
51	6113	DOM02	6 960	947	681
52	6114	DOM03	1 144	344	274
53	6115	DOM04	8 405	1 307	953
54	6691	PCR01	2 540	657	517
55	6692	PCR02	2 089	603	485
56	6693	PCR03	2 270	607	482
57	6694	PCR04	6 053	1 113	830
58	6695	PCR05	5 047	989	741
59	6696	PCR06	4 165	928	709
60	6697	PCR07	3 189	773	603
61	6698	PCR08	19 889	2 151	1 493
62	6699	PCR09	3 638	769	585
63	6700	PCR10	4 971	975	714
64	6701	PCR11	4 550	985	763
	Total		**394 017**	**11 598**	**6 602**

FIDEL : enquête auprès d'agents EDF à propos de la fidélisation au chauffage électrique.

CLIM_PAV : enquête auprès de propriétaires de pavillons ou d'appartements équipés d'une climatisation réversible individuelle.

CLIM_HLM : enquête auprès de locataires d'un HLM d'une ville du Midi, équipés d'une climatisation réversible individuelle.

CHAU_HLM : enquête auprès de locataires d'un HLM de la banlieue parisienne équipés d'un chauffage électrique individuel avec programmation. NB : les numéros 1,2,3,4 et 15 sont des responsables de l'opération (concepteur, maître d'ouvrage, gestionnaire et agent EDF).

DOM : enquête auprès de propriétaires d'habitations équipés d'un gestionnaire programmable d'équipements électriques.

PCR : enquête auprès de propriétaires d'habitations neuves équipées d'un plancher chauffant-rafraichissant avec pompe à chaleur, gestionnaire programmable et électricité.

Annexe 2

Etude des spécificités du vocabulaire du groupe des usagers

Vocables significativement suremployés au seuil de 1%
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : HLM, Cannes, Annecy, F4, Tremblay

Verbes : être, avoir, mettre, venir, arriver, chauffer, rentrer, marcher, trouver, servir, allumer, programmer, ouvrir, laisser, fermer, arrêter, éteindre, entendre, remettre, régler, obliger, sentir, aimer, tourner, laver, toucher, préférer, habiter, manger, lever, suffire, brancher, aérer, cuisiner, tromper, rallumer, visiter, refroidir, reprogrammer, admettre, attraper, rebaisser, attendre, descendre, enlever, repasser, regarder, souffler, étouffer, cailler, déranger, savoir, profiter, tenter, devoir, appuyer, partir, contrôler, surveiller, nettoyer, sécher, supporter

Substantifs : heure, fois, radiateur, chambre, hiver, mois, température, truc, appartement, froid, soir, matin, pièce, cuisine, clim, été, tout, route, fenêtre, journée, horloge, air, nuit, salle, enfant, chaud, bain, programmation, machine, bruit, début, voisin, thermostat, plaque, télé, week-end, four, endroit, couloir, prise, midi, lumière, entrée, mur, chaîne, zone, bébé, dimanche, lave-vaisselle, parent, vacance, picot, horaire, grille, linge, musique, boîtier, halogène, notice, fille, soufflerie, lampe, école, gamin, cité, terre, lendemain, toit, commande, avril, tranche, cuisinière, balcon, console, store, gel, signal, parking, plat, chat, radio, chevet, lessive, manuel, prospectus, récepteur, gendre, trappe, fusible, réveil, sèche-linge, volet, après-midi, quartier, meuble, vaisselle, espèce, demie, vent, janvier, micro-onde, carton, minitel, cassette, hifi, gadget, chauffe-eau, douche, frigo, style, utilité, novembre, aération, courant, bouton, peinture, fer, dessous, lit, somme, micro-ondes, distance, voisine, histoire, position, programme, minute

Adjectifs : vrai, petit, chaud, pareil, froid, grand, creux, frais, ouvert, agréable, normal, vieux, automatique, frileux, sympa, tranquille, allumé, malade, digital, manuel, calme, programmable, intermédiaire, surpris, continu, rouge, pratique, moderne, rapide, sain, sale, allergique, vert, seul, demi, ambiant, indispensable

Pronoms : je, ce, il, ça, le, moi, en, tout, quoi, rien, un, celui, celui-là, tu

Adverbes : pas, ne, oui, non, là, plus, bon, même, ici, vraiment, déjà, franchement, là-bas, automatiquement, juste, normalement, carrément, apparemment, manuellement, constamment, dehors, beaucoup, demi, assez, tard, jamais, trop, au-dessus, tôt

Déterminants : tout, deux, mon, six, onze, trois, seize

Conjonctions et prépositions : que, mais, quand, parce que, donc, comme, voilà, sinon, près, jusque, vu, malgré, après, avant, selon, dès, hors

Vocables significativement sousemployés au seuil de 1%
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : Kiterm, EDF, PAC, Tempo, B, Optimia, Bagneux, B et B, X, EJP, PI, Gaz de France, Cruppet, CIAT

Verbes : vendre, falloir, convaincre, proposer, placer, parler, satisfaire, aller, développer, conseiller, traiter, présenter, vouloir, voir, perdre, coûter, répondre, intéresser, créer, investir, poser, former, passer, donner, travailler, fidéliser, ressentir, exister, rencontrer, obtenir, accepter, informer, apporter, rendre, intervenir, considérer, agir, décider, mesurer, imposer, prendre, gagner, fournir, suivre, construire, concerner, rénover, représenter, dépasser, motiver, réagir, intégrer, climatiser, connaître, changer, tenir, démontrer, défendre, pousser, rattraper, penser, évoluer, démarcher, impliquer, percevoir, prouver, bloquer, remplacer, amener, porter

Substantifs : client, clientèle, agent, fidélisation, gens, pompe, produit, qualité, conseil, service, installateur, électricité, entreprise, gaz, tarif, plancher, prix, moyen, chauffage, énergie, centre, installation, étude, conseiller, rafraîchissement, fuel, plan, formation, investissement, agence, démarche, domaine, résistance, gestionnaire, chose, action, marché, inertie, image, information, kilowatt, dossier, compresseur, matériel, panneau, rôle, offre, accueil, confort, label, ouvrage, maître, femme, prime, régulation, argument, situation, cas, technicien, coût, rénovation, opération, contre-référence, travail, satisfaction, gamme, chaudière, poste, activité, but, convecteur, nombre, professionnel, résultat, organisation, aide, terrain, locataire, partie, solution, objectif, contact, contrainte, électricien, évidement, partenaire, architecte, relation, condition, mise, bilan, circulateur, départ, direction, piscine, question, gestion, abonnement, enquête, tuyau, fichier, constructeur, sol, structure, plombier, métier, chef, politique, bureau, isolation, terme, avis, dalle, réunion, manière, place,

affaire, suivi, tarification, forme, plateau, documentation, stage, diagnostic, outil, fabricant, part, mesure, comportement, taux, prêt, type, appareil, commercial, rendement, problématique, culture, logiciel, quantité, mode, monsieur, bâtiment, collecteur, phase, enjeu, tertiaire, motivation, partenariat, maison, inconvénient, vente, communication, jambe, usage, point, confiance, demande, notion, liste, catastrophe, développement, exploitation, bord, garage, appel, réponse, référence, réseau, promoteur, devis, montant, étudiant, négociateur, propane, directeur, vision, réalité, balle, recherche, utilisation, raison, connaissance, tas, norme, écart, retour, compétence, année, idée, possibilité, concurrence, priorité, profession, décision, sentiment, tube, message, pression, madame, main, ingénieur, collègue, an

Adjectifs : électrique, technique, chauffant, commercial, insatisfait, rayonnant, cher, neuf, différent, faux, moyen, national, meilleur, global, tarifaire, interne, important, prêt, précis, content, difficile, spécifique, administratif, nouveau, fort, performant, lourd, pur, actuel, chauffé, réversible, mixte

Pronoms : qui, nous, lui, leur, que, on, vous, certain, cela, quel, tel, lequel, dont, autre, chacun,

Adverbes : aujourd'hui, actuellement, également, forcément, alors, tant, plutôt, notamment, mal, presque, là-dessus, effectivement, tout, essentiellement, probablement, aussi, auprès, vis-à-vis, certainement, peut-être

Déterminants : le, un, certain, notre, son, mille, leur, ce, autre, cent, soixante, quelque, votre, vingt, quel, tel, cinquante

Conjonctions et prépositions : de, avec, sur, dans, à, derrière, par

Phrases les plus spécifiques en valeur absolue (avec leurs scores)

Il aura tourné pour rien, à partir du moment où je ne suis pas là, même s'il me sert à moi dix minutes, il aura tourné une heure pour réchauffer la pièce dans le vide, ce qui n'est pas intéressant, moi je préfère qu'il chauffe quand je suis là tant qu'à faire, c'est pour ça que je le fais tourner seulement le soir quand je sais que je vais… — soit je vais aller fumer dehors, parce que je vais promener la chienne et je ne vais pas me recoucher, hop je ferme — je mets le chauffage, je me regarde mon film et puis je l'éteins et je vais me coucher, et la température ambiante…

Oui donc l'après-midi de treize heures à quatorze heures, c'est heures creuses et à deux heures du matin je crois quelque chose comme ça, mais bon c'est vrai qu'il faudrait que je … comme ils viennent assez souvent il faudrait que je demande et puis que je contrôle, mais là lui m'avait dit pour le, parce que la première fois, il était venu je crois que c'est juin juillet, celui qui s'occupe du contrôle donc il vient souvent, il m'avait dit, parce que moi je l'avais pas mis encore, je ne l'avais pas mise en route, il m'a dit ben ça serait dommage que vous ne mettiez pas parce que il n'y a pas grande différence, il m'a dit sur la totalité de l'été ça vous fera une différence à peu près de quatre cent francs enfin c'est à peu près l'estimation qu'il m'a faite au niveau froid.

Ben, c'est vrai que c'était quand même pratique, parce que bon je veux dire quand il fait très chaud, moi je trouve que c'est quand même agréable, on se sent quand même beaucoup mieux, d'autant plus que comment vous dire on peut la régler, il y avait bon moi je sais que mon mari l'hiver il a très froid, moi j'aime régler à peu près à 21, je trouve que c'est une température qui chauffe pas trop, qui fait pas trop de différence en extérieur et où ça marque pas trop sur le corps non plus.

Je veux dire vraiment les miens, il faut qu'ils marchent pour six, si je mets le thermostat à six, on sent la chaleur dans l'appartement non, ici il faut que ça marche à six, alors c'est sûr que quand il y a les deux chambres, c'est vrai que une fois que c'est lancé dans les chambres je suis obligée d'éteindre parce qu'il fait très chaud, donc là je les laisse parce que on a quand même, je laisse toutes les portes du couloir tout ça, mais dans les chambres, je suis obligée de fermer à un moment donné, et de le remettre vraiment quand ils se couchent parce que la nuit c'est vrai qu'il fait frais.

Non parce qu'en fin de compte moi je sais que deux personnes, la première personne qui m'avait mis mon horloge s'était trompée, bon ! c'est ce qu'elle m'avait dit la personne qui me l'a remis, il s'était trompé au niveau manuel c'étaient des petits plots qu'il fallait baisser, c'était pas ça, elle les avait abaissé pour la nuit en fin de compte ça marchait la journée et la deuxième personne qui me l'a fait c'était pareil ça ne marchait pas du tout, alors j'ai dit : c'est pas la peine de chaque fois de déranger les personnes , autant le faire moi-même.

En deux fois, parce que en fait, c'était un système tout nouveau, donc le système Moulinex, au niveau de la console, on n'a pas eu de problème, mais on avait sur les cinq, on avait trois prises qui marchaient pas, c'est-à-dire qu'elles se mettaient en - mince comment ça s'appelle, parce que les prises ont trois positions, donc la position arrêt, la position marche, marche normale et programmation, et la position hors gel voilà - donc elles étaient en, il y avait trois prises qui étaient en permanence hors gel, elles avaient soit la position arrêt, soit la position hors gel, mais une semaine après ils nous ont tout installé, c'était au niveau des cartes électroniques, il y avait eu un problème.

Non mais bon, je passe ma journée, mes poussières, je suis désolée, elles ne sont pas faites tous les jours à cause de ça, je ne peux pas parce que ça me… tout ça, mes yeux et tout, donc ou je le fais fenêtres ouvertes, généralement je le fais la nuit … je ne suis pas comme tout le monde … la poussière reste …

Ou même à la maison, c'est comme mon fils c'est pareil, il est tout petit mais dès qu'il bouge et tout, il transpire énormément, donc c'est vrai que mon mari lui il aime bien quand il fait bien froid, quand il fait bien frais, il préfère à la limite se mettre autre chose et puis que ce soit froid, moi je trouve que pour la santé c'est pas trop bon donc c'est pour ça que je ne suis pas d'accord.

Mais le problème c'est que si vous avez fait une programmation et qu'à titre exceptionnel, vous souhaitez - le fait de couper l'horloge l'arrête, et donc vous allez tout vous décaler dans le temps parce que l'horloge elle va, elle, elle est avec ses petits picots, elle n'a pas dans sa petite tête qu'il est telle heure, elle ne sait pas, puisque vous devez lui dire, quand on la programme : on part du jour où on met, de l'heure où on met pour qu'elle parte, enfin il y a aucun endroit où je vois l'heure sur cette horloge là, donc pour moi elle ne sait pas quel jour on est, elle ne sait pas, on met des jours deux, trois, et donc en fonction de ça, il faut programmer, et enfin à mon sens ça devait être, donc c'est pour ça qu'on arrive finalement à une gestion manuelle en disant : ben non, c'est juste maintenant que je voudrais avoir un petit peu plus chaud.

Alors je ne sais pas pourquoi ça affiche auto, je ne lui ai rien demandé à ce couillon … alors utilisez les avantages de la programmation, je m'en sers comme un manche moi, alors c'est vrai que ça m'arrive par contre, ça m'est arrivée de l'utiliser au tout début de la chaleur c'est-à-dire que je programmais la clim pour quatre heures de l'après-midi vous savez au tout début, quand il a fait, au début, je l'ai programmé - là actuellement elle reste toute la journée parce qu'il fait chaud - mais au début je la programmais pour trois heures de l'après-midi, c'est-à-dire que le matin non elle marchait pas, hein.

Ben parce qu' au début, bon, je vous dis comme je suis quelqu'un qui adore la chaleur je mettais pas, bon je l'ai mis assez tard la clim, il y avait eu une personne qui était déjà venue plusieurs fois et bon, au niveau froid il n'y avait pas tellement de différence de consommation, et puis après comme il a commencé vraiment à faire assez chaud, je l'ai réglée, je vous dis je l'ai réglée entre vingt trois et vingt cinq degrés donc ce qui fait que c'était assez haut, ça ne fait pas plus de consommation au niveau chaleur l'hiver que …

Et puis, après, c'est vrai que, comme il y avait une position, enfin il marchait n'importe comment, c'est-à-dire qu'il chauffait quand on n'était pas là, et donc comme on avait juste cette petite feuille photocopiée qui était à peine lisible, je n'arrivais pas à comprendre, je me suis dit plutôt, j'ai lu, ce que j'ai retenu c'est que, ben, si on voulait pas que ça fonctionne, on avait qu'à le mettre en fonctionnement autonome c'est-à-dire continu.

Ben, au départ, au niveau climatisation, bon, pas trop au niveau froid puisque'il m'avait dit que, je l'avais vu moi avant de la mettre en route, il m'a dit il n'y aura pas trop de différence, mais c'est vrai que je me suis un peu méfié au niveau chaleur, je vous ai dit j'aime bien la chaleur, je sais que les convecteurs ça coûte très cher, la climatisation quand on connaît pas, on ne sait pas, on ne sait pas trop, mais, bon, je sais déjà que je suis le plus, celui qui a consommé le plus au niveau chaleur quoi, ça c'est sûr, quand ils ont fait leur … ils ont contrôlé au niveau déjà des compteurs, déjà ils voyaient au niveau de, ils sont venus de Paris ou je ne sais pas quoi là les …

On peut la réduire, comme je vous dis avec ce système voilà de thermostat qui ferme cette trappe, si là je mets à vingt trois et que je mets par exemple dans une chambre à vingt degrés, il est évident que ça va couper le chauffage, l'arrivée du chauffage, ça passe quand même parce que une trappe, c'est une petite trappe en fer qui est derrière les grilles, la chaleur va venir, donc elle viendra moins dans les autres pièces, c'est ce qui se passait

d'ailleurs au début, c'est que comme sur les chambres, je n'ai pas fait attention à la température, je n'avais pas touché aux thermostats, donc elle était supérieure à la température que j'avais là, les trois chambres en fait étaient beaucoup plus ventilées, ce qui fait qu'ici il y avait pratiquement, il y avait une sortie, mais je veux dire c'était nettement plus froid que les chambres.

Oui … mais enfin moi personnellement, c'était plus, bon là j'avais, où l'habitais avant j'avais un plafonnier lampes ventilateur, et puis dans la salle à manger des trucs comme ça, je n'aime pas trop le vent parce que j'attrape la crève, c'est systématique, la ventilation j'aime pas trop, mais sinon c'était ça, quand il faisait vraiment très chaud.

J'étais un petit peu surpris, parce que niveau HLM c'est vrai que c'est pas habituel, en plus c'est vrai qu'on est assez bien situés, c'est bien fréquenté, c'est classe, c'est bien, c'est bien et puis, non, mais c'est vrai que c'est un bon système, par rapport déjà ça prend pas de place, déjà il n'y a pas de radiateurs on n'a plus de problème.

Vraiment je n'ai pas souffert de la chaleur l'été la nuit, c'était suffisant bon maintenant j'aime la chaleur, je ne suis pas… je n'en souffre pas non plus, je ne suis pas une personne qui souffre de la chaleur, enfin bon ça va, mais le samedi et le dimanche quand j'étais là, je mettais toute la journée quand même, pour faire le ménage c'est bien agréable quand il fait chaud hein … pour passer l'aspirateur c'est bien agréable.

Je regroupe c'est-à-dire qu' au lieu de repasser une heure tous les jours et de brancher un fer pour une heure, je le branche quatre heures admettons d'affilée, quatre à cinq heures et même des fois six heures, moi je fais six heures de repassage, mais je repasse tout, je mets à plat et je n'ai pas à refaire après pour ne pas que mon fer marche au fait, au lieu de marcher pendant six heures, repasser, plier, il marche pendant trois heures de repassage et après je plie, mais le fer est éteint donc ça, c'est un souci d'économie parce que je fais du repassage à domicile d'un côté, de l'autre c'est un souci pour moi parce qu'on est quand même cinq, donc le linge dans la maison, il faut trouver le moyen de … de regrouper, moi je ne sais pas, c'est ma façon de faire à moi, et je pense que je fais des économies comme ça aussi …

Ah ben, j'ai arrêté la clim, c'est ce que j'ai fait, tout était ouvert, toutes les fenêtres, alors comme je suis, j'ai deux, comment dire, je n'ai pas tout sur la même façade, ça faisait un courant d'air entre la cuisine et les chambres et les gamins ne sont pas tombés malades et on ne s'en est pas porté plus mal parce que j'ai toujours vécu comme ça …

Donc il retire la grille, il m'a montré comment faire d'ailleurs, il m'a dit : "si jamais un jour vous avez un problème, que vous trouvez qu'il y a de la poussière qui sort ou quoi", donc il m'a montré, il enlève la grille, il retire le filtre, il aspire, hop ! au revoir, j'ai : "ça je peux le faire, oh oui ça je peux le faire largement, mais bon il se trouve qu'il est là autant qu'il le fasse".

Oh ben ça dépend si les fenêtres sont ou pas ouvertes, c'est sûr que si je les laisse ouvertes, bon ! après quand j'ai fini, que je ferme, il caille assez oui.

Phrases les plus spécifiques en valeur relative

Elle, elle a froid, elle allume, moi j'ai trop chaud.

Oh oui là j'éteins tout quand je pars en week-end.

Oui mais je suis tranquille, il fait tellement chaud que je suis tranquille quand j'arrive hein.

Ah oui l'horloge moi je ne m'en sers pas.

Ah oui moi, pour moi, je me serais mis trois pulls.

Donc c'est pour ça que je ne sais pas trop.

Je ne comprends rien, je ne l'utilise pas, j'ai juste regardé comme ça mais …

Oui ben, oui, je ne savais pas comment ça marchait mais quand je l'ai su … ben j'ai tout arrêté.

Mais je ne peux pas dire que j'avais froid non plus, je ne peux pas dire que j'avais froid.

Ah oui vraiment ça se sent, il fait pas chaud …

Non ben non, parce que je n'y touche pas, je préfère pas y toucher donc …

Je monte mes radiateurs quand j'ai froid, je les redescends quand j'ai chaud enfin et puis voilà.

Oui oui, je sais que je peux le faire, mais bon ! je préfère pas parce que je me dis que si je mets mes radiateurs et puis que j'augmente je ne sais pas quoi, je préfère pas.

Non, ça ils me l'ont pas fait, bon il est vrai que je ne l'ai pas demandé non plus, peut-être que c'est pour ça.

Mais moi, bon ! comme je suis célibataire tout ça, quand je rentre c'est à mon envie quoi …
Ah oui fantastique, c'était très pratique, et puis bon ! moi il est vrai que je préfère, ça consomme quand même moins.

Annexe 3

Les spécificités du vocabulaire des agents d'EDF par rapport au français contemporain.

Vocables significativement suremployés au seuil de 1%
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : EDF, Tempo, Optimia, Bagneux, PI, Gaz de France, Poitiers, HLM, SRC, Angers, Clio, Cholet, EDF-GDF, Godin, TGC, Promotelec, Saumur, France Télécom, Japonais, Lille, Renault, Saint-Pierre

Verbes : aller, falloir, pouvoir, savoir, voir, mettre, croire, passer, vendre, payer, proposer, coûter, expliquer, poser, conseiller, satisfaire, convaincre, chauffer, placer, construire, gérer, acheter, consommer, traiter, envoyer, développer, installer, isoler, montrer, porter, intégrer, ressentir, investir, fidéliser, motiver, concerner, persuader, refaire, démarrer, apprécier, renseigner, rénover, baisser, démarcher, rattraper, téléphoner, remonter, déplacer, détecter, plaindre, contacter, pratiquer, maîtriser, prêter, louer, calculer, mesurer, devoir, faire, écouter, recevoir, surprendre, comparer, décevoir, orienter, appeler, imaginer, perdre, exister, apercevoir, communiquer, inquiéter, demander, entrer, adapter, former, fonctionner, attendre, importer, intéresser, commencer

Substantifs : client, chauffage, gens, gaz, électricité, clientèle, agent, heure, niveau, fait, produit, fidélisation, service, facture, qualité, centre, prix, installation, moyen, maison, conseil, tarif, installateur, entreprise, énergie, franc, agence, nombre, conseiller, radiateur, appareil, isolation, système, confort, fuel, limite, domaine, avis, électricien, accueil, consommation, logement, action, plan, démarche, organisation, coût, kilowatt, convecteur, téléphone, argument, mise, image, type, offre, marché, eau, métier, matériel, main, collègue, manière, voiture, étude, rénovation, investissement, commercial, contre-référence, immeuble, chef, locataire, instant, activité, appartement, hiver, contact, chaleur, pavillon, technicien, gestionnaire, compteur, satisfaction, appel, partenaire, gamme, panneau, programmation, professionnel, prime, propriétaire, label, rendez-vous, concurrence, charge, abonnement, utilisation, chaudière, fichier, pièce, référence, promoteur, matière, stage, bilan, intervention, mode, froid, suivi, diagnostic, norme, outil, contrat, tas, bois, plateau, culture, taux, demie, motivation, courrier, avenir, puissance, minimum, prêt, kilo, tarification, négociateur, usage, bon, montant, quantité, contrainte, étudiant, présent, centime, marketing, retard, mètre, constructeur, régulation, placement, publicité, interne, enquête, tertiaire, oeuvre, mixité, fenêtre, rendement, loyer, documentation, plage, foire, devis, propane, bien-être, câble, balle, facturation, existant, gazier, sol, construction, général, réforme, machin, terrain, réseau, message, rue, gestion, dommage, structure, bas, maximum, développement, communication, mois, compte, fonctionnement, phase, compétence, fonction, année, programme, réponse, effort

Adjectifs : électrique, vrai, commercial, petit, cher, neuf, technique, insatisfait, fort, chaud, performant, rayonnant, individuel, supplémentaire, tarifaire, global, froid, interne, téléphonique, pur, creux, moyen, mécontent, radiant, mixte, excessif, bon, pareil, spécifique, naturel, content

Pronoms : on, il, qui, nous, vous, lui, leur, nous-même, quel

Adverbes : ne, oui, bien, même, peu, puis, aussi, alors, tout, maintenant, déjà, aujourd'hui, effectivement, mal, forcément, actuellement, plutôt, certainement, cher, auprès, notamment, correctement, généralement, également, demain, surtout, systématiquement, malheureusement, par-là, directement, peut-être

Déterminants : : un, cent, mille, dix, quinze, votre, cinq, six

Conjonctions et prépositions : en, donc, parce que, quand, si, sur, voilà, chez, derrière, soit, puisque, vers, que, par, mais, ou

Vocables significativement sousemployés au seuil : 1%. (Classement par catégories grammaticales et spécificité décroissante)

Noms propres : A, Allemagne, B, C, CAP, M, OK, Paris, SNCF, T, X, Y, Z, RPR, France, Chirac, Europe, Mitterrand, PS, CSN, Montréal, Québec, V

Verbes : vérifier, élever, établir, étonner, étudier, évaluer, éviter, évoluer, être, estimer, faciliter, fatiguer, favoriser, fermer, finaliser, fixer, forcer, gagner, garantir, garder, habiter, habituer, impliquer, inciter, influencer, inscrire, interroger, intervenir, jeter, juger, lancer, libérer, lier, limiter, lâcher, maintenir, manquer, marcher, marquer, mener, modifier, monter, mourir, nommer, offrir, organiser, paraître, permettre, poursuivre, abandonner, aborder, accorder, admettre, adresser, agir, ajouter, améliorer, analyser, animer, annoncer, apporter, argumenter, assurer, attacher, attaquer, atteindre, augmenter, bloquer, bouger,

casser, changer, circuler, commander, compliquer, comprendre, compter, conduire, confronter, considérer, constater, consulter, correspondre, couper, courir, créer, descendre, diminuer, diriger, distribuer, durer, déborder, débrouiller, découler, découvrir, défendre, dépasser, déranger, déterminer, embaucher, emmener, empêcher, engager, enlever, sensibiliser, pousser, prioriser, procéder, produire, prouver, préférer, présenter, raconter, rajouter, ramasser, rappeler, rapporter, rapprocher, rechercher, regrouper, remarquer, remettre, remplacer, rendre, repartir, reprendre, reprocher, représenter, ressortir, retomber, retourner, retrouver, revoir, rire, risquer, réagir, réaliser, récupérer, réduire, réfléchir, répondre, répéter, résoudre, réunir, traîner, signaler, situer, soigner, souffrir, souhaiter, soutenir, subir, suffire, suggérer, supposer, tirer, tomber, tourner, venir, tuer, utiliser, valoir, avoir, régler, vivre, boire, voter, partir, trouver, négocier, battre, penser, sortir, écrire, identifier, amener, fumer, regarder, marier, élire, supporter, embarquer, aider, décider, déposer, devenir, jouer, prendre, asseoir, plaire, entendre, partager, participer, suivre, parler, échanger, exprimer, chercher, tenter, lire, réussir, travailler, retenir, rencontrer, frapper, appliquer, discuter, aimer, quitter, revenir, souvenir, terminer, avancer, profiter, dépenser, ressembler, donner, ramener, rentrer, imposer, préparer, signer, ouvrir, refuser, tendre, taper, accepter, laisser, rester, exclure, divorcer, rejoindre, énerver, dormir, recommencer, assister, cacher, définir, respecter, apprendre, sentir, occuper, toucher, retirer, arrêter, continuer

Substantifs : week-end, zone, échec, économie, écoute, élu, élément, épouse, été, envie, erreur, espace, espèce, exigence, explication, exploitation, expression, extérieur, face, facteur, faute, fil, fin, fonctionnaire, force, formation, forme, formule, février, fête, garantie, gars, genre, gros, habitude, haut, horaire, horreur, impact, implication, importance, influence, information, instance, intention, janvier, journée, langage, langue, ligne, long, machine, magasin, mai, maire, mairie, majorité, mal, maman, marre, mars, mentalité, million, minute, mission, moitié, mot, naissance, nature, nom, note, notion, nuit, objectif, obligation, octobre, oeil, option, opération, ordre, orientation, origine, page, paie, papier, paquet, parole, part, partage, partenariat, participation, particulier, pas, passé, patron, perception, personnel, petit, peur, pied, plaisir, plancher, plupart, population, possibilité, poste, aberration, absentéisme, air, aise, ajustement, ambition, ami, amélioration, aménagement, arrière, arrêt, aspect, assurance, attente, attention, avance, avant, avantage, avril, base, bien, bureau, bébé, bêtise, cahier, campagne, capacité, carte, cas, chance, chiffre, coeur, coin, commerce, commun, concept, condition, conduite, conscience, constat, conséquence, contexte, contraire, contrôle, conviction, courant, course, crédibilité, dame, danger, dehors, dernier, devoir, dialogue, difficulté, différence, dimanche, dimension, direction, donnée, doute, définition, départ, département, détail, effet, endroit, engagement, enjeu, ensemble, entretien, secteur, sein, pression, preuve, priorité, prise, profession, propos, proposition, préparation, présence, présentation, quart, quotidien, reconnaissance, recul, relais, responsable, reste, retour, risque, route, réaction, réalité, réflexe, réflexion, région, résultat, réunion, salle, samedi, septembre, situation, société, soir, sortie, souci, style, sécurité, tableau, technique, tendance, tentative, terre, titre, tort, tournée, tout, train, vente, truc, tâche, télé, urgence, vendredi, vision, ville, négociation, femme, enfant, homme, syndicat, mari, comité, employeur, convention, vie, politique, employé, famille, élève, table, entente, article, processus, mandat, rencontre, classe, solution, grief, membre, confiance, porte-parole, parti, méthode, texte, partie, établissement, pouvoir, fille, assemblée, jeune, thème, mère, argent, parent, intérêt, temps, collège, caractère, école, pays, gamin, responsabilité, idée, succursale, affaire, garçon, travail, copain, monde, brainstorming, discussion, incompatibilité, clause, président, équipe, prof, individu, sorte, emploi, père, violence, sujet, santé, maîtresse, façon, critère, représentant, séance, cours, climat, grève, piste, boisson, ministre, mésentente, éducation, session, mouvement, gouvernement, période, humeur, professeur, coup, règlement, quartier, infidélité, ex-mari, mariage, milieu, position, délégué, résolution, principe, maladie, caisse, droit, respect, conciliateur, quota, job, bord, député, alcool, monsieur, gauche, discours, droite, exercice, sens, début, lycée, décembre, adulte, date, boss, relation, exécutif, conseillère, accord, loi, liste, ancienneté, infirmière, fils, changement, étape, élection, problème, raison, facilitateur, côté, politique, préoccupation, alcoolisme, expérience, liberté, enseignant, ménage, médecin, débat, suite, bloc, médiation, dispute, conciliation, jeu, semaine, prévention, projet, document, interprétation, juin, groupe, chemin, évaluation, application, directeur, fin, vice-président, ministère, mécanisme, approche, âge, ouverture, photo, conflit, problématique, allocation, fédération, historique, point, madame, valeur, foyer, fond, confrontation, moment, fois, agriculture, instabilité, stratégie, gardien, protection, opinion, congé, tour, intervenant, remue-méninges, ras-le-bol, hôpital, regroupement, usine, attitude, couple, demande, décision, libération, dette, éducateur, agriculteur, caissier, domicile, dépôt, peuple, sous-comité, goût, lien, mécanique, administration, assistante, gosse, institution, divorce, anglais, augmentation, absence, manque, exemple, enquêtée, université, création, guerre, bout, méfiance, réserve, front, accréditation, pratique, ressource, égalité, vote, amie, temporaire, tête, salaire, signature, étranger, statut, ambiance, citoyen, jalousie, sexe, sport, ado, choix, journal, foi, chômage, novembre, liaison, vacance, règle, travailleur, cadre, procédure, lieu, boulot, intérieur, somme, crise, parité, comportement, retraite, mesure, échelle, histoire, volonté, jour

Adjectifs : âgé, énorme, évident, essentiel, extérieur, facile, fait, faux, final, fondamental, formidable, fou, gagnant, grave, gros, haut, heureux, important, impossible, indépendant, intelligent, intéressant, jaloux, juste, large, libre, logique, lourd, malheureux, meilleur, même, national, normal, nouveau, négatif, ouvert, particulier, pauvre, physique, pire, plein, possible, absent, acceptable, actuel, adapté, administratif, agréable, ancien, bas, beau, certain, clair, complet, compliqué, compétent, concret, correct, difficile, disponible, drôle, entendu, sensible, pratique, primaire, privé, proche, précis, précédent, prêt, psychologique, public, raisonnable, rare, responsable, réel, seul, simple, supérieur, sérieux, sûr, terrible, total, unique, valable, vieux, raisonné, syndical, politique, patronal, traditionnel, autre, partiel, collectif, jeune, capable, familial, donné, violent, social, conjoint, général, municipal, professionnel, demi, majeur, sexuel, objectif, scolaire, régulier, long, nombreux, monétaire, malade, dur, français, positif, méchant, salarial, paritaire, féminin, instable, prochain, court, économique, différent, principal, mutuel, dangereux, égal, temporaire, spécial, dernier, local, humain, grand, utile

Pronoms : eux-mêmes, lui-même, je, mien, moi-même, même, nôtre, ceci, cela, dont, quelqu'un, soi, tu, autre, se, nous autres, eux autres, un, chacun, toi, ça, en, ils, rien, moi, personne, le, ce, celui, que, quoi, y, tout

Adverbes : énormément, éventuellement, environ, essentiellement, exactement, extrêmement, facilement, finalement, fort, franchement, globalement, heureusement, hier, longtemps, lors, là-dessus, moins, normalement, particulièrement, partout, personnellement, petit, pourtant, absolument, ailleurs, au fur et à mesure, auparavant, automatiquement, autrement, beau, carrément, ci, clairement, combien, complètement, d'abord, d'autant, dedans, dehors, dessus, différemment, encore, presque, probablement, quasiment, quelquefois, rapidement, relativement, réellement, régulièrement, seulement, si, simplement, suffisamment, super, tant, tard, tôt, uniquement, vite, ensemble, jamais, là-dedans, nécessairement, toujours, évidemment, tantôt, juste, plus, assez, ainsi, autant, trop, pourquoi, enfin, où, bon, très, beaucoup, loin, présentement, non, là, tellement, vraiment, pas, sûrement, ensuite, autour, comment

Déterminants : zéro, notre, cinquante, deuxième, douze, second, quarante, quatorze, quelque, treize, trente, sept, soixante, vingt, ton, mon, son, troisième, quatre, seize, sixième, plusieurs, premier, ce, deux, quel, cinquième, même, quatrième, le, trois, chaque, un, leur, tel, aucun

Conjonctions et prépositions : hors, jusque, lorsque, malgré, plein, après, avant, devant, dès, près, sinon, sous, tandis que, vu, dans, entre, avec, sauf, comme, sans, voici, et, car, de, pour, ni, selon, à

Phrases les plus spécifiques en valeur absolue

"Si on veut vendre du chauffage électrique, il faut montrer que ça coûte, bon, peut-être pas aussi cher que de faire un chauffage à eau chaude en gaz ou en fuel, mais que ça peut coûter cher aussi d'investissement, au départ, si on veut réellement avoir une installation qui soit performante et donc, là, il faut peut-être essayer de vendre du chauffage électrique de qualité, de vendre peut-être du chauffage électrique avec la climatisation, des appareils réversibles, de vendre des choses qui tiennent debout, abandonner un peu cet état d'esprit qu'on a eu à une époque, de faire du chauffage électrique à tout va parce que ça demandait pas beaucoup d'investissement pour le promoteur.

Donc il faut les impliquer aussi : si on a un client qui veut changer d'énergie, qu'on n'arrive pas à le garder à l'électricité, il faut l'envoyer vers un chauffagiste de notre réseau et lui faire comprendre qu'aujourd'hui on lui envoie quelqu'un, s'il a quelqu'un qui est en électricité qui veut un devis eau chaude à ce moment là, il vaut peut-être mieux qu'il nous en parle, il faut jouer un petit peu la contrepartie, c'est une manière de fonctionner.

Ah oui, quand on est une journée complète au téléphone, mais c'est ça en fait : on ne voit que les problèmes, on ne voit que les clients qui ne peuvent pas payer, on ne voit que les clients qui sont mécontents de leur facture, on ne voit que les clients qui ont un problème.

Déjà on dit : il faut quand même sur une facture d'électricité, voir qu'il n'y a pas que le chauffage, bon le chauffage est peut-être un peu plus cher, mais si on compare ce qui est comparable, on peut quand même arriver à avoir des arguments aussi au niveau du chauffage électrique.

Parce que, si on veut exploiter fidélisation à fond, c'est un client qu'on veut raccrocher en chauffage électrique, on veut lui faire comprendre donc que le chauffage électrique peut être très performant, mais effectivement qu'il y a des produits derrière qui sont de qualité, qu'il y a un tarif adapté.

Pareil, il faut des constructions de qualité, il faut une bonne isolation, il faut des bons produits, il faut des clients informés qui ne mettent pas tout en disant : tiens je préfère faire une belle terrasse ou une belle véranda, et puis le chauffage on verra après.

Pour pouvoir faire du chauffage électrique, il faut avoir une chaîne globale de qualité, un environnement global qui puisse garantir que vous ne ferez pas une contre-référence demain, donc des installations…

Maintenant, le client nous téléphone, vient nous voir - avant, on se disait bon, qu'est ce qu'on fait ? on va lui vendre de l'électricité ou du gaz ? on lui proposait les deux bien évidemment — le client va être devant nous, il va nous dire : bon, vous vous êtes quoi ? qu'est ce qui est le mieux ? on va lui dire : "ben, écoutez le mieux c'est l'électricité", et le client, il revient nous voir, quand sa maison a été terminée, avec une facture de 20000 francs, il va nous dire : "mais attendez : vous vous êtes foutus de moi".

C'est pareil, quand vous expliquez sur un pavillon neuf au client, en général avec les taux, quoique les taux vont baisser maintenant, vous payez au minimum deux fois votre maison, ben quand vous avez remboursé deux fois votre chaudière, il est temps de la changer, parce qu'une installation de chauffage fuel n'est pas la même qu'une installation de chauffage électrique, même de qualité.

Et je veux compléter ça en disant : aujourd'hui — quand on parlait de qualité, c'est ça l'idée que j'avais oubliée tout à l'heure — il y a des possibilités en matière d'électricité qui n'existent pas en gaz, en matière de modulation, en matière de régulation, en matière, je dirais, d'usage aussi, puisque vous avez des choses qui sont en finalisation par exemple, globalement des offres qui peuvent être beaucoup plus performantes en électricité qu'en gaz, qui sont certainement des offres qu'il faut bien différencier en matière de marketing parce qu'on ne vendra pas la Rolls-Royce à un smicard.

Alors bon, donc avec Tempo, il faut maintenant avoir carrément des gestionnaires de trucs, de machins et tout le cirque… les gens voudront jamais investir 5 000 balles pour économiser 200 francs parce qu'en fait, on en est là, quand on voit les simulations sur une carte, c'est, par rapport à heures creuses, Tempo fait économiser, dans certains cas, 200 francs par an, et il faudrait que les gens investissent 6 000 balles de matériel, de main d'oeuvre et tout ça, pour économiser 200 francs par an ?

Oui c'est ça, c'est-à-dire que, quand on est en groupe fonctionnel 4, on n'a qu'un seul travail, qu'une seule fonction, et soit anciennement réceptionniste, gestionnaire de contrats, quand on est en 5, on est réceptionniste plus gestionnaire de contrats ou gestionnaire de comptes, quand on est en 5, pardon, quand on est en 6 on est réceptionniste, gestionnaire de comptes, gestionnaire de contrats et quand on est en 7, maintenant, et en 8, on est chef du groupe voilà, et, moi, c'est tout un autre métier.

Ben oui il faut quand même … il faut quand même vendre des kilowatts aussi de l'autre côté, il faut tout voir, mais il faut vendre des bons kilowatts je dirais.

Phrases les plus spécifiques en valeur relative

Une maison neuve ou un appartement neuf : possibilité d'électricité.

Il faut quand même qu'on ait un minimum de commercial dans EDF.

Donc on veut bien mettre nos compétences en avant, mais il faut aussi savoir nous reconnaître derrière.

On le calcule aussi bien en électricité qu'en gaz d'ailleurs.

Ben oui évidement, mais ça c'est un message qui passe mal.

On a un client qui téléphone : détecter si chauffage électrique ou pas, lui proposer un certain nombre de tarifs, lui donner quelques conseils.

On veut retravailler un peu nos agents sur la culture services.

Alors le client il nous dit : "soit on va pas chauffer ou soit ça nous coûte une fortune".

On a tellement vécu sur un produit qui se vendait tout seul, sur une énergie nucléaire qu'on imaginait qu'elle était bon marché, bon ben voilà.

Bibliographie

BAULIEU F. B. 1989, "A Classification of Presence/Absence Basec Dissimilarity Coefficients", *Journal of Classification*, 6, p 233-246.

BARTHELEMY Jean-Pierre, GUENOCHE Alain, 1988, *Les arbres et les représentations des proximités*, Paris, Masson.

BARTHELEMY Jean-Pierre, LUONG Xuan, 1998, "Représenter les données textuelles par des arbres", in Sylvie MELLET (ed), *4e journées internationales d'analyse statitistique des données textuelles*, Université de Nice, 1998, p. 49-71.

BENZECRI Jean-Paul, 1980, *L'analyse des données. 1. La taxinomie*, Paris, Dunod.

BERGERON Jean-Guy, LABBE Dominique, 2000, "L'évaluation de la négociation raisonnée par les acteurs : une analyse lexicométrique ", XVIe congrès de l'Association Internationale des Sociologues de Langue Française, Québec (à paraître aux Presses de l'Université Laval).

BRUGIDOU Mathieu, LABBE Dominique, 1999, *Le discours syndical français contemporain (CGT, CGT, FO en 1996-98)*, Grenoble-Paris, CERAT-EDF(GRETS).

BRUNET Etienne 1988, "Une mesure de la distance intertextuelle : la connexion lexicale", *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, Université de Liège.

GOUGENHEIM Georges et Al, 1964, *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

HUBALEK Zdenek, 1982, "Coefficients of Association and Similarity, Based on Binary (Presence Absence) Data : an evaluation", *Bio. Rev.*, 57, p 669-689.

JACCART P. (1908), "Nouvelles recherches sur la distribution florale", *Bull. Soc. Vand. Sci. Nat*., 44.

LABBE Dominique, 1990a, *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

LABBE Dominique, 1990b, *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation nationale des sciences politiques.LABBE Dominique, HUBERT Pierre, 1998, "La connexion des vocabulaires", *Communication aux 4e Journées d'analyse des données textuelles*, Nice.

LABBE Cyril, LABBE Dominique, 1994, "Que mesure la spécificité du vocabulaire ?", Grenoble, CERAT. A paraître dans : *Lexicometrica*, 3, 2001.

LABBE Dominique, MONIERE Denis, 2000, « La connexion intertextuelle. Application au discours gouvernemental québécois », Martin RAJMAN et Jean-Cédric CHAPPELIER (eds), *Actes des 5^{e} journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, vol 1, p 85-94.

LAFON Pierre, 1984, *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.

LUONG Xuan, 1994, « L'analyse arborée des données textuelles : mode d'emploi », *Travaux du cercle linguistique de Nice*, 1994, 16, p 25-42.

MULLER Charles, *Principes et méthodes de statistique lexicale*, Paris, Hachette université, 1977.

ROBERTS F.S. et Al 1971, *Measurement Theory*, Addison-Wesley, Reading.

SNEATH P.H., SOKAL R.R. 1973, *Numerical Taxonomy*, Freeman, San Francisco.

TOMASSONE Richard et Al 1988, *Discrimination et classement*, Paris, Masson, 1988.