Normalisation et lemmatisation d'une question ouverte

Dominique Labbé

Journées d'études
Grenoble, 8 juin 2001
MSH Alpes
Société Française de Statistique

Traitement des questions ouvertes dans les enquêtes et sondages

Normalisation et lemmatisation d'une question ouverte

Les femmes face au changement familial

Dominique LABBE
(CERAT – Institut d'Etudes Politiques de Grenoble)
dominique.labbe@iep-grenoble.fr

Publiée dans : *Journal de la Société Française de Statistique*, 142, 2001-4, p 37-57.

Résumé :
La normalisation consiste à réduire les majuscules des noms communs, à uniformiser les orthographes multiples des noms propres, des dates et des chiffres ou de certains mots communs, à déployer les abréviations, etc. La lemmatisation associe à ces graphies normalisées un lemme correspondant à l'entrée du dictionnaire et une catégorie grammaticale. Ces tâches sont confiées à un automate dont l'efficacité est testée sur les réponses à une question ouverte dans une enquête sur les causes de divorce. Par rapport aux formes graphiques brutes, les données lemmatisées réduisent le nombre de mots différents et permettent de retrouver les principaux thèmes. Elles mettent également à jour certaines déformations produites par la manière dont les enquêteurs retranscrivent les réponses.

Nous remercions l'Institut National d'Etudes Démographiques qui a accepté de mettre à notre disposition l'enquête sur Les femmes face au changement familial. Cette étude n'aurait pas été possible sans l'aide du Groupe de Recherche Energie, Technologie et Société, département de la Division Recherche et Développement d'Electricité de France.

Tout texte en langue naturelle peut être représenté comme une collection d'évènements rares et très inégalement répartis. Peu de mots dépassent le seuil de 1% de fréquence relative et ce ne sont probablement pas les plus intéressants puisque, selon un adage classique la quantité d'information véhiculée par un mot est inversement proportionnelle à sa fréquence d'apparition…

Dès lors, comment réduire le nombre des mots différents sans perdre trop d'informations ?

La question est banale en sciences humaines. Par exemple, cela ne choque personne de voir les statuts sociaux et professionnels de la population française réduits à une grille d'une trentaine de PCS, voire à 6 groupes, alors que les métiers sont au moins aussi divers que le vocabulaire de la langue usuelle. Une semblable réduction serait-elle possible sur les mots et notamment sur ceux des réponses aux questions ouvertes dans les sondages ? La parole des enquêtés accéderait enfin au statut de variable indépendante et ne viendrait plus simplement en illustration des conclusions obtenues à partir des seules variables sociologiques éventuellement croisées avec quelques questions fermées (sur le traitement des questions ouvertes : Lebart 1994).

Nous pensons que la normalisation et la lemmatisation apportent une partie de la solution. Après avoir brièvement expliqué en quoi consistent ces deux opérations, nous donnerons un exemple d'application.

I. La Normalisation et la lemmatisation

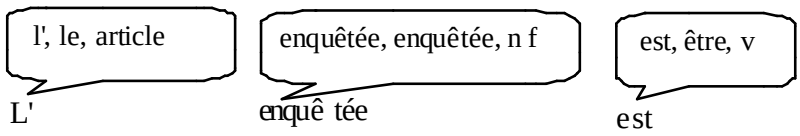
Au départ, il faut rappeler deux évidences. Premièrement, la plupart des mots sont susceptibles d'avoir plusieurs graphies : majuscules ou minuscules, élisions, abréviations… Deuxièmement, des mots différents peuvent s'écrire de la même manière. Comme nous le verrons plus bas, dans tout texte français, cette homographie touche plus du tiers des mots. L'énoncé du problème contient les solutions : normaliser les graphies (un mot, une seule orthographe) et résoudre les homographies en ajoutant aux mots ambigus une marque quelconque levant ces homographies.

Les travaux pionniers de G. Gougenheim, Ch. Muller, A. Julliand ou G. Engwall indiquent une voie évidente : utiliser la nomenclature des "dictionnaires de langue", nomenclature bien connue des usagers du français et devenue en quelque sorte "naturelle" au même titre que l'ordre alphabétique ou les PCS des sociologues. Certes, il existe parfois des débats entre spécialistes ou de légères différences entre les dictionnaires, mais la lexicographie, comme branche spécialisée de la lexicologie, est maintenant une science appliquée dont on aurait tort de se priver (sur ce point, voir les ouvrages classiques de C. et J. Dubois, J. Picoche ou A. Rey). Dès lors qu'il existe une nomenclature relativement solide, pourquoi ne pas l'apprendre à l'ordinateur ? C'est en nous inspirant de cette idée simple que, il y a une quinzaine d'années, nous avons mis au point une chaîne de traitement du français contemporain dans le cadre de nos travaux sur le discours politique (Labbé 1990a).

La nomenclature des mots français, telle qu'on l'a apprise à l'ordinateur, est systématique (par exemple, en français, les substantifs se distinguent par le genre, donc tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots doivent y trouver leur place), elle exclut tout double compte (pour la réduction des graphies multiples : voir la liste dressée par le CLIF), elle ne comporte pas de catégorie ad hoc ou fourre-tout, etc. Elle a été élaborée à l'aide du *Dictionnaire général* de A. Hatzfeld et A. Darmesteiter (préconisé par Ch. Muller pour la langue du XVIIe), du *Larousse*, du *Littré*, du *Robert* et du *de Villers*. Le principe général consiste à regrouper les flexions d'un même mot sous une "forme vedette" ou "lemme" auquel est associé une catégorie grammaticale. Ainsi, il existe un

accord général pour regrouper les conjugaisons d'un même verbe sous son infinitif ou les pluriels du substantif sous le singulier ou encore les féminins et pluriels de l'adjectif sous le masculin singulier. Par exemple, "être v." regroupe toutes les formes conjuguées de ce verbe, tandis que "être n. m." ne se rencontre que sous le singulier et le pluriel… Comme on le voit, cette opération oblige à lever les "homographies" (par exemple « est » : verbe "être" ou point cardinal ?). Il a donc fallu également apprendre à l'automate les principales règles de la syntaxe française. Nous nous sommes appuyé pour cela essentiellement sur les grammaires de M. Arrivé et Al., C. Blanche-Benveniste, R.-L. Wagner et J. Pinchon ainsi que sur certains articles du "Bon usage".

En résumé, à l'aide du vocabulaire et de la syntaxe du français, l'automate associe à chaque mot du texte - dont il a, au préalable, normalisé la graphie - un lemme formé d'une forme canonique et d'une catégorie grammaticale. Cette opération obéit à quelques règles simples dont la première, d'un évident bon sens, est de ne pas altérer le texte original mais d'ajouter à chaque mot une "étiquette". Par exemple :



Dans les étiquettes, en première position, on trouve la graphie normalisée : réduction des majuscules des mots communs, des graphies multiples, des abréviations, contrôle de l'orthographe des noms propres et de certains mots communs ; en seconde position : l'entrée de dictionnaire et, en troisième, la catégorie grammaticale ;

Une bonne lemmatisation doit être exhaustive et sans ambiguïté : chaque mot du texte se voit attribuer un lemme et un seul. De plus, elle est réversible, c'est-à-dire qu'on peut retrouver le texte original à partir du fichier des lemmes (nous donnons plus bas une illustration de l'utilité de ce principe). Enfin elle ne comporte pas d'erreur (nous revenons plus bas sur ce point).

L'exemple ci-dessus comporte trois homographes (le : pronom ou article ; enquêtée : substantif ou participe passé ; est : nom ou verbe). Lorsque l'automate ne parvient pas à trancher en toute certitude, il interroge l'opérateur en lui offrant les solutions possibles avec le rappel des règles (nous évoquons plus bas les raisons pour lesquelles la lemmatisation ne peut être totalement automatique).

Voici un exemple extrême d'homographie : "tout" et ses flexions, véritable "bonne à tout faire" de la langue française. Le problème est résumé dans le tableau suivant :

	déterminant	pronom	adverbe	nom
tout	x	x	x	x
toute	x	x	x	
toutes	x	x	x	
tout	x	x		

L'automate applique un nombre fini de règles qui vont lui permettre de résoudre le maximum de cas. Dans l'exemple ci-dessus, les 12 cases se ramènent pratiquement aux 4 règles fondamentales suivantes :

1) "tout" est déterminant (adjectif indéfini) quand il est employé dans un groupe nominal et qu'il est accordé aux autres éléments du groupe (tout le monde, tous deux).

2) "tout" est pronom lorsqu'il est employé seul ou associé à un groupe verbal (il a tout su).

3) "tout" est adverbe lorsqu'il est placé devant un adjectif ou employé dans une locution adverbiale ou prépositive (il est tout seul, une femme toute nue).

4) "tout" est substantif quand il est précédé d'un déterminant ou d'une préposition et suivi d'autre chose que d'un substantif ou d'un adjectif (le tout pour le tout).

L'algorithme bute sur des cas impossibles à résoudre parce que dépendant d'une interprétation méta-textuelle : "elles sont toutes contrites" peut signifier : "toutes (pronom) sont contrites" ou "elles sont extrêmement (adverbe) contrites". De même pour "ils ont tous leurs cadeaux" (à l'oral, on prononce le "s" terminal quand il s'agit d'un pronom mais à l'écrit, cette aide n'existe pas). Certaines homographies peuvent toutefois être résolues par la prise en compte de locutions (tout à coup, tout de même, après tout, en tout...)

En fin de compte, l'analyseur syntaxique bute sur environ 5% des occurrences de ce vocable du fait de ces télescopages entre pronom, déterminant et adverbe et, si l'on vise le "zéro défaut", il doit interroger l'opérateur.

Certes "tout" représente un cas extrême, dans la mesure où par exemple le déterminant peut ne pas s'accorder avec le nom qu'il détermine (j'ai lu tout Les Plaideurs), alors que l'adverbe, théoriquement invariable, pourra s'accorder pour des raisons euphoniques (elle est toute heureuse). Même si l'on récuse "l'absurdité de notre orthographe" (Valéry), il est donc irréaliste de vouloir résoudre automatiquement et sans erreur tous les cas d'homographie. En voici un autre exemple tiré de l'application qui va être présentée dans un instant : "je suis une femme" sera probablement "être" si c'est une femme qui parle mais "suivre" si la phrase est prononcée par un dragueur. Toutefois, ce critère du genre du locuteur est dangereux : il peut s'agir d'une assistante sociale qui parle d'un cas qu'elle suit… Dans les entretiens ou les questions ouvertes "tout" ou "suis" figurent toujours parmi les 100 premières formes les plus fréquentes. Cependant, comme on va le voir, il ne faut pas s'exagérer l'importance de ces difficultés : les interrogations que l'automate, d'ailleurs perfectible, renvoie à l'opérateur portent sur 1 à 2% des mots comme nous allons le voir à l'aide d'un exemple.

II Une application

A la demande de la CNAF (Caisse Nationale d'Allocations Familiales), l'Institut National d'Etudes Démographiques, associé à l'INSEE, a mené en 1985 une enquête nationale sur "Les femmes face au changement familial" (Festy et Valetas ; Garnier et Guérin-Pace). Un échantillon représentatif de 2.329 femmes, séparées de leur mari au cours des 15 années précédant l'enquête, ont été interrogées sur les raisons de leur divorce et ses conséquences. Cette enquête comportait une question ouverte ainsi formulée :

"Quelles étaient les raisons à l'origine de votre mésentente ?"

Si l'enquêtée ne livrait pas spontanément les motifs de la rupture, l'enquêteur devait la relancer en lui demandant :

"Qu'est-ce qui a effectivement provoqué la séparation ?"

L'INED a bien voulu remettre ces données à M. Becue pour une exploitation secondaire (présentée notamment dans le mémoire de A.-B. Pizarro-Diaz et M. Trujillo-Palomo). M. Becue a effectué un "nettoyage" et une correction orthographique approfondie de ces réponses et nous les a confiées. Nous les avons importées du format "Spad" vers nos propres outils puis réexportées, après normalisation et lemmatisation, dans un format utilisable par ce logiciel.

En quelque sorte, il s'agissait de mettre en place une dérivation permettant d'extraire les réponses à la question ouverte, de les envoyer vers le lemmatiseur puis, de rediriger les réponses normalisées et lemmatisées vers le logiciel de traitement de données. Au cours de cette boucle, il est évidemment

possible d'effectuer certains traitements lexicométriques dont les résultats viennent compléter les résultats des analyses standard (nous donnons ci-dessous quelques exemples).

Naturellement, chacune de ces étapes pose quelques problèmes.

En premier lieu, lors de l'importation des données, il faut isoler tout ce qui n'est pas le texte par des balises. En l'occurrence, les données importées de Spad se présentaient de la manière suivante ;

----0004
Points de vues différents : éducation enfants, relations avec amis.
r - Egoïsme, il s'achetait des objets chers, pour lui-même, sans subvenir aux besoins des enfants.
++++

Le programme d'importation place des balises pour isoler le début de la question et son numéro (----0004), l'indication de la relance éventuelle de l'enquêteur (r-), la fin de la question (++++). Le reste est considéré comme du texte et fait l'objet de la normalisation et de la lemmatisation.

De même, lors de l'exportation des réponses normalisées et lemmatisées, il faut tenir compte de ce que les logiciels de traitement de données ne travaillent que sur les formes graphiques et ne peuvent lire les étiquettes. On a généré deux fichiers lisibles par Spad. Le premier comportait les formes graphiques normalisées. Dans l'exemple ci-dessus : "points" est mis en minuscule, "lui_même" est soudé en une forme, etc. Le second fichier comportait les lemmes auxquels on a attaché la catégorie grammaticale : "ilPro sePro acheterVer dePre leDet objetNm", etc. D'autres formats de données sont concevables, notamment ceux des "syntagmes répétés" : en s'appuyant sur les catégories grammaticales, le programme élimine les "mots outils" et conserve les groupes nominaux et verbaux (Pibarot-Labbé).

Au cours de chacune des phases de l'opération, divers problèmes sont survenus.

En premier lieu, l'automate n'est pas parvenu à reconnaître quelques unes des 55600 formes brutes. Voici la liste de ces échecs.

Echecs de la normalisation (formes non reconnues, en gras les fautes d'orthographe)

àu	**1**	flambeur	1	piquette*	1
bringueur	1	foirard	1	pompette	1
ca	**5**	**foutiste**	**1**	radinerie	1
cavaleur	9	frites*	1	rouleur	1
coquard	1	insupportait	1	**strip**	1
crane	**1**	mac	1	**strip tease**	1
d2	1	**méditerranée**	**1**	**teaseuse**	1
delirium	**1**	nénette	1	traditionalisme	1
etaient	**7**	**oedipe**	**1**		
etais	**2**	parano	2		

* l'homographie figure dans la table mais le verbe n'est pas reconnu.

Les mots suivants manquaient dans la nomenclature : bringueur, cavaleur, coquard, foirard, mac, nénette, parano, pompette, radinerie, rouleur et les verbes friter, piquetter (mais pas les substantifs frite et piquette). Ceci vient de ce que nous avions travaillé jusqu'à maintenant sur du français "soutenu"… Il y avait également une faute d'orthographe dans une des tables (traditionalisme était écrit avec deux « n ») et un préfixe oublié pour le verbe "supporter".

Le texte comportait aussi quelques fautes d'orthographe (qui avaient échappé à la correction préalable déjà considérable). En plus de celles figurant en gras dans le tableau, et détectées lors de la phase de normalisation, on a d'ailleurs relevé ultérieurement une demi-douzaine de prépositions « à » sans accent (confondues avec le verbe « avoir »), quelques « dû » (participe passé de devoir) écrits sans l'accent, etc. Ces échecs doivent être rapportés à l'effectif total soit 55 606 mots. A ce stade et sous réserve d'une correction orthographique préalable, le taux d'échec est d'environ un pour mille ou encore : à l'issue de la première phase, on a identifié 99,9% des mots. Le tableau ci-dessous donne le détail de ces résultats.

	55 606 mots = 100	%
Nombre d'échecs		0,09
Nombre de formes identifiées		99,91
Nombre d'homographies potentielles		36,70
Dont:		
Au sein d'une même catégorie		2,49
Entre substantifs et adjectifs (autres que participes)		1,82
Entre verbes et substantifs ou adjectifs		15,55
Entre verbes et autres catégories		1,18
Entre substantifs et autres catégories		4,15
Entre autres catégories que verbes et substantifs		11,39
Dont déterminants et pronoms		8,85
pronoms-prépositions		0,90
adjectifs-pronoms		0,02

A l'issue de cette première phase, 37% des mots peuvent être rattachés à plus d'une entrée de dictionnaire. Attention, suivant les conventions lexicographiques (et les conseils de Muller), la nomenclature est très synthétique. Naturellement, des nomenclatures plus fines (incluant les personnes et les temps des verbes, le partitif "de", les homographies entre prépositions et adverbes…) pourraient conduire à considérer que pratiquement la moitié des mots sont homographes…

La deuxième phase consiste en une résolution automatique de ces homographies, le logiciel utilisant le « contexte » (les mots situés devant et derrière l'homographe). Lorsqu'il ne peut conclure avec certitude, le logiciel propose à l'opérateur les différentes solutions possibles. Dans le cas présent, il a dû le faire à 667 reprises (pour 55 606 mots traités). Le taux de reconnaissance automatique est donc de : 98,8 %. Ce taux est dans la moyenne de ce que l'on a obtenu sur les autres corpus traités à ce jour. Cependant, étant donné la faible étendue du vocabulaire des réponses, on aurait pu s'attendre à ce qu'il soit un peu meilleur.

L'annexe I. résume les cas non-résolus (jusqu'à la fréquence 2). Cette liste donne une idée des principales difficultés rencontrées par l'automate qui, rappelons-le, a été conçu pour l'analyse d'un français "soutenu". En fait, la principale difficulté réside dans l'écriture télégraphique de certaines réponses. Parmi les problèmes les plus fréquents, outre « que » (pronom/conjonction), on trouve :

• la phrase construite sans verbe du type : « **pas** de raison » : substantif (« (un) pas de danse ») ou adverbe ?

• l'impossibilité de choisir entre groupe nominal et groupe verbal :

— "**nouvelle rencontre**" : (adjectif ou substantif) + (substantif ou verbe)... Au passage, l'écriture télégraphique des réponses a permis de révéler une faiblesse de l'analyseur syntaxique : « rencontre » ne peut être verbe qu'avec un COD devant ("il la rencontre") ou derrière ("cette nouvelle rencontre le scepticisme")...

— « **manque** d'argent » : « (le)» ou « (il) manque d'argent » ?

Enfin, une relecture attentive de l'ensemble du fichier lemmatisé a permis de déceler 10 erreurs dans l'analyse automatique, entraînant 16 mots codifiés de manière erronée dont un « que » et deux « tout ». Voici les principales erreurs :

• Les participes passés sans auxiliaire :

— "jamais **été** là" ("été" est analysé comme un substantif puisque, en français soutenu le participe passé est toujours précédé d'un auxiliaire…)

— "**rendu** compte de rien", "**parti** d'un seul coup" : "rendu" et « parti » passent adjectifs pour les mêmes raisons.

• « c'était **le supporter** ou partir ». « le supporter » est analysé comme un couple article+nom (sur le patron « c'était le pouvoir ou rien »). Les règles concernant cette construction doivent donc être affinées.

• adverbe et adjectif : dans "pas de sortie **même** en famille " "même" est codé adjectif (rattaché à "sortie") au lieu d'adverbe.

• nom et verbe : « rester comme ça **le reste** de mes jours » : "le" est codé « pronom » et reste : « verbe » (ça le reste). Il manque donc une règle concernant la locution « comme ça » dans laquelle "ça" ne se comporte pas en pronom par rapport à ce qui suit...

Le lemmatiseur n'est donc pas parfait. Mais, sous réserve, d'une correction orthographique préalable puis d'une intervention manuelle de l'opérateur pour un peu plus d'un mot sur cent, on obtient une lemmatisation sûre à plus de 99,97 %. Il ne s'agit pas d'un "zéro faute", mais les erreurs relevées étaient évitables moyennant un enrichissement des règles d'analyse syntaxique. Rappelons que le programme est expérimental et qu'il a été conçu il y a 15 ans. De meilleurs ratios sont certainement possibles. En revanche, une codification entièrement automatique devrait faire appel, pour 1 à 2% des mots, à des décisions probabilistes, type "chaînes de Markov", dont le rendement s'avère décevant : le nombre d'erreurs serait donc certainement significatif. En l'état actuel de la recherche, il semble préférable d'avoir recours à l'opérateur en encadrant rigoureusement ses décisions…

III. Impact et intérêts de la lemmatisation

Ces traitements ont d'abord permis de mesurer l'impact de la lemmatisation. Le tableau ci-dessous récapitule les principales dimensions caractéristiques issues des trois dépouillements possibles : les "formes brutes" (après correction orthographique), les formes normalisées, les lemmes.

Principales dimensions caractéristiques avec les trois normes de dépouillement

	Formes brutes	Formes Normalisées	Lemmes
Taille (N)	55 606	55 399	56 107
(Nombre total de mots)	100	99,6	100,9
"Vocabulaire" (V)	4 324	3 780	2 785
(Nombre de mots différents)	100	87,4	64,4
Mots de fréquence > 23	299	266	254
(% V)	6,9	7,0	9,1
	100	101,8	131,9
Surface du texte total	43 522	45 321	47 594
(% de N)	78,3	81,8	84,8
	100	104,5	108,4
Hapax :	2 165		1 224
(fréquence 1)	100		56,5
% V	50,1		43,9
	100		87,8

En premier lieu, le tableau permet de mesurer l'impact des opérations sur la taille du fichier. Dans les formes brutes, "aujourd'hui" ou "parce que" donnent deux formes qui sont agglutinées en une seule dans le fichier des formes normalisées. Dès lors il est logique que la normalisation réduise légèrement le nombre des mots. En revanche, cet effectif remonte dans le fichier lemmatisé du fait des formes "contractes". Par exemple, "du" possède deux étiquettes ("de le") : cette opération est logique si l'on songe qu'aucun dictionnaire ne comporte d'entrée à "du"… Il ne faut donc pas s'exagérer l'impact de ces opérations qui entraînent des fluctuations de ± 1%. En revanche, il est évident qu'il faudrait toujours préciser l'unité utilisée quand on annonce la longueur d'un corpus.

Le nombre de "mots différents" (V) est lui, beaucoup plus nettement affecté par la normalisation (qui réduit le vocabulaire de 13%) et par la lemmatisation (-35,5). Ces gains ne sont pas négligeables, même s'ils peuvent paraître un peu limités. En fait, comme nous l'avons déjà suggéré, les enquêteurs ont déjà réalisé une partie du travail en utilisant le "style télégraphique". Il faut également tenir compte de la taille du corpus. En effet, le "rendement" de la normalisation et de la lemmatisation est proportionnel à cette taille : V est réduit de moitié sur des corpus de 150.000 mots environ.

Le cadre suivant du tableau s'explique par le fait que les auteurs du mémoire déjà cité (Pizarro-Diaz et Trujillo-Palomo) ont choisi de limiter leur analyse aux mots dont la fréquence est supérieure à 23. Pour les formes graphiques brutes, ce seuil revient donc à ne considérer que 6,9% du vocabulaire, proportion qui passe à 9,1 % avec les lemmes (soit un gain de 32%). On considère alors près de 85% de la surface

du texte (sous réserve de ce que nous avons dit plus haut à propos de la quantité d'information véhiculée par ces mots les plus fréquents).

Enfin, les effectifs des basses fréquences, et notamment des hapax, sont fortement réduits. C'est là un avantage important puisque le statisticien n'a rien à dire sur un fait unique, sinon qu'il est unique…

Examinons un peu plus en détail la distribution des fréquences (tableau et graphique ci-dessous). La première ligne du tableau signale un phénomène général : les gains les plus significatifs apportés par la normalisation des graphies et par la lemmatisation se situent toujours dans les plus hautes fréquences. Ce phénomène s'explique aisément : c'est parmi les formes les plus fréquentes qu'on rencontre le plus de graphies différentes pour un même mot (majuscules initiales de phrase, élisions, etc.). On pourra le vérifier en consultant le premier tableau de l'annexe II. Nous y avons mis en gras les "fantômes" que la normalisation des graphies permet de chasser des textes : cette simple opération concerne environ un cinquième des mots les plus fréquents. Mais le gain essentiel est apporté par le regroupement sous une même entrée des flexions de la préposition "de", de l'article "le" ou des verbes usuels comme l'indique la comparaison entre les deux tableaux de l'annexe II.

Dans toutes nos expériences antérieures, nous avons également constaté des gains importants au bas du dernier quartile, c'est-à-dire, dans ce corpus, dans les fréquences comprises entre 5 et 10. Naturellement, plus la taille du corpus augmente, plus cette limite sera située haut dans la distribution…

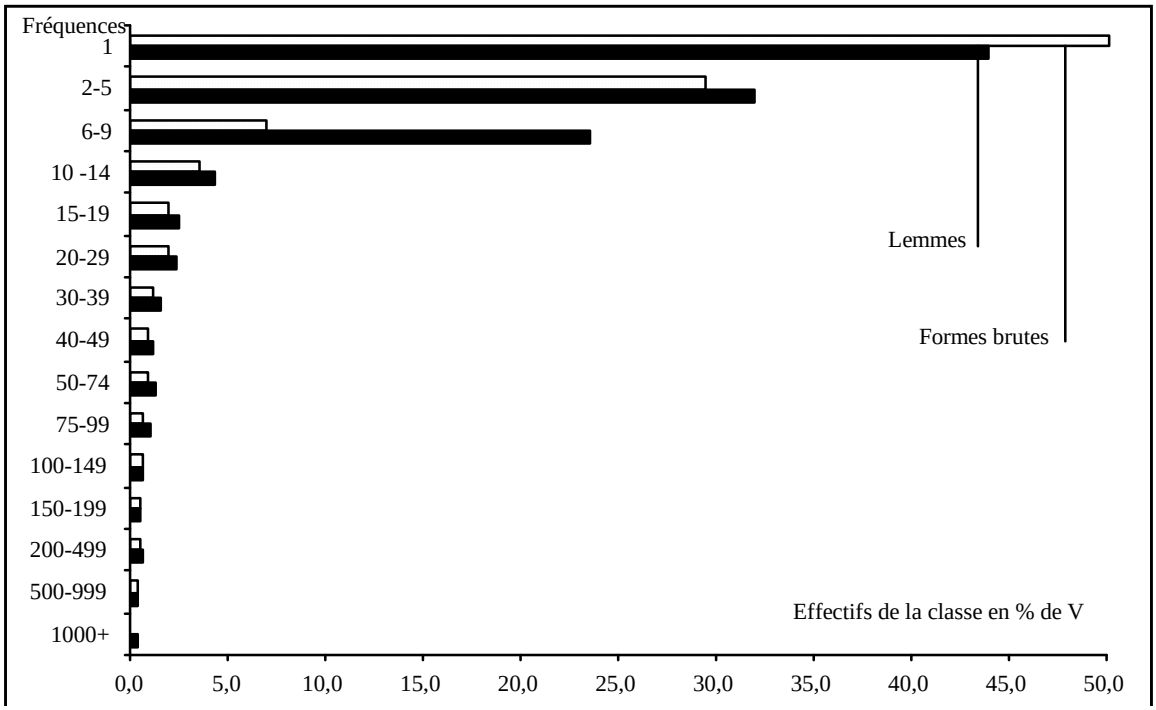
Comparaison des distributions des fréquences (réponses "brutes" et réponses lemmatisées).

Fréquences	Formes (% de V)**	Lemmes (% de V)*	Gains de la lemmatisation (%)
1000+	0,1	0,4	+ 242
500-999	0,4	0,4	- 6
200-499	0,6	0,6	+ 10
150-199	0,5	0,6	+ 13
100-149	0,6	0,6	+ 7
75-99	0,6	1,0	+ 67
50-74	0,9	1,3	+ 40
40-49	0,9	1,1	+ 21
30-39	1,2	1,6	+ 37
20-29	1,9	2,4	+ 22
15-19	2,0	2,5	+ 30
10 -14	3,6	4,3	+ 21
6-9	7,0	23,6	+ 236
2-5	29,5	32,0	+ 8
1	50,1	43,9	- 12

* 2 785 lemmes différents

** 4 324 formes brutes différentes.

Effectifs relatifs des différentes classes de fréquences (en % de V) : formes brutes et lemmatisées



Naturellement, le graphique montre que le profil de la distribution des fréquences reste caractéristique des données textuelles : le vocabulaire est surtout constitué d'un grand nombre de mots rares couvrant un faible pourcentage de la surface totale mais véhiculant probablement une part importante de l'information.

La normalisation des graphies est une opération de bon sens sur laquelle nous ne reviendrons pas. En revanche, qu'apporte la lemmatisation ? Pour le savoir, les auteurs du mémoire déjà cité ont comparé les résultats obtenus avec les formes graphiques normalisées puis avec les lemmes. Au préalable, ils se sont livrés à une classification traditionnelle sur les variables "lourdes" (âge, durée du mariage, nombre d'enfants du couple, niveau d'études…) utilisant la réponse ouverte comme illustration des différentes classes obtenues à l'aide de ces variables. A titre d'exemple, le tableau ci-dessous reproduit le vocabulaire caractéristique de la première des classes ainsi formées (formes graphiques normalisées et lemmes).

On constate plusieurs recouvrements entre les deux colonnes, notamment les substantifs "maîtresse", "alcoolisme", "alcool", de même que "partir" et "buvait/boire". La lemmatisation permet d'obtenir une liste plus courte et d'éliminer certains accidents comme "ai", "dit" ou "voulait" (les infinitifs correspondants n'apparaissent pas dans la colonne de gauche). Elle permet surtout d'affirmer que le premier thème caractéristique de ce groupe est le verbe "partir" auquel on peut probablement associer le pronom "il" et le mot "maîtresse" (les syntagmes répétés dans cette classe permettent de contrôler cette association), le second thème est formé de : "boire + alcoolisme + alcool + boisson" et "battre" forme sans doute le troisième thème (ce dernier étant absent de la liste des formes).

Classe 1. Femmes nées entre 1938 et 1947, couple avec enfant(s), durée du mariage moyenne, bas niveau d'instruction. Classement des mots selon l'indice de spécificité.

	Formes graphiques normalisées	Lemme
1	ai	partir (v)
2	il	il (pro)
3	maîtresse	boire (v)
4	dit	temps (n m)
5	si	là (adv)
6	voulait	je (pro)
7	la	alcoolisme (n m)
8	temps	si (conj)
9	partir	maîtresse (n f)
10	alcoolisme	toujours (adv)
11	me	jamais (adv)
12	toujours	pendant (pré)
13	là	battre (v)
14	jamais	alcool (n m)
15	buvait	boisson (n f)
16	pris	pour (pré)
17	pendant	marre (n f)
18	partie	—
19	alcool	—
20	parti	—

De manière générale, l'utilisation des lemmes permet de retrouver les verbes caractéristiques au-delà de flexions circonstancielles, d'obtenir des listes plus courtes et plus éclairantes. C'est également la conclusion de l'étude citée selon laquelle les lemmes "illustrent mieux les thèmes caractéristiques des réponses. C'est-à-dire qu'on retrouve bien les causes de divorce caractéristiques de chaque classe en lisant les réponses caractéristiques de cette classe mais c'est un travail laborieux et subjectif que la lemmatisation rend inutile en livrant les thèmes sans avoir besoin de relire les réponses" (p 79).

Au-delà de ces gains évidents, les données lemmatisées apportent beaucoup d'autres informations intéressantes. Elles facilitent notamment les recherches directes dans les réponses. Par exemple, à la question : "qui est parti du mari ou de la femme ?", on peut fournir des éléments de réponse en recherchant les concordances des principaux mots correspondant à ce thème, notamment celles des verbes, comme "partir", "quitter" ou des substantifs comme "départ"… Nous donnons en annexe la première page de la concordance du verbe "partir" qui apparaît en tout 362 fois dans le corpus. Ces concordances permettent d'identifier la personne qui est partie dans 304 cas. Pour 228 d'entre eux (75%), c'est le mari qui est parti et dans 76 cas c'est la femme. "Quitter" permet d'identifier 40 cas (mais certains étaient déjà connus grâce à "partir") : dans 27 d'entre eux, c'est le mari qui a quitté (le domicile conjugal plus souvent cité que la personne de l'épouse…) et dans 13 c'est la femme qui l'a quitté (dans ce cas c'est l'homme qui est quitté, pas le domicile…)

Mais surtout, la normalisation et la lemmatisation rendent les données textuelles comparables d'une enquête à l'autre, du moins s'il n'y a pas d'erreurs dans la normalisation et la lemmatisation… Le cumul raisonné de diverses enquêtes permettrait de constituer une sorte de "base de comparaison" à partir de laquelle nous pourrions savoir en quoi la parole de ces femmes est ou non singulière. Pour illustrer l'intérêt de cette démarche, voici les résultats obtenus en utilisant comme base de comparaison un large

échantillon du français oral contemporain[1].. L'expérience met à jour certaines caractéristiques singulières des réponses (tableau ci-dessous) et permet d'identifier les thèmes spécifiques développés dans l'enquête (annexe IV).

Densité comparée des catégories grammaticales employées dans les réponses comparées au "français oral contemporain".

	A (Corpus de référence)	B (Corpus étudié)	B-A
Noms propres	0.6	0.2	- 68.8
Verbes	19.3	20.4	+ 6.0
Formes fléchies	*13.4*	*13.1*	*- 2.5*
Participes passés	*2.6*	*4.3*	*+ 70.0*
Participes présents	*0.1*	*0.2*	*+ 64.5*
Infinitifs	*3.2*	*2.8*	*- 11.4*
Substantifs	14.0	18.1	+ 29.9
Adjectifs	3.5	4.4	+ 25.7
Adj participe passé	*0.3*	*0.4*	*+ 27.3*
Pronoms	19.2	17.1	- 11.1
Pronoms personnels	*10.0*	*12.4*	*+ 23.6*
Determinants	13.0	14.0	+ 7.3
Articles	*9.4*	*8.9*	*- 5.8*
Nombres	*1.5*	*0.8*	*- 42.7*
Possessifs	*0.7*	*2.9*	*+ 320.4*
Démonstratifs	*0.4*	*0.2*	*- 48.1*
Indéfinis	*1.0*	*1.1*	*+ 10.2*
Adverbes	11.1	9.7	- 12.3
Prépositions	11.5	11.8	+ 2.1
Conjonctions	7.0	4.2	- 39.4

Par rapport au corpus de référence, l'écart dans la densité des catégories grammaticales est très significatif. Ces écarts peuvent provenir de deux causes : la nature particulière des questions posées ou la transcription des réponses par certains enquêteurs. Par exemple, l'excédent considérable du participe passé est probablement dû au contenu même de la question : on demande à l'enquêtée de raconter les circonstances de la mésentente, il est donc normal que la réponse soit formulée au passé (on en trouve

[1] Cette base d'entretiens a été constituée grâce à des travaux antérieurs. Il s'agit de 190 transcriptions d'enregistrements oraux, réalisées en suivant les mêmes règles, et comportant au total 1.250.600 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), 55 entretiens sur les relations professionnelles et la négociation collective au Québec (Bergeron-Labbé, 2000), 64 entretiens sur "le confort électrique" remis par la Division Recherche et Développement d'Electricité de France (dont une analyse a été présentée récemment dans cette même salle), etc. Même s'il s'agit du plus grand corpus étiqueté existant sur le français oral, il n'atteint pas encore la taille et la diversité qu'exigerait un véritable échantillon représentatif.

confirmation dans l'annexe IV : l'auxiliaire "avoir" est le verbe le plus spécifique de ces réponses par rapport au français oral). De même, l'excédent des pronoms personnels (*il* et *je*) et des possessifs (*mon* et *son*) signale la principale caractéristique des réponses — le face-à-face des époux — encore soulignée par le suremploi de l'indéfini "autre" ("une autre femme", "une autre personne", "une autre rencontre").

En revanche, le déficit considérable en conjonctions ou en pronoms relatifs est certainement dû à la transcription : pour gagner du temps, l'enquêteur saute assez systématiquement les mots de liaison et coupe les phrases complexes. Il privilégie, de ce fait, le groupe nominal comme l'indique l'excédent des substantifs, des adjectifs et de la préposition "de". Pour les mêmes raisons, il omet assez systématiquement les noms propres, les dates, les chiffres et tout ce qu'il considère comme du "détail"…

Enfin le vocabulaire caractéristique permet de détecter un autre biais fréquent : la transcription à la troisième personne d'une réponse certainement faite à la première personne. Certains enquêteurs utilisent même des formules comme l'"enquêtée" (présent dans 44 questionnaires, ce qui explique sa présence dans l'annexe IV (à la troisième ligne des substantifs caractéristiques entre "dispute" et "alcoolisme")… A titre d'exemple, cette réponse : "Cruauté mentale (je souligne à la demande de l'enquêtée qui regrette que cette notion ne figure pas dans les motifs de divorce en France)" (0384).

La comparaison avec le corpus de référence permet aussi d'isoler les phrases les plus caractéristiques de l'enquête. Par exemple, voici les huit phrases dont tous les mots sont caractéristiques :

2999	Il ne s'occupait jamais de son ménage ni des enfants.
2976	Un mari très possessif, trop jaloux, pas assez de maturité.
1975	Il ne prenait pas de responsabilités, ne prenait pas de décision.
1789	Trop de sorties, trop de dépenses et trop d'alcool.
1767	Il sortait avec d'autres femmes et ne travaillait pas.
1617	Il ne travaillait pas et ne cherchait pas de travail.
0429	Liaison du mari, incompatibilité de caractères, découverte de personnalités différentes.
0949	Il buvait, ne travaillait pas et il me battait, il avait une maîtresse et un enfant.

En conclusion, nous voudrions souligner trois points.

En premier lieu, l'exploitation des questions ouvertes dans les sondages repose sur deux préalables. D'une part, leur transcription doit être effectuée par les enquêteurs avec le même soin qu'ils mettent à documenter les questions fermées. D'autre part, la saisie doit également être améliorée et systématiquement suivie d'une correction orthographique sérieuse.

Deuxièmement, nous n'avons évoqué qu'un petit nombre des voies ouvertes par la normalisation et la lemmatisation, mais nous espérons avoir montré qu'elles sont des techniques intéressantes et respectueuses du matériel original. En tous cas, elles sont préférables à d'autres méthodes comme l'élimination des accents ou la troncature des mots pour les réduire à des racines supposées communes…

Troisièmement, normalisation et lemmatisation ne sont qu'un premier pas. Pour reprendre la comparaison qui ouvre cet exposé, ce travail préalable aboutit à quelque chose de comparable au "répertoire des métiers" à partir duquel sont constituées les PCS (dont le niveau le plus fin compte d'ailleurs quelque 480 postes). Puisque des agrégations successives se révèlent possibles sur l'infinie diversité des activités humaines, pourquoi ne le seraient-elles pas sur le vocabulaire usuel de la langue orale ? Les outils de la lexicologie — la synonymie, l'hyperonymie, l'antonymie, etc. — ne demandent qu'à être automatisés. Appliqués à de grandes bases de données étiquetées, ils permettront de reconstituer les principaux champs lexicaux à partir desquels il sera possible de donner aux réponses aux questions ouvertes la même puissance explicative qu'aux variables socio-démographiques.

Annexe I.
Les cas non-résolus par l'analyseur syntaxique

Lemme	Catégorie	Effectifs
que	conjonction	46
pas	adverbe	45
que	pronom	45
manque	nom masc	41
partir	verbe	34
adultère	nom masc	26
dispute	nom fem	20
le	article	16
rencontre	nom fem	16
violent	adj	16
en	pronom	12
avorter	verbe	8
fait	adj	7
jaloux	nom masc	7
marié	adj	7
parti	adj	7
tout	pronom	7
tout	déterminant	7
vivre	verbe	6
départ	nom masc	5
égoïste	adj	5
être	verbe	5
cause	nom fem	4
célibataire	nom masc	4
découverte	nom fem	4
fréquent	adj	4
méchant	adj	4
parler	verbe	4
radin	adj	4
sortie	nom fem	4
sortir	verbe	4
tout	adverbe	4
autre	pronom	3
battu	adj	3
bien	adverbe	3
couple	nom masc	3
coureur	nom masc	3
ensemble	adverbe	3
faire	verbe	3
fils	nom masc	3
jeune	adj	3
malade	adj	3
manger	verbe	3
ménage	nom masc	3
alcoolique	nom masc	2
amie	nom fem	2
amoureux	adj	2
après-midi	nom masc	2
avant	préposition	2
avare	nom masc	2
bonne	nom fem	2
claque	nom fem	2
conjoint	adj	2
coureur	adj	2
divorce	nom masc	2
élevé	adj	2
enceinte	adj	2
enquêtée	nom fem	2
entre	préposition	2
fainéant	adj	2
fatigue	nom fem	2
fou	adj	2
fugue	nom fem	2
handicapé	adj	2
impossible	adj	2
le	pronom	2
mort	nom fem	2
multiple	adj	2
parent	nom masc	2
pêche	nom fem	2
restaurant	nom masc	2
savoir	verbe	2
secrétaire	nom fem	2
sérieux	nom masc	2
soit	conjonction	2
un	pronom	2
voiture	nom fem	2
vue	nom fem	2

Annexe II.A
Les 100 formes graphiques brutes les plus fréquentes
(en italiques, les variantes graphiques d'une même forme)

Formes	Fréquence	mon	316	ça	159
de	1814	du	302	se	158
il	1690	enfants	300	ce	148
Il	***1169***	qui	294	elle	147
et	1117	sa	292	avais	145
pas	1073	tout	276	maison	142
ne	906	son	268	ses	142
a	824	femme	264	sur	142
était	813	lui	263	toujours	140
la	792	***m***	***256***	étais	139
à	756	***c***	***251***	même	138
d	***719***	***Je***	***250***	eu	136
est	683	argent	221	enfant	133
le	655	buvait	205	***L***	***133***
un	644	parti	199	ma	129
n	***638***	suis	198	mais	129
l	***630***	très	193	vivre	129
mari	623	au	193	***On***	***126***
une	613	on	192	rien	126
je	612	nous	192	par	121
avait	591	***J***	***189***	été	116
plus	515	dans	187	***quelqu***	***115***
les	514	voulait	185	mère	106
avec	504	fait	184	***Incompatibilité***	***106***
me	493	vie	183	sans	105
ai	485	***Mon***	***179***	parents	101
en	484	moi	179	ex	101
que	430	trop	176	humeur	100
des	425	***C***	***174***	famille	97
j	***413***	jamais	172	beaucoup	97
s	***383***	travail	168	travaillait	96
pour	377	***Le***	***167***	caractère	96
qu	***344***	***La***	***166***	boisson	94
autre	319	y	160		

Annexe II. B

Les 100 vocables les plus employés (lemme et catégorie grammaticale)

Rang	Vocable	Fréquence
1	le (det)	3717
2	de (pré)	3306
3	il (pro)	3050
4	avoir (v)	2461
5	je (pro)	2209
6	être (v)	2160
7	ne (adv)	1569
8	un (det)	1268
9	et (cj)	1204
10	pas (adv)	1139
11	à (pré)	1039
12	son (det)	797
13	mon (det)	731
14	que (cj)	565
15	mari (n m)	556
16	ce (pro)	552
17	se (pro)	535
18	plus (adv)	521
19	avec (pré)	511
20	enfant (n m)	437
21	pour (pré)	392
22	partir (v)	362
23	le (pro)	355
24	vouloir (v)	342
25	en (pré)	339
26	femme (n f)	339
27	faire (v)	323
28	on (pro)	318
29	qui (pro)	294
30	lui (pro)	270
31	boire (v)	251
32	nous (pro)	251
33	autre (det)	238
34	argent (n m)	222
35	très (adv)	208
36	ça (pro)	203
37	trop (adv)	194
38	vie (n f)	192
39	caractère (n m)	189
40	dans (pré)	187
41	vivre (v)	187
42	travailler (v)	185
43	travail (n m)	180
44	jamais (adv)	179
45	en (pro)	176
46	tout (pro)	176
47	moi (pro)	174
48	dire (v)	170
49	que (pro)	162
50	y (pro)	160
51	rien (pro)	152
52	tout (det)	152
53	sur (pré)	148
54	toujours (adv)	148
55	rentrer (v)	146
56	mais (cj)	145
57	maison (n f)	143
58	aller (v)	141
59	incompatibilité (n f)	137
60	pouvoir (v)	137
61	quand (cj)	134
62	par (pré)	127
63	autre (pro)	124
64	entendre (v)	123
65	ce (det)	122
66	jour (n m)	121
67	quelqu'un (pro)	115
68	maîtresse (n f)	114
69	battre (v)	109
70	beaucoup (adv)	109
71	sans (pré)	108
72	mésentente (n f)	107
73	supporter (v)	107
74	boisson (n f)	106
75	prendre (v)	106
76	violent (adj)	102
77	humeur (n f)	101
78	sortir (v)	101
79	connaître (v)	98
80	infidélité (n f)	98
81	savoir (v)	98
82	an (n m)	95
83	ex-mari (n m)	95
84	voir (v)	95
85	copain (n m)	91
86	problème (n m)	91
87	aimer (v)	89
88	bien (adv)	89
89	occuper (v)	89
90	après (pré)	88
91	deux (num)	88
92	fille (n f)	88
93	mettre (v)	88
94	marier (v)	87
95	mariage (n m)	86
96	chez (pré)	85
97	raison (n f)	85
98	seul (adj)	85
99	notre (det)	84
100	famille (n f)	83

Annexe III. Concordances du verbe : partir
Fréquence du vocable : 362

0008	Délaissement évident de la part du mari, j'ai demandé qu'il	parte	: il est parti. Même caractère : d'où heurts, absence
0008	vident de la part du mari, j'ai demandé qu'il	parte	: il est parti . Même caractère : d'où heurts, absences inexpliquées
0015	ient … Il buvait et j'ai appris qu'il avait une maîtresse. À	partir	de ce moment là, tout a été très vite, tout était fi
0015	, tout a été très vite, tout était fini entre nous et il est	parti	rejoindre sa maîtresse. Problèmes d'argent : restaura
0021	mes d'argent : restaurant fait ensemble. Brutalement, il est	parti	avec une personne de connaissance. Cette personne. No
0030	s'entendre il y avait des hauts et des bas, un jour je suis	partie	je ne pouvais plus, j'étais dépressive, ma vie était
0042	e quelques mois en montagne et moi je ne voulais pas, il est	parti	mais sans l'idée de divorcer et c'est moi qui ai déci
0049	nous étions deux étrangers. Absences prolongées sans motif,	partait	sans laisser de nouvelles, plus d'argent, dettes. P
0062	s, buts, idéaux communs, en réalité pas de disputes. Je suis	partie	. J'étais malade mentalement, il ne comprenait pas l
0086	, il, vivait pour lui seul, souvent hors du domicile. Il est	parti	vivre ailleurs. Instabilité chez mon mari. Je suis pa
0087	est parti vivre ailleurs. Instabilité chez mon mari. Je suis	partie	. Pas les mêmes goûts. Il ne voulait pas d'enfant et
0089	ts. Il ne voulait pas d'enfant et j'en attendais un. Il est	parti	. Je n'aimais pas la Réunion, je ne me suis jamais ha
0110	devenir mac, le lendemain j'ai pris mes affaires et je suis	partie	. Il jouait aux courses. Elle a épousé une personne
0129	trompait. Il me le cachait et je m'en suis rendue compte, à	partir	de là j'ai demandé le divorce. La boisson, il buvait
0132	ute. Pas d'argent et il s'est mis à jouer au casino et c'est	parti	. Il jouait la paye et les allocations familiales, j'
0132	supporter pendant trois ans, et au bout de trois ans je suis	partie	, si on n'avait eu un logement au départ on serait t
0132	en Suisse, il dépensait son argent avec des secrétaires, il	partait	au milieu du repas. Plus de fric, plus de contacts
0135	ais beaucoup trop jeune, je me suis mariée à quinze ans pour	partir	de chez moi, j'étais enceinte, manque de maturité de
0138	personne à une réunion de famille le premier janvier et est	parti	avec, plantant là son épouse. Tout, nous n'allions pa
0154	était assez désintéressé après le décès de mon père, c'est à	partir	de là, et en grande partie à cause de ça, je ne veux
0156	a rencontré une autre femme qui lui plaisait mieux, il est	parti	avec pas de mésentente. Coup de foudre de mon mari po
0177	ngées par son travail, et il s'est mis à boire. Mon mari est	parti	parce qu'il a connu quelqu'un d'autre : il est parti
0177	ari est parti parce qu'il a connu quelqu'un d'autre : il est	parti	du jour au lendemain oui, du jour au lendemain. Le co

0183	mêmes idées, il voulait toujours commander. Mon ex-mari estparti	avec une autre femme avec laquelle il a eu un enfant.	
0264	manque de responsabilités. Jamais ensemble les week-ends, il	partait	souvent à la chasse et sortait beaucoup avec ses co
0303	pour des problèmes sexuels (homosexuel). Infidélité, il estparti	avec une femme qui avait deux ans de plus que notre f	
0376	it comme il respirait, il avait quelqu'un. C'est lui qui est	parti	avec une fille qu'il connaissait depuis neuf mois. Je
0398	rosse scène de jalousie, plus boisson, plus violent, je suis	partie	. Questions financières, je n'avais jamais d'argent,
0399	s en plus, le médecin m'a fait un certificat et conseillé de	partir	. La voisine. Action en divorce parce que j'en avais
0428	t sa femme comme une esclave. Alcoolisme du mari lui, il est	parti	sur Paris en soixante dix, moi je suis restée à Toulo
0430	e ça va s'arranger, puis ça ne s'est pas arrangé. Les jeunes	partent	tout de suite, mais moi je n'ai pas voulu, j'ai éle
0430	vé mes filles, elles me reprochent maintenant de ne pas être	partie	avant. Je ne sais pas, mon mari est parti sur un cou
0433	nt de ne pas être partie avant. Je ne sais pas, mon mari est	parti	sur un coup de tête. Instabilité dans tous les domain
0440	l'ai mis dehors, il est retourné chez papa maman. Il devaitpartir	, il est parti, il y avait une autre femme, depuis l	
0440	s, il est retourné chez papa maman. Il devait partir, il est	parti	, il y avait une autre femme, depuis le mois de mars
0440	ait une autre femme, depuis le mois de mars quatre vingt, il	partait	, il revenait. En novembre j'en ai eu marre. Et mon
0446	'ai plus supporté. Va te promener, je lui ai dit mais il est	parti	après le divorce, il était handicapé à cent pour cent
0475	is prête à l'accepter à cause des enfants, c'est lui qui est	parti	. Il est parti avec une autre femme. Mon mari était t
0475	ccepter à cause des enfants, c'est lui qui est parti. Il est	parti	avec une autre femme. Mon mari était très coléreux. À
0497	re et faisait subir sa mauvaise humeur à toute la famille. À	partir	de la naissance du deuxième enfant, il a commencé à
0505	s, je n'ai jamais voulu. N'ayant jamais voulu céder, j'ai dû	partir	. Ne s'entendaient sur aucun point, erreur de s'être
0541	de pensées et caractères qui se heurtaient. Son mari voulait	partir	depuis longtemps (la facilité matérielle et son affe

Annexe IV.

Vocabulaire spécifique de l'enquête par rapport à un corpus de référence du français parlé contemporain (suremplois caractéristiques au seuil de 1% classés par catégories grammaticales et par spécificités décroissantes)

Verbes :

avoir, partir, vouloir, boire, vivre, travailler, rentrer, entendre, battre, supporter, sortir, connaître, occuper, aimer, marier, tromper, rencontrer, frapper, rester, quitter, devenir, revenir, séparer, dépenser, laisser, décider, accepter, taper, divorcer, attendre, apprendre, répondre, jouer, préférer, reprocher, ramener, disputer, entraîner, habiter, courir, fréquenter, élever, compter, emmener, provoquer, refuser, dégrader, rejoindre, tuer, assumer, casser, menacer, épouser, crier, prévenir, coucher, arranger, subir, retrouver, gagner, durer, prendre, choisir, reprendre, plaire, chercher, apercevoir, garder.

Substantifs :

mari, enfant, femme, argent, vie, caractère, travail, maison, incompatibilité, jour, maîtresse, mésentente, boisson, humeur, infidélité, ex-mari, copain, fille, mariage, raison, famille, parent, personne, liaison, mère, manque, coup, dispute, ***enquêtée***, alcoolisme, ménage, violence, alcool, instabilité, fils, soir, différence, ras-le-bol, père, cause, couple, dette, divorce, départ, café, adultère, responsabilité, jalousie, situation, séparation, rencontre, amie, naissance, absence, belle-mère, domicile, foyer, marre, ami, liberté, sortie, paie, vacance, conjoint, nuit, salaire, beaux-parents, maladie, goût, scène, décision, jeu, commun, chômage, âge, prison, garçon, profession, union, lendemain, désaccord, épouse, bonne, hôpital, dépense, belle-famille, coureur, comportement, divergence, déplacement, crise, accident, évolution, frère, gosse, mort, lassitude, malade, week-end, époux, maturité, brutalité, erreur, mensonge, dégradation, bêtise, dépression, aventure, arrêt, médecin, amour, tort, soeur, commerce, participation, indépendance, maman, jupon, célibataire, décès, éducation, peur, vue, état, entente, horaire, envie, début, charge, porte.

Adjectifs :

violent, seul, jeune, sexuel, différent, mauvais, instable, jaloux, méchant, commun, familial, malade, enceinte, personnel, professionnel, financier, conjugal, absent, brutal, total, égoïste, caractériel, dépensier, nerveux, sérieux, coureur, coléreux, impossible, spécial, nombreux, alcoolique, soûl, parti, dû, stable, dépressif, autoritaire, mental, infidèle, gentil, aîné, ivre, insupportable, marié, intellectuel, continuel, grave, libre.

Pronoms :

il, je, se, lui, nous, tout, rien, autre, quelqu'un, cela, moi-même, lui-même

Adverbes :

ne, pas, plus, très, trop, jamais, toujours, souvent, ensemble, petit, dehors, tard, vis-à-vis, régulièrement, ainsi, assez, ensuite

Déterminants :

un, son, mon, autre, aucun, plusieurs, deuxième

Annexe V.
Bibliographie

ARRIVE Michel, GADET Françoise, GALMICHE Michel, 1986, *La grammaire d'aujourd'hui. Guide alphabétique de linguistique française*, Paris, Flammarion.

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).

BLANCHE-BENVENISTE Claire et Al, 1964, *Grammaire Larousse du français contemporain* (réed 1988).

CONSEIL INTERNATIONAL DE LA LANGUE FRANÇAISE, 1988, *Pour l'harmonisation orthographique des dictionnaires*, Paris, CLIF.

DUBOIS Claude, DUBOIS Jean, 1971, *Introduction à la lexicographie*, Paris, Larousse.

ENGWALL Gunnel, 1984, *Vocabulaire du roman français (1962-1968) Dictionnaire des fréquences*, Stockholm, Almqvist-Wicksell International.

FESTY P., VALETAS M.-F., 1988, "Le divorce en plus : ruptures et continuités, *Société française*, 26.

GARNIER Bébédicte, GUERIN-PACE France, 1998, "La statistique textuelle pour traiter une question ouverte suivie d'une relance" in MELLET Sylvie (ed), *aux 4e Journées d'analyse des données textuelles*, Université de Nice-Sophia Antipolis, p 315-324.

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.

GREVISSE Maurice, *Le bon usage*, Gembloux, Duculot, (réed 1986).

HATZFELD Adolphe, DARMEISTETER Arsène, THOMAS Antoine, 1898, *Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours, Paris, Delagrave*, 1898 environ.

JUILLAND Alphonse, BRODIN Dorothy, DAVIDOVITCH Catherine, 1970, *Frequency Dictionnary of French Words*, La Haye, Mouton.

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.

LEBART Ludovic, 1994, "Traitement des questions ouvertes" in GRANGE D. et LEBART L. (éd.), *Traitement statistique des enquêtes*, Paris Dunod.

LITTRE Emile, 1863-1877, *Dictionnaire de la langue française*, Paris, Hachette.

MULLER Charles, 1967, *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, (réédition : Genève-Paris, Slatkine-Champion, 1979).

PIBAROT André, LABBE Dominique, 1998, "Les syntagmes répétés dans l'analyse des commentaires libres", in MELLET Sylvie (ed), *aux 4e Journées d'analyse des données textuelles*, Université de Nice-Sophia Antipolis, p 507-516.

PICOCHE Jacqueline, 1977, *Précis de lexicologie française*, Paris, Nathan.

PIZARRO DIAZ Ana Belén, TRUJILLO PALOMO Monica, 2001, *Las mujeres frente al cambio familiar : razones del divorcio tal como expresan mujeres separadas*, Diplomatura de Estadistica, Universitat Politecnica de Catalunya, Barcelone.

REY Alain, 1977, *Le lexique. Images et modèles du dictionnaire à la lexicologie*, Paris, A. Colin.

ROBERT Paul, *Dictionnaire alphabétique et analogique de la langue française*, Paris, 1953-1971 et 1985.

VILLERS Marie-Eva de, 1992, *Multi-dictionnaire des difficultés de la langue française*, Montréal, Québec-Amérique.

WAGNER R.-L., PINCHON J., 1962, *Grammaire du français classique et moderne*, Paris, Hachette.