CORPUS

Corpus

2 | 2003
La distance intertextuelle

Comparaison des structures induites sur un ensemble de réponses ouvertes par le choix de l'unité statistique

Mónica Bécue-Bertaut

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/28
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 décembre 2003
ISSN : 1638-9808

Référence électronique
Mónica Bécue-Bertaut, « Comparaison des structures induites sur un ensemble de réponses ouvertes par le choix de l'unité statistique », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 20 avril 2019. URL : http://journals.openedition.org/corpus/28

Comparaison des structures induites sur un ensemble de réponses ouvertes par le choix de l'unité statistique

Mónica Bécue-Bertaut

1. Introduction

1 Le choix de l'unité statistique de segmentation du corpus est fondamental lorsqu'on analyse un corpus de réponses ouvertes dans une enquête socio-économique. La structure induite sur l'ensemble de réponses, ou ensemble de distances interindividuelles, dépend de cette unité ainsi que, bien sûr, de la définition de la distance.

2 On présente ici l'application d'une méthodologie particulière, l'analyse factorielle multiple pour tableaux de contingence AFMTC (Bécue & Pagès, 1999, 2003) à la comparaison des structures induites sur un même corpus par différentes unités lexicales. On montre aussi que cette méthodologie permet de définir une distance « compromis » entre les distances initiales et donc de prendre en compte, simultanément, plusieurs unités statistiques.

2. Corpus-exemple

3 Pour mieux connaître les raisons pour lesquelles les femmes vivant en France sont amenées à divorcer, la CNAF (Caisse Nationale d'Allocations Familiales) a demandé à l'INED (Institut National d'Etudes Démographiques) de mener une enquête nationale sur « les femmes face au changement familial » (Festy & Valetas, 1988 ; Garnier & Guérin-Pace, 1998). Un échantillon de 2329 femmes, séparées de leur mari au cours des 15 années précédant l'enquête, a répondu à un questionnaire qui comprenait, entre autres, des questions sur leur statut socio-économique et la question ouverte suivante : « Quelles sont les raisons à l'origine de votre mésentente ? ». Si l'enquêtée ne livrait pas spontanément les motifs de la rupture, l'enquêteur devait insister et lui demander : « Qu'est-ce qui a effectivement provoqué la séparation ? ».

4 Nous avons regroupé, pour chaque femme, les réponses aux deux questions – réunies comme s'il s'agissait des réponses à une seule question, ce qui est légitime puisque certaines femmes donnaient les motifs de leur divorce en réponse à la première question, d'autres en réponse à la deuxième.

3. Découpage du corpus en unités statistiques3.1 Unités statistiques simples ou complexes

5 Les unités statistiques les plus fréquemment choisies sont la forme graphique et le lemme. Nous considérons ici non seulement la forme graphique correspondant au corpus « normalisé » (*forme normalisée*), c'est-à-dire nettoyé des erreurs orthographiques, des notations normatives, comme les majuscules correspondant au début d'une phrase, et des abréviations ambiguës (Labbé, 1990), mais aussi la forme graphique définie sur le corpus « brut » (*forme brute*), tel qu'il est fourni à l'issue de son informatisation dont la correction est très variable, en fonction des normes imposées lors de la saisie. Nous considérons cette unité car il est fréquent que les corpus soient analysés tels quels, sans corrections préalables, en particulier les corpus volumineux, étant donné le coût de la normalisation qui ne peut être complètement automatisée.

6 On peut aussi choisir une unité complexe, comme le *segment répété* (Salem, 1984), le *quasi-segment* (Bécue, 1993) ou le *syntagme répété* (Pibarot & Labbé, 1998). Le segment répété est une séquence de formes (ou lemmes) identiquement répétée dans le corpus, par exemple « *il est parti avec* » ou « *je suis partie* ». Le quasi-segment est une séquence de formes (ou lemmes) d'une même phrase, mais non nécessairement contigus, identiquement répétée dans le corpus, par exemple « *avec une femme* » est un quasi-segment détecté quand l'enquêté a dit « *avec une femme* » ou bien « *avec une autre femme* ». Le *syntagme* est une séquence formée de deux lemmes, dont la catégorie syntaxique doit être adjectif, verbe ou substantif, éventuellement séparés par un mot-outil (préposi-tion, adverbe, etc.), par exemple « *partir_rejoindre* » ou « *partir_vivre* ». Ces unités complexes permettent de prendre en compte les mots selon leur contexte d'emploi, ce qui se révèle particulièrement intéressant dans le cas des questions ouvertes.

3.2 Tableaux lexicaux et tableaux lexicaux multiples

7 Toutes les différentes formes graphiques du corpus « Divorce » brut sont identifiées et leur fréquence calculée, afin de construire le tableau lexical Individus X Formes brutes.

8 Après une normalisation du corpus, selon des règles classiques, on construit le tableau lexical Individus X Formes normalisées.

9 On dispose aussi du corpus lemmatisé, grâce à l'analyseur morpho-syntactique mis au point par Dominique Labbé, ce qui permet d'obtenir le tableau lexical Individus X Lemmes. On peut consulter Labbé (2001) pour une information détaillée sur l'opération de lemmatisation et la qualité des résultats ainsi qu'une comparaison systématique de « Divorce » sous les formes normalisée et lemmatisée.

10 Pour ces trois tableaux, on ne conserve que les unités employées au moins 24 fois dans l'ensemble des réponses. Cette sélection des formes ayant une fréquence minimum est liée aux méthodologies employées, l'analyse de corres-pondances et l'analyse factorielle multiple pour tableaux de contingence, qui fournissent des résultats peu robustes s'il existe des fréquences relatives marginales trop faibles.

11 Dans la section 5, on utilise le tableau multiple juxtaposant, en ligne, les trois tableaux lexicaux croisant les individus avec, respectivement les formes brutes, les formes normalisées et les lemmes.

Tableau 1. Taille et vocabulaire du corpus selon les trois normes de dépouillement[1]

	Formes brutes	Formes normalisées	Lemmes
Taille Vocabulaire	56334 3879	56758 3963	56107 2785
Surface conservée Mots de fréquence sup. à 24	45096 268	45398 267	47594 254

Le tableau 1 présente la taille et le vocabulaire du corpus selon les normes de dépouillement.

3.3 Analyse des tableaux

12 Une méthodologie classique pour l'analyse des tableaux lexicaux est l'analyse de correspondances, dont les principes de base et les principaux résultats sont rappelés dans la section 4.1, appliquée à l'un ou l'autre des tableaux selon le choix d'unité effectué.

13 Ici, le but est de comparer les résultats obtenus avec les différentes unités. Pour cela, on emploie une méthodologie qui généralise l'approche offerte par l'analyse des correspondances, l'analyse factorielle multiple pour tableaux de contingence AFMCT (présentée section 4.2), avec deux objectifs différents : comparer les résultats qu'offrent les trois analyses de correspondances séparées des trois tableaux, d'une part, et combiner les analyses séparées (dans ce cas, seulement les deux analyses effectuées sur les formes normalisées et sur les lemmes car conserver le bruit introduit par la non normalisation ne présente guère d'intérêt) dans une analyse globale qui intègre l'information contenue dans chaque type d'unité.

4. Distance et structure au sens de l'analyse des correspondances4.1 Structure et analyse de correspondances

14 On peut s'intéresser à la structure du nuage des points-individus, dont rend compte l'ensemble des distances interindividuelles et/ou à la structure du nuage des points-unités, c'est-à-dire à l'ensemble des distances entre unités statistiques. Ces deux nuages sont interdépendants et la structure d'un tableau de contingence est aussi concernée par la relation entre les lignes-individus et les colonnes-unités, profondément liée à la distance choisie.

15 Rendre compte de la structure d'un tableau de contingence au moyen de l'analyse des correspondances signifie choisir la distance du χ^2 entre les profils-lignes, d'une part, et entre les profils-colonnes, d'autre part. La représentation superposée obtenue permet de rendre compte de la relation entre l'ensemble des individus et l'ensemble des unités.

16 Nous rappelons brièvement les principes de base de cette méthode.

17 Le tableau de contingence, croisant I réponses et J unités, a pour terme général f_{ij}, i=1,..., I ; j=1,..., J, fréquence relative avec laquelle l'individu i emploie l'unité j. L'analyse des correspondances décrit la structure de dispersion de l'ensemble des lignes (et, symétriquement, de l'ensemble des colonnes), c'est-à-dire l'ensemble des distances entre lignes (et symétriquement entre colonnes). Ceci équivaut à comparer les I profils-réponses ou fréquence relative de chacune des unités dans la réponse, $\{f_{ij}/f_{i}$; j= 1,..., $J\}$) (et, respectivement, comparer les J profils-unités ou distribution des unités entre les différentes réponses, $\{f_{ij}/f_{.j}$; i=1,...,I $\}$).

18 Les proximités entre les lignes-individus sont mesurées par la distance de X^2. La distance au carré entre les lignes *i* et *l* est égale à :

$$d^2(i,l) = \sum_{j \in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2 \qquad (1)$$

19 De façon symétrique, les distances entre colonnes sont obtenues en échangeant le rôle des indices *i* et *j* dans (1).

20 L'analyse des correspondances visualise l'ensemble des distances par projection sur les axes de dispersion maximum, appelés axes principaux. La proximité entre deux points-unités correspond à une similitude de leur profil, donc à un emploi fréquent ou bien dans les mêmes réponses (association syntaxique ou association thématique) ou bien dans des contextes voisins (relation de substitution qui peut indiquer une synonymie). La proximité entre deux points-réponses indique que ces individus emploient un vocabulaire similaire.

21 On peut consulter, pour une présentation détaillée de l'analyse des correspondances appliquée aux données textuelles, Benzécri (1981), Bécue & Lebart (2000) et Lebart & Salem (1994).

4.2 AFMTC, distance-compromis et comparaison de structures4.2.1 Brève présentation de l'AFMTC

22 L'analyse factorielle multiple pour tableaux de contingence AFMTC (Bécue & Pagès, 1999, 2003) constitue un outil pour la comparaison de tableaux de contingence ayant la dimension-ligne (et/ou la dimension-colonne) en commun. On peut trouver dans les références citées les principes de la méthode et les règles d'interprétation des résultats, présentée très brièvement dans cette section.

23 Cette méthode reprend les principes de base de l'analyse de correspondances internes (Escofier & Drouet, 1983 ; Cazes & Moreau, 1991, 2000) et de l'analyse factorielle multiple (Escofier & Pagès, 1988-1998, 1994). Cette méthodologie permet d'analyser un tableau de contingence multiple, en conservant autant que possible l'approche et les propriétés de l'analyse des correspondances tout en équilibrant l'importance de chacun des tableaux. Cette méthode offre des résultats :

- analogues à ceux de l'AC, principalement une représentation globale des lignes-individus et des colonnes-mots.
- spécifiques des tableaux multiples, principalement, la représentation superposée des structures induites par chacun des groupes de mots sur l'ensemble des individus (ou catégories d'individus) – appelées structures partielles – et la représentation des facteurs obtenus lors des analyses séparées.

24 La lecture des résultats est facilitée par de nombreuses aides à l'interprétation, héritées de l'analyse factorielle multiple, l'AFM.

25 On s'intéresse ici à deux possibilités offertes par cette méthodologie appliquée à un tableau lexical multiple dans lequel chaque sous-tableau correspond à une unité lexicale différente :

- définition d'une distance de compromis entre les distances définies à partir de différentes unités ;
- comparaison des structures induites par les différentes unités, facilitée par la représentation superposée des structures partielles.

4.2.2 Définition d'une distance « compromis »

26 Le terme général du tableau de contingence multiple, obtenu par la juxtaposition en ligne de *T* tableaux de contingence, noté f_{ijt}, *i*=1,..., *I* ; *j*=1,..., *J* ; *t*=1,..., *T*, est la fréquence relative avec laquelle l'individu *i* emploie le mot *j* dans le tableau *t*.

27 On peut voir l'AFMTC comme une méthode aux axes principaux qui offre la meilleure représentation simultanée possible de l'ensemble de distances entre lignes, définie par (2), et l'ensemble des distances entre colonnes, définie par (3). L'équilibre entre les tableaux est obtenu au moyen de la surpondération des colonnes du tableau *t* par le facteur $1/\lambda_1^t$, inverse de la première valeur propre de l'analyse des correspondances séparée du tableau *t*.

28 La distance « compromis » entre individus est la distance issue de l'analyse globale. La distance entre les lignes *i* et *l*, calculée à partir de l'ensemble des colonnes et considérée comme distance « compromis » est donnée par (2) :

$$d^2(i,l) = \left[\sum_t \frac{1}{\lambda_1^t} \sum_{j \in J_t} \left(\frac{f_{ijt}}{f_{i..}} - \frac{f_{ljt}}{f_{l..}} \right)^2 \cdot \frac{1}{f_{.jt}} \right] - \left[\sum_t \frac{1}{\lambda_1^t \cdot f_{..t}} \left(\frac{f_{i.t}}{f_{i..}} - \frac{f_{l.t}}{f_{l..}} \right)^2 \right] \qquad (2)$$

29 Dans l'expression (2), excepté la pondération des colonnes de chaque tableau par l'inverse de la première valeur propre de l'analyse de correspondances séparée de ce tableau,

- le premier terme correspond à la distance (entre les profils *i* et *l*) dans l'analyse de correspondances du tableau juxtaposé ;
- le second terme correspond à la distance (entre les profils *i* et *l*) dans l'analyse des correspondances de la table contenant les sommes par ligne et par sous-tableau. Le terme général *i.t* de ce tableau est la somme de la ligne *i* dans le tableau *t*.

30 La distance au carré entre la colonne *j* (du tableau *t*) et la colonne *k* (du tableau *r*), est donnée par :

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{ikr}}{f_{.kr}} \right) - \left(\frac{f_{i.t}}{f_{..t}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2$$

Les proximités entre colonnes peuvent s'interpréter comme une association similaire avec les lignes.

4.2.3 Comparaison des structures

31 L'AFMTC offre une représentation de la structure « globale » des individus, c'est-à-dire de la structure définie sur les individus par l'ensemble des colonnes, ou ensemble des distances interindividuelles selon la distance donnée par (2).

32 Cette méthode permet aussi de mettre en évidence les ressemblances et différences entre structures au moyen d'indices synthétiques et aussi de la comparaison des axes de dispersion. En particulier, on recherche si les axes de dispersion globaux correspondent

ou non à des directions de dispersion importantes dans chacun des groupes. D'autre part, le calcul des coefficients de corrélations entre les facteurs globaux et les facteurs obtenus lors des analyses de correspondances séparées permet de comparer les axes de dispersion globaux et les axes de dispersion séparés.

5. Résultats

33 Le tableau multiple à analyser est la juxtaposition, en ligne, des trois tableaux lexicaux croisant les Individus X Formes brutes, Individus X Formes normalisées et Individus X Lemmes. On a donc trois groupes de colonnes, selon le sous-tableau auquel elles appartiennent, avec, respectivement, 268, 267 et 254 colonnes pour les trois groupes (voir Tableau 1). Dans ce qui suit, on emploiera indistinctement groupe ou sous-tableau ; les sous-tableaux seront dénommés « brut », « normalisé » et « lemmatisé ».

34 L'AFMTC, méthodologie dont les principes de base ont été rappelés à la section 4.2, permet donc de comparer les structures induites sur les individus-lignes par les trois types d'unité statistique, c'est-à-dire, par les trois groupes de colonnes. Nous verrons, section 5.1, les principaux indices qui permettent de porter un jugement sur la ressemblance ou la dissemblance entre ces structures. A la section 5.2, on montrera comment cette même méthodologie permet de définir une distance « compromis » entre distances définies à partir de différentes unités.

5.1 Comparaison des structures induites par les formes brutes, les formes normalisées et les lemmes5.1.1 Dimensionalité des groupes et liaison globale entre les trois groupes

35 Les indices L_g (Escofier & Pagès, 1998, pp. 161-167) et RV (Robert & Escoufier, 1976 ; Lebart & *al.* 1995, p. 343) fournissent des mesures globales de la liaison entre les trois groupes (Tableau 2).

36 L'indice de liaison $L_g(t,r)$ est d'autant plus grand que les deux groupes ont des directions de dispersion importantes en commun ; il vaut 0 si les groupes *t* et *r* ne partagent aucune direction de dispersion. Cet indice peut être interprété comme une approximation du nombre de directions de dispersion communes aux deux groupes (Escofier & Pagès, 1998, p. 167).

37 L'indice L_g appliqué à un seul groupe (sur la diagonale du tableau correspondant) est égal à la somme des carrés des valeur propres du groupe (en tenant compte de la pondération de l'AFMTC), et peut s'interpréter comme un indicateur de dimensionalité. La dimensionalité du groupe « brut » est légèrement plus faible que celle des deux autres groupes, ce qui indique que la non-normalisation fait perdre une direction de dispersion qui correspond, sans doute, à l'indication contenue dans une ou plusieurs formes non prises en compte car elles n'atteignent pas le seuil requis.

38 Le coefficient RV, lié à l'indice L_g par l'expression

$$d^2\left(j \in t, k \in r\right) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{ikr}}{f_{.kr}} \right) - \left(\frac{f_{i.t}}{f_{..t}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2$$

a une valeur comprise entre 0 et 1 ; cet indice vaut 1 si les groupes *t* et *r* sont parfaitement homothétiques.

39 Les trois groupes ont des structures voisines, comme le montre le coefficient RV qui vaut respectivement RV(brut, normalisé) = 0.88, RV(brut, lemmatisé) = 0.69, RV(normalisé, lemmatisé) = 0.71. On constate néanmoins que les deux premiers groupes (brut et

normalisé) sont plus proches entre eux qu'ils ne le sont du troisième groupe (lemmatisé). Dans ce cas-ci, la lemmatisation modifie davantage la structure que la normalisation. Dans cette enquête – ce qui est fréquent dans ce type de corpus –, certaines formes fléchies conservent une information que la lemmatisation perd (par exemple, *parti* et *partie* réfèrent deux faits différents et importants étant donné le thème de l'enquête. La forme *parti,* issue de l'expression *il est parti* indique le départ du mari tandis que *partie*, employé dans *je suis partie*, montre que la femme a pris l'initiative de la rupture).

Tableau 2. Liaison entre les groupes

	L_g				RV			
	brut	norm.	lem.	glob.	brut	norm.	lem.	glob.
brut	18.4				1			
normalisé	16.9	19.8			0.88	1		
lemmatisé	13.0	13.9	19.5		0.69	0.71	1	
Global	16.6	17.4	16.0	17.2	0.93	0.94	0.87	1

5.1.2 Comparaison des axes des analyses séparées

40 La corrélation entre les axes principaux de rang 1 issus des analyses séparées est forte dans le cas des groupes 1 et 2 (0.86) et dans le cas des groupes 2 et 3 (0.91) et seulement égale à 0.63 dans le cas des groupes 1 et 3 (Tableau 3). La dispersion des individus sur ces axes présente une certaine similitude, assez proche de l'homotéthie pour les groupes 1 et 2 (brut et normalisé) et surtout 2 et 3 (normalisé et lemmatisé).

41 Des corrélations entre les axes suivants, on peut conclure qu'il serait très difficile de comparer directement les 3 structures, telles qu'elles sont décrites par les axes principaux qui, à partir du deuxième, ne coïncident pas.

Tableau 3. Corrélation entre les facteurs des analyses séparées

		Groupe 1 (brut)				Groupe 2 (normalisé)			
		axe 1	axe 2	axe 3	axe 4	axe 1	axe 2	axe 3	axe 4
Gr. 2 (normalisé)	axe 1	**0.86**	-0.46	0.03	-0.02				
	axe 2	-0.44	**-0.83**	0.23	0.02				
	axe 3	-0.08	-0.21	**-0.93**	-0.12				
	axe 4	-0.01	0.00	0.04	-0.09				
Gr.3 (lemmatisé)	axe 1	0.63	-0.72	-0.09	-0.04	**0.91**	0.31	0.20	-0.02
	axe 2	0.28	0.32	-0.26	-0.06	0.09	-0.48	0.16	0.03
	axe 3	0.46	0.47	-0.42	-0.01	0.18	-0.73	0.26	-0.01
	axe 4	0.40	0.26	0.78	0.09	0.26	-0.23	**-0.88**	0.00

5.1.3 Les facteurs de l'analyse globale

42 La première valeur propre, égale à 2.9, confirme que les trois premiers axes séparés sont très voisins. La décroissance des valeurs propres de l'analyse globale est assez lente, semblable à la décroissance observée dans les analyses de correspondances séparées. On peut identifier douze facteurs dominants, qui représentent 13.6% de l'inertie globale, valeur comparable au pourcentage d'inertie conservé sur les 12 premiers axes des analyses séparées.

43 Les trois groupes d'unités fournissent un apport équilibré à l'inertie des premiers axes. Les corrélations entre le premier facteur global et les projections de ce facteur sur les trois nuages-mots, chaque nuage correspondant à un des états du corpus, sont très fortes sur tous les axes considérés (voir Tableau 4 pour les 8 premiers axes). On en conclut que les douze premiers facteurs sont communs aux trois nuages-mots et constituent des directions importantes d'inertie pour chacun des états du corpus, cependant non confondues avec les principales directions de dispersion de chacun d'eux, pris séparément.

Tableau 4. Corrélations entre la projection du nuage global et celle de chacun des trois nuages partiels associés aux trois unités (8 premiers axes).

	F1	F2	F3	F4	F5	F6	F7	F8
Formes brutes	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98
Formes normalisées	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99
Lemmes	0.99	0.98	0.98	0.99	0.98	0.97	0.97	0.96

5.1.4 Corrélation des facteurs des analyses séparées et des facteurs globaux

44 Le Tableau 5 montre les corrélations entre les quatre premiers facteurs de chacune des analyses des correspondances séparées et les quatre premiers axes de l'AFMTC (ou axes globaux). Le premier axe global est très corrélé aux trois premiers axes séparés, surtout aux premiers axes des groupes normalisé et lemmatisé, ce qui corrobore que la direction de dispersion globale est proche des premiers axes principaux de chaque analyse séparée. Le deuxième axe du groupe 3 n'est corrélé à aucun des quatre premiers axes de l'AFMTC.

5.1.5 Conclusion sur la comparaison des structures

45 Les structures des trois nuages sont très semblables, mais les analyses séparées les mettent en évidence de façon très différente. L'analyse des correspondances appliquées aux questions ouvertes fournit des axes souvent fragiles. L'AFMTC privilégie les directions de dispersion communes aux trois groupes et ne peut qu'offrir des résultats plus robustes. Dans le cas de cette analyse, l'étude des unités les plus contributives à chacun des premiers axes globaux montre que celles-ci proviennent des trois groupes (ce qui correspond bien au fait que les axes représentent des axes de dispersion communs aux trois groupes) et que, d'un groupe à l'autre, elles correspondent aux mêmes thèmes.

46 Dans certains cas, on retrouve des unités pratiquement « équivalentes » car elles correspondent à une forme brute, que l'on retrouve sous sa forme normalisée puis sous le

lemme qui lui correspond. Ainsi, les unités les plus contributives au premier axe global sont *alcoolisme* (forme normalisée), *alcoolismeNm* (lemme) et *ALCOOLISME* (forme brute).

Tableau 5. Corrélations entre les axes globaux et les axes des analyses de correspondances séparées

		Axes globaux			
		F1	F2	F3	F4
Axes partiels Groupe « brut »	F1	**0.78**	-0.54	0.25	0.01
	F2	-0.61	**-0.74**	0.24	-0.02
	F3	-0.48	0.35	**0.91**	-0.09
	F4	-0.03	0.02	0.09	**0.98**
Axes partiels Groupe « normalisé »	F1	**0.98**	-0.13	0.13	0.00
	F2	0.15	**0.97**	-0.11	0.01
	F3	0.12	-0.12	**-0.97**	-0.03
	F4	-0.01	0.00	0.01	**-0.74**
Axes partiels Groupe « lemmatisé »	F1	**0.97**	0.16	-0.11	-0.01
	F2	0.05	-0.52	-0.10	-0.04
	F3	0.11	**-0.80**	-0.16	0.03
	F4	0.12	-016	**0.95**	0.02

47 Dans d'autres cas, les unités « équivalentes » sont complétées par une unité qui n'apparaît que sous forme de lemme, car les unités correspondantes brute et normalisée n'atteignent pas le seuil de sélection. Ce lemme peut apporter une précision intéressante. Ainsi, pour l'axe 8, les unités les plus contributives sont *INSTABILITE*, *instabilité*, *instabilitéNf* et *professionelAdj*. L'adjectif *professionnel* laisse entrevoir que l'*instabilité*, donnée comme cause de divorce, est surtout *l'instabilité professionnelle*, ce qui devra être vérifié en recherchant, par exemple, toutes les concordances de *instabilité*.

5.2 Utilisation conjointe de différentes unités au moyen de la distance de compromis

48 L'AFMTC offre aussi la possibilité de prendre en compte simultanément plusieurs unités, au moyen de la distance-compromis que constitue la distance entre individus calculée à partir de l'ensemble des groupes.

49 La forme brute n'offre d'autre intérêt que de permettre l'expérimentation effectuée à la section précédente. Par contre, la forme normalisée et le lemme apportent chacune une information différente sur le corpus et on peut être intéressé par la prise en compte simultanée des deux unités, ce qu'offre justement l'AFMTC qui calcule les distances entre individus à partir à partir des deux groupes d'unités.

50 Les individus-réponses sont alors comparés en fonction des formes normalisées qu'ils emploient mais aussi en fonction des lemmes, ce qui comporte un effet de lissage dû au rapprochement des individus qui emploient un même lemme, plus nombreux que ceux qui emploient une même forme. Cela est tout particulièrement utile lorsqu'on recherche des groupes d'individus proches quant au langage employé, ce que nous fournira une

classification des individus à partir des coordonnées sur les premiers axes principaux globaux.

5.3 Distance entre unités de groupes différents

51 L'analyse globale définit aussi une distance entre les unités des différents groupes. On peut ainsi s'intéresser non seulement aux groupes d'individus similaires mais aussi aux petites structures locales formées par les unités statistiques proches, qu'elles soient des formes normalisées ou des lemmes. Ces « patrons » d'unités proches recouvrent fréquemment les différents thèmes employés dans les réponses. En effet, on trouve dans une même classe des unités employées dans une même expression, comme par exemple tout ou partie des unités correspondant à *il est parti avec une amie* et des unités qui peuvent substituer certaines unités de cette expression (par exemple : *maîtresse* peut se substituer à *amie*).

52 La lecture des différentes classes d'unités statistiques permet de détecter clairement certains thèmes des réponses qui peuvent correspondre à une seule classe ou à plusieurs classes sémantiquement proches. Parmi les principaux motifs de divorce, on peut citer : les problèmes avec la belle-famille ; le départ de la femme ; le départ du mari avec une maîtresse, une amie ou une autre femme ; le jeu, les dépenses et les dettes ; le chômage et l'instabilité professionnelle ; le refus de travailler du mari ; la mésentente du couple ; le mari coléreux, violent ou méchant ; le mari devenu difficile à supporter ; le mari battait, frappait et buvait ; l'infidélité ; le mariage trop jeune ; les goûts différents ; l'alcool et la boisson, l'alcoolisme ; la différence de caractère ou d'éducation ; le peu de choses en commun ; le divorce par accord ; les problèmes sexuels ; l'incompatibilité d'humeur.

Tableau 6. Quatre classes d'unités statistiques

amie avait avec cette lendemain maîtresse parti trouvé une vivait amieNf avecPre coureurNm fréquenterVer lendemainNm maîtresseNf nouveauAdj quitterVer unDet.
argent dettes donnait dépensait jeu jouait salaire argentNm detteNf dépenserVer gagnerVer jeuNm jouerver salaireNm
coléreux, méchant, violent coléreuxAdj, méchantAdj, violentAdj
Assez bonne c homme parce possible supporter supporté toute était aimerVer assezAdv bonneNf cePro commeCnj hommeNm jePro lorsqueCnj parce_queCnj possibleAdj supporterVer êtreVer

53 Cette lecture montre aussi que certains thèmes sont exprimés de façon brève et synthétique (comme le thème de la méchanceté dans la troisième classe du Tableau 6) tandis que d'autres sont accompagnés d'un discours plus complexe et plus varié (comme celui de la dernière classe du Tableau 6, *la femme aimait un autre homme*, thème accompagné d'explications et de justifications variées – et donc souvent non conservées – mais que l'on entrevoit à partir, par exemple, des unités *parce_queCnj*, *possibleAdj).*

54 Pour aller plus avant vers un outil d'aide à la post-codification des réponses, il serait nécessaire d'identifier les différentes combinaisons d'unités qui permettent d'attester la présence d'un thème dans une réponse. On pourrait alors compléter la chaîne « repérage des thèmes – identification des unités indiquant la présence d'un thème – assignation des

réponses aux thèmes employés ». Pour améliorer les résultats, deux voies sont intéressantes :

- l'utilisation d'unités complexes, telles celles qui ont été définies à la section 3.1, permettrait de prendre en compte des unités en contexte (comme *il-est-parti* ou *je-suis-partie*) et ainsi à la fois identifier plus clairement les thèmes sous-jacents et faciliter le lien thème-combinaison d'unités ;
- le filtrage des unités en fonction de leur catégorie grammaticale, pour ne retenir que les unités, éventuellement complexes, porteuses d'information.

6. Conclusion

55 L'analyse des textes passe par le choix d'une unité statistique qui permette le découpage du texte. Les unités les plus fréquemment choisies sont la forme graphique et le lemme. La méthodologie présentée ici permet d'utiliser ces unités simultanément, ce qui permet de profiter à la fois de l'information véhiculée par chacune d'elles. Elle permet aussi de comparer les structures induites par chacune des unités.

56 Une démarche similaire à la démarche proposée ici pourrait s'appliquer à d'autres distances, même non euclidiennes au moyen de méthodes issues du *multi-dimensionnal scaling*, qui offrent des possibilités similaires. On pourrait ainsi mettre en évidence les textes qui se situent très différemment, relativement aux autres textes, selon la distance choisie, ce qui peut révéler des traits de grand intérêt.

BIBLIOGRAPHIE

Bécue M. & Lebart L. (2000). « Analyse statistique de réponses ouvertes », in J. Moreau, P.A. Doudin & P. Cazes (éds.) *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*. Berlin-Heidelberg : Springer, 59-83.

Bécue M. & Pagès J. (1999). « Intra-Sets Multiple Factor Analysis. Application to textual data », in J. Jansen *et al.* (éds.) *Proc. of the 9th International Symposium on Applied Stochastic Models and Data Analysis*. Lisbonne : Universidade de Lisboa, 51-60.

Bécue M. & Pagès J. (2003, sous presse). « A principal axes method for comparing contingency tables : MFACT », *Computational Statistics and Data Analysis*.

Benzécri J.P. (1981). *Pratique de l'analyse des données*, T. III, *Linguistique & Lexicologie*. Paris : Dunod.

Cazes P. & Moreau J. (1991). « Analysis of a contingency table in which the rows and the columns have a graph structure », in E. Diday & Y. Lechevallier (éds.) *Symbolic-numeric data analysis and learning*. New York : Nova Science Publishers, 271-280.

Cazes P. & Moreau J. (2000). « Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique », in J. Moreau, P.A. Doudin & P. Cazes (éds.) *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*. Berlin-Heidelberg : Springer, 87-103.

Escofier B. & Drouet D. (1983). « Analyse des différences entre plusieurs tableaux de fréquence », *Les Cahiers de l'Analyse des Données* 8 (4) : 491-499.

Escofier B. & Pagès J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation.* Paris : Dunod.

Escofier B. & Pagès J. (1994). « Multiple Factor Analysis : AFMULT package », *Computational statistics and data analysis* 18 : 121-140.

Festy P. & Valetas M.F. (1988). « Le divorce en plus : ruptures et continuités », *Société française* 26.

Garnier B. & Guérin-Pace F. (1998). « La statistique textuelle pour traiter une question ouverte suivie d'une relance », in S. Mellet (éd.) *JADT 1998, 4èmes Journées Internationales d'Analyse statistique de Données Textuelles.* Nice : Université de Nice, 315-324.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Lebart L. & Salem A (1994). *Statistique textuelle*. Paris : Dunod.

Lebart L., Morineau A. & Piron M. (1995). *Statistique exploratoire multidimensionnelle*. Paris : Dunod.

Pibarot A. & Labbé D. (1998). « Les syntagmes répétés dans l'analyse des commentaires libres », in S. Mellet (éd.) *JADT 1998, 4èmes Journées Internationales d'Analyse statistique de Données Textuelles.* Nice : Université de Nice, 507-515.

Robert P. & Escoufier Y. (1976). « An unifying tool for linear multivariate methods. The RV coefficient » *Applied Statistics* 25 (3) : 257-265.

Salem A. (1984). « La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes », *Les Cahiers de l'Analyse des Données* 9 (4) : 489-500.

NOTES

1.. Les corpus brut et normalisé sont légèrement différents des corpus brut et normalisé traités dans Labbé, 2001.

RÉSUMÉS

Le choix de l'unité statistique de segmentation du corpus, ainsi que celui de la distance entre les réponses, induit une structure sur l'ensemble des réponses. Nous proposons d'appliquer une méthodologie statistique, l'analyse factorielle multiple pour tableaux de contingence, AFMTC, pour, d'une part, comparer les structures induites sur un même corpus par différentes unités statistiques et, d'autre part, définir une distance entre textes capable de prendre en compte différentes unités et de profiter ainsi de l'information véhiculée par chacune d'elles. Pour présenter cette méthodologie, généralisation de l'analyse des correspondances aux tableaux de contingence multiples, on utilise une enquête auprès de femmes divorcées qui comporte une question ouverte sur les raisons de leur divorce.

Comparison between the structures that are induced by the choice of the statistical unit

The election of the statistical unit, as well as that of the distance between responses, induces a structure on the whole of the responses. We propose to apply a statistical methodology, the

multiple factorial analysis for contingency table, MFACT, in order to, on the one hand, compare the structures induced on a same corpus by different statistical units and, on the other hand, define a distance between texts able to take into account different units and so take advantage of the information that they convey. To present this methodology, a generalisation of correspondence analysis to multiple contingency tables, we use a survey carried out on divorced women, that includes an open-ended question about the reasons of their divorce.

INDEX

Mots-clés : analyse de correspondances, analyse factorielle multiple pour tableaux de contingence, analyse statistique de textes, questions ouvertes d'enquête, tableaux de contingence multiples

AUTEUR

MÓNICA BÉCUE-BERTAUT

Département EIO/ UPC (Barcelona, Espagne ; monica.becue@upc.es)