

J.-P. BENZÉCRI

**Notes de lecture : sur l'analyse des réponses
libres dans une enquête internationale**

Les cahiers de l'analyse des données, tome 17, n° 3 (1992),
p. 353-358

http://www.numdam.org/item?id=CAD_1992__17_3_353_0

© Les cahiers de l'analyse des données, Dunod, 1992, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

NOTES DE LECTURE: SUR L'ANALYSE DES RÉPONSES LIBRES DANS UNE ENQUÊTE INTERNATIONALE

[RÉP. LIBRE]

J.-P. BENZÉCRI

Trois enquêtes, menées parallèlement à Tokyo, New-York et Paris sur les préférences alimentaires, comportaient des réponses libres dont l'analyse a fourni à H. AKUTO et L. LEBART des résultats suggestifs présentés dans ce même cahier (cf. [REPAS IDÉAL]). Dans le §4 de l'article, les auteurs ont bien voulu considérer certaines critiques dont nous leur avons fait part. La franchise de leur exposé permettra au lecteur d'apprécier, par lui-même, ce qui pourrait, légitimement, subsister de nos critiques; sur lesquelles nous ne reviendrons pas directement. Notre objet est ici de rappeler, en termes généraux, des principes que les auteurs connaissent; mais qu'ils peuvent apprécier autrement que nous.

1 Formes, mots et morphèmes

H.A et L.L. posent d'abord:

L'unité statistique choisie sera la forme graphique, séquence de caractères séparés par des caractères délimiteurs (blancs, virgules, points, parenthèses, etc.). À cette fin, les réponses en japonais ont été romanisées.

Ce parti pris est justifié dans la pratique; mais il ne peut être tenu pour un idéal fondé sur la science linguistique: celle-ci, en effet, ne parvient qu'imparfaitement à délimiter des unités au sein du discours. L'autorité de l'écriture peut donner l'illusion que la délimitation des mots est hors de conteste; mais les incertitudes de l'orthographe suggèrent le contraire; où doit-on mettre les traits d'unions? Quant au sens, "compte rendu" est-il plutôt un mot ou une locution? l'usager du français comprend-il pourquoi on écrit "il s'en va" avec "en" comme pronom; et "il s'enfuit" avec "en" comme préfixe? est-il logique que, par passage au féminin, l'unique mot "du" produise les deux mots "de la"?... Un texte en langue japonaise comporte des locutions écrites en caractères chinois: à chacun de ceux-ci, on peut convenir d'attribuer un sens, même si la fonction de ce que nous appelons mot est remplie par des suites de plusieurs caractères; dont chacun serait plutôt l'analogie d'un morphème, mais avec une autonomie supérieure à celle de nos préfixes, racines ou désinences.

A. SALEM a consacré d'utiles recherches au dénombrement des segments répétés. Certains de ceux-ci, qu'il appelle "contraints", sont d'un usage général et jouent, dans le discours ordinaire, le même rôle qu'un mot; d'autres ne prennent ce rôle qu'à l'intérieur d'une spécialité, ou seulement d'un texte particulier. Les gastronomes de Paris citent la *coquille Saint Jacques*, sans songer à l'Apôtre, ni même à l'insigne des pèlerins de Compostelle... Dans le corpus de l'enquête, ni "Saint" ni "Jacques" ne figurent par eux-mêmes, non plus qu'"Honoré", ou "Émilion"; et sans doute s'agit-il toujours, ou presque toujours, de la même espèce de coquille.

La forme "blanc" peut être nom ou adjectif: on peut parler de deux formes homonymes, appartenant à deux mots distincts. Le nom "blanc" lui-même a des sens différents dans "blanc de poulet" et "blanc d'œuf"; bien que ce ne soit pas l'usage de dire que "blanc" est polysémique; comme l'est le verbe "voler", dit de l'oiseau et du voleur.

Si des formes peuvent relever de mots ou de sens bien distincts, la conjugaison, la déclinaison, la suffixation..., au contraire, imposent ou suggèrent communément à rattacher plusieurs formes à un même mot ou *lemme*. La *lemmatisation* est pour un grand corpus, une opération coûteuse; aussi H.A et L.L. se réjouissent-ils de voir "que les pluriels et les singuliers ne sont pas souvent dans les mêmes classes"; trouvant ainsi, dans l'analyse distributionnelle, un argument contre "la lemmatisation aveugle". Ils n'en concluent pas moins en comptant loyalement, au nombre des "difficultés ... recensées, sinon maîtrisables", le "choix des unités statistiques pertinentes", et le "degré de lemmatisation" à adopter.

Disons franchement que ces *difficultés* ne nous font pas peur! D'une part, nous sommes convaincu que dans la langue, comme dans toute structure naturelle ou seulement produite par les vivants avec quelque liberté, il n'y a pas de stratification hiérarchique: les différents niveaux communiquent, en quelque sorte, par des passages horizontaux; et un objet présent sur plusieurs niveaux peut, à l'un de ceux-ci, tenir les deux rôles de composé et de composant.

Mais nous croyons, d'autre part, qu'après avoir considéré le plexus de la structure, pour analyser, le statisticien doit trancher: adopter un principe de dénombrement qui corresponde aux moyens techniques dont il dispose; et, s'il obtient des résultats suggestifs, s'assurer, ensuite, de la validité de ceux-ci, en expérimentant avec d'autres règles de dénombrement; tout aussi naturelles et non moins praticables que celles qu'il avait prises d'abord.

Puisque la notion de la *forme graphique* ne relève d'aucune théorie parfaite, il n'y a pas lieu d'en respecter strictement la définition: on peut la modifier. La véritable rigueur consiste ici, non à invoquer des *tests*, mais à vérifier, par l'expérience, que la part d'arbitraire que comporte la mise en œuvre des principes qu'on croit justes n'a pas d'effet sensible sur les résultats interprétés.

2 Thème, genre et style

Il est commun d'appeler *mots pleins* les noms, adjectifs, verbes et adverbes qui représentent, dit-on en bref, respectivement, des objets, des qualités, des actions dans le temps et des modalités des actions; tandis qu'on appelle *mots outil* les autres mots, conjonctions, prépositions, pronoms,... qui mettent en relation les mots pleins ou se substituent génériquement à ceux-ci.

Comme celle entre *mot* et *locution*, la distinction entre *mot outil* et *mot plein* est imprécise. Ainsi, dans maints contextes, la locution prépositionnelle "par voie de", équivaut à "par": le sens du nom "voie" s'efface. Le verbe "être", en dehors de son acception proprement ontologique, est un outil, liant le sujet au prédicat; outil peu usité dans les langues, tel le russe, où la phrase nominale est reçue; et il y a, en français, de nombreux substituts du verbe "être" qui, tels "sembler" ou "demeurer", n'ajoutent à la fonction de copule logique qu'une connotation quant à la certitude ou à la durée d'une assertion.

Mais, de même que la statistique doit, d'une manière ou d'une autre, dénombrer des mots ou des formes, de même il convient ordinairement qu'elle prenne en compte la distinction entre *mot plein* et *mot outil*. Selon notre expérience, attestée dans plusieurs articles (cf. [LING. TRI], [TEXTES GRECS], [TEXTES LATINS]), l'analyse d'un corpus requiert le choix d'un lexique 'Δ' des formes retenues pour la construction du tableau de correspondance principal. Voici comment nous avons procédé dans le cas des textes latins:

Le choix de 'Δ' repose sur la liste, ordonnée par fréquence croissante, de toutes les formes se rencontrant dans le corpus. Il est clair que les hapax (formes qu'on ne trouve qu'une fois) et même les formes de faible fréquence n'ont pas de place dans Δ; mais, une fois choisi un seuil de fréquence minima, il reste à choisir parmi les mots situés au-dessus du seuil. Ici, (à la différence de ce qu'on a fait dans [LING. TRI 2] pour des textes bibliques en langue grecque,) on a retenu exclusivement des formes de mots outil: éliminant non seulement les mots pleins proprement dits: *deus*, 'Dieu', *hominum*, 'des hommes' (génitif: cas du complément de nom)...; mais aussi *res*, 'chose', *ait*, 'dit-il', ainsi que toute forme, verbale ou pronominale, renvoyant à la première ou à la deuxième personne.

En bref, d'après les outils, on reconnaît le style de l'auteur et aussi le genre de l'œuvre (poétique, historique, rhétorique, didactique...); les deux étant liés. Tandis qu'en dénombrant les mots pleins, on accède au thème. Toutefois, ainsi que l'a noté A. SALEM, dès le début des applications de l'analyse des correspondances en linguistique, le cas des formes renvoyant à la 1-ère et à la 2-ème personne est particulier: même s'il s'agit d'outils, la présence affirmée de l'auteur ou d'un interlocuteur, donne des présomptions non seulement quant au genre (e.g. théâtral), mais quant au contenu.

Dans les réponses recueillies à Tokyo et à Paris, H.A et L.L. notent explicitement la présence de nombreux mots outils; il apparaît, au contraire, qu'à New-York le style est plus télégraphique (à moins que les enquêteurs n'aient noté les réponses autrement): seuls sont cités {AND, WITH}, respectivement aux rang 4 et 5 dans la liste des fréquences.

Le §4.2.2, rend compte d'une classification de 173 mots français, d'après leur correspondance avec 1000 sujets. La partition retenue comporte une classe 16, {AVEC, DES, DU, ET, UN}, où les auteurs voient, à juste titre "un exemple étonnant de squelette de réponse, formé uniquement de mots-outil". Plusieurs mots outil vont avec des mots pleins avec lesquels ils forment des locutions usuelles: ainsi 'AU', avec {BEURRE, CHOCOLAT, CITRON}. D'autre part, les classes 34 à 40, qui se séparent à un niveau élevé de la hiérarchie ne contiennent guère, outre des outils, que des mots pleins que les auteurs appelleraient, à juste titre, génériques: {PAS, LES, TOUS, EQUILIBRE, REPAS, PLATS, CUISINE, SPECIALITES, NSP (ne sait pas)}.

Selon nous, de l'analyse des réponses libres recueillies dans une enquête, on ne peut attendre une typologie stylistique (les auteurs notent toutefois que: "les femmes de plus de 50 ans répondent souvent NSP"); il convient donc, *a priori*, d'éliminer les mots outil. On peut s'interroger quant à des mots génériques comme {bœuf, viande,...}: c'est par l'expérience qu'on en appréciera le rôle dans l'analyse. Ici, comme au §1, nous concluons en recommandant d'user de pragmatisme dans l'application des règles générales.

Mais la question se pose maintenant du choix de l'ensemble des unités de texte avec lequel mettre en correspondance le lexique des mots retenus; c'est ce que nous ferons sous le titre du §3.

3 Conjonction des langues

Dans [LING. TRI], [TEXTES GRECS], [TEXTES LATINS] (in CAD, Vol. XV, n° 1, Vol. XVI, n°s 1, 2, 4), nous avons pris pour règle de considérer des fragments de texte contenant, chacun, environ 500 occurrences. Dans l'enquête, le petit texte que constituent les réponses libres d'un sujet compte en moyenne moins de 20 occurrences: il n'est donc pas surprenant que les auteurs notent (cf. §4.2.2) que: "l'analyse factorielle d'un tableau clairsemé, [glane] quelques petits conglomerats à la surface d'une gigantesque hypersphère..."

Analyser (comme le font les auteurs: cf. §§1, 2 & 3,) la correspondance entre le lexique et 6 catégories combinées du signalement, âge-sexe, revient en somme, à répartir l'ensemble du corpus en 6 textes ou fragments, dont chacun comprendrait, mis bout à bout, comme autant de paragraphes, les réponses fournies par les sujets rentrant dans la catégorie correspondante. Plus généralement, on peut de la même manière associer un texte composite à toute modalité concernant l'ensemble I des sujets; qu'il s'agisse d'une modalité du signalement; ou d'une modalité de réponse à une question fermée.

Prenons l'exemple de l'enquête de Paris: à une modalité d'âge-sexe correspond un texte de près de 2000 occurrences: un tel nombre est plus qu'il ne faut pour éviter d'avoir un *tableau clairsemé*: on pourrait donc subdiviser ces modalités; soit en restreignant les tranches d'âge; soit en combinant âge et sexe avec d'autres éléments du signalement, tels que revenu ou profession. Par le fait, la configuration des modalités combinées acquerrait une structure plus complexe que celle en deux chapelets (3+3).

Demander que cette structure soit, de quelque manière, représentée sur les plans issus de l'analyse factorielle ou dans la CAH, imposerait une contrainte plus forte: obtenir une telle représentation, fort improbable *a priori*, serait un résultat positif dont la validité ne pourrait être mise en doute.

Une deuxième vérification s'offre ici. Notons L , un lexique des quelque 100 mots, choisis comme on l'a expliqué au §2; M un ensemble de catégories combinées, qu'on peut supposer mutuellement exclusives et dont le nombre total serait de l'ordre de 20. Au tableau $L \times M$, on peut adjoindre en colonnes supplémentaires l'ensemble I des quelque 1000 sujets interrogés, chacun décrit par son profil lexical. Ceci fait, dans l'espace issu de l'analyse factorielle de $L \times M$, l'ensemble I se projette; et l'on peut, selon la méthode usuelle de l'analyse discriminante, affecter chaque individu i à la modalité m dont il est le plus proche. Comme, initialement, chaque individu rentre dans une modalité m , (e.g. sa catégorie combinée âge-sexe-profession,) on peut mettre en forme de matrice de confusion les résultats de cette analyse discriminante: $k(m, m')$ étant le nombre des i dont la véritable catégorie est m et qui ont été attachés à m' . L'analyse de ce tableau $M \times M$ montrerait, en bref, ce qui, dans le comportement verbal, subsiste de la structure de M .

D'un point de vue complémentaire, on peut encore regarder la présence, ou l'absence, d'un mot dans le discours d'un sujet i comme une modalité descriptive, au même titre qu'une modalité du signalement (ou une modalité de réponse à une question fermée). Si J désigne un ensemble disjonctif complet de modalités, on peut construire un tableau de BURT sur $\{J \cup M^+ \cup M^-\}$, (M^+ et M^- désignant, respectivement, les ensembles de modalités de présence et d'absence des mots). On peut encore dire que chaque mot m a un profil sur J : $k(m, j)$ étant le nombre total des sujets faisant usage du mot m et rentrant dans la modalité j .

Venons maintenant au titre du présent §3. Avec un J , choisi de quelque manière qu'il plaira, on peut croiser la réunion des 3 lexiques M_1, M_2, M_3 afférents à Tokyo, Paris et New-York. L'analyse du tableau $(M_1 \cup M_2 \cup M_3) \times J$ montrerait dans quel contexte de modalités de J s'insèrent chacun des mots des trois lexiques; on verrait si {HAMBAGA, HAMBAGU} sont, dans l'espace rapporté aux axes factoriels, des voisins de {HAMBURGER, HAMBURGERS}; si le POULET tient le rôle du CHICKEN; et le FROMAGE, du CHEESE...

Les auteurs ont noté qu'à Paris, "comme à Tokyo, la cuisine étrangère intéresse plutôt les jeunes"; avec pour "autres points de convergence... l'attrait des jeunes femmes pour la cuisine italienne, et les mentions par les personnes les plus âgées de plats nationaux traditionnels, (représentés par le cassoulet, le bœuf bourguignon)". De telles remarques prendraient une plus grande précision. On pourrait même, recourant, ici encore, à l'analyse discriminante, déterminer, pour tout mot d'un lexique, le mot qui en est le plus proche dans chacun des deux autres. Au risque de trouver que d'un ragoût l'équivalent social est un entremets...

Dans le cas où les échantillons I_1 , I_2 et I_3 ont sur J des profils différents, il convient d'adjoindre ces profils en lignes supplémentaires à l'analyse du tableau $(M_1 \cup M_2 \cup M_3) \times J$: en effet la place d'un mot m_i est à interpréter relativement à celle de l'échantillon I_i afférent au même pays. Mais cette difficulté ne se rencontre pas si J ne comprend que des modalités du signallement dont les poids sont imposés par les quotas de l'enquête; ce qui doit être approximativement réalisé pour les modalités d'âge-sexe.

D'ailleurs, en multipliant chacune des colonnes j par un coefficient approprié, on peut se ramener au cas où les 3 échantillons I_i ont même profil sur J . De façon précise, notons $k(I_i, j)$ le nombre des sujets de I_i rentrant dans la modalité j ; et $k(I, j)$ la somme $k(I_1, j) + k(I_2, j) + k(I_3, j)$. Dans chacun des tableaux $M_i \times J$, on multipliera la colonne j par le coefficient $(k(I, j)/k(I_i, j))$; puis on procédera à l'analyse du tableau $(M_1 \cup M_2 \cup M_3) \times J$, construit par superposition des 3 tableaux $M_i \times J$ ainsi modifiés.