L'APPORT DE L'INFORMATIQUE DANS L'ANALYSE CONJOINTE DES DONNÉES NUMÉRIQUES ET TEXTUELLES

Isabelle MARAMOTTI, Javier SANCHEZ

INTRODUCTION

Les dernières évolutions technologiques et de récents travaux en linguistique informatique [1], en particulier en ce qui concerne les capacités de traitement ainsi que l'élaboration de nouvelles méthodologies de recherche linguistique, ont contribué à banaliser l'utilisation des nouveaux logiciels. En psychologie, par exemple, le traitement des enquêtes et des entretiens - outils courants du psychologue - bénéficie aujourd'hui des toutes dernières techniques pour la gestion et l'exploitation des données numériques et textuelles. Ainsi, notamment, le C.I.S.I.A. [2] propose actuellement deux versions du logiciel SPAD (Système Portable d'Analyse des Données) : SPAD-N [3] permettant le traitement des données numériques (questionnaires) et SPAD-T [4] permettant l'analyse des données textuelles (entretiens ou questions ouvertes de questionnaire). Mais l'aspect qui constitue l'évolution la plus importante en psychologie réside dans l'intérêt méthodologique offert par les procédures de couplage des analyses numériques et textuelles réalisable grâce à ces deux logiciels. En effet, jusqu'à présent cette tâche était méthodologiquement impossible à réaliser et le psychologue ne

1 Sanchez, J. (1993) : Qu'est-ce que l'analyse relationnelle informatique des textes ? Revue Informatique et Statistique dans les Sciences Humaines, Université de Liège, 29, 135-165. Sanchez, J. (1994) : De la désambiguïsation assistée par ordinateur à l'élaboration des grammaires textuelles. Revue Literary and Linguistic Computing, Oxford University Press, vol.9, n°3, 195-202.

2 Centre International de Statistique et d'Informatique Appliquée (C.I.S.I.A.), 1, Avenue Herbillon, 94160 Saint-Mandé (France).

3 SPAD-N (Système Portable d'Analyse des Données - Numériques).

4 SPAD-T (Système Portable d'Analyse des Données - Textuelles).

disposait pas des outils nécessaires pour effectuer à la fois une analyse qualitative et une analyse quantitative à partir d'un nombre important de questionnaires et d'entretiens.

Nous présenterons brièvement ici, ces possibilités de traitement et de couplage statistique des enquêtes et des entretiens qui constituent désormais pour les étudiants et les chercheurs psychologues, grâce au progrès informatique, une nouvelle voie de formation et de recherche dont il faudra tenir compte dans nos filières universitaires. D'autant que, depuis 1986, les étudiants en sciences humaines sont, grâce au plan d'initiation à l'informatique (I.P.T.), formés précocement à l'utilisation de l'ordinateur, ce qui rend possible la pratique, à partir de logiciels spécialisés, des différentes méthodes statistiques enseignées et utilisées en psychologie. Nous illustrerons notre propos à partir de nos derniers travaux [5] - [6], où nous avons étudié et appliqué les possibilités de traitement et de couplage des données numériques et textuelles.

2. LE TRAITEMENT DES DONNÉES NUMÉRIQUES

Après diverses explorations de logiciels, nous avons retenu SPAD-N d'une part parce qu'il permet d'effectuer des analyses multidimensionnelles telles que l'analyse factorielle des correspondances et les techniques de classification automatique, et d'autre part, parce qu'il est capable de se coupler avec SPAD-T qui permet le traitement des données textuelles issues des entretiens et/ou des questions ouvertes des questionnaires. SPAD-N et SPAD-T sont donc des outils bien adaptés aux applications développées en psychologie.

Ce logiciel, dont les commandes sont parfaitement intégrées dans des écrans spécialisés ou dans des menus déroulants, présente deux phases préparatoires importantes permettant la construction des

---

5 Maramotti, I. (1994): Approche psychologique des facteurs de l'environnement: application à la multi-exposition au bruit et aux vibrations. Thèse de Doctorat Nouveau Régime. Université de Paris X-Nanterre.
Maramotti, I. (1994): La perception des vibrations et leur rapport avec la gêne en milieu d'habitation en vue de la construction d'une norme européenne. Contrat avec le Ministère de l'Environnement. 1991-94.

6 Sanchez, J. (1992): Méthodologie et outils de l'analyse relationnelle informatique des textes. Analyse textuelle et nouvelles technologies, Centre de Recherche de l'Université de Paris VIII, Saint-Denis, 350 p.
Sanchez, J. (1994): SPAD-N au service d'une méthodologie pour l'analyse des données textuelles. Revue Enseignement Public et Informatique (EPI), 74, 219-236.

"fichiers libellés" et des "fichiers données" qui serviront de base aux différents traitements statistiques :

1) Le codage des variables et des modalités de réponse.
2) La saisie des données numériques.

2.1. Le codage des variables et des modalités de réponse

La première phase consiste à créer une codification des variables et des modalités de réponses. Ainsi, par exemple, une de nos dernières études comportait un total de 146 variables et de 374 modalités de réponse, et une population de 176 individus, ce qui constitue un tableau de contingence de 25696 entrées. Notons d'ailleurs que, malgré la grande dimension du tableau des données, le temps de traitement est de quelques secondes (avec des ordinateurs de type 486 et Pentium), ce qui rend le travail d'analyse très convivial, en particulier en ce qui concerne les nombreux essais statistiques nécessaires pour affiner certaines analyses (définitions des variables les plus représentatives, tris par filtre ou seuil, partition et constitution de classes, etc.).

Afin de permettre l'identification des variables, il est possible de reformuler les questions pour que les informations essentielles soient présentes dans les résultats statistiques (calculs et plans factoriels) car la longueur des libellés est souvent très limitée en nombre de caractères.

Parallèlement, il est possible de déterminer le nombre de modalités de réponse pour chaque question, de définir la variable numérique associée à ces différentes modalités, ainsi que l'identificateur court (4 caractères) de la modalité (pour les représentations graphiques) et l'identificateur long de chaque modalité (pour rendre plus compréhensible le sens des modalités représentées sur les graphiques). Cela permet donc de coder de manière différente et sous forme mnémotechnique les identificateurs des modalités de réponse, afin que les graphiques factoriels soient lisibles pour l'interprétation.

2.2. La saisie des données numériques

La phase de codage préparatoire achevée, la seconde étape consiste à saisir les valeurs numériques représentant chaque modalité de réponse pour les différents individus. Préalablement à cette saisie, nous avons également la possibilité de construire un code pour chaque individu sous la forme d'une chaîne alphanumérique afin que la lecture des plans factoriels soit la plus claire possible.

Un sous-écran spécialisé de saisie intégré au logiciel présente en colonne les numéros des variables et en ligne les individus avec leur identificateur. On peut ainsi visualiser le tableau de contingence qui servira à effectuer les différentes analyses statistiques.

2.3. Les traitements statistiques

Nous réalisons généralement deux types de traitements statistiques concernant les données numériques du questionnaire : en premier lieu, des **analyses statistiques descriptives** permettant d'effectuer des tris à plats sur les résultats de chacune de nos études et fournissant une première approche ; en second lieu, des **analyses statistiques comparatives** en utilisant les méthodes d'**A.F.C.** (Analyses Factorielles des Correspondances) et de **classification** automatique afin d'interpréter au mieux les résultats dans le cadre des hypothèses.

2.3.1. L'Analyse Factorielle des Correspondances

La réalisation des A.F.C. passe par un certain nombre d'étapes statistiques qui reposent sur un enchaînement de procédures informatiques. Pour les enquêtes, nous utilisons l'Analyse des Correspondances Multiples (A.C.M.), lorsqu'elles ne comportent que des variables nominales (qualitatives). SPAD-N propose plus d'une quarantaine de procédures correspondant à des actions statistiques différentes, parmi lesquelles nous avons choisi celles qui nous ont semblé adaptées à notre type de traitement. A titre indicatif, les noms des quatre procédures retenues sont, dans l'ordre chronologique, les suivantes : SELEC $\Rightarrow$ CORMU $\Rightarrow$ DEFAC $\Rightarrow$ GRAPH.

Il faut commencer par définir les variables actives et les variables illustratives, ce choix déterminant l'objectif de chaque analyse. Cette première opération est réalisée grâce à la procédure SELEC (SELECtion des variables). Ces variables étant établies, les principaux calculs de l'A.C.M. sont ensuite exécutés par la seconde procédure appelée CORMU (CORrespondances MUltiples), qui fournit différentes indications listées pour l'aide à l'interprétation. D'autre part, cette aide est complétée par la troisième procédure DEFAC (DEscription des FACteurs) qui se révèle utile notamment pour l'interprétation des axes car elle propose les points considérés comme les plus représentatifs à partir de la "valeur test".

Enfin, les graphiques factoriels sont construits au moyen d'une dernière procédure GRAPH (construction des GRAPHiques). Nous soulignons que le C.I.S.I.A. propose actuellement SPAD-GF (GraFic) qui

est un excellent outil d'exploitation graphique des plans factoriels, et par conséquent, d'aide à l'interprétation.

2.3.2. La classification

Pour pousser au-delà de l'interprétation des deux premiers axes obtenus avec l'analyse factorielle, le logiciel propose la méthode de classification automatique, qui présente l'avantage d'établir des classes sur l'ensemble des axes factoriels. Le but de cette méthode est de découvrir l'existence de structures cachées de l'ensemble des individus, ces structures étant des groupes et ces groupes comportant les mêmes modalités de réponse. Ainsi, une classe est composée par un ensemble d'individus et un ensemble de réponses qui la caractérisent. C'est un point particulièrement intéressant pour l'interprétation des résultats et la validation des hypothèses. Les quatre procédures de classification sont les suivantes : RECIP ⇒ PARTI ⇒ DECLA ⇒ GRAPH.

La première procédure RECIP (ou voisins RECIProques) construit le dendrogramme des classes établies à partir des individus caractérisés par leurs coordonnées factorielles, puis la procédure PARTI (PARTItion) permet de couper l'arbre selon les paramètres définis par l'utilisateur afin d'obtenir une partition statistiquement valable en nombre de classes. La liste et le contenu de chacune des classes obtenues par la troisième procédure DECLA (DEscription des CLAsses) aident à définir le nombre des partitions et à interpréter le contenu de chaque classe. Tout comme pour l'A.F.C., la procédure GRAPH permet d'obtenir les représentations factorielles des classes.

3. LE TRAITEMENT DES DONNÉES TEXTUELLES

Comme nous l'avons déjà indiqué, le logiciel SPAD-T permet le traitement des données textuelles - les entretiens pour les psychologues - et l'importation des données numériques de SPAD-N en vue d'effectuer le couplage avec les données textuelles. C'est-à-dire que le psychologue qui a utilisé une méthode d'enquête par questionnaire et par entretien, va pouvoir mettre en relation les réponses au questionnaire de chaque individu avec le contenu de son discours lors des entretiens. Cela permet donc d'atteindre une analyse à la fois quantitative et qualitative.

Le première difficulté de cette analyse textuelle est la préparation des données afin d'aboutir à l'équivalent d'un tableau de contingence qui puisse nous permettre de réaliser les différents traitements statistiques. Ainsi comme la parole de chaque sujet ne comporte pas la même forme

(choix et ordre des mots) et ne développe pas les mêmes thèmes (sens des mots et thèmes traités) notre principale préoccupation, en tant que psychologue, est de pouvoir définir à partir du discours de chaque sujet les différents thèmes que chaque individu a développés, pour ensuite construire un fichier qui permet, grâce à un codage spécifique, de l'utiliser comme un tableau de contingence.

Néanmoins, afin de pallier certains problèmes d'analyse contextuelle, nous avons mis en place d'autres procédures complémentaires à celles proposées par SPAD-T pour étudier les différents mots présents dans les discours retranscrits. Mais ce choix personnel n'est pas obligatoire pour pouvoir effectuer l'analyse textuelle avec SPAD-T. Nous insisterons seulement sur le fait que nous avons opté pour des procédures lexicométriques contextuelles afin de préserver les informations nécessaires pour déterminer le sens des mots. En effet, nous n'avons pas voulu nous contenter de l'examen des données textuelles hors contexte (liste de mots isolés) car ces listes ne permettent pas une interprétation correcte du point de vue sémantique. Donnons l'exemple du mot "bruit" : il peut faire référence à un bruit positif ("j'aime le bruit de l'eau de la fontaine") ou à un bruit négatif ("je n'aime pas le bruit des voitures"). Si nous réalisons l'analyse thématique hors contexte les deux sens de la forme bruit seront représentés par un même point dans le plan factoriel, ce qui induit une erreur d'interprétation. Par contre si cette analyse est réalisée en contexte, le sens de "bruit" est bien défini et nous pouvons ainsi créer deux "items" différents de bruit, l'un positif et l'autre négatif, qui seront traités et représentés à part sur les plans factoriels.

Les procédures du traitement statistico-lexicométrique sont fondamentales pour bien comprendre les principes des analyses textuelles. Les analyses sont surtout basées sur l'utilisation de logiciels et la mise au point de principes lexicométriques (mesure et description informatique du lexique) qui supposent la maîtrise d'un certain nombre de définitions dont nous présenterons les plus importantes :

1) Les unités de traitement : forme, occurrence,
2) Les listes des formes : index et concordances.

La forme graphique est l'unité de base que l'ordinateur est capable de traiter, c'est-à-dire, une chaîne de caractères séparée par deux blancs (avant et après) et accompagnée ou non d'une ponctuation. Autrement dit, il s'agit d'une chaîne de caractères alphanumériques non-délimiteurs (lettres et chiffres) qui est entourée par d'autres caractères différents appelés délimiteurs (blancs, points, virgules, deux points, etc.). A partir

de là nous pouvons définir les occurrences qui correspondent tout simplement au nombre de fois où une forme graphique apparaît dans un texte, un entretien ou une question ouverte. Ainsi on dira que la forme graphique "bruit" comporte 48 occurrences si cette forme apparaît 48 fois dans le texte.

En psychologie, la récurrence de certains thèmes étant importante, il est donc intéressant d'étudier le poids en nombre d'occurrences de chaque "item", c'est-à-dire, de croiser toutes les fréquences des "items" avec l'ensemble des individus afin de caractériser l'utilisation des formes qui sont sémantiquement le support du message et de l'information. Pour cela il faut se munir des outils linguistico-informatiques d'aide à l'interprétation : les index et les concordances (ou index contextuels) qui présentent de façon différente mais complémentaire le lexique du corpus. L'index présente l'ensemble des formes du corpus, classées soit par ordre alphabétique soit par ordre hiérarchique (ordre de fréquence). La concordance (ou index contextuel) présente les mêmes informations que l'index mais en contexte, c'est-à-dire, que contrairement aux index, les mots listés dans les concordances sont présentés dans leur environnement contextuel qui permet de les analyser. Un index ne donne que des indications sur l'existence et le poids des mots en nombre d'occurrences alors que les concordances permettent de réaliser les analyses contextuelles qui sont à la base de toute étude sérieuse visant à atteindre le contenu de l'information textuelle informatiquement retranscrite. Les concordances sont par conséquent des outils qui nous aident, dans nos analyses de contenu, à dégager les différents thèmes que les sujets ont traités dans les entretiens et dans les questions ouvertes.

3.1. Le protocole d'analyse des données textuelles

Il faut tout d'abord commencer par la retranscription des entretiens et/ou la saisie des questions ouvertes. A partir des bandes sonores enregistrées, l'ensemble des entretiens est regroupé dans un même fichier afin d'être utilisé comme données sources selon une codification spéciale qui permet de construire un tableau de contingence théorique qui sera ensuite associé aux individus déclarés sous SPAD-N. Cette dernière opération constitue **la codification des données textuelles**. Pour cela chaque sujet est codé de la même façon que dans SPAD-N mais on associe également des codes spécifiques aux différentes questions des entretiens.

A titre d'exemple, nous donnons un petit échantillon de codification d'un individu selon la méthode SPAD-T :

----0029
mon village, blotti le long de la forêt, qui lui fait une couleur verte ou ocre, suivant les saisons, est ce que j'aime dans mon environnement. cette coulée de verdure, avec un soleil couchant qu'on ne trouve que dans notre village. la forêt, pleine d'oiseaux qu'on voit évoluer avec les saisons (...).
++++
je déteste ces maisons prétentieuses qu'on érige sans ordre et sans génie, ces décharges sauvages, ces dépôts d'herbe tondue que d'abominables citadins, qui se croient devenus des ruraux, essaiment dans la forêt. je déteste ces voitures, sans parler des cyclomoteurs, des tondeuses, des camions et tous ces cancres qui tolèrent le boucan des moteurs mais gueulent si un coq chante, ces avions, ces télés, tous ces objets de consommation qui nous rendent cons et nous empêchent de rêver (...).
++++
oui, je suis préoccupé par les promoteurs immobiliers qui ne peuvent pas voir un espace vierge sans vouloir le violer. j'ai peur de tous ces citadins avides qui prétendent aimer la campagne mais voudraient avoir le métro à leur porte. j'ai peur d'un avenir où la mégalopole nous rejoindra (...).
++++
fermer Roissy. interdire tout nouveau permis de construire dans la commune. interdire les camions et surtout réprimer le peu de cas qu'ils font des interdictions de circuler, qui existent dans le village. mettre fin à l'exploitation du gypse. obliger les agriculteurs à respecter la nature en luttant contre les pesticides, les nitrates, etc. éduquer les gens, depuis nos voisins jusqu'aux directeurs de l'aéroport de paris, pour que chacun respecte la liberté des autres et la propriété collective, places, routes, etc. (...)

L'échantillon que nous venons de présenter reproduit le format du contenu du "fichier données textuelles" qui a été saisi et où l'on peut constater un certain nombre de séparateurs informatiques. En effet, il s'agit d'un échantillon sur un individu ayant répondu à quatre questions. Celui-ci est identifié par "----0029" (car c'est le sujet n°29) et les questions sont séparées par "++++". SPAD-T peut ainsi reconnaître chaque numéro d'individu toujours précédé de quatre tirets consécutifs. De la même façon, le logiciel peut soit travailler sur l'ensemble du discours de chaque individu, soit dissocier le lexique de chaque question grâce au délimitateur "++++". La fin du corpus est marquée par le codage suivant : = = = =.

Pour chaque sujet, il est donc impératif de garder la même structure afin d'éviter les erreurs de codification qui fausseraient les analyses statistiques. D'autre part, il est possible également d'introduire des séparateurs de groupes d'individus à partir du code suivant : ****. Cela permet d'effectuer des études du point de vue individuel et/ou collectif.

A ce codage spécifique qui sert de tableau de contingence théorique des données textuelles, on pourra y associer les données numériques

issues de SPAD-N, rendant ainsi réalisable l'étude conjointe des variables nominales et textuelles des individus.

3.2. Le traitement statistique des données textuelles

Le protocole que nous venons de présenter constitue donc la phase initiale de l'analyse des données textuelles qui rend possible les analyses statistico-linguistiques. En effet, la codification du corpus permet d'exploiter les données à partir de la création des listes de formes (index et concordances) en proposant un classement des informations textuelles recueillies qui seront analysées de façon thématique. Nous pouvons ainsi étudier la liste des formes, leur poids statistique (fréquence) et les différents thèmes présents dans les discours.

Nous passerons en revue rapidement les différentes procédures employées pour le traitement des corpus textuels ainsi que le couplage avec les données numériques issues des questionnaires.

Afin de ne pas être redondant par rapport à la présentation des procédures de traitement statistique de SPAD-N, nous définirons sommairement les différentes phases utilisées avec SPAD-T :

1) Archivage des données numériques à coupler avec les données textuelles (ARDON),
2) Archivage des données textuelles (ARTEX),
3) Sélection questions/individus : définition des numéros des questions (ouvertes ou entretiens) à traiter et éventuellement des filtres pour la sélection des individus (SELOX),
4) Définition des formes graphiques, calcul des fréquences, création du tableau lexical, tri alphabétique et hiérarchique hors contexte (NUMER),
5) Analyse factorielle des correspondances simples du tableau de contingence croisant en ligne les individus et en colonne les formes graphiques. Calcul des valeurs propres, coordonnées factorielles pour les plans factoriels (ASPAR),
6) Positionnement des variables nominales en illustratif sur les axes factoriels des données textuelles (POSIT),
7) Création de l'arbre hiérarchique (voisins réciproques) et partition en classes (RECIP / PARTI),
8) Caractérisation des classes par les mots-items les plus représentatifs (MOCAR).

Toutes ces procédures ouvrent donc la possibilité de coupler, pour l'analyse, les résultats des questionnaires avec le discours contenu dans les entretiens. On obtient ainsi des graphiques comportant à la fois les informations issues des données numériques saisies dans SPAD-N et les différents aspects des discours qui apparaissent dans les entretiens et qui sont thématisés et comptabilisés dans l'analyse. Ainsi le psychologue peut en même temps étudier les variables fermées qu'il a définies dans son questionnaire et les informations d'ordre textuel (à partir des entretiens) beaucoup plus ouvertes qu'il ne peut pas toujours inventorier a priori et qui constituent des éléments importants qui viennent s'ajouter à l'étude (perception, représentations, évaluations, opinions, attitudes, etc.).

Afin d'illustrer au mieux cette démarche, nous proposerons une représentation schématique moins techniques de la méthodologie que nous pratiquons pour l'analyse et le couplage statistique des données numériques et textuelles :

Traitement des données :

Données numériques (SPAD-N) :
Création et codification des variables et des modalités de réponse,
Saisie des données numériques,
Analyses factorielles des correspondances multiples,
Classifications hiérarchiques.

⇓

Données textuelles (SPAD-T) :
Saisie et codification des entretiens et des questions ouvertes,
Création des index et des concordances,
Constitution d'items sémantiques,
Analyse factorielle des items sémantiques,
Méthode de classification automatique.

⇓

Couplage des données textuelles et numériques (SPAD-T et données SPAD-N) :
Analyses factorielles et classification des items sémantiques et des variables nominales,
Interprétation.

4. EN CONCLUSION

A travers l'exemple de l'analyse conjointe des questionnaires et des entretiens, nous pouvons nous rendre compte du rôle important des dernières évolutions informatiques en matière de traitement de l'information numérique et textuelle grâce à l'amélioration des performances des ordinateurs et à la mise en oeuvre de nouvelles techniques d'analyse des données textuelles (procédures lexicométriques et analyses contextuelles).

Dès lors que l'on dispose de moyens informatiques performants, il devient possible d'analyser des données de très grande dimension, telles que celles recueillies par les méthodes d'enquête. A l'analyse statistique des questionnaires, utilisée depuis déjà longtemps, vient s'ajouter une nouvelle perspective qui offre deux possibilités extrêmement intéressantes. En premier lieu, celle de pouvoir effectuer des analyses de contenus de nombreux entretiens (plusieurs centaines, par exemple) difficilement réalisables manuellement. Or, nous savons combien le discours des individus est riche d'informations, et lorsque l'on s'intéresse à la perception, aux représentations, à l'évaluation, etc., comme c'est le cas en psychologie, il devient très utile d'employer ce type de méthode sur de grands échantillons. Mais surtout, la seconde possibilité, qui consiste à coupler informatiquement deux formats de données d'origine différente, est particulièrement attractive dans la mesure où elle rend possible la mise en relation des informations recueillies par questionnaires et de celles issues d'entretiens.

Cette méthode a déjà fourni des résultats dans le cadre de nos études psychologiques sur les stress de l'environnement et leurs conséquences sur l'homme. Ainsi, le couplage des thèmes des entretiens avec les résultats des questionnaires a permis d'expliquer certains phénomènes d'interactions entre l'homme et son environnement, tel que le lien entre le niveau de gêne exprimée dans un site bruyant et la représentation positive ou négative du cadre de vie. On a pu constater, par exemple, que plus les sujets appréciaient leur cadre de vie et moins ils toléraient l'intrusion de nuisances.

Nous ne nous attarderons pas davantage sur les résultats de nos études, mais nous conclurons en rappelant que la banalisation de ce type d'outils, auparavant réservés à des spécialistes (statisticiens), ouvre aujourd'hui, moyennant une formation méthodologique pluridisciplinaire (informatique, linguistique, statistique), de nouvelles perspectives non

seulement en psychologie mais également dans bien d'autres disciplines qui travaillent à partir du texte ou du discours.

Isabelle MARAMOTTI,
Département de Psychologie
Université de Paris X-Nanterre
Javier SANCHEZ
Université de Limoges