Le multimédia et le couplage statistique des données numériques et textuelles : nouvelles perspectives dans l'enseignement et dans la recherche en sciences humaines

Javier SANCHEZ et Isabelle MARAMOTTI

Abstract. The increasingly frequent use of computers in the field of human and social sciences has leas to the treatment of large amounts of textual and numerical data. In this perspective, we will present in this article the résults of an interdisciplinary collaboration between linguists and psychologists on the methodological choice of automatic tools for the processing of surveys. These tools will enable the user to process the textual data corresponding to each interview and the numerical data corresponding to each questionnaire in order to provide a statistical analysis based on the combination of the results yielded by both kinds of data. We will also discuss some of the recent progress made in the field of multimedia which, besides the development of paedagogical applications, have enabled us to make use of voice recognition softwares, which, together with macrocommand programming, have made it possible to analyse a written or transcribed text (interview) from voice commands.

Keywords: Linguistics, psychology, computer science, statistics, textual analysis.

Mots-clés : Linguistique, psychologie, informatique, statistique, analyse textuelle.

1. Introduction

Depuis le début de cette décennie, on a pu observer une banalisation importante de l'utilisation de l'informatique dans les secteurs des sciences

✉ Javier SANCHEZ; Département d'Espagnol (Linguistique) de la Faculté des Lettres de Limoges; 39E, rue Camille Guérin; F–87036 Limoges Cedex (France).
Fax : +33 55 43 56 03

✉ Isabelle MARAMOTTI; Département de Psychologie de l'Université de Paris X-Nanterre; 200, avenue de la République; F–92001 Nanterre Cedex (France).
Fax : +33 1 40 97 71 58

humaines et sociales ayant besoin de traiter d'importantes masses d'informations. C'est le cas notamment des disciplines travaillant sur les données textuelles, et tout particulièrement de la linguistique. D'autres disciplines peuvent également faire appel à l'informatique pour l'étude du texte, comme la psychologie lorsqu'elle utilise la méthode d'enquête par entretiens (nous ne faisons pas allusion ici à l'entretien clinique, mais aux entretiens recueillis lors d'enquêtes sur le terrain, et par conséquent auprès d'un grand nombre d'individus). Ainsi, qu'il s'agisse du texte écrit, qui constitue l'objet d'étude en lui-même pour le linguiste, ou du discours oral qui représente davantage un moyen d'étude pour le psychologue, et bien que se situant dans des perspectives de recherche différentes, l'impératif méthodologique est le même : *comment approcher le discours de très grande dimension afin de l'analyser et d'en réaliser une interprétation exhaustive?* Une première réponse à ce besoin méthodologique a été le développement des logiciels lexicométriques que nous utilisons dans nos recherches depuis de nombreuses années[1].

Dans cette optique, nous présenterons ici le fruit d'une collaboration interdisciplinaire, en tant que linguiste et psychologue, concernant le choix méthodologique d'outils informatiques pour le traitement statistique d'enquêtes, comportant à la fois un recueil de données textuelles par entretien et un recueil de données numériques par questionnaire. En effet, les nouveaux outils qui sont proposés pour ce couplage statistique ont comme principal objectif la réalisation d'études approfondies permettant d'atteindre une analyse à la fois qualitative et quantitative des données d'enquête de grande dimension. Ce couplage statistique des questionnaires fermés avec la parole libre (entretiens) des sujets interrogés peut intéresser non seulement les psychologues mais également les chercheurs en sciences du langages travaillant, à partir d'enquêtes linguistiques, sur l'usage des formes selon les registres de langue, les différences régionales, les domaines de spécialité (terminologies), etc. Quant au psychologue, l'intérêt réside dans la possibilité de pouvoir mettre en relation les données contenues dans les discours des individus avec de nombreuses variables recueillies par un autre mode d'investigation (questionnaires, données expérimentales, mesures physiques et physiologiques, etc.).

Avant de développer ces aspects, nous présenterons les évolutions de l'informatique multimédia qui peuvent être utilisées non seulement dans l'enseignement mais aussi dans ce type de recherches. En effet, l'oral et l'écrit

[1] SANCHEZ (Javier) : 1992, *Méthodologie et outils de l'analyse relationnelle informatique des textes.* Collection : Analyse textuelle et nouvelles technologies, n° 2, Université de Paris VIII, Saint-Denis, 370 pp.; ARMANET (Véronique) : 1977, *Des structures syntaxiques de l'espagnol à l'analyse relationnelle des textes.* Honoré-Champion, Paris.

constituent deux médias à part dans le sens où en linguistique textuelle l'acquisition (enregistrement) des données se réalise actuellement très aisément grâce à l'utilisation du scanner et des logiciels de reconnaissance optique des caractères, alors que dans les disciplines travaillant exclusivement sur le discours oral (entretiens), comme en psychologie, le chercheur ne dispose pas, pour l'instant, de moyens de saisie automatique de la parole. Ce vide méthodologique implique une difficulté opératoire importante dans la mesure où la retranscription des données textuelles s'effectue manuellement au clavier.

Mais ce décalage entre l'oral et l'écrit commence actuellement à se réduire depuis l'apparition de l'informatique multimédia grâce à la naissance de composantes telles que les cartes sonores, le CD-ROM, et le développement des logiciels pouvant utiliser simultanément les différents médias : oral, écrit, images (fixes ou mobiles). En effet, depuis peu on commence à obtenir des résultats encourageants dans le développement des logiciels de reconnaissance et de saisie de la parole même si pour l'instant ils ne sont pas encore accessibles au grand public et méritent beaucoup d'améliorations. Pour notre part, dans le cadre de ces recherches interdisciplinaires, c'est en nous penchant sur ce problème de la saisie automatique du discours oral que nous nous sommes intéressés à l'univers de l'informatique multimédia qui nous a permis, du point de vue méthodologique, d'ouvrir de nouvelles perspectives de recherche. Bien que nos besoins méthodologiques dans le domaine de la reconnaissance vocale n'aient pas encore trouvé de solution pour l'enregistrement du discours, nous avons pu dégager, malgré tout, d'autres types d'applications multimédias qui constituent pour nous de nouveaux outils de recherche. L'utilisation des logiciels de reconnaissance de la parole associés à la programmation de macrocommandes, ouvre à la recherche la possibilité d'analyser et de coder le texte retranscrit à partir de commandes vocales, c'est-à-dire que le chercheur n'utilise plus le clavier pour exécuter certaines tâches, comme par exemple les codifications éditoriales, structurelles et syntaxiques en linguistique textuelle ou l'analyse des items sémantiques en psychologie. D'autre part, comme nous l'avons déjà précisé, le multimédia nous a permis de découvrir ses puissantes capacités d'interactivité et nous nous penchons actuellement sur l'élaboration d'outils pédagogiques surtout dans le domaine de l'apprentissage des langues, en particulier de l'espagnol.

Nous présenterons tout d'abord les aspects de l'informatique multimédia, nouveau point de rencontre entre la recherche et l'enseignement, afin d'introduire les méthodes récentes de couplage statistique des données numériques et textuelles pour lesquelles nous utilisons certaines des techniques multimédias (reconnaissance de la parole).

2. Le multimédia dans l'enseignement et dans la recherche

L'axe principal de réflexion, à l'origine même du terme multimédia[2], est centré sur l'utilisation de différentes voies de dialogue entre l'utilisateur et l'ordinateur, mais l'aspect le plus important est la possibilité de créer une situation de dialogue préprogrammée que l'on appelle interactivité. En effet, dans le cadre des programmes d'enseignement, cette interactivité permet de réaliser une progression cohérente et évolutive de l'apprentissage, dans la mesure où l'outil multimédia veille à ce que l'utilisateur soit guidé dans un univers qui présente un ensemble de connaissances préenregistrées ainsi que des applications pratiques à des fins didactiques. L'exploitation de cette interactivité, comme nous le verrons plus tard, peut être également d'une grande utilité au niveau de la recherche.

2.1. Le multimédia dans l'apprentissage des langues

C'est dans l'apprentissage des langues[3] que le développement des systèmes multimédias semble le plus prometteur car il offre de nouvelles méthodes d'enseignement interactives individuelles et collectives. Ainsi grâce aux outils multimédias on a vu naître des logiciels qui mettent à la disposition des étudiants textes, sons et images. C'est-à-dire que l'enseignement gagne en qualité puisque l'on peut intervenir dans les deux niveaux d'apprentissage de la langue : l'oral et l'écrit. De plus, le renouvellement des méthodes traditionnelles a permis un nouvel élan pédagogique grâce à ces nouveaux supports interactifs qui permettent d'élaborer des programmes plus structurés et plus conviviaux.

En effet, les progrès réalisés sur la numérisation du texte, du son et des images, grâce à l'interactivité proposée avec des commandes simples, ont apporté une meilleure aide à l'apprentisssage de la grammaire aussi bien à l'oral qu'à l'écrit. Il est vrai que depuis longtemps le texte ou les images ainsi que les enregistrements sonores ou audiovisuels sont utilisés par les professeurs de langues vivantes. Mais l'originalité des logiciels multimédias est d'accéder directement et simultanément aux diverses composantes de l'apprentissage des langues, en proposant dans les logiciels des simulations de

---

[2] LES DOSSIERS DE L'INGÉNIERIE ÉDUCATIVE : 1995, *Multimédia : enseignement, formation, téléformation* CNDP (France) / Les Publications du Québec, 168 p.

[3] SANCHEZ (Javier) : 1995, « L'ordinateur dans l'enseignement de la langue et de la linguistique hispaniques », *Revue Enseignement Public et Informatique (EPI)*, n° 77, Paris (sous presse).

situations de communication les plus proches du fonctionnement réel de la langue. Ainsi par exemple, l'élève prend connaissance des instructions orales ou écrites, réalise des exercices et l'ordinateur corrige les fautes en proposant les règles à la demande. Cette interactivité et cette convivialité mettent en puissance le travail de l'élève, car l'étudiant est plongé dans un univers attractif dont il fait partie et dans lequel il joue un rôle important.

C'est ainsi que nos recherches linguistiques (en langue espagnole) sont orientées actuellement vers la création de logiciels multimédias interactifs. La présentation du matériel pédagogique doit mettre en éveil chez l'apprenant sa façon d'inférer, de généraliser et de mettre en relation toutes les données afin d'activer les automatismes linguistiques de base : *écouter et lire (décodage)*, *parler et écrire (codage)*. La répétition nous semble donc être un des points essentiels dans l'acquisition des automatismes en langue vivante.

Nous dirons en conclusion, qu'il nous semble que l'évolution la plus marquante du multimédia dans l'enseignement des langues concerne la correction phonétique et la pratique de l'expression orale, car les cartes sonores ont permis d'introduire un enseignement interactif de l'oral. Ainsi, l'apprenant, plongé dans une situation de communication fictive, écoute le modèle proposé par l'ordinateur et répond ou lit les mots ou les phrases affichées à l'écran. Si l'ordinateur a des difficultés à reconnaître la réponse, l'utilisateur passe au test de prononciation en essayant d'imiter le modèle jusqu'à la réussite de l'exercice phonétique. L'objectif principal de ce système est de favoriser un travail de production orale fondé sur l'imitation d'un modèle dans une langue donnée.

2.2. Application du multimédia dans la recherche : la codification vocale

Comme nous l'avons vu, le multimédia apporte de nouvelles perspectives d'enseignement dans les disciplines en sciences humaines et sociales, mais il ouvre également de nouvelles possibilités méthodologiques au niveau de la recherche proprement dite.

En effet, actuellement nous exploitons dans nos recherches textuelles, une partie des ressources du multimédia pour la codification linguistique du discours (en linguistique) ou l'analyse d'items sémantiques (en psychologie). La technique consiste à utiliser la reconnaissance vocale afin de constituer un inventaire de commandes orales permettant des applications de codage sur le texte retranscrit. Il est ainsi possible d'associer des actions préprogrammées par le chercheur, ce qui évite un travail plus ou moins long au clavier.

Donnons quelques exemples. Supposons que nous voulions coder un texte interlocutif pour une recherche d'ordre linguistique. Le linguiste préprogramme les commandes orales et ses actions afin de réaliser automatiquement l'insertion des références textuelles nécessaires aux codages structuraux et éditoriaux du corpus. Le chercheur parcourt ce corpus et lorsqu'il y a nécessité de réaliser un codage, prononce le nom de la commande et l'action est immédiatement exécutée. Cela permet d'éviter un grand nombre de manipulations au clavier et rend le travail plus convivial et plus rapide.

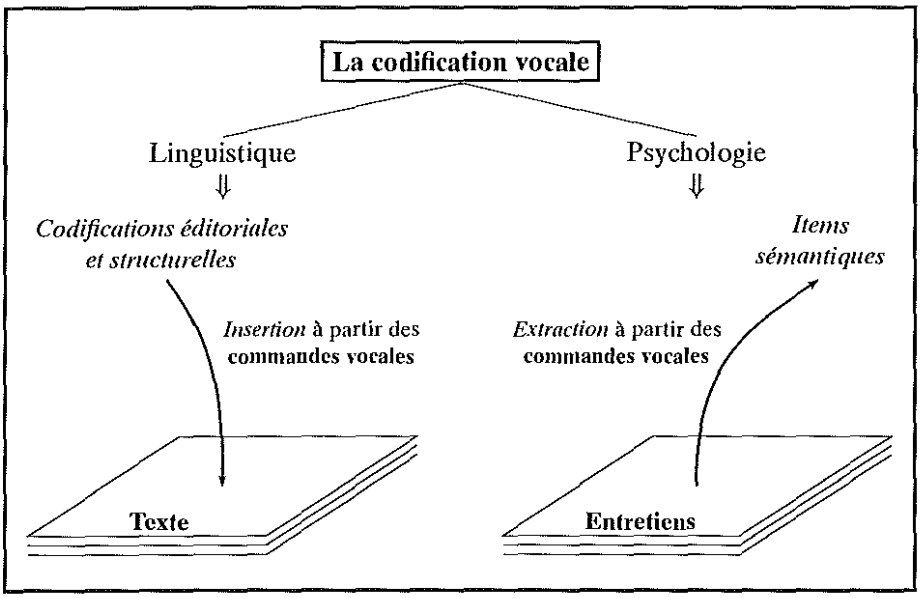

Fig. 1.– La codification vocale en linguistique et en psychologie

Un autre exemple, en psychologie, correspond à l'analyse du discours au moyens d'items sémantiques qui doivent être traités statistiquement. Dans ce cas, le chercheur a besoin de transformer la structure superficielle du texte en une série d'items qui représentent la structure profonde. Le psychologue constitue progressivement les commandes vocales correspondant aux items sémantiques et lorsqu'un thème apparaît, il prononce la commande correspondant à l'item. Les actions peuvent être simples, dans le sens où la commande orale se limite à insérer le nom de l'item sémantique dans un fichier spécifique, ou plus complexes avec la création d'une base de données comportant à la fois les items sémantiques et le contenu même du discours. Le chercheur obtient

ainsi une véritable aide à la constitution des bases de données interrogeables et/ou à la création des fichiers en vue d'une analyse statistique.

3. Le couplage statistique des données numériques et textuelles

Si le multimédia permet d'ouvrir de nouvelles voies d'application, nos récents travaux en linguistique-informatique[4] ont essayé de banaliser l'utilisation des procédures de description du lexique et d'analyse statistique multidimensionnelle pour l'étude des corpus. Mais parce que nous souhaitions élargir cette banalisation au-delà des applications en linguistique, nous nous sommes penchés sur certains aspects méthodologiques du traitement des enquêtes et des entretiens qui ont donné lieu à une collaboration interdisciplinaire, en particulier avec les psychologues.

Après avoir illustré les applications multimédias possibles dans la recherche et dans l'enseignement des langues, nous allons maintenant présenter de nouvelles méthodes informatiques pluridisciplinaires dans le cadre d'une analyse conjointe d'entretiens et de questionnaires. Nous donnerons un exemple à partir de l'un de nos derniers travaux en psychologie, et plus précisément en psychologie de l'environnement[5]. L'étude qui nous servira d'illustration a été réalisée dans le cadre d'une recherche commandée par le Ministère de l'Environnement français, ayant comme objectif la construction des normes concernant les principaux facteurs de gêne existant dans l'environnement. Nous présenterons brièvement cette étude afin de montrer l'intérêt du choix méthodologique. Il s'agissait d'examiner l'impact de deux facteurs environnementaux importants, le bruit et les vibrations, sur le bien-être des habitants dans quatre sites différents et notamment de déterminer l'existence de seuils à partir desquels un facteur donné devenait une source de nuisance pour l'individu.

La problématique de cette recherche nécessitait d'enquêter « sur le terrain » afin de conserver l'ensemble des paramètres présents dans les situations

---

[4] SANCHEZ (Javier) : 1993, « Qu'est-ce que l'analyse relationnelle informatique des textes ? » *Revue. Informatique et Statistique dans les Sciences Humaines*, Université de Liège, 29e année, nos 1 à 4, pp. 135–165.

[5] MARAMOTTI (Isabelle) : 1994, *Approche psychologique des facteurs de l'environnement : application à la multi-exposition au bruit et aux vibrations*, (Université de Paris X-Nanterre. Thèse de Doctorat Nouveau Régime); MARAMOTTI (Isabelle) : 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).

de la vie quotidienne. Il était donc indispensable de pouvoir mettre en évidence les différentes évaluations des personnes, mais également tous les processus perceptivo-cognitifs (perception, représentation, etc.) qui concourent à donner une signification au monde dans lequel nous évoluons, et qui médiatisent plus ou moins l'évaluation que nous faisons de notre cadre de vie. Ces aspects, étant difficiles à atteindre par simple questionnaire à choix multiple, nous avions opté pour l'entretien, qui permet aux personnes de s'exprimer librement et qui offre une information plus riche et plus exhaustive dans la mesure où l'on ne dirige pas les réponses de l'enquêté.

Par ailleurs, la perception et l'évaluation étant elles-mêmes dépendantes de nombreuses variables individuelles, situationnelles et contextuelles nous avions besoin de recueillir ces paramètres pour chacun des individus interrogés, ce qui impliquait que chaque personne ayant subi un entretien devait également répondre à un questionnaire. On peut comprendre pourquoi il était indispensable de pouvoir associer ce que les habitants exprimaient dans les entretiens à ce qu'ils avaient répondu dans le questionnaire.

Pour le couplage nous avons choisi d'utiliser deux logiciels : *SPAD-N*[6] permettant le traitement des données numériques et *SPAD-T*[7] permettant l'analyse des données textuelles. *SPAD-N* effectue des analyses multidimensionnelles telles que l'analyse factorielle des correspondances et les techniques de classification automatique[8]. Pour notre étude la codification des variables et des modalités de réponses du questionnaire comportait un total de 146 variables et de 374 modalités de réponse sur une population de 176 individus répartis sur quatre sites, ce qui constitue un tableau de contingence de 25 696 entrées. Le logiciel *SPAD-T* rend possible le traitement des données textuelles mais surtout l'importation du tableau des données numériques du questionnaire construit avec *SPAD-N* en vue d'effectuer le couplage des données numériques et textuelles. Cette phase permet au chercheur de mettre en relation les réponses du questionnaire avec le contenu des entretiens et d'atteindre donc l'analyse quantitative et qualitative dont nous avons parlé.

La première difficulté de cette analyse textuelle est la préparation des données afin d'aboutir à l'équivalent d'un tableau de contingence qui puisse

---

[6] *SPAD-N* (Système Portable d'Analyse des Données — Numériques), Centre International de Statistique et d'Informatique Appliquée (C.I.S.I.A.), 1, Avenue Herbillon, F–94160 Saint-Mandé (France).

[7] *SPAD-T* (Système Portable d'Analyse des Données — Textuelles).

[8] SANCHEZ (Javier) : 1994, «*SPAD-N* au service d'une méthodologie pour l'analyse des données textuelles» *Enseignement Public et Informatique (EPI)*, n° 74 (Paris) pp. 219–236.

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 est de définir à partir du discours les différents thèmes que chaque individu a développés, pour ensuite construire un fichier qui servira à l'analyse statistique. C'est à ce niveau que le multimédia contribue à l'avancement de notre méthodologie avec la codification vocale que nous avons déjà évoquée.

Néanmoins, afin d'améliorer la définition des items sémantiques, nous pouvons utiliser également 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 effectuer l'analyse textuelle avec *SPAD-T*. Nous insisterons seulement sur le fait que nous pouvons opter pour des procédures lexicométriques contextuelles[9] afin de préserver les informations nécessaires pour déterminer le sens des mots. En effet, nous ne nous contentons pas 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-lexicométriques[10] sont fondamentales pour bien comprendre les principes des analyses textuelles proposées par *SPAD-T*. Les analyses sont surtout basées sur l'utilisation des principes des logiciels lexicométriques (mesure et description informatique

---

[9] SANCHEZ (Javier) : 1994, « De la désambiguïsation assistée par ordinateur à l'élaboration des grammaires textuelles » *Literary and Linguistic Computing*, (Oxford University Press) vol. 9, n° 3, pp. 195–202.

[10] ARMANET (Véronique) et SANCHEZ (Javier) : 1991, « Théorie, méthode et techniques de recherche en linguistique informatique hispanique » *Revista de Estudios Pluridisciplinares sobre as Sociedades Mediterrânicas,* n° 1 (Lisbonne) pp. 163–178.

du lexique) qui supposent la maîtrise d'un certain nombre de définitions dont nous rappelons 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 de traitement, et elle se définit comme 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.). À 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 256 occurrences si cette forme apparaît 256 fois dans le texte.

En psychologie, par exemple, 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. Pour cela l'utilisation des index et des concordances constituent des outils linguistico-informatiques d'aide à l'interprétation (*SPAD-T* fournit ces types de listes). Rappelons que 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.

3.1. Le protocole d'analyse des données textuelles

Pour l'étude qui nous sert d'illustration, nous avions réalisé 176 entretiens semi-directifs qui comportaient plusieurs questions. La première étape consiste à réaliser la retranscription des entretiens à partir des bandes sonores enregistrées. Ainsi, l'ensemble du corpus textuel est regroupé dans un fichier afin d'être utilisé comme données sources selon une codification spéciale qui

permet à *SPAD-T* de générer les index et les concordances qui serviront d'aide pour la définition des items sémantiques. Le chercheur, grâce à la codification vocale, peut ensuite construire un autre fichier du même format contenant les différents items sémantiques. Ce dernier fichier servira de tableau de contingence qui sera associé aux données issues des questionnaires déclarés sous *SPAD-N*. Pour cela chaque sujet est codé avec le même numéro d'ordre que dans *SPAD-N* mais on associe également des codes spécifiques aux différentes questions des entretiens.

À titre d'exemple, nous donnons un petit échantillon de codification du texte source d'un individu selon la méthode *SPAD-T* :

```
----0135
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 source des données textuelles » qui a été saisi et où l'on

peut constater un certain nombre de séparateurs informatiques. Il s'agit ici d'un échantillon sur un individu ayant répondu à quatre questions. Celui-ci est identifié par « – – – – 0135 » (car c'est le sujet n° 135) et les questions sont séparées par « + + + + ». *SPAD-T* reconnaît ainsi 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 (comparaison de groupes).

3.2. Exemple d'application

Le protocole que nous venons de présenter très sommairement constitue donc la phase initiale de l'analyse des données textuelles qui rend possibles les traitements statistico-linguistiques. L'objectif est d'obtenir une série de graphiques factoriels comportant à la fois les informations numériques issues du questionnaire et les données textuelles des entretiens, qui ont été préalablement préparées pour l'analyse. Les différentes procédures employées par *SPAD-T* pour le traitement des corpus textuels couplés avec les données numériques des questionnaires pourraient être schématisées comme représenté ci-contre.

À titre d'illustration, nous proposerons un graphique factoriel (Fig. 2) issu du couplage de données textuelles et numériques dans le cadre de l'étude que nous avons précédemment présentée. Il concerne l'influence de l'évaluation du cadre de vie sur le jugement des nuisances environnementales. Dans cet exemple, nous avions comparé ce que les individus appréciaient dans leur environnement (données recueillies par entretien) et le degré de gêne due au bruit et aux vibrations (données recueillies par questionnaire). Le nuage de points en caractère minuscule représente les items sémantiques provenant des réponses à l'une des questions posées lors des entretiens semi-directifs. Les points en caractères majuscules correspondent aux différents degrés de gêne exprimée vis-à-vis des facteurs environnementaux, issus des données numériques du questionnaire. La superposition de ces deux nuages nous permet de faire une analyse à la fois quantitative et qualitative des données d'origine différente.

Archivage des données numériques et des données textuelles
⇓
Sélection questions/individus :
définition des numéros des questions (ouvertes ou entretiens) à traiter
et éventuellement des filtres pour la sélection des individus,
⇓
Définition des formes graphiques ou des « items sémantiques »,
calcul des fréquences, création du tableau lexical,
tri alphabétique et hiérarchique hors contexte
⇓
Analyse factorielle des correspondances simples du tableau
de contingence croisant en lignes les individus
et en colonnes les formes graphiques.
Calcul des valeurs propres, coordonnées factorielles
pour les plans factoriels,
⇓
Positionnement des variables nominales en illustratif
sur les axes factoriels des données textuelles
⇓
Création de l'arbre hiérarchique (voisins réciproques)
et partition en classes
⇓
Caractérisation des classes par les mots-items les plus représentatifs
⇓
Interprétation des résultats.

Ce couplage nous a permis de constater que les personnes qui se déclaraient gênées et très gênées par les vibrations (3-4EV, 5-6EV, 7+EV) se situaient à l'intérieur de la zone comprenant les évaluations positives concernant leur cadre de vie (signe +), de même que les gens dérangés par le bruit (GB). À l'inverse, les individus qui exprimaient peu de gêne à l'égard des vibrations et du bruit étaient localisés en dehors de cette zone. En d'autres termes, les habitants qui s'estimaient satisfaits de leur cadre de vie exprimaient plus de gêne vis-à-vis de ces deux facteurs environnementaux, confirmant le fait que l'évaluation de l'environnement joue un rôle dans le jugement que l'on porte sur les nuisances, dans le sens où plus les gens apprécient leur environnement et moins ils tolèrent l'intrusion de certaines nuisances. Ce résultat a pu être mis en évidence grâce au couplage statistique des entretiens et des questionnaires.

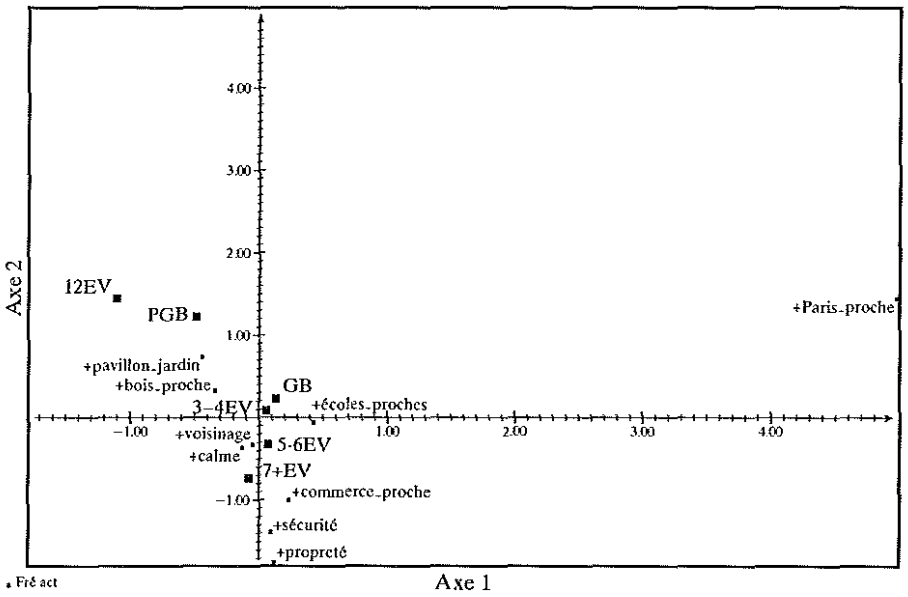

Fig. 2.– Exemple de couplage de données numériques et textuelles

Pour conclure notre propos, nous proposons un schéma récapitulatif et comparatif entre les procédures plus traditionnelles utilisées pour l'analyse linguistique des textes et les nouvelles procédures pour l'analyse des enquêtes en psychologie (Fig. 3).

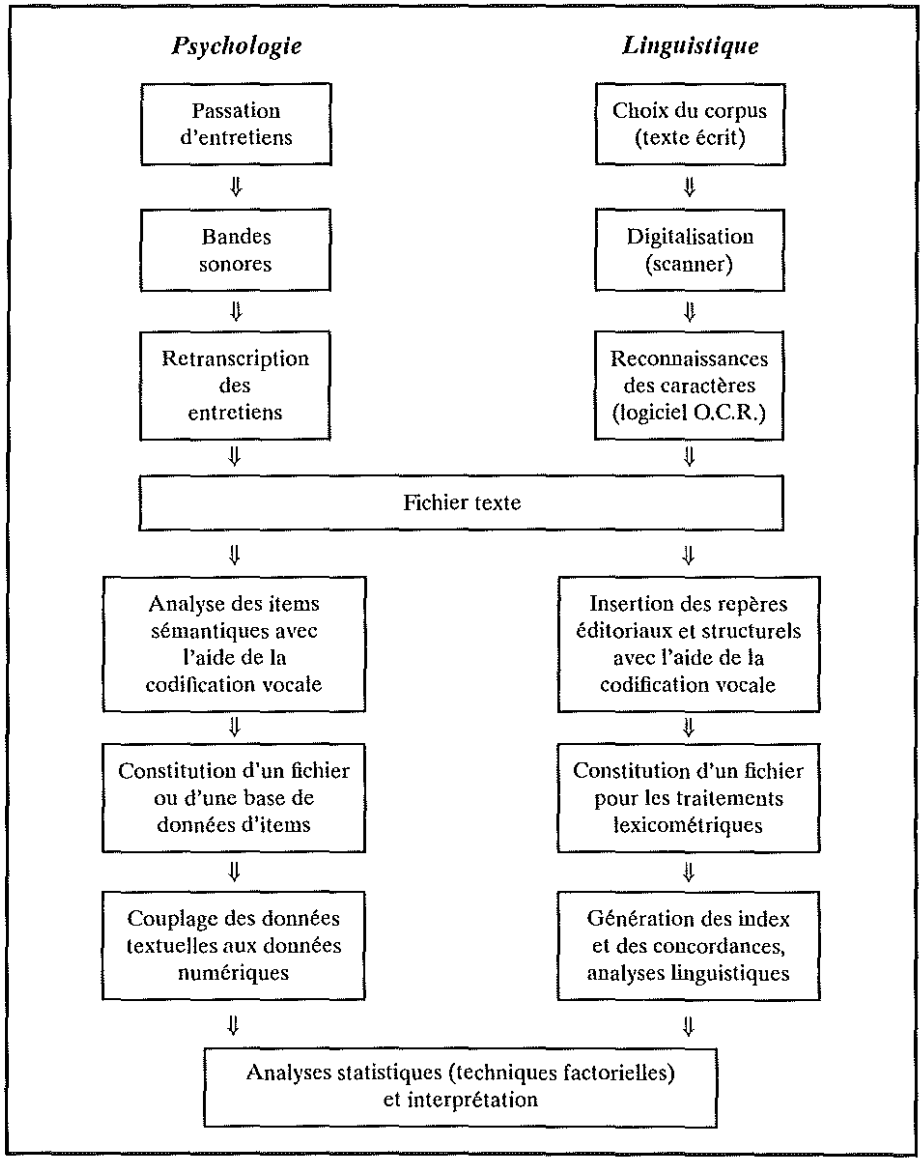

Fig. 3.– Mise en parallèle des procédures d'analyse textuelle en psychologie et en linguistique