http://www.sylvette-denefle.fr
Sylvette Denèfle, "Trois logiciels, trois interprétations ? Étude comparative d'un même corpus", Actes des Secondes Journées Internationales d'Analyse Statistique des Données Textuelles, Montpellier, 1993, p.102-112

Trois logiciels, trois interprétations ? Analyse comparative d'un même corpus

SUMMARY

To investigate computer assisted interpretation of social sciences data, we processed a corpus of semi-directive interviews, very common in our field of study, using three textual data analysis programs : **Alceste**, **Hyperbase** and **SPAD.t.** The goal of our investigative approach was to identify conceptual fields, not expressed in this form, common to our various interviewees.
The three programs provided different or even complementary interpretation components. While **Hyperbase** proved problematic for the interpretative use of available nominative and qualitative variables outside the corpus, **SPAD.t** dit provide a solution. However, we found the latter more statistically demanding than the others and it failed to produce the promising results revealed by **Alceste**.
We are therefore considering joint use of these tools, which offer differing potential.

Chercheurs en sciences sociales (histoire, sociologie), nous envisageons l'utilisation des logiciels d'analyse des données textuelles dans le cadre du travail d'interprétation des données qualitatives recueillies : documents d'archives de tous types, entretiens et réponses à des questions ouvertes de questionnaires.
Notre approche est celle d'utilisateurs d'outils qui n'en maîtrisent ni la conception ni la complexité et qui explorent les possibilités qui leur sont accessibles, compte tenu d'une connaissance limitée en statistique.
Les textes sur lesquels nous travaillons se caractérisent par leur nombre, ils sont non littéraires, généralement assez longs, très souvent transcrits de la langue parlée et diluent à travers le vocabulaire le plus courant l'expression des jugements qui retiennent notre attention. Par ailleurs, la connaissance que nous avons de la position de nos interlocuteurs dans le champ social détermine de façon très conséquente toutes nos interprétations.

Il nous faut donc explorer nos corpus, en comparer entre eux les différents éléments, les mettre en relation avec des données extérieures aux textes de façon à en analyser et en interpréter le contenu.
Nous avons considéré que les moyens informatiques devaient nous permettre d'appréhender ces textes, par le truchement de l'analyse lexicométrique, de façon à faire émerger d'éventuels champs sémantiques diffus dans le corpus, non immédiatement accessibles à la lecture habituelle de l'analyse de contenu et qui traverseraient la variation sociologique.
Pour ce faire, nous disposions, dans notre laboratoire, des logiciels **Alceste** (version 3.2), **Hyperbase** (version 1.2) et **SPAD.t** (version T.40) sur Macintosh. Le corpus de textes choisi est de forme très commune dans nos données. Il s'agit de 16 entretiens d'environ une heure (400K soit 75000 mots environ dont 6000 formes différentes) transcrits à partir d'enregistrements et portant sur les valeurs morales et les choix idéologiques de personnes, 8 hommes et 8 femmes, tous adultes en 1968 et tous issus de la région nantaise, se déclarant sans-religion.
L'une des principales interrogations de recherche concernait la mise en évidence éventuelle de systèmes idéologiques partagés par des personnes se situant dans un premier temps négativement par rapport à la religion catholique.

Saisie des données

Nous ne traiterons pas ici des difficultés inhérentes à la retranscription du discours parlé mais des aspects particuliers d'écriture liés à l'utilisation des logiciels d'analyse des données textuelles.
Nous sommes partis d'un corpus (questions des enquêteurs et réponses qui leur étaient faites) saisi à l'aide d'un logiciel de traitement de texte et d'un ensemble de variables nominatives ou quantitatives caractérisant sociologiquement les personnes interrogées.
Le logiciel **Alceste** accepte des partitions dans l'ensemble du texte, réduit en minuscules le texte à l'exception des mots écrits totalement en majuscules qu'il traite par la suite en éléments supplémentaires dans les analyses statistiques ; il supprime les accents et permet de spécifier les formes que l'on souhaite. Il autorise l'introduction d'éléments hors corpus pour chaque sous-texte (éléments également traités en éléments supplémentaires par la suite). C'est sous cette forme qu'ont été entrées les variables dont nous disposions. Nous avons choisi de mettre en majuscules les questions des enquêteurs pour éviter les interférences avec les discours que nous voulions analyser, sans cependant perdre les informations qu'elles pouvaient contenir.
Par ailleurs, Alceste opère une codification automatique de formes particulières qui n'interviennent pas dans l'analyse statistique (adverbes,

verbes modaux, organisateurs du discours, etc...), ce qui permet d'affiner l'interprétation.
Les limitations du logiciel à ce stade concernent le nombre des éléments hors corpus et la dimension totale du corpus (1 000 K).
Le logiciel **Hyperbase** autorise la partition du corpus, permet la reconnaissance de formes spécifiées mais interdit toute introduction directe d'éléments hors corpus telles que les variables sociologiques bien qu'on puisse, à la rigueur, traiter une variable à la fois en adoptant ses modalités comme partition. Nous avons donc éliminé toutes les questions en choisissant les enquêtés comme déterminants de la partition des textes.
Le logiciel **SPAD.t** distingue et analyse toutes les formes et nous a contraint à éliminer les questions posées par les enquêteurs. Mais il permet évidemment de traiter toutes les formes spécifiées et ne pose pas de limite théorique à la taille du corpus. On verra toutefois que cette qualité est problématique. Il admet l'introduction de variables sans limitation dans la mesure où on les écrit sous un format particulier dans deux documents parallèles au texte.

Pour notre corpus particulier, **Alceste** et **SPAD.t** répondent aux exigences sociologiques pour le traitement des variables nominatives ou quantitatives et aucun des trois logiciels ne posent de problèmes trop délicats de saisie des données bien que leurs façons différenciées de gérer la graphie puissent introduire quelques nuances.

La lemmatisation ou la réduction des formes

Alceste est le seul logiciel testé (compte tenu des versions dont nous disposions) qui met en œuvre une lemmatisation automatique. Mais ce traitement n'est pas incontrôlable et l'utilisateur peut modifier ce qu'il désire. Il suffit d'éditer le premier dictionnaire constitué avec un traitement de texte puis de le corriger ou de le modifier. Sur environ 3500 formes lemmatisées, 250 corrections ont été effectuées dans notre corpus. Ce travail de contrôle de la lemmatisation nous a paru plus accessible que l'écriture de la procédure de **SPAD.t** qui permet d'opérer des fusions de formes graphiques mais qui n'est utilisable que pour des opérations ponctuelles. **Hyperbase** quant à lui ne procède à aucune lemmatisation. Soulignons enfin qu'**Alceste** recompose les textes en formes réduites qui peuvent être directement traités par les autres logiciels.

Fonctions des trois logiciels et résultats de nos traitements

* Après la construction d'un dictionnaire, répertoriant l'ensemble des formes selon leur fréquence et leur mode d'intervention dans l'analyse, Alceste propose d'effectuer une classification descendante hiérarchique dont les paramètres sont modifiables. Le tableau de base traité est constitué en lignes par les Unités de Contexte Elémentaires (éléments obtenus par un découpage du texte en unités de 240 caractères environ) et en colonnes par les formes réduites analysées issues des textes. Cette classification, double ou simple au choix, se lit et s'interprète à l'aide du dendrogramme et des profils des classes.
Pour chaque classe, on dispose en fin d'analyse, des listes des formes et des couples significativement présents et absents. Sont également constitués un concordancier qui propose la liste des UCE de la classe contenant chaque forme réduite significative, des listes des segments répétés et des UCE significatifs par classe. Tous ces éléments permettent de déterminer le contenu sémantique des classes. L'AFC effectuée sur ces classes les positionne les unes par rapport aux autres.

Les profils et l'AFC font figurer les formes illustratives dans leur résultat. En fin de liste des classes, après les formes analysées, on peut lire les mots en majuscules (dans notre cas il s'agit des questions posées par l'enquêteur que nous avons pu conserver sous cette forme), puis les formes particulières reconnaissables par leur code et les variables. Il est possible d'opérer sur ces dernières des tris croisés qui permettent d'obtenir pour chaque modalité d'une variable, un profil de texte. Par contre, Alceste contrairement à SPAD.t n'offre pas la possibilité de faire une AFC sur un tableau où les variables seraient en colonnes et les formes en lignes ce qui nous permettrait d'en savoir plus sur le rapport des variables avec le texte. Les formes particulières affectées d'un code peuvent elles aussi faire l'objet d'un traitement statistique qui aide à l'interprétation de la classification obtenue.

Plusieurs analyses ont été effectuées après la modification du dictionnaire. La double classification la plus satisfaisante statistiquement qui analysait plus de 70% des UCE a donné quatre classes. Mais une classification en 5 classes, à la résolution statistique moins performante, permettait une autre interprétation intéressante des textes. Les personnes interrogées ont développé des points de vue autour de trois thèmes qui apparaissent de façon récurrente dans les différentes classifications : l'éducation religieuse, la religion et les phénomènes paranormaux. La lecture des fichiers des profils et des anti-profils a permis de déterminer ces champs thématiques.
L'aide des fichiers de contextualisation des formes est nécessaire pour caractériser la position précise des individus sur ces thèmes. Tous nos enquêtés ont eu une éducation catholique conformiste dans des familles non-pratiquantes. Leur rupture avec cette éducation se situe à peu près à l'âge adulte. Après un mariage à l'Église par exemple, ils choisissent de ne pas faire

baptiser leurs enfants (classe 1). On voit apparaître alors un certain anticléricalisme lié à l'opposition au caractère dogmatique des religions. Ils prônent la tolérance et reconnaissent que certaines personnes religieuses sont exemplaires dans leur comportement (classe 2). Ils ne cherchent pas refuge dans d'autres croyances et adoptent une position hostile à l'égard du paranormal mais participent quand même à l'un ou l'autre de ces faits en constatant par exemple qu'autour d'eux un tel est allé voir un guérisseur dont les pouvoirs sont incontestables (classe 3). Ces trois thèmes ont en commun la même logique de construction. Tout d'abord les interrogés affirment leur position à l'égard du sujet abordé et mettent des distances. Ensuite ils nuancent en faisant référence à la pratique ou à des contre-exemples.
La quatrième classe issue de l'analyse statistiquement la plus performante se caractérise par un contenu faisant référence aux valeurs morales et sociales qui s'imposent dans la vie en société. En revanche, la classification aboutissant à cinq classes en génère deux nouvelles. L'une fait référence à la métaphysique et à la mort, les enquêtés s'interrogeant sur la création du monde et sur l'existence de l'Au-delà. L'autre renvoie à la science et à l'explication rationnelle du monde qu'ils opposent à Dieu, en contrepoint.
Alceste fait donc apparaître des champs d'idées qui étaient transversaux dans les textes et peu évidents à isoler car peu explicites compte tenu de la non-linéarité de l'expression thématique dans un entretien. Cependant nous avons eu des difficultés à faire des catégorisations en fonction des individus et des variables qui les caractérisent. Les tris croisés par exemple n'ont donné aucun résultat, car finalement ce sont des personnes très peu différenciées dont le discours est très homogène. Pour vérifier cela nous avons effectué une expérience sur un corpus de 8 entretiens de femmes, quatre étant issues du corpus précédemment testé et quatre autres d'un ensemble d'entretiens de femmes ouvrières réalisés par ailleurs. Nous avons obtenu alors des résultats différenciés où les individus caractérisaient les classes. Ce test montre à notre sens la pertinence de l'analyse et illustre notre recherche de mise en évidence de champs conceptuels communs à travers des discours homogènes.

* Hyperbase produit un dictionnaire des formes graphiques interactif qui se présente selon un ordre alphabétique mais qu'il est possible de transformer en une liste des fréquences décroissantes. Ce logiciel repose sur une analyse comparative de la fréquence des formes graphiques. La première comparaison se rapporte à un corpus externe, un extrait du Trésor de la Langue Française, la seconde au corpus lui-même.
Les premières statistiques produisent un tableau de la richesse lexicale de chaque entretien (norme du T.L.F.). Celui-ci permet de situer le niveau de langue des locuteurs mais, dans notre corpus, cela s'avère peu probant car

du point de vue des données sociologiques disponibles (niveau de diplôme, métier, origine sociale, etc...), les enquêtés sont relativement proches. Et on peut se demander si ces variables suffisent pour caractériser le rapport à la langue des individus.

Hyperbase produit trois types de fichiers de vocabulaire spécifique. Le premier établit les particularités du corpus par rapport au T.L.F. ; il met en valeur les vocables religieux ou se rapportant au paranormal qui constituent en fait le thème de l'entretien. Par contre, il s'avère impossible d'interpréter les listes de vocabulaires propres à chaque texte, ou individu, parce qu'il n'y a pas de discours opposés et que tous utilisent des idées et des mots communs. Enfin, la répartition des mots spécifiques n'indique pas de tendance nette : par exemple, une recherche sur l'utilisation des pronoms personnels ne permet pas de distinguer les personnes qui privilégient le "je" ou le "moi" par rapport à "nous" ou "ils".

On peut effectuer, sous Hyperbase, des A.F.C. sur des formes graphiques par l'intermédiaire du programme *ADDAD*. Elles portent sur des listes limitées, pré-établies de différentes façons : par les fréquences, par la longueur des formes, à partir du vocabulaire spécifique, par thèmes, etc... Dans notre cas, faute de vocabulaire spécifique interprétable, on a choisi de construire manuellement des listes thématiques à partir du dictionnaire : "paranormal", "mort/métaphysique", "discours scientifisant", "biographie", "valeurs morales", "religions traditionnelles". Ici encore, l'exploitation des résultats d'analyse s'avère difficile : les tableaux de contribution et les graphes factoriels opposent des vocables appartenant aux mêmes champs sémantiques. D'autre part, il n'est pas possible d'expliquer les écarts à partir de l'édition des concordanciers et de la lecture des entretiens désignés par l'A.F.C.

L'ultime essai, à ce jour, s'inspire des travaux rapportés par A. Salem ("Analyse factorielle et lexicométrie : synthèse de quelques expériences", *in Mots*, n°4, p. 147-167, 1982) : les principales discriminations entre les textes sont repérables dans l'analyse des formes graphiques les plus fréquentes jusqu'au seuil de 40% des occurrences. Cette méthode présente l'avantage d'éliminer une grande part de subjectivité mais, dans notre cas, n'apporte pas d'éléments explicatifs nouveaux.

À notre sens, les difficultés rencontrées dans l'analyse de notre corpus à l'aide d'Hyperbase proviennent d'une partition inadaptée : la population et les discours sont trop homogènes. De nouveaux essais vont être tentés en agrégeant les différents textes selon une variable qui semble plus opératoire que la distinction individuelle : le comportement religieux des parents des locuteurs.

* Les premières utilisations du logiciel SPAD.t ont été marquées par une phase de mise au point technique, la version PC, utilisée dans le cadre de notre travail, étant limitée aux 640 Ko de mémoire vive des machines même si elles sont équipées de la mémoire étendue, ce qui s'est révélé insuffisant pour le traitement de notre corpus. D'autres problèmes sont survenus avec l'utilisation de la version Mac, notamment d'adéquation de la pile Hypercard avec le système 7.x.
Une fois ces problèmes résolus, les premiers traitements ont porté sur notre corpus considéré comme un ensemble de textes indépendants, sans ajouter de fichier numérique pour décrire les individus. Dans ce cas, on s'intéresse surtout aux comparaisons entre textes. La recherche des mots spécifiques n'a pas donné de résultats intéressants. Ce mode d'utilisation se rapproche de celui d'Hyperbase et ne permet pas de mettre en valeur toutes les possibilités d'utilisation de SPAD.t.
Notre corpus a donc été archivé une nouvelle fois : les entretiens ont été considérés comme une seule réponse et des variables sociologiques introduites. À la suite de cette phase d'archivage, sont produits les dictionnaires et la liste du vocabulaire à partir duquel on travaille. L'utilisateur peut faire varier un certain nombre de seuils – nombre de mots retenus, fréquence des mots retenus, longueur des mots, etc... – mais les conséquences de ces variations ne sont pas toujours perceptibles et il n'est pas facile de savoir comment ces paramètres agissent sur les résultats fournis par les autres procédures. Les premiers essais que nous avons réalisés ne retenaient que les formes graphiques qui avaient une fréquence égale ou supérieure à 20. Par la suite, ce seuil a été abaissé à 5, nous permettant de retenir 900 formes graphiques.
La recherche des mots caractéristiques a porté sur des sous-populations définies d'après les catégories socio-professionnelles, le niveau d'instruction et la situation matrimoniale ce qui a rendu nécessaire l'abaissement des seuils de rétention des formes graphiques pour que la recherche des mots caractéristiques porte sur un vocabulaire plus riche. Nous n'avons pas utilisé la procédure de recherche des réponses caractéristiques (RECAR) car nos entretiens nous semblent trop longs et trop peu nombreux.
Les listes de vocabulaire établies ont été soumises aux procédures d'A.F.C. Le logiciel offre au chercheur diverses possibilités de traitements, à partir de l'ensemble des individus, de sous-populations ou des variables contenues dans le fichier des données. Nous avons établi la typologie des individus par rapport aux formes graphiques, que ce soit en considérant l'ensemble de nos individus ou en sélectionnant une partie d'entre-eux seulement puis mis en évidence la disposition des variables par rapport aux formes graphiques. Les résultats fournis par ces procédures ont toujours été complétés par les résultats obtenus par les procédures de classification ascendante hiérarchique.

SPAD.t permet aussi de travailler sur les segments répétés. Dans cette étude, nous avons retenu tous les segments répétés ayant une fréquence égale ou supérieure à 5, mais les segments de longueur 2 et 3 n'ont été retenus qu'au-dessus d'une fréquence égale ou supérieure à 20. Au total, 736 segments répétés ont été repérés et inventoriés. Le fichier contenant ces segments répétés a été soumis à une A.F.C. de la même manière que les formes graphiques. Les résultats sont cependant plus intéressants car ils permettent de mieux mettre à jour des combinaisons imprévisibles dans la langue comme les contraintes sémantiques exercées par la négation appliquée aux formes verbales comme « croire », « penser», etc...
Il est à noter que cette procédure est gourmande en mémoire vive et pose problème avec des textes longs. Il faut envisager ce genre de traitements sur des machines équipées en conséquence et paramétrer le logiciel en augmentant sa réservation d'espace mémoire.
À la suite de ces diverses expériences, nous pensons que SPAD.t offre au chercheur un grand nombre de possibilités de traitements assez fins sur des corpus vastes plutôt de questions ouvertes et des individus nombreux. Cependant, le chercheur devra prévoir un temps assez long pour apprendre à bien utiliser ce logiciel, avant de pouvoir passer à l'interprétation des résultats fournis par les diverses procédures, résultats bien difficiles à élaborer dans le cas du vocabulaire très large des entretiens.

Aisance de manipulation des logiciels

La documentation fournie avec le logiciel Alceste qui mériterait d'être complémentaire de la simple reproduction sur support papier des cartes d'aides comprises dans le logiciel, devrait insister sur les conséquences des modifications des paramètres des plans d'analyse sur les résultats obtenus. Cette documentation pourrait offrir des commentaires sur les pages de résultats fournis par les analyses.
L'utilisation du logiciel Hyperbase est à la portée des chercheurs peu familiers des environnements informatiques et la documentation présente bien les fonctionnalités du programme et les résultats obtenus.
SPAD.t offre une documentation qui explique bien les différents paramétrages des procédures de ce logiciel somme toute assez ardu à mettre en œuvre pour un utilisateur moyen en informatique et en analyse des données textuelles. Cependant, au-delà de la simple présentation des procédures, ce document pourrait être plus explicite à propos des nombreux résultats fournis par le programme. Pour trouver l'explication de certaines notions d'analyse des données, il nous a fallu nous reporter à la documentation du logiciel SPAD.n. Il nous semblerait utile de baliser le terrains des utilisateurs dans la profusion des résultats fournis par SPAD.t.

Conclusion

Notre perspective de départ qui était double, d'une part explorer des logiciels d'analyse des données textuelles et d'autre part faire ressortir des champs conceptuels cohérents au travers de discours libres autour de thèmes communs, nous amène aux constats suivants :

Alceste qui exige une connaissance peu étendue de la statistique nous a permis de distinguer des univers conceptuels non immédiatement perceptibles à l'analyse de contenu, notamment à propos de l'élaboration par nos interlocuteurs de leur position par rapport à la science et plus largement à ce qui peut être prouvé. Cette position se construit de façon très diffuse dans les entretiens par quelques mots ici et là ou par quelques exemples que le logiciel a rassemblé de façon intéressante. Cependant nous avons regretté de ne pas disposer de suffisamment d'éléments pour élaborer une interprétation en fonction des variables dont nous disposions.

Hyperbase qui est aisé d'accès pour des chercheurs en sciences sociales nous a permis d'appréhender nos textes par une lecture différente de celle que nous pratiquons habituellement mais ne nous a pas permis de mettre en évidence des organisations conceptuelles particulières dans le corpus du fait, nous a-t-il semblé, de la relative homogénéité des entretiens que nous avons étudiés et probablement de la non-pertinence de la partition du corpus que nous avions choisie. Par ailleurs, nous aurions aimé disposer des procédures de listes de segments répétés et d'une possibilité de prendre en compte des variables extérieures au corpus.

SPAD.t nous a paru receler les possibilités les plus élaborées en matière d'analyse des données et en particulier pour les relations éventuelles entre les variables et les textes. Toutefois, sa mise en oeuvre, du moins avec la version dont nous disposions, nous a semblé nécessiter une connaissance plus que moyenne en analyse des données et en informatique. Il nous a demandé un lourd investissement en temps d'apprentissage du logiciel lui-même qui s'est fait au détriment de la lecture interprétative de nos données. Nous avons en particulier sorti un très grand nombre de fichiers reflétant des analyses très diverses que nous permettait la souplesse du paramétrage des procédures mais que nous avons eus du mal à interpréter. Une bonne conception statistique aurait pu éventuellement nous dispenser de ces errements. Cependant, il nous semble que nos difficultés tenaient aussi au type de textes que nous avons choisi, l'entretien, qui, par sa longueur et l'importance du nombre des formes à analyser, engorge les procédures.

Lorsque, après nos premières approches, nous envisageons nos perspectives de recherche en termes d'aide informatique à l'interprétation, nous sommes

amenés à penser l'utilisation conjointe de ces logiciels pour les analyses lexicales des uns, statistiques ou transversales des autres. Mais, en tout état de cause, l'occasion des rencontres de Montpellier nous a permis de constater la nécessité d'un long travail de préparation avant toute utilisation vraiment efficace des logiciels d'analyse des données textuelles pour obtenir des interprétations pertinentes des données.