Chapitre 11

Petite histoire des ressources logicielles au service de la sociologie qualitative

11.1. Introduction

Dans ce chapitre, l'outil informatique est envisagé en tant que ressource pour l'analyse qualitative de corpus de textes en sciences humaines et sociales. On se concentrera sur l'analyse informatisée de corpus de *textes*[1], en occultant les outils permettant d'analyser des documents non strictement textuels comme, par exemple, les objets confiés par les informateurs, les souvenirs que l'observateur retient de son contact avec le terrain ou les enregistrements (sonores, photographiques ou vidéographiques) d'interactions ou de milieux de vie. Les outils permettant d'analyser ces matériaux non exclusivement textuels (comme *Aquad*, *Anvil*, *Atlas*, *Porphyry*[2], *Transana*, *Transcriber* et *Videograph*) ne sont donc pas analysés ici. Ce travail méthodologique procède à l'ouverture de certaines boîtes noires de l'analyse textuelle. Un tel éclaircissement passe par le rappel des circonstances de l'élaboration des différentes techniques[3] et permet de constater que beaucoup d'outils reprennent, sous des noms différents, des fonctionnalités très similaires. Pour cette raison, je préfère dégager des familles de fonctionnalités plutôt que des familles d'outils (de telles typologies sont disponibles par ailleurs [JEN 96, KLE 01, POP 97, WEIT 95]). A ceux que la technique effraie, ce panorama voudrait montrer

1. Un corpus est un ensemble de documents. Pour ce chapitre, c'est un ensemble de textes.
2. Voir le chapitre 1.
3. Je me concentre sur la dimension méthodologique de ces outils, la question de l'interprétation étant traitée dans la première partie du volume 2 de ce traité.

Chapitre rédigé par Christophe LEJEUNE, à paraître dans Humanités Numériques (Hermes).

que ces outils sont proches de leur façon quotidienne de travailler sans logiciel. Aux enthousiastes, ce chapitre rappelle qu'aucune des techniques proposées ne recèle ni magie, ni boîte de Pandore ou intelligence, car il s'agit bien d'outils. Enfin, ceux de mes lecteurs qui utilisent déjà certains outils mentionnés vivront peut-être cette mise en perspective comme une épreuve de force. En effet, comme toute liste, cette mise à plat traite ensemble d'outils développés dans des univers fort différents. J'espère qu'elle aura la vertu de permettre le dialogue entre utilisateurs provenant de différentes communautés.

11.2. Quel outil pour quelle analyse ?

Comme l'anthropologie des sciences l'a brillamment montré pour d'autres disciplines, les scientifiques transforment les matériaux observés en inscriptions. De traductions en traductions, les mesures liminaires sont déplacées (mobiles) du champ phénoménal au laboratoire [LAT 95]. L'inscription (sur un support) garantit leur stabilité durant ce transfert. Etant donné que leur faculté de représenter le phénomène n'est pas altérée, ils sont dits immuables. Au mouvement convergent du champ phénoménal vers le laboratoire s'adjoint ensuite un mouvement qui s'en éloigne : le chercheur mobilise les inscriptions dans l'assemblage de dispositifs argumentatifs que sont les articles [CAL 86]. C'est la qualité (à la fois mobile et immuable) de ces inscriptions qui garantit la fidélité aux observations de départ. Par vocation, ces nouvelles inscriptions entament un déplacement centrifuge par rapport au laboratoire. En sociologie, la série de transformations en « mobile immuable » débute par le recueil (et l'enregistrement) du témoignage de l'informateur. Celui-ci est ensuite rapporté au laboratoire ; les scientifiques y opèrent d'autres transformations (traductions) qui produisent de nouvelles inscriptions. Parmi elles, la transcription des entretiens occupe prototypiquement la position liminaire de traduction, rendant possibles les transformations suivantes. Même lorsque aucun autre outil n'est mobilisé ensuite, cette opération est en effet nécessaire à la citation (sans laquelle l'assise empirique des arguments sociologiques est largement fragilisée). Une fois les entretiens transcrits, le chercheur opte souvent pour l'analyse de ces inscriptions intermédiaires.

Les paragraphes qui suivent passent en revue des dispositifs mobilisés comme adjuvants à cette tâche. Les fonctionnalités présentées seront le feutre (paragraphe 11.2.1), le traitement de texte (paragraphe 11.2.2), le système d'exploitation – et, en particulier, les expressions régulières – (paragraphe 11.2.3), le cotexte – en particulier, les concordances – (paragraphe 11.2.4), la cooccurrence (paragraphe 11.2.5), l'analyse benzécriste des données (paragraphe 11.2.6), les segments de texte (paragraphe 11.2.7) et les dictionnaires (paragraphe 11.2.8).

11.2.1. *Le feutre*

L'équipement minimal du chercheur en sociologie pour analyser son matériau qualitatif (textuel) est le feutre (ou, si l'on préfère, la plume, le crayon, le stylo). Sans que ce point de départ ne soit l'occasion de discuter le lien entre science et écriture [GOO 79, SER 93], il rappelle néanmoins que les sources empiriques manipulées par le chercheur sont dotées d'une matérialité. Face aux éléments empiriques que sont des transcriptions d'entretiens, l'analyste procède au repérage des thèmes qui lui apparaissent pertinents. Le feutre est ici la technologie minimale permettant cette annotation. Vu que je m'intéresse à l'instrumentation informatique, ce cas limite me sert de borne ouverte (ce qui signifie que ce chapitre n'étudie pas plus avant le travail au feutre pour lui-même)[4].

11.2.2. *Le traitement de texte*

Certains chercheurs proposent de tirer parti de la diffusion et de la familiarité des traitements de texte et de les mobiliser pour analyser des corpus de texte [LAP 04, MOR 91]. Cet usage spécifique d'un outil bureautique consiste à délimiter des segments de texte et à les associer à des étiquettes choisies par le chercheur[5]. Deux fonctionnalités sont principalement mobilisées à cet effet. Il s'agit tout d'abord des marques invisibles. Initialement conçues comme adjuvant à la rédaction, elles permettent à l'usager de commenter son texte (sans que cela apparaisse à l'impression) ; elles sont donc particulièrement mises à profit dans le cas de rédaction à plusieurs. Mobilisées dans le cadre d'une analyse sociologique, ces marques de révision permettent d'annoter le texte, transposant – sur l'écran – l'usage du feutre. Selon la stratégie du chercheur, ces annotations peuvent aller d'un simple relevé des dires à une interprétation. Comme le montre la figure 11.1, les tableaux peuvent également être mobilisés avec un rôle homologue : les annotations ne sont plus alors dans des marques invisibles mais dans une colonne dédiée (ce qui rappelle également le travail sur papier mais, cette fois, plutôt au niveau des commentaires apportés dans la marge)[6]. L'usage du traitement de texte pour analyser du matériel qualitatif est attesté et donc possible. Cependant, l'expérience montre que ce détournement n'offre pas l'outil le plus adéquat[7].

---

4. Je montre au paragraphe 11.2.7 que certains outils logiciels mobilisent la technologie du feutre en tant qu'usage de référence.

5. Proche du travail « au feutre », cette déterritorialisation du traitement de texte s'inscrit dans la logique d'outils très répandus chez les Anglo-saxons (que je présente au paragraphe 11.2.7).

6. L'illustration présentée ici recourt au traitement de texte libre *Open Office,* http://www.openoffice.org/.

7. Sur de gros corpus, les fichiers tabulaires deviennent en effet peu aisés à manipuler.

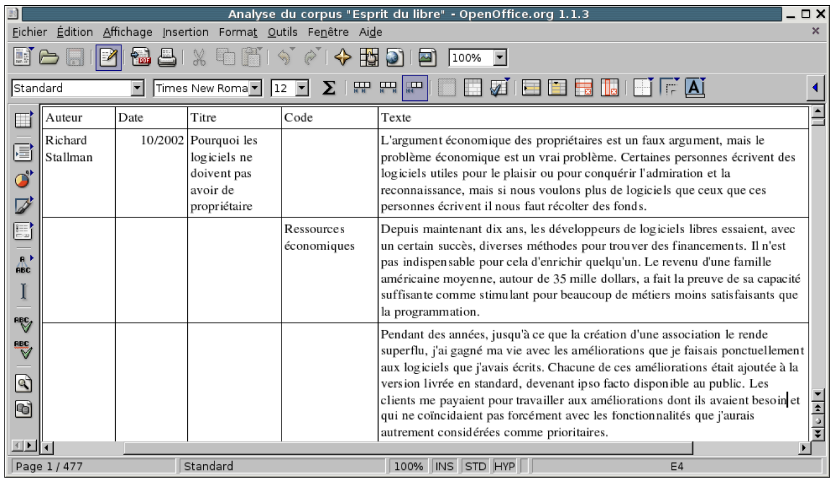

Figure 11.1. *Un traitement de texte (*Open Office*)*
(les codes du chercheur apparaissent dans la quatrième colonne)

11.2.3. *Le système d'exploitation*

Dans la pratique, les chercheurs qui recourent effectivement au traitement de texte comme outil d'analyse profitent de la disponibilité d'autres fonctionnalités des traitements de texte. L'exploration et le codage de leur matériau textuel mobilisent la localisation au sein du texte de mots (ou de groupes de mots ou de parties de mots). En tant que telles, ces opérations de recherche de chaînes de caractères relèvent plus du savoir-faire de chacun que de l'analyse scientifique de la sociologie qualitative. La recherche de segments alphanumériques ne se limite pas aux chaînes de caractères typographiques. Elle regroupe les célèbres troncatures – qui permettent, par exemple, de localiser toutes les formes conjuguées d'un même verbe (ou, de manière plus générale, toutes les flexions[8] d'un même lexème [LYO 90]) – ainsi que des opérateurs, comme le « ou » logique. La palette des motifs qui peuvent être ainsi recueillis est infinie. Une telle localisation de patrons génériques (ou recherche de motifs – *pattern matching*) fait appel à ce que les informaticiens appellent « expressions régulières » [FRI 03]. Celles-ci sont disponibles sur tous les systèmes d'exploitation depuis la naissance de la micro-informatique. Elles naquirent cependant de l'étude scientifique des neurones. Dans les années 1940, le neurophysiologiste Warren McCulloch parcourait diverses

8. Parmi les transformations morphologiques, les linguistes distinguent la flexion d'un même lexème en différentes formes (dont les déclinaisons en allemand et la conjugaison des verbes en français sont des exemples) de la dérivation d'un nouveau lexème (passer d'un nom comme « territoire » à un verbe comme « territorialiser »).

disciplines afin de comprendre comment la pensée vient aux hommes. Sa rencontre avec le talentueux logicien Walter Pitts débouche sur un article [MCC 43] qui tente de modéliser le système nerveux. Le calcul propositionnel est appliqué à ces machines (communiquant entre elles par impulsion électrique) que sont les neurones. Pour McCulloch, ce travail s'insère dans un incessant questionnement sur le fondement (matériel) de l'esprit humain et sur la volonté de donner à cette question une réponse scientifique (et non métaphysique) grâce à une psychologie expérimentale aidée de la biologie. Ces dernières disciplines oublieront ce travail[9].

Les travaux de McCulloch sont ensuite repris en mathématiques. Stephen Kleene leur adjoint le concept d'automate fini et les inclut dans son algèbre des ensembles réguliers[10]. Après plusieurs développements en mathématiques, les expressions régulières s'introduisent peu à peu dans les préoccupations des informaticiens. Le co-concepteur d'*Unix* les introduit à la fin des années 1960 à la fois dans la littérature scientifique [THO 68] et dans une série d'applications informatiques (d'abord l'éditeur *qed*, puis *ed*, qui les popularise). Dans ces éditeurs, il fallait saisir la commande suivante pour exécuter une correspondance de motifs :

```
g/Regular Expression/p
```

Cette fonction donna son nom à l'application *grep* (pour *global regular expression print*). L'histoire se prolonge par la montée en puissance des expressions régulières [FRI 03]. À la fin des années 1980, le langage *Perl*[11] joue un rôle déterminant dans leur diffusion. Or, le concepteur de ce langage de programmation, Larry Wall, est linguiste [SCH 02]. Les expressions régulières peuvent donc s'enorgueillir de leurs origines interdisciplinaires. Aujourd'hui, elles font partie de l'arsenal de l'analyste de corpus de textes. Ce paragraphe suggère donc que le chercheur peut donc se contenter du système d'exploitation pour analyser ses données, sans recourir à d'autres outils plus spécifiques que les fonctions de recherche[12].

---

9. L'aura qui règne autour de l'article de 1943 – considéré comme précurseur des réseaux de neurones et des sciences cognitives [AND 92, DUP 99] – invite à quelques réserves : avec une telle force d'attraction, il peut être cité sans être nécessairement la référence la plus pertinente pour la fondation de la recherche par motifs. Son histoire montre néanmoins comment diverses préoccupations (scientifiques, philosophiques et épistémologiques) ont pu converger dans l'élaboration d'un outil générique.

10. En français, « regular » est parfois traduit par « rationnel », ce qui conduit certains chercheurs français [SIL 00] à parler d'expressions rationnelles [FRI 03].

11. http://www.perl.org/.

12. Jocelyne Le Ber déterritorialise ainsi les commandes *grep*, *freq* et *diff* du système d'exploitation afin d'analyser des œuvres littéraires, en particulier l'*Antigone* de Jean Cocteau [LEBE 06].

11.2.4. *Le cotexte*

En quittant l'utilisation des outils bureautiques conventionnels, le présent panorama livre une classe d'outils plus spécifiques. Proches des précédentes fonctionnalités de recherche, ces outils servent à identifier des termes, des expressions ou des motifs. Ils ajoutent cependant la localisation des éléments recherchés dans les textes explorés. Le regroupement de la phrase ou du paragraphe dans lequel apparaît la cible permet au chercheur de rapporter chaque occurrence à un contexte d'apparition (que les linguistes appellent environnement topologique immédiat ou cotexte [KER 02, MAI 98, WIL 01]).

Ce mode de visualisation des segments de texte qui précèdent et suivent chacune des occurrences est appelé « concordances » [LEBA 94] ou « index de mots-clés en contexte » [WEIT 95]. La présentation des concordances[13] procède typiquement de l'alignement sur la cible (ou forme-pôle), de sorte que les différents cotextes peuvent être aisément triés, rapprochés et comparés [PIN 06] (voir figure 11.2).

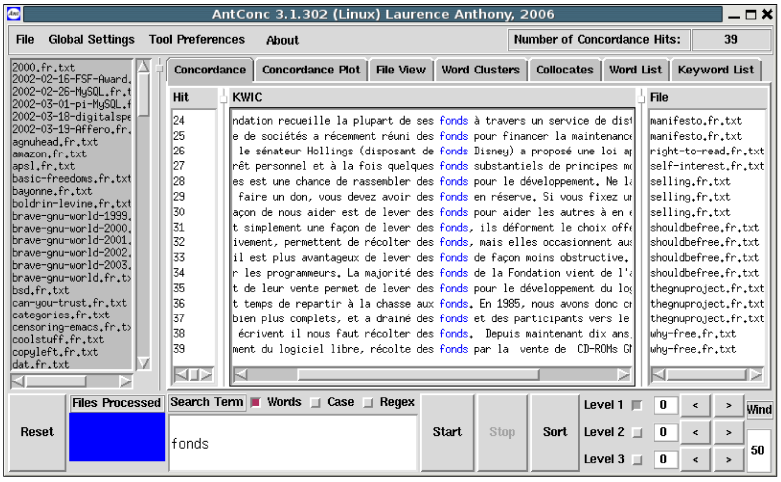

Figure 11.2. *Un concordancier (*AntConc*)*

L'histoire des index remonte au début de notre ère. Les premiers index organisés selon l'ordre alphabétique dateraient du IV^e^ siècle de notre ère. Les Grecs auraient dressé des index (non alphabétiques) de noms géographiques au VI^e^ ou VII^e^ siècle. Mais l'invention des concordances et des index de mots-clés en contexte s'insère

13. *AntConc* de Laurence Anthony (http://www.antlab.sci.waseda.ac.jp/) ; *Glossanet* de Cédrick Fairon (http://glossa.fltr.ucl.ac.be/) [FAI 06].

dans la tradition des exégètes de textes sacrés. Une division des textes bibliques en chapitres (appelée « capitulation ») est déjà attestée au IVe siècle [MEY 89]. Le texte du *Nouveau Testament* carolingien comporte également, en marge, une capitulation ainsi que des renvois aux passages analogues dans les autres évangiles. La multiplication des livres et la réduction de la taille des volumes laissent moins de place à ces informations. Parallèlement, les exégètes produisent des outils assurant cette intertextualité. Au Xe siècle, les Massorètes – les « maîtres de la tradition » juifs – créent des listes alphabétiques de mots accompagnés de leur cotexte immédiat[14]. Ils préfigurent (voire inventent) alors les mots-clés en contexte [WEIN 04]. A l'aube du XIIIe siècle, Etienne Langton introduit le chapitrage dans la bible latine[15]. Entre 1238 et 1240, les moines dominicains du couvent Saint Jacques à Paris composent, sous la houlette de Hugues-de-Saint-Cher, une liste alphabétique de concordances mobilisant ces nouvelles références. Celles-ci comprennent le numéro du chapitre et d'une lettre (de A à G) indiquant la position du mot dans le chapitre. L'environnement cotextuel n'étant pas disponible dans cette édition, le dispositif en question préfigure pour sa part les index de mots-clés *hors* contexte. Dans les éditions suivantes, la proposition encadrante est ajoutée comme troisième élément. Ce sont précisément les débats entre Chrétiens et Juifs qui favorisent le développement, par Rabbi Nathan Mardochee, d'une concordance hébraïque, imprimée pour la première fois en 1523 [SEK 95].

Dans les années trente du XXe siècle, l'étude scientifique du cotexte apparaît aux Etats-Unis. Le contexte n'est plus religieux, mais politique : les commanditaires de ces études sont aux prises avec des langues amérindiennes aussi nombreuses que peu connues. D'inspiration naturaliste, la linguistique structurale se développe alors. Elle distingue l'environnement de chaque élément d'un ensemble d'énoncés. L'ensemble des environnements d'un élément est appelé distribution [HAR 64]. Bien que distincte des concordances, l'analyse distributionnelle marque l'entrée du cotexte parmi les outils scientifiques.

Revenons aux concordances. Les premières générations automatisées de concordances sont proposées à la fin des années cinquante[16]. En 1959, Hans Peter Luhn propose une présentation des concordances centrées sur la forme recherchée, qu'il qualifie selon l'acronyme *KeyWords In Context* ou KWIC [LUH 60, LUH 66]. Quelques années plus tard (en 1963) apparaissent les premiers index de mots-clés réalisés selon ce principe (les mots provenaient des titres de dix ans de parution de l'*Association of Computer Machinery* [YOU 63]). Après des siècles d'existence, la

14. Ces listes n'étaient pas indexées, vu l'absence de numérotation des chapitres ou des versets.

15. Les versets, plus tardifs, sont insérés par Robert Estienne en 1551.

16. En 1951, Roberto Busa, auteur de la postface de ce traité, réalise, à l'aide d'outils mécanographiques, une concordance de quatre poèmes de Thomas d'Aquin.

construction des concordances est désormais automatisée. Le concordancier est l'un des outils dont la tradition remonte au plus loin. Son caractère interdisciplinaire n'a rien à envier aux expressions régulières. Il connaît même actuellement une extension hors de l'étude de matériau textuel, avec des applications spécifiques en génétique et en chimie. Les concordances servent d'appuis à de nombreuses inférences d'exploration du corpus. Leurs propriétés graphiques les rendent particulièrement adéquates pour découvrir la récurrence de locutions ou d'expressions (constituées de plusieurs mots simples).

11.2.5. *La cooccurrence*

La cooccurrence est une modalisation de la classe d'outils reposant sur le cotexte. Différents algorithmes existent en ce domaine. Le principe repose cependant toujours sur la logique suivante. Il s'agit de repérer la « proximité »[17] de deux termes. Cette proximité n'est ni sémantique, ni syntaxique, ni pragmatique. Comme les outils cotextuels, elle renvoie à une dimension topologique. Deux termes sont d'autant plus proches qu'ils sont séparés par peu de caractères. Cette proximité se mesure donc le long de l'axe syntagmatique[18]. Pour passer de cette mesure à la cooccurrence, on agrège les proximités de chaque apparition (ou occurrence) des deux termes. Contrairement à l'étude des environnements qui, grâce à la notion opératoire de distribution, trouve une justification dans la linguistique structuraliste américaine, la proximité sur laquelle se greffe la cooccurrence fait l'objet de relativement peu d'études linguistiques au sens strict. Les travaux attestés sont l'œuvre de Maurice Tournier. Sa conception de la proximité s'inspire des découvertes psychologiques liées au conditionnement pavlovien et accompagne épistémologiquement l'outil *Lexico* d'André Salem.

La cooccurrence est donc une fonctionnalité très intimement liée aux outils d'analyse textuelle. La variabilité des algorithmes dépend de la mesure de la proximité. Certains prennent effectivement en compte le nombre de caractères, d'autres reposent sur un comptage des mots séparant les deux termes ; les uns considèrent que la proximité peut être mesurée à travers l'ensemble du texte, les autres limitent la pertinence à un chapitre, à un paragraphe, à une phrase, voire à une proposition (ces différentes unités de contexte étant la plupart du temps définies selon des critères typographiques). A la différence de la concordance, la cooccurrence ne tient le plus souvent pas compte de l'ordre des mots ; on parle de paires de cooccurrents ou du réseau des mots associés à un pôle. Certains logiciels

17. On parle également de l'attirance [HEI 98] ou de l'attraction mutuelle de deux termes.
18. Les linguistes [SAU 95] distinguent l'axe (syntagmatique) horizontal de succession des mots dans une phrase de l'axe (paradigmatique) vertical d'association des mots qui « vont ensemble » (par exemple, les mots qui ont la même sonorité, signification ou distribution).

(comme *Tropes*[19] [MAR 98] et *Weblex*[20] [HEI 04]) proposent néanmoins des fonctionnalités qui distinguent l'association selon l'ordre d'apparition des mots ; on parle alors de couples (orientés) de cooccurrents. Contrairement à la concordance, la cooccurrence ne connaît pas de présentation typique de ces résultats. Elle peut être présentée sous forme de listes ou de graphes. Ces derniers sont souvent non orientés [OSG 59, TEI 91]. Le recours aux graphes orientés est néanmoins attesté lorsque l'ordre des mots est pris en compte. Tout comme les concordances, les cooccurrences servent à l'élaboration d'inférences. Dans une logique proche de la catégorisation automatique, certains outils utilisent cette mesure pour construire des agrégats d'éléments fortement liés entre eux. Ces agrégats (ou *clusters*) offrent une vue synthétique sur le corpus et font l'objet d'une interprétation par l'analyste. En sociologie, de telles cartes de thèmes sont mobilisées par l'anthropologie des sciences, notamment au sein du logiciel *Candide* [COU 95, MAL 95, NOY 95, TEI 91, TEI 95] qui repose lui-même sur les développements de *Leximappe* [CAL 91, LAW 88, VAN 92][21].

11.2.6. *L'analyse de données*

Les travaux du mathématicien Jean-Pierre Benzecri ont donné naissance à une série de techniques comme l'analyse en composante principale et l'analyse factorielle. Pierre Bourdieu en a produit l'utilisation qui a le plus marqué la sociologie française. Chez l'auteur de *La distinction*, le plan factoriel est devenu un dispositif de représentation de l'espace social comme champ de forces [BOU 79]. Deux axes (orthogonaux) traversent cet espace et tracent quatre cadrans ouvrant la combinatoire désormais célèbre de répartition des capitaux culturels et économiques. Outre le coup de force (procédant du rapprochement du tableau et de la théorie des classes sociales), la rencontre de l'auteur et de l'analyse des données introduit en sociologie une inscription graphique de ce que donne un champ de positions relatives les unes par rapport aux autres. Cette conception relationnelle, chère à Bourdieu, se matérialise dans un espace à deux dimensions sur lequel se retrouvent aussi bien des individus que des variables, aussi bien des classes sociales que des pratiques [BOU 94]. Les analyses de Pierre Bourdieu portent sur des données chiffrées et, de façon générale, l'analyse des données appartient à l'arsenal des méthodes quantitatives. Cet outil fut toutefois initialement développé pour servir la linguistique. Il n'est donc pas étonnant qu'il ait inspiré le développement de logiciels communément assimilés aux outils qualitatifs. C'est pour cette raison, et par respect du champ sociologique associé, que ces outils sont évoqués ici, même s'il s'agit de l'interface avec les statistiques.

19. http://www.acetic.fr/.
20. http://weblex.ens-lsh.fr/.
21. Ces outils s'appuient sur des recherches en informatique [CHA 88, MIC 88].

En France, un des outils sociologiques[22] qui a su tirer au mieux parti des enseignements de Benzecri est *Alceste*, conçu par Max Reinert[23]. Bien plus proche des automates que du feutre, le cœur d'*Alceste* repose sur la classification descendante hiérarchique des formes lemmatisées[24] des mots pleins[25] du corpus analysé. Celle-ci débouche sur une série de classes construites de manière formelle. Sont ensuite mobilisés des modules d'analyse ascendante (pour mettre en évidence les mots les plus typiques pour chaque classe) et d'analyse factorielle des correspondances (essentiellement pour ses vertus graphiques). Outre les plans factoriels, les résultats sont exprimés sous la forme typique de graphes orientés appelés dendogrammes. Cette représentation arborescente fournit au chercheur une illustration d'agrégats imbriqués [KRI 04]. Comme dans le cas des cooccurrences, ceux-ci peuvent servir de base aux inférences de l'analyste. Le concepteur d'*Alceste*, Max Reinert, défend que l'intérêt de ces classes est essentiellement exploratoire et heuristique. Ce faisant, il insiste sur la nécessaire complémentarité d'une connaissance intime du corpus et des outils informatiques (qui aident à formuler les hypothèses plus qu'à instrumenter l'administration de la preuve).

11.2.7. *Les segments de texte*

La cooccurrence et l'analyse des données nous ont donné à voir des modes d'étiquetage et de segmentation automatique des corpus de textes. L'annotation manuelle de segments de texte emprunte une stratégie tout à fait différente. Très répandus chez les chercheurs anglo-saxons, les outils fondés sur le codage des segments de texte portent le nom de logiciels d'analyse qualitative (*Computer-Assisted Qualitative Data Analysis Software* ou CAQDAS)[26]. Se réclamant souvent de la théorie de l'émergence (*grounded theory*), les CAQDAS prônent que le chercheur s'imprègne du corpus à analyser[27]. Les textes appartenant au corpus

22. C'est dans les sciences du langage, plus qu'en sociologie, que se trouvent les applications les plus orthodoxes des analyses factorielles des correspondances, comme les outils *Data and Text Mining* développés à la suite de *SPAD-T* par Ludovic Lebart (http://ses.enst.fr/lebart/).
23. *Alceste* est distribué par la société *Image* (http://www.image.cict.fr/).
24. La lemmatisation consiste à ramener à une racine commune différentes formes fléchies ou dérivées. Cette opération répond aux phénomènes morphologiques évoqués précédemment.
25. Les approches automatiques écartent parfois une série de mots de leur analyse. Cette liste comprend les articles, les prépositions, les connecteurs, les pronoms et les conjonctions ou, plus simplement, les mots les plus fréquents du corpus. Ces mots sont qualifiés de « vides » ; par opposition, les mots sur lesquels portent les traitements sont dits « pleins ».
26. Le recours au traitement de texte, présenté au paragraphe 11.2.2, en constitue un cas particulier.
27. *WeftQDA* d'Alex Fenton (http://www.pressure.to/qda/) ; *TamsAnalyser* de Matthew Weinstein (http://tamsys.sourceforge.net/).

sont lus par le chercheur. Celui-ci thématise les passages qu'il désire : cette annotation se fait par sélection (de nos jours, le plus souvent avec la souris) et association à une étiquette choisie par l'analyste (figure 11.3). Cette façon de procéder est très proche, tant dans son ergonomie que dans sa philosophie, de l'usage du feutre. Au niveau du geste, la sélection rappelle le surligneur et l'étiquetage, les annotations dans la marge. Nombre de logiciels font apparaître un espace à droite ou à gauche du texte qui reproduit la disposition de la feuille de papier[28]. Au niveau de la stratégie d'analyse, ces outils réservent le codage au chercheur. Cette façon de procéder est sans conteste plus fine que les automatismes décrits dans les paragraphes précédents. Cependant, elle requiert un travail exigeant, voire laborieux lorsqu'il s'agit de revenir sur l'ensemble du corpus suite à une modification du cadre d'analyse. Ses limites sont, à en croire le témoignage de ses praticiens, une tendance à l'obsession du codage au détriment de l'analyse, de l'interprétation et de la théorisation. Certains chercheurs envisagent cependant cette limitation au relevé thématique comme un gage de scientificité ; ils se rapprochent sur ce point de l'ethnométhodologie radicale qui proscrit toute interprétation [GAR 02].

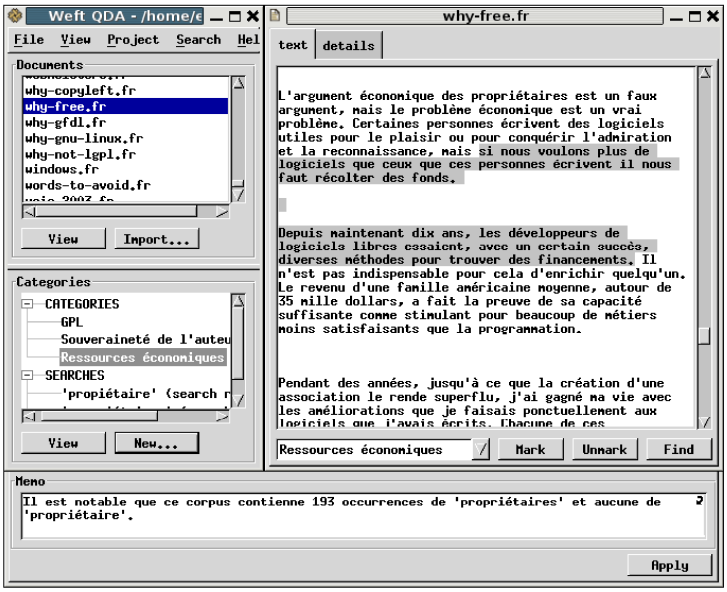

Figure 11.3. *Segments de texte surlignés dans le texte (* WeftQDA*)*

28. *NVivo* de Lyn et Tom Richards (http://www.qsrinternational.com/).

Si le parcours du matériau est similaire à la lecture sur papier, ces logiciels offrent (évidemment) des avantages liés à la numérisation du corpus. Il devient ainsi aisé de localiser un passage, tant grâce à son contenu qu'à l'étiquette qui lui a été attribuée[29]. De la même manière, les limitations inhérentes au papier ne s'appliquent plus ; le chercheur peut ainsi annoter à loisir un passage qui l'inspirerait particulièrement, même si cette note doit s'avérer longue. Comme je l'ai mentionné, ces outils se réfèrent à l'analyse « à la main » et s'inscrivent, ce faisant, dans les techniques d'analyse de contenu. Les spécialistes identifient comme un événement précurseur de ces techniques la controverse qui accompagna dans la Suède des années 1640 la publication de quatre-vingt-dix cantiques [KRI 04]. Bien qu'ayant l'aval du censeur suédois, les « chansons de Sion » inquiétèrent l'organisation luthérienne. Le contenu du recueil fit l'objet d'une controverse entre défenseurs et critiques : pour chacun des thèmes, on comptabilisa la fréquence et le traitement à la fois dans les cantiques incriminés et dans les sources classiques. Opérée de manière contradictoire (par les différents protagonistes de la polémique), cette confrontation est considérée comme une préfiguration de l'analyse de contenu. Les techniques d'analyse de contenu proprement dites sont apparues aux Etats-Unis à la fin du XIX^e^ siècle dans l'étude quantitative, diachronique et comparative des moyens de communication, en particulier de la presse. Une des premières enquêtes de ce type fut menée sur un corpus couvrant les éditions du *New York Times* sur plus de dix ans. L'ampleur des différents sujets traités était alors mesurée en pouces ou en centimètres (longueur des articles) [BER 52]. L'auteur de cette étude déplorait la tendance des journaux à donner une importance exagérée aux faits divers et aux articles à sensation au détriment des articles de fond sur la politique, la littérature et la religion.

Les sciences politiques s'emparèrent ensuite de ces techniques, notamment pour étudier la propagande diffusée lors des deux guerres mondiales de la première moitié du XX^e^ siècle. Ce faisant, l'outil fut acéré afin de rencontrer les critères d'une discipline scientifique. C'est de cette époque que datent les florissantes études des discours politiques. C'est également à cette période qu'apparaissent les premiers automates destinés à systématiser les opérations de codage. Dès les débuts de l'informatique, des chercheurs en sciences humaines ont développé des outils de ce type [STO 66]. Bien entendu, ils furent aidés dans cette entreprise par des ingénieurs pionniers en informatique (ceux-là mêmes dont j'ai montré qu'ils avaient forgé notamment les expressions régulières)[30]. Les CAQDAS ne sont donc pas tout jeunes ! En outre, dès les années 1960, la plupart des questions soulevées par l'analyse qualitative avaient été formulées, allant des critères de sélection des

29. De telles fonctionnalités font appel aux outils présentés au paragraphe 11.2.3.

30. Développé au début des années 1960, *General Inquirer* est toujours en activité aujourd'hui (http://www.wjh.harvard.edu/~inquirer/). Son concepteur, Philip Stone, est décédé le 31 janvier 2006.

passages pertinents et du partage de ceux-ci entre analystes jusqu'à la nécessité d'interpréter les codages et la validité de ces interprétations, en passant par la congruence de l'analyse avec le contenu des textes.

Comme je l'ai exposé dans ce paragraphe, la vertu centrale des segments de texte est l'adéquation entre le codage et le sens que l'interprète est en mesure d'extraire d'une connaissance intime du corpus. L'investissement à consentir en retour est une lecture exhaustive, approfondie, voire (souvent) répétée. Certains outils proposent un pari sur l'économie de ce travail conséquent. C'est le cas des outils à base de cooccurrence ou de statistique benzécriste. Ici, ce n'est pas du temps qui est sacrifié, mais une certaine finesse, puisque les subtilités de l'expression des acteurs et le contexte de chaque énonciation se trouvent la plupart du temps noyés par de tels automatismes. Il existe une voie moyenne entre, d'une part, la lecture (et l'annotation) de l'ensemble du corpus par le chercheur et, d'autre part, le report de cette tâche déterminante sur un automate. Cette solution fait appel à des dictionnaires.

11.2.8. *Les dictionnaires*

Les dictionnaires permettent d'automatiser le codage tout en ne sacrifiant pas nécessairement la finesse (et l'indexicalité) du sens commun étudié. Le recours aux dictionnaires repose sur le pari qu'il existe une signification stable sur laquelle on peut s'appuyer. Ces lexiques peuvent regrouper des mots ou des locutions selon leur catégorie grammaticale, leur (quasi) synonymie ou leur pertinence à l'égard de la théorie de l'analyste. De manière prévisible, les outils qui recourent aux catégories grammaticales sont principalement développés au sein des sciences du langage[31]. Ceux qui regroupent des synonymes, des registres argumentatifs ou des logiques d'actions se rencontrent plutôt dans les sciences de la culture (comme la sociologie, l'histoire ou l'anthropologie). Ces catégories s'articulent alors dans un cadre d'analyse. La question de leur adéquation au corpus, prégnante, est bien entendu variable selon que ces répertoires sont fournis tels quels dans l'outil ou qu'ils sont construits en regard de la réalité idiomatique de la recherche en cours. Il arrive que ces listes opèrent une sélection des unités sur lesquelles porte l'analyse (c'est le cas des listes de « mots vides » ou lorsqu'une seule catégorie grammaticale – le plus souvent, celle des substantifs – est prise en considération).

31. Par exemple en ingénierie linguistique, *Unitex* de Sébastien Paumier (http://www-igm.univ-mlv.fr/~unitex/) ou *Nooj* de Max Silberstein (http://www.nooj4nlp.net/) ; en linguistique de corpus, *Xaira* de Lou Burnard (http://www.xaira.org) et, en analyse statistique des données textuelles, *Hyperbase* version catégorisée d'Etienne Brunet (http://ancilla.unice.fr/~brunet/pub/hyperbase.html) ou *Weblex* de Serge Heiden.

11.3. Conclusion : tirer parti des logiciels

Lors d'une analyse sociologique, les fonctionnalités passées en revue dans ce chapitre sont rarement mobilisées indépendamment l'une de l'autre. C'est le plus souvent en les combinant que les logiciels permettent de seconder effectivement le chercheur. Par exemple, les dictionnaires sont mobilisés en tant que classes d'expressions régulières (cas des formules de *Prospéro*[32] de Francis Chateauraynaud [CHA 03] ou de *Nooj* de Max Silberstein). Concordances et cooccurrences sont envisagées comme accès complémentaires au cotexte (dans les lexicogrammes de *Weblex*). Des analyses benzécristes mobilisent des dictionnaires [BOL 84], des segments de texte (les CAQDAS permettent souvent d'exporter des résultats intermédiaires vers des logiciels statistiques) ou sont combinées avec des modules de cooccurrences (comme dans *T-Lab*[33] de Franco Lancia ou *Hyperbase* d'Etienne Brunet). Les inférences procèdent donc par croisement, rapprochement et recoupement. Au sortir de cet éventail de possibles, je souhaite avoir sinon répondu aux questions de lecteur, du moins avoir dissipé l'ombre planant autour des mystérieux logiciels d'analyse sociologique[34]. Quelle que soit la stratégie pour laquelle on opte, j'espère avoir montré qu'il n'est pas de technique qui garantisse la scientificité ou l'originalité d'une recherche. Qu'elle soit ou non informatisée, la méthode nécessite de la rigueur, et, dans tous les cas de figure, la qualité de l'interprétation revient toujours au scientifique[35]. En sociologie qualitative, l'atout des logiciels d'analyse réside en définitive dans la facilitation offerte pour échanger et discuter entre chercheurs[36]. Aussi, plutôt que de brandir l'outil, tel un bouclier face à la critique, il importe d'ouvrir les boîtes noires et de partager les expériences[37].

---

32. http://www.prosperologie.org/.

33. http://www.tlab.it/.

34. Ce chapitre est une invitation à s'emparer des outils disponibles : non seulement des logiciels domestiques (présentés aux paragraphes 11.2.2 et 11.2.3), mais également des logiciels libres, dont la logique (de diffusion et de modification) est congruente avec l'esprit scientifique.

35. Même lorsque l'on recourt aux ordinateurs, la construction d'une interprétation qui dépasse l'analyse thématique reste un réel défi. Celui-ci ne peut être relevé qu'en évitant les excès opposés de la théorisation, qui oublie (ou écrase) le terrain, et de l'empirisme naïf, qui néglige le travail interprétatif.

36. La première partie du volume 2 de ce traité poursuit cette discussion. Voir aussi [DEM 06].

37. Je remercie Gaëlle Lortal, Goritsa Ninova, Bénédicte Pincemin et Bernard Reber pour leurs relectures attentives, ainsi qu'Aurélien Benel, Sacha Mandelcwajg et Michel Marcoccia pour leurs conseils avisés. Je dédie ce chapitre à la mémoire de mon professeur de méthodologie et directeur de thèse à l'Institut des sciences humaines et sociales de l'Université de Liège, René Doutrelepont, décédé le 1er avril 2005.

11.4. Bibliographie

[AND 92] ANDLER D. (DIR.), *Introduction aux sciences cognitives*, Gallimard, Paris, 1992.

[BER 52] BERELSON B., *Content Analysis in Communication Research*, The Free Press, Glencoe, 1952.

[BOL 84] BOLTANSKI L., DARRE Y., SCHILTZ M.-A., « La dénonciation », *Actes de la Recherche en sciences sociales*, vol. 51, p. 3-40, 1984.

[BOU 79] BOURDIEU P., *La distinction. Critique sociale du jugement*, Editions de Minuit, Paris, 1979.

[BOU 94] BOURDIEU P., *Raisons pratiques. Sur la théorie de l'action*, Seuil, Paris, 1994.

[CAL 86] CALLON M., « Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc », *L'Année sociologique*, vol. 36, p. 169–208, 1986.

[CAL 91] CALLON M., COURTIAL J.-P., TURNER W.A., BAUIN S., « From translations to problematic networks : An introduction to co-word analysis », *Information sur les sciences sociales*, vol. 22, n° 2, p. 191-235, 1991.

[CHA 88] CHARTRON G., Analyse des corpus de données textuelles, sondage de flux d'informations, Thèse de doctorat, Université Paris VII, Paris, 1988.

[CHA 03] CHATEAURAYNAUD F., *Prospéro : Une technologie littéraire pour les sciences humaines*, CNRS, Paris, 2003.

[COC 95] COCAUD M. (DIR.), *Histoire et Informatique. Base de données, recherche documentaire multimédia*, PUR, Rennes, 1995.

[COU 95] COURTIAL J.-P., KERNEUR L., « Contribution de l'analyse des mots associés au suivi du développement d'un champ scientifique », dans M. Cocaud (dir.), *Histoire et Informatique. Base de données, recherche documentaire multimédia*, PUR, Rennes, 1995.

[DEM 06] DEMAZIERE D., BROSSAUD C., TRABAL P., VAN METER K.M., *Les logiciels d'analyse textuelle en actions – Usages, résultats, productions dans une perspective sociologique comparative*, PUR, Rennes, 2006.

[DUP 99] DUPUY J.-P., *Aux origines des sciences cognitives*, La Découverte, Paris, 1999.

[FAI 06] FAIRON C., SINGLER J., « I'm like, "hey, it works!" : Using Glossanet to find attestations of the quotative (be) like in English- language newspapers », dans A. Renouf et A. Kehoe (dir.), *The Changing Face of Corpus Linguistics*, Rodopi, Amsterdam/New York, p. 325-336, 2006.

[FRI 03] FRIEDL J., *Maîtrise des expressions régulières*, O'Reilly, Paris, 2003.

[GAR 02] GARFINKEL H., *Ethnomethodology's Program : Working Out Durkheim's Aphorism*, Rowman and Littlefield, Boston, 2002.

[GOO 79] GOODY J., *La raison graphique*, Editions de Minuit, Paris, 1979.

[HAR 64] HARRIS Z.S., « Distributional structure », dans J.A. Fodor et J.J. Katz (dir.), *The Structure of Language. Readings in the Philosophy of Language*, Prentice-Hall, New Jersey, p. 33-49, 1964.

[HEI 98] HEIDEN S., LAFON P., « Cooccurrences. La CFDT de 1973 à 1992 », *Des mots en liberté, Mélanges Maurice Tournier*, vol. 1, p. 65-83, 1998.

[HEI 04] HEIDEN S., « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex », dans G. Purnelle, C. Fairon et A. Dister (dir.), *Le pouvoir des mots. Actes des 7e Journées internationales d'Analyse statistique des Données Textuelles*, PUL, Louvain, p. 577-588, 2004.

[JEN 96] JENNY J., « Analyse de contenu et de discours dans la recherche sociologique française : pratiques micro-informatiques actuelles et potentielles », *Current Sociology*, vol. 44, n° 3, p. 279-290, 1996.

[KER 02] KERBRAT-ORECCHIONI C., « Contexte », dans P. Charaudeau et D. Maingueneau (dir.), *Dictionnaire d'analyse du discours*, Seuil, Paris, p. 134-136, 2002.

[KLE 01] KLEIN H, « Overview of Text Analysis Software », *Bulletin de Méthodologie Sociologique*, vol. 70, p. 53-66, 2001.

[KRI 04] KRIPPENDORFF K., *Content Analysis. An introduction to Its Methodology*, Sage, Thousand Oaks, 2004.

[LAP 04] LA PELLE N., « Simplifying Qualitative Data Analysis Using General Purpose Software Tools », *Field Methods*, vol. 16, n° 1, p. 85-108, 2004.

[LAT 95] LATOUR B., *La Science en action*, Gallimard, Paris, 1995.

[LAW 88] LAW J., BAUIN S., COURTIAL J.-P., WHITTAKER J., « Policy and the mapping of scientific change : a co-word analysis of research into environmental acidification », *Scientometrics*, vol. 14, p. 251-264, 1988.

[LEBA 94] LEBART L., SALEM A., *Statistique textuelle*, Dunod, Paris, 1994.

[LEBE 06] LE BER J., « L'adaptation comme contraction. Une analyse informatique d'Antigone », C. Duteil-Mougel et B. Foulquié (dir.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes électroniques du colloque international d'Albi « Langages et Signification »*, p. 257-268, 2006.

[LUH 60] LUHN H.P., « Keyword-In-Context Index for Technical Literature », *American Documentation*, vol. 11, n° 4, p. 288-295, 1960.

[LUH 66] LUHN H.P., « Keyword-in-Context Index for Technical Literature (KWIC Index) », dans D.G. Hays (dir.), *Readings in Automatic Language Processing*, Elsevier, New York, p. 159-167, 1966.

[LYO 90] LYONS J., *Sémantique linguistique*, Larousse, Paris, 1990.

[MAI 98] MAINGUENEAU D., *Analyser les textes de communication*, Dunod, Paris, 1998.

[MAL 95] MALINGRE M.-L., « Une application de CANDIDE, logiciel d'analyse textuelle pour une histoire de la traduction littéraire », dans M. Cocaud (dir.), *Histoire et Informatique. Base de données, recherche documentaire multimédia*, PUR, Rennes, 1995.

[MAR 98] MARCHAND P., *L'analyse du discours assistée par ordinateur*, Armand Colin, Paris, 1998.

[MCC 43] MCCULLOCH W., PITTS W., « A logical calculus of the ideas immanent in nervous activity », *Bulletin of Math. Biophysics*, 5, 1943.

[MEY 89] MEYNET R., *L'Analyse rhétorique. Une nouvelle méthode pour comprendre la Bible. Textes fondateurs et exposé systématique*, Editions du Cerf, Paris, 1989.

[MIC 88] MICHELET B., L'analyse des Associations, Thèse de doctorat, Université Paris VII, Paris, 1988.

[MOR 91] MORSE J., « Analysing unstructured interactive interviews using the Macintosh computer », *Qualitative Health Research*, vol. 1, n° 1, p. 117-122, 1991.

[NOY 95] NOYER J.-M., « Utilisation d'un outil infométrique, "candide" dans le contexte d'une réflexion stratégique. Les réseaux de simulation distribuée de l'armée américaine : émergence et description de l'émergence », dans J.-M. Noyer (dir.), *Les sciences de l'information. Bibliométrie, scientométrie, infométrie*, PUR, Rennes, 1995.

[OSG 59] OSGOOD C., « The representational model and relevant research methods », dans I. De Sola Pool (dir.), *Trends in Content Analysis*, University of Illinois Press, Urbana, p. 33-88, 1959.

[PIN 06] PINCEMIN B., ISSAC F., CHANOVE M., MATHIEU-COLAS M., « Concordanciers : Thème et variations », dans J.-M. Viprey, A. Lelu, C. Condé et M. Silberztein (dir.), *Actes des 8 e Journées internationales d'Analyse statistique des Données Textuelles*, Presses Universitaires de Franche-Comté, Besançon, vol. 2, p. 773-784, 2006.

[POP 97] POPPING R., « Computer Programs for the Analysis of Texts and Transcripts », dans C.W. Roberts (dir.), *Text Analysis for the Social Sciences. Methods for Drawing Statistical Inferences From Texts and Transcripts*, Lawrence Erlbaum, New Jersey, p. 209-221, 1997.

[SAU 95] DE SAUSSURE F., *Cours de linguistique générale*, Payot, Paris, 1995.

[SCH 02] SCHWARTZ R., PHOENIX T., *Introduction à Perl*, O'Reilly, Paris, 2002.

[SEK 95] SEKHRAOUI M., Concordances : Histoire, Méthodes et Pratique. Thèse de doctorat, Université de la Sorbonne nouvelle Paris 3 et École normale supérieure de Fontenay St-Cloud, Paris, 1995.

[SER 93] SERRES M., *Les origines de la géométrie*, Flammarion, Paris, 1993.

[SIL 00] SILBERZTEIN M., *INTEX, Manuel de l'utilisateur*, http://intex.univ-fcomte.fr/, 2000.

[STO 66] STONE P.J., DUNPHY D.C., SMITH M.S., OGILVIE D.M., *The General Inquirer : A Computer Approach to Content Analysis*, MIT Press, Cambridge, Etats-Unis, 1966.

[TEI 91] TEIL G., « Candide™ : un outil de veille technologique basé sur l'analyse des réseaux », dans D. Vinck (dir.), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*, Armand Colin, Paris, 1991.

[TEI 95] TEIL G., LATOUR B, « The hume machine. can association networks do more than formal rules ? », *Stanford Humanities Review*, vol. 4, n° 2, 1995.

[THO 68] THOMPSON K., « Programming techniques : Regular expression search algorithm », *Communications of the Association of Computer Machinery*, vol. 11, n° 6, p. 419-422, ACM Press, New York, 1968.

[VAN 92] VAN METER K.M., TURNER W.A., « A Cognitive Map of Sociological AIDS Research », *Current Sociology*, vol. 40, n° 3, p. 123-134, 1992.

[WEIN 04] WEINBERG B.H., « Predecessors of Scientific Indexing Structures in the Domain of Religion », *Second Conference on the History and Heritage of Scientific and Technical Information Systems*, p. 126-134, 2004.

[WEIT 95] WEITZMAN E., MILES M., *A Software Source Book. Computer Programs for Qualitative Data Analysis*, Sage, Londres/Thousand Oaks/New Delhi, 1995.

[WIL 01] WILMET M., « L'architectonique du "conditionnel" », *Recherches linguistiques*, vol. 25, p. 21-44, 2001.

[YOU 63] YOUDEN W.W., « Index of the Journal of the Association of Computer Machinery », vol. 1-10 (1954-1963), *Journal of the Association of Computer Machinery*, vol. 10, n° 4, p. 583-646, 1963.