UNIVERSITE FRANÇOIS-RABELAIS DE TOURS

Ecole Doctorale : Santé, Sciences, Technologies

Année Universitaire : 2004-2005

THESE POUR OBTENIR LE GRADE DE
DOCTEUR DE L'UNIVERSITE DE TOURS

Discipline : Informatique

Présentée et soutenue publiquement

par :
Hubert MARTEAU

Le
24 Novembre 2005

UNE METHODE D'ANALYSE DE DONNEES TEXTUELLES POUR LES SCIENCES SOCIALES BASEE SUR L'EVOLUTION DES TEXTES

Directrices de thèse :
Nicole VINCENT
Sylvette DENEFLE

JURY :

Mme	DENEFLE	Sylvette	Professeure	*Co-Directrice*	Tours
Mr	KODRATOFF	Yves	D.R. CNRS	*Examinateur*	Paris 11
Mme	SEDES	Florence	Professeure	*Rapporteur*	Toulouse 3
Mr	TOUSSAINT	Yannick	Chargé de Recherche	*Examinateur*	INRIA–LORIA
Mme	VINCENT	Nicole	Professeure	*Co-Directrice*	Paris 5
Mr	ZIGHED	Djamel	Professeur	*Rapporteur*	Lyon 2

Sommaire

Des Sciences Sociales à l'Informatique

Dans l'optique de définir les Sciences Sociales, on peut proposer trois types de définition allant de la plus usuel à la plus spécifique.

Dans un dictionnaire quelconque, les sciences sociales sont définies comme des sciences humaines qui ont pour objet l'étude des phénomènes sociaux.

Wikipédia [WWW WIK] est une encyclopédie libre, gratuite, universelle et multilingue, écrite bénévolement par des volontaires et basée sur un site Web. Cela en fait, une encyclopédie dynamique qui évolue selon les critiques des experts propre à chaque domaine. La définition qui y est faite des sciences sociales est plus complète que celle énoncée précédemment :
« Les Sciences sociales étudient les hommes au travers de leurs relations. Elles regroupent un ensemble de disciplines au sein des sciences humaines: sociologie, anthropologie ou ethnologie, histoire, géographie, économie, droit et psychologie (dans certains cas seulement). Critique littéraire et philosophie en sont exclues: elles revendiquent une portée plus universelle, et ne rapportent pas leurs conclusions au contexte (politique, social, économique...) de la société étudiée.
En France, l'expression de « sciences sociales » est utilisée par les professeurs de SES notamment qui souhaitent réunir au moins sociologie et économie. Cependant, la plupart des sociologues et la plupart des économistes ignorent l'autre discipline.
Les termes de regroupement reflètent des conflits méthodologiques et souvent idéologiques entre auteurs et entre disciplines. Ainsi, certains sociologues se considèrent plus proches des économistes que des historiens (s'ils estiment être plus proches du modèle des sciences dures), d'autres estiment que l'étude des individus en société doit faire une si grande place à leurs valeurs que la sociologie obéit plus au modèle de « l'interprétation » (histoire, critique littéraire) que de « l'explication » scientifique. ».

Enfin, toujours dans l'optique de définir la sociologie et son utilité, on peut s'intéresser à quelques citations de Pierre Bourdieu, un éminent sociologue. Les deux premières citations ont pour but de définir la sociologie et les deux suivante ont pour but de définir l'utilité de la sociologie.

« La sociologie ne mériterait peut-être pas une heure de peine, si elle avait pour fin seulement de découvrir les ficelles qui font mouvoir les individus qu'elle observe, si elle oubliait qu'elle a affaire à des hommes, lors même que ceux-ci, à la façon des marionnettes, jouent un jeu dont ils ignorent les règles, bref, si elle ne se donnait pour tâche de restituer à ces hommes le sens de leurs actes. » [BOU 62]
« La sociologie n'a pas pour fin d'épingler les autres, de les objectiver, de les mettre en accusation parce qu'ils sont par exemple « fils de tel ou tel ». Tout au contraire, elle permet de comprendre le monde, d'en rendre raison ou, pour utiliser une expression de Francis Ponge que j'aime beaucoup, de le « nécessité » - ce qui n'implique pas qu'il doive être aimé

ou conservé comme tel. Comprendre pleinement la conduite de l'agent agissant dans un champ, comprendre la nécessité sous laquelle il agit, c'est rendre nécessaire ce qui apparaît d'abord comme contingent. C'est une manière non de justifier le monde, mais d'apprendre à accepter des foules de choses qui autrement paraîtraient inacceptables. » [BOU 92]

« Aujourd'hui, parmi les gens dont dépend l'existence de la sociologie, il y en a de plus en plus pour demander à quoi sert la sociologie. En fait, la sociologie a d'autant plus de chances de décevoir ou de contrarier les pouvoirs qu'elle remplit mieux sa fonction proprement scientifique. Cette fonction n'est pas de servir à quelque chose, c'est-à-dire à quelqu'un.
Demander à la sociologie de servir quelque chose, c'est toujours une manière de lui demander de servir le pouvoir. Alors que sa fonction scientifique est de comprendre le monde social, à commencer par les pouvoirs. Opération qui n'est pas neutre socialement et qui remplit sans aucun doute une fonction sociale. Entre autres raisons parce qu'il n'est pas de pouvoir qui ne doive une part – et non la moindre – de son efficacité à la méconnaissance des mécanismes qui le fondent » [BOU 80]
« Mon but est de contribuer à empêcher que l'on puisse dire n'importe quoi sur le monde social. Schoenberg disait un jour qu'il composait pour que les gens ne puissent plus écrire de la musique. J'écris pour que les gens, et d'abord ceux qui ont la parole, les porte-parole, ne puissent plus produire, à propos du monde social, du bruit qui a les apparences de la musique. Quant à donner à chacun des moyens de fonder sa propre rhétorique, comme dit Francis Ponge, d'être son porte-parole vrai, de parler au lieu d'être parlé, cela devrait être l'ambition de tous les porte-parole, qui seraient sans doute tout à fait autre chose que ce qu'ils sont s'ils se donnaient le projet de travailler à leur propre dépérissement. On peut bien faire rêver, pour une fois… » [BOU 84]

La sociologie, dans la mesure où elle se veut science, s'inscrit dans un déterminisme selon lequel tout dans la nature obéit à des lois rigoureuses, y compris les conduites humaines. L'ensemble contenant ces lois est fini. Mais le nombre de contraintes qui constituent cet ensemble est si grand qu'il paraît infini. De ce point de vue, la sociologie a pour but de découvrir un maximum de ces contraintes pour tenter « *de restituer aux hommes le sens de leurs actes »*.

Afin de décrypter les lois de la variation sociale, les sociologues disposent d'outils qualitatifs et quantitatifs. Dans le cadre de ce travail, ces deux types d'outils s'imbriquent.
L'outil qualitatif est utilisé, premièrement, afin de confirmer ou d'infirmer les hypothèses émises et, deuxièmement, afin de proposer des hypothèses. Cet outil permet de formuler des hypothèses sur le sens des actes à partir d'une réalité terrain. Ces hypothèses sont utilisées pour constituer l'outil quantitatif.
L'outil quantitatif permet de confirmer ou d'infirmer les hypothèses formulées sur un échantillon aussi grand que possible.

Dans le contexte de ce travail, l'outil qualitatif utilisé est la série d'enquêtes et l'outil quantitatif utilisé est le questionnaire. Les enquêtes sont menées oralement et enregistrées, puis elles sont retranscrites manuellement sous la forme de fichiers textes informatiques. La

série d'enquêtes est basée sur une liste définie de thèmes pour lesquels des hypothèses ont été émises. Chacun des entretiens aborde, au moins une fois, chacun des thèmes.
Le questionnaire est une série de questions, le plus souvent fermées, élaborés à partir des connaissances issues des entretiens.
L'analyse de données quantitatives, de par sa nature, est très proche des statistiques et des analyses de données numériques. L'intérêt de l'outil informatique y est reconnu et approuvé depuis longtemps [CIB 84].
L'analyse de données qualitatives, quant à elle, n'a, compte tenu de ses besoins, qu'une utilisation très limitée de l'outil informatique. Pourtant des solutions existent tant pour le traitement d'enquêtes semi-directives [LOG SPA], (…) que pour les enquêtes non-directives [LOG ALC], [LOG LEX], [LOG ALT], [LOG NVI], (…). Mais ces outils ne répondent qu'à certains aspects de la demande.

Il paraît donc important d'essayer d'améliorer la démarche méthodologique des sociologues en proposant de nouveaux outils. Il ne s'agit nullement de « découvrir » une méthode magique qui, à base d'ordinateurs, résoudrait toutes les difficultés des sciences humaines. Il faut explorer de nouveaux outils qui pourraient améliorer les techniques et la fiabilité des méthodes qualitatives. La lexicométrie, s'appuyant sur l'analyse statistique des données, connaît un développement très important depuis une petite dizaine d'années. Elle permet de traiter, à l'aide d'un outil relativement objectivant, des données d'enquête, de quelque nature qu'elles soient, à partir du moment où elles sont transformées en textes, dans le sens le plus large qu'on puisse donner à ce mot (discours parlés, écrits, séquences gestuelles, bruits, mimiques, etc.) qui deviennent « textes » par la retranscription qu'en fait le chercheur.
L'outil informatique permet, si on accepte l'hypothèse de la possibilité de percevoir le sens d'un discours à travers le vocabulaire utilisé et l'organisation du propos, de faire apparaître des champs sémantiques spécifiques au texte étudié. On en vient donc, d'une certaine façon, à une analyse de contenu sans qu'intervienne, à toutes les étapes, la subjectivité du chercheur.

Depuis une vingtaine d'années, des travaux ont été entrepris par des linguistes et des statisticiens sur l'analyse lexicale du discours à l'aide de la méthode d'analyse statistique des données mise au point en France dans les années 60 par J. P. Benzécri.
Avec les possibilités de calcul des ordinateurs et les méthodes statistiques d'analyses multidimensionnelles, on se retrouve en situation d'analyser statistiquement tous les textes en eux-mêmes ou dans leur contexte (indiqué par des variables). Il y a là un outil nouveau susceptible de modifier complètement le rapport à l'analyse de textes, à la condition toutefois que cette analyse soit productrice d'autres choses que de comptage de mots et permette de percevoir du sens.

Le travail présenté s'inscrit dans cette problématique. L'attention a été tout spécialement portée sur la question de l'analyse de données textuelles. De nombreuses méthodes ont vu le jour. Elles s'appuient principalement et, pour rester simple, sur des méthodes statistiques ou sur des méthodes linguistiques. Mais l'élaboration de modèles reste un problème difficile.
Ce travail s'intéresse à l'étude des textes longs constitués d'entretiens qui ont été élaborés par des sociologues. Ces textes présentent des caractéristiques communes : un certain nombre de textes traitent du même sujet, la variété des thèmes abordés et des participants est si grande qu'il est difficile, pour les maîtriser, d'employer des dictionnaires spécialisés, ces textes sont

en langue naturelle parlée avec une grande variabilité de niveaux et dérogent largement aux normes grammaticales et syntaxiques. Cette particularité rend difficile, voire impossible, l'application de méthodes qui reposent sur des constructions linguistiques précises.

Ce travail s'est donc plus particulièrement intéressé à la représentation de textes longs et du sens qu'ils portent.

Ce manuscrit s'articule en cinq chapitres. Le premier chapitre s'intéresse aux méthodes existantes, de toutes les sortes, pouvant apporter une solution au problème posé. Les chapitres deux et trois étudient les étapes de la classification de données numériques. En effet, au-delà de la représentation, ce travail constitue une étape dans la classification des entretiens d'une même série.
Le quatrième chapitre s'intéresse à diverses représentations. Comme écrit précédemment, il s'agit de l'essentiel de ce travail de recherche. On s'est intéressé à plusieurs niveaux de représentation : le contenu, la structure et l'évolution des textes… Nous proposons des méthodes statistiques et des méthodes basées sur des automates. Nous proposons aussi le concept d'évolution textuel et montrons comment cela peut être un apport dans le cas de textes dépourvus de contenu et de structure.
Enfin, le chapitre cinq présente un corpus d'application réel et les résultats qui ont été obtenus.

Chapitre 1 – Les Solutions : Du traitement manuel des données au traitement automatique

Le but de ce projet consiste donc à analyser et à classer des entretiens sociologiques oraux retranscrits. Ces entretiens sont des données textuelles qualitatives. Leur analyse permet de dénombrer, dans le cas d'une logique déterministe, un maximum de contraintes sociologiques.
L'analyse et la classification peuvent être effectuées de manière complètement manuelle. Dans ce cas, l'outil informatique, s'il est utilisé, ne sert que pour des besoins de sauvegarde. Cette solution est fortement limitée et très contraignante.
Sinon, l'analyse et la classification peuvent être effectuées de manière semi-automatique. Il incombe alors au sociologue une partie des traitements. Selon le niveau auquel souhaite s'impliquer le sociologue, il peut utiliser un logiciel d'Analyse de Données Assistée par Ordinateur (ADAO) ou un logiciel d'Analyse Textuelle Assistée par Ordinateur (ATAO). Dans le premier cas, le sociologue s'est attaché à former un index de valeurs numériques et symboliques. Le logiciel fournit les outils d'analyse des données numériques et symboliques (ACP, AFC, CAH, ...). Dans le second cas, le logiciel, outre les précédents outils d'analyse des données, aide à la constitution de l'index.
Enfin, l'analyse et la classification peuvent être effectuées de manière complètement automatique. Le logiciel s'occupe, alors, de toutes les étapes de traitement. Il ne reste plus au sociologue qu'à comprendre les résultats de l'analyse.

Ce chapitre propose tout d'abord un aperçu non-exhaustif des solutions logicielles proposées, à l'heure actuelle, pour le traitement de données textuelles sociologiques. Puis, ce chapitre présente les possibilités d'informatisation d'une méthode manuelle. Ensuite, une partie est consacrée aux méthodes informatiques pouvant apporter une solution à ce travail. Enfin, une partie conclut ce chapitre en détaillant la méthode qui a été retenue.

1. Les Solutions Logicielles

Cette partie s'intéresse, de manière non-exhaustive, aux offres logicielles existant et permettant de résoudre tout ou une partie du problème posé. Il est évident que le nombre de logiciels existant, même s'il reste grandement inférieur au nombre de méthodes, est très important. Cette partie tente donc simplement de présenter un éventail aussi large que possible des logiciels, d'un point de vue méthode. Dans une première partie, les logiciels dits semi-automatiques, c'est-à-dire nécessitant l'action de l'utilisateur, seront décrits. Puis une étude sera faite sur les logiciels entièrement automatiques. Les informations qui suivent proviennent des développeurs, d'utilisateurs ayant fait un rapport ou de simples déductions à partir d'articles. Pour une étude plus approfondie des logiciels, de leurs méthodes, de leur utilisation et de leur critique d'un point de vue sociologique, il est conseillé de se rendre sur la page personnelle de Jacques Jenny, chercheur en sociologie au CNRS, [WWW Jenny].

1.1. Solutions existantes semi-automatiques

Les anglophones ont un label pour désigner les méthodes qualitatives informatisées nécessitant une lecture humaine : le CAQDAS, c'est-à-dire « **C**omputer **A**ssisted **Q**ualitative

Data **A**nalysis **S**oftware ». Ce label regroupe tous les logiciels permettant un traitement semi-automatique des données textuelles. Ces logiciels ne répondent pas à la demande de ce travail, mais leur présentation permet d'acquérir un certain point de vue. Cette partie présente quatre logiciels de ce type.

1.1.1. MODALISA

[LOG MOD] Interviews, module d'extension du Logiciel Modalisa, appartient à la catégorie des analyses de contenu thématique. Il possède plusieurs atouts qui en font probablement l'un des plus complets pour des besoins de recherche classique et l'un des plus souples d'utilisation de sa catégorie. Par contre, il ne traite que les séries d'enquêtes directives.
Les procédures de préparation du corpus (découper, surligner, commenter, annoter,...), de navigation hypertextuelle, de construction d'index, de consultations lexicales en contexte, correspondent à des modes usuels de lecture. C'est-à-dire que le logiciel fournit un ensemble d'outils permettant l'exploration des entretiens, mais que le travail reste un travail manuel. Il revient à l'utilisateur d'indiquer les mots ou les groupements de mots qui représentent un thème. L'analyse finale permet de croiser de plusieurs façons les groupements effectués manuellement entre eux ou avec les caractéristiques d'âge, de sexe, …

1.1.2. SATO

[LOG SAT] SATO est développé par François Daoust [DAO 90], au Centre A.T.O. de l'Université de Québec à Montréal. SATO est avant tout un gestionnaire de textes proposant une boîte à outils génériques. Un texte est représenté sous sa forme brute ou sous la forme de son lexique. SATO attribue, ainsi, aux lexèmes une catégorie "socio sémantique" pertinente pour la recherche. Le logiciel est assisté de l'analyseur morphosyntaxique Deredec. Il offre, ainsi, la possibilité de filtrer le corpus (patrons de fouilles). Enfin, SATO ne propose aucune décomposition factorielle ou classification automatique (caractéristique dominante de la statistique textuelle française), mais propose d'autres types d'analyse multidimensionnelle en rapport direct avec les hypothèses de recherche : comparaison, comptage, distance, lisibilité, participation et tamisage.
Le comptage nécessite une segmentation préalable du corpus. Il correspond à un ensemble de cinq valeurs : la moyenne, l'écart type, l'indice de répartition, le Chi-2 et la valeur discriminante.
Pour un lexème donné, la valeur discriminante est calculée de la façon suivante : *Fqmax * ln (1/répartition)* où *Fqmax* est la plus grande des fréquences relatives du lexème, calculées pour chacun des segments, et *répartition* est le nombre de contextes où apparaît le lexème divisé par le nombre total de contextes.
La distance est basée sur le Chi-2. Elle a pour but de mesurer l'originalité et la différenciation dans l'utilisation du vocabulaire.
L'indice de lisibilité correspond à la formule de Gunning. Son principal intérêt est de fournir plusieurs données quantitatives sur les éléments constitutifs du corpus.
L'analyseur «Participation» permet d'apprécier les différences significatives du pourcentage d'utilisation des lexèmes spécifiés par un filtre dans les différentes subdivisions d'un corpus.
L'analyseur «Tamisage» permet de visionner les lexèmes les plus fréquemment associés à un lexème donné, et ce dans un contexte choisi à volonté (contexte numérique, contexte de phrases, de strophes, de poèmes, etc.), détail important puisque la sémantique des relations entre les mots diffère selon la longueur des contextes.

Il s'agit donc, comme pour Modalisa, d'un logiciel de traitement aidé par ordinateur. Le travail nécessite l'intervention permanente de l'utilisateur.

1.1.3. ATLAS

[LOG ATL] Le système ATLAS.ti est issu d'un projet de recherche du département de psychologie de l'Université Technique de Berlin (1989-1992). Depuis 1993, il est commercialisé par son auteur, Thomas Muhr, en tant qu'atelier d'analyse qualitative de documents.
Son élément central est la citation, fragment défini par le lecteur sur un document primaire. Chaque citation peut être reliée à d'autres par l'usage d'hyperliens, et décrite par des codes. Ces codes, communs à plusieurs citations, peuvent être reliés à d'autres par des liens typés (cause, équivalence, généralisation, …). Un autre objet, le mémo, est un petit texte permettant de commenter un code, une citation ou un document primaire. Mémos, documents primaires et codes peuvent être regroupés dans plusieurs familles. Enfin, le supercode se distingue du code par une définition en intention (en fonction d'autres codes, supercodes ou familles) des citations qu'il décrit.
Ce logiciel se distingue par la liberté totale qu'il offre pour construire et agencer des catégories (ou boîtes de classement des données) selon des relations qui vont bien au-delà de la simple structure hiérarchique, ainsi que dans la représentation graphique qu'il est possible de faire de ces relations. D'ailleurs, en plus des relations programmées par les concepteurs du logiciel, le chercheur peut créer à volonté d'autres relations qui lui conviennent pour donner du sens à ses données. Par ailleurs, ce logiciel permet de représenter par un graphique les relations tissées entre les différents éléments. Il peut s'agir de catégories d'analyse (codes), de documents entiers, d'extraits de documents, des idées ou des commentaires.
Mais cela reste un logiciel semi-automatique, c'est-à-dire que l'intervention de l'utilisateur est régulièrement nécessité lors du traitement des textes. En d'autres termes, cela signifie que sans les multiples interventions de l'utilisateur, aucun travail n'est de traitement réalisé automatiquement.

1.1.4. NViVo

[LOG NVI] (le cahier d'accompagnement [BOU 05-2]) NViVo fonctionne par découpage manuel des textes du corpus. Les découpages permettent d'identifier des nœuds. Un nœud est un contenant dans lequel NVivo emmagasine un thème particulier. Les nœuds essaient donc de représenter les idées et la théorie portées par les textes. Ces nœuds permettent de classifier et de représenter des processus, des faits, des concepts abstraits, des lieux ou des individus. Un nœud peut indexer un nombre illimité de parties de textes et un texte peut être codé par un nombre illimité de nœuds.
Les attributs emmagasinent des informations quantitatives réduites. Ces informations peuvent servir à caractériser et classifier des individus (âge, genre, score à un test…), des lieux (type, nombre d'employés…) ou des documents (date de l'entrevue, interviewer…). Ces données sont plus aisément manipulées sous forme tabulaire.
Fréquemment, le type de conceptualisation qui émerge de données qualitatives riches nécessite plus que la simple utilisation du codage. Des liens se tissent entre les différentes parties du projet qui dépassent l'unidimensionnalité le plus souvent associée aux opérations de catégorisation. S'inspirant du modèle d'hypertextualité rendu familier par le web, NVivo

fournit plusieurs types de liens visant à connecter différentes parties du projet entre elles, et avec des documents externes.
Les ensembles permettent de regrouper, de manière temporaire ou permanente, certains nœuds ou certains documents dans un objectif précis. Les ensembles sont intégrés à tous les niveaux de NVivo. Ils peuvent servir tant au codage qu'à l'exploration des données à l'aide de l'outil de recherche. De plus, ils offrent de vastes possibilités d'organisation des données et des concepts d'un projet.
L'utilisation des modèles en recherche qualitative permet d'explorer, de conceptualiser et de communiquer de manière interactive les relations qui se développent au cours de l'analyse. Chaque item (nœud, document, attribut ou même un autre modèle) du projet peut être représenté dans un modèle puis mis en lien visuel avec les autres.
Ce logiciel propose donc de nombreux outils de marquage qui permettent une étude « sous toutes les coutures » des textes. L'ensemble des marquage est réalisé par les utilisateurs. C'est-à-dire qu'à nouveau, ce logiciel nécessite une intervention permanente de l'utilisateur.

1.2. Solutions existantes automatiques

Cette partie s'intéresse plus particulièrement aux logiciels offrant un traitement automatique des textes du corpus. Il est entendu par automatique le fait que le passage des textes aux matrices numériques représentatives se fait de manière automatique. Cette partie regroupe des logiciels effectuant une segmentation du corpus et des logiciels qui n'en font pas.
Ces logiciels sont de type lexicométrique. C'est-à-dire qu'ils s'attachent à comparer des profils lexicaux à l'intérieur d'un corpus ou entre corpus textuels, avec des options différentes concernant les opérations de pré-traitement.

1.2.1. SPAD-T

[LOG SPA] (Annexe A de [LEB 04]) SPAD-T (Système Portable pour l'Analyse des Données Textuelles) est une extension "qualitative-textuelle" d'un logiciel classique de traitement d'enquête par questions codées et numériques (SPAD).
L'entrée sollicitée par SPAD-T est une liste, marquée par différents séparateurs, des réponses du corpus. Cela signifie qu'initialement ce logiciel ne traite que des enquêtes directives (c'est-à-dire que les questions sont identiques pour toutes les enquêtes de la série). Cependant, ce logiciel peut être utilisé à la suite d'une classification thématique des entretiens. Dans ce cas, chaque thème est considéré comme une question.
Quoiqu'il en soit, SPAD-T apparaît comme une "boîte noire". L'interprétation des résultats produits comporte une part mal contrôlée d'arbitraire et de subjectivité. Il paraît alors préférable de varier les paramètres et les protocoles d'utilisation optionnels de SPAD-T ou de diversifier les points de vue à l'aide d'autres logiciel. L'essentiel est que les résultats puissent apparaître comme convergents ou complémentaires, mais jamais contradictoires !

1.2.2. ALCESTE

[LOG ALC] (Annexe A de [LEB 04], [GAR 98]) Inspirée du courant de l'analyse des données de Benzécri, la méthodologie ALCESTE (Analyse des Lexèmes Co-occurrents dans les Enoncés Simples d'un Texte) a été mise au point par Max Reinert [REI 86]. Plus qu'une comparaison des distributions statistiques des mots dans différents corpus, Reinert préconise l'étude de la structure formelle de leurs cooccurrences dans les énoncés d'un corpus donné.

De manière générale, ALCESTE permet d'identifier des univers de discours ou des classes d'énoncés, interprétés en fonction de la nature du corpus et des objectifs de l'analyse.
Les textes sont découpés en segments de taille homogène, appelés Unités de Contexte Elémentaires ou UCE, de longueurs variables. Chaque texte est appréhendé comme un ensemble d'unités, et chaque unité est décrite par les mots qu'elle contient (à l'exclusion des mots grammaticaux). On obtient ainsi une matrice binaire, comprenant des 1 si le mot apparaît dans l'UCE, et des 0 dans le cas contraire. L'objectif est de construire sur la base de cette matrice une typologie des segments de textes, ce qui est mis en œuvre à partir d'un processus de classement itératif. Chaque UCE appartient à une classe, et les typologies obtenues sont ensuite comparées ; seules sont conservées les classes les plus stables.
On obtient ainsi une matrice de répartition lexicale, sur laquelle le logiciel procède aux calculs d'Analyse Factorielle des Correspondances (A.F.C.) et de Classification Ascendante Hiérarchique (C.A.H.). Les résultats de ces calculs sont exprimés sous forme de graphiques qui appellent une interprétation, car c'est aux chercheurs qu'il appartient de nommer les zones des diagrammes factoriels et les branches des arbres de classification.
Alceste propose des calculs complémentaires. Tout d'abord, il y a les calcul de spécificités. Ils indiquent les formes lexicales sur-représentées et sous-représentées, selon telle ou telle catégorie de locuteur (grâce à des encodages péritextuels, par exemple de type sociologique "objectif"). Alceste propose aussi un inventaire des expressions figées, ou "segments répétés". Tous les deux sont caractéristiques notamment de ce qui est appelé la "langue de bois" et le "figement de la pensée".
Enfin, connaissant le dictionnaire complet des "mots" et des segments répétés, et c'est une fonction qu'Alceste partage avec la plupart des logiciels d'analyse textuelle, on peut obtenir l'affichage/impression des usages énonciatifs en contexte (au sens propre de contexte contigu d'emploi: les concordances) et des "mots associés" (à telle ou telle distance avant ou après: les co-occurrences) de n'importe quel élément simple ou composé du dictionnaire.

1.2.3. LEXICO

[LOG LEX] Lexico est développé par André Salem (Annexe A de [LEB 04], [LEI 95]) au laboratoire Lexicométrie et textes politiques de l'E.N.S. de Fontenay-Saint-Cloud. Par principe, Lexico-1 refuse toute transformation des formes graphiques du texte, sauf exceptions dûment motivées, et ne procède à la partition du corpus que sur critères "externes", pour distinguer par exemple différents auteurs ou catégories d'auteurs ou phases chronologiques. Comme pour Alceste, on obtient une matrice de répartition lexicale, sur laquelle le logiciel procède aux A.F.C. et C.A.H.
Les résultats, ici aussi, appellent une interprétation des chercheurs. Lexico propose, d'une manière générale, les mêmes outils complémentaires qu'Alceste (calculs de spécificités, …). La grande distinction entre Lexico et Alceste tient avant tout dans la représentation qui est faite du texte.

1.2.4. HYPERBASE

[LOG HYP] Hyperbase est développé à Nice par Etienne Brunet (INALF, Université de Nice). Il permet le dépouillement et le traitement statistique des textes. Hyperbase compte les mots d'après leur forme graphique (le mot est ici appréhendé comme une chaîne graphique) et met en évidence les termes les plus spécifiques d'une œuvre ou d'un texte par rapport à l'ensemble du corpus entré.

Il est ainsi possible de traiter son propre corpus ; différents outils sont disponibles : analyse factorielle, concordancier, richesse du vocabulaire, évolution des termes, etc.

1.2.5. SAS

[LOG SAS] SAS (Statistical Analysis System ou Système d'Analyse Statistique) est apparu sur le marché du progiciel à la fin des années 70 ; il présente l'avantage de proposer un grand nombre de fonctions, qui en font un logiciel des plus complets.
En outre, c'est un outil complet, qui assure à la fois les fonctions de gestion, de préparation et de traitement des données. Le langage de programmation SAS dispose d'un vocabulaire étendu et précis et d'une syntaxe facile d'emploi. Le volume des données peut être très important : SAS ne connaît d'autres limites que celles de l'ordinateur de l'usager. Le logiciel permet le traitement des tableaux observations/variables, ce qui correspond à la majorité des tableaux statistiques.
SAS permet non seulement de traiter les résultats obtenus à partir de l'analyse du corpus, mais encore ceux obtenus avec l'enquête par questionnaire.

1.2.6. 3AD

[LOG 3AD] Le tout récent logiciel 3AD est développé par l'équipe CRISTAL-GRESEC ([TIM 97] et [TIM 98])de l'Université Stendhal de Grenoble. Ce logiciel se distingue par ses 3 étapes.
Tout d'abord, un pré-traitement regroupe : une segmentation syntaxique, une analyse morphologique et une désambiguïsation. Ensuite, le traitement crée le dictionnaire du corpus et la liste des occurrences des formes puis identifie la ventilation des traits morphologiques et le degré de la stéréotypie du corpus. Enfin, le dépouillement calcule la distance entre les énoncés en fonction des variations du seuil et du poids et édite, par suite, les classes d'équivalence de paraphrases.
L'une des principales applications du logiciel 3AD est donc de contribuer quasi-automatiquement au dépouillement de questionnaires (codage des réponses en langage naturel à des questions ouvertes), et à l'indexation documentaire (mise à jour de thesaurus, extraction terminologique).

1.3. Bilan des logiciels

Aussi performants et intéressants soient-ils, les logiciels de type CAQDAS ne peuvent en aucun cas être considérés comme des logiciels de traitement automatique. Ce sont des logiciels de Traitement de Textes Assisté par Ordinateur. Cela signifie que l'informatique ne reste qu'un support. Ces logiciels offrent des résultats compréhensibles, car ils sont issus de l'interaction permanente entre le logiciel et l'analyste. Cependant l'aide qu'ils fournissent est insuffisante pour un traitement à l'échelle.

Les logiciels automatiques offrent, certes, un traitement intégralement automatique, cependant les résultats nécessitent de nombreux efforts d'analyse et d'interprétation.
La démarche du passage à l'automatisation est souhaitée afin de concentrer l'activité humaine sur l'analyse sociologique. Cette analyse sociologique est appuyée sur des représentations graphiques, ou autres, facilement compréhensibles.

Aucun logiciel semble être adapté à des traitements spécifiques. C'est certainement la raison pour laquelle aucun d'entre eux ne s'est imposé face aux autres. Ces logiciels sont, néanmoins, à garder comme exemples de traitements et de résultats.

2. Informatisation d'une méthode manuelle

De nombreuses méthodes d'analyse manuelle existent. Ces méthodes dépendent du sociologue qui l'emploie. Pour ce travail, il a été choisi l'approche de Sylvette Denèfle, la sociologue qui co-dirige ce travail. En effet, jusqu'à présent, Sylvette Denèfle traite ses entretiens, manuellement, à partir d'une méthode qu'elle a créée. Cette partie présente, tout d'abord, cette méthode manuelle. Puis, les possibilités d'informatisation de cette méthode sont détaillées et étudiées.

2.1. Méthode manuelle

Comme énoncé précédemment, la méthode manuelle présentée ici a été créée et est utilisée par Sylvette Denèfle, Professeure de Sociologie à l'Université François-Rabelais de Tours. Cette méthode, totalement nouvelle, se caractérise par son côté systématique, balayant et gérable d'un point de vue statistique.
Ces caractéristiques sont primordiales car elles permettent un traitement aussi objectif et régulier que possible des données. Or, le non-respect de l'objectivité et de la régularité est le plus grand défaut des méthodes manuelles. La gestion d'un point de vue statistique permet, quant à elle, d'offrir la possibilité d'innombrables traitements, analyses et représentations reconnus.

La méthode est, en fait, basée sur une recherche de variables sociologiques dans des entretiens non-directifs oraux retranscrits. Dans ces entretiens, une liste de points directeurs (thèmes), commune à tous les entretiens et fixée initialement, est abordée. La recherche d'informations se faisant de manière manuelle, cette méthode ne peut être considérée comme complètement objective.
Cette méthode permet de passer d'informations textuelles non structurées à une table de correspondances sociologiques, indiquant, pour chaque individu et pour chaque thème, la présence ou l'absence d'une caractéristique sociologique.

Le passage des données brutes aux données structurées se fait par l'application de deux segmentations successives suivies d'une agrégation sélective.

La première étape consiste à effectuer une classification thématique. Le but de cette étape est de retrouver chaque point directeur et de regrouper l'ensemble des informations. Cette classification est faite par blocs. La taille d'un bloc est très variable pour un même texte, depuis la phrase jusqu'aux groupes de paragraphes. Le but est de garder un maximum de connexité. De plus, le manque d'informations relatives à un thème est évité au maximum. Pour chaque classe, il est préféré la présence d'informations superflues au manque d'informations utiles.
A la fin de la segmentation, les différentes classes formées peuvent donc partager des informations communes. Cela signifie, en d'autres termes, que l'intersection entre les classes

peut être non nulle. Si cette remarque est tournée comme une possibilité, il faut signaler que c'est souvent le cas. Dans cette étape, le doute est soulagé par une acceptation par défaut.

Dans l'exemple qui suit, extrait de l'entretien 203 du corpus Le Corbusier, les thèmes abordés sont indiqués entre crochets. L'exemple n'aborde que trois thèmes : Sociabilité **[SO]**, Vie Familiale **[VF]** et Aménagement **[AM]**. Les questions posées par l'enquêteur sont mises en police grasse. Enfin, les réponses de plus de deux lignes sont écourtées par le sigle **[…]**.

[SO]
Et donc là, c'était déjà des proprios qui étaient là avant. Tu ne l'as pas acheté à L.A.H. ?
Non, c'était déjà des propriétaires qui étaient, eux, locataires avant mais dans un autre appart. ***[…]*** *enfin, bon, ils m'avaient raconté tout ça, quoi. Voilà.*
[/SO]
Et tu sais pourquoi ils sont partis ?
Oui, parce qu'ils ont acheté une maison.
D'accord. Et quand tu es arrivée, l'appartement était déjà un peu transformé ?
Oui.
[AM]
Il y avait déjà cette pièce là qui était faite ?
Voilà, cette pièce là qui était faite avec la porte vitrée là-bas ***[…]*** *Donc, d'un T4 c'est quand même devenu un T2-T3, quoi.*
[/AM]
[VF]
Mais du coup, pour une ou deux personnes, ça fait un grand appart, quoi.
Voilà. C'est spacieux. Pour deux, c'est impeccable.
[/VF]

Cet exemple trouve son intérêt dans la succession des thèmes abordés. Il peut être observé que les thèmes sont des parties très succinctes de textes. Les réponses, écourtées pour la présentation, n'excèdent pas 10 lignes. De plus, certaines parties ne sont pas classées. Cela signifie que l'union des classes est incluse dans le texte mais ne le recouvre pas complètement.

La seconde étape permet de créer une première table de correspondances. A partir de cette étape et jusqu'à la création de la table finale de correspondances, chaque thème est traité indépendamment des autres thèmes abordés.
Cette étape consiste à extraire, pour chaque individu, un maximum de variables abordées dans le thème. Cette recherche est basée sur les hypothèses émises. Mais d'autres variables, non suggérées lors de la formulation des hypothèses, peuvent apparaître. La liste des variables à chercher est donc ouverte. C'est à cause du manque d'exhaustivité de cette liste qu'il paraît impossible d'extraire toutes les variables. Les sociologues déterministes émettent l'hypothèse que ces variables sont dénombrables et limitées mais elles restent très nombreuses.
Pour extraire les variables, chaque phrase ou groupe de phrases du thème est reformulé. Le but de cette reformulation consiste à apporter aux phrases un niveau d'information pouvant être retrouvé dans d'autres entretiens. Pour cela, une même phrase peut être reformulée de

plusieurs façons et chaque version est alors gardée en tant que variable. L'exemple qui suit présente une réponse à une question lors d'un entretien (112), puis il présente 5 variables qui en ont été extraites. L'extraction consiste, ici, à un découpage en expressions et à une reformulation de certaines expressions, celles pouvant apporter du sens, dans un français plus proche de l'expression écrite.

> *Si je garde des contacts avec les anciens, ceux qui sont partis on s'écrit, on se téléphone, mais y'a beaucoup de mouvement alors les nouveaux bah…c'est des nouvelles amitiés mais y'a moins de convivialité quand même qu'avant, ça c'est clair, y'a moins de vie collective, c'est-à-dire y'a moins de gens qui s'investissent aussi comme moi dans l'association, tu vois on aime bien…moi je donne du temps, d'ailleurs j'ai quelqu'un qui me le reproche, ce sont mes enfants !*

je garde des contacts avec les anciens : avec ceux qui ont quittés Le Corbusier (on s'écrit, on se téléphone)
les nouveaux deviennent des nouvelles amitiés
il y a moins de convivialité qu'avant
il y a moins de vie collective qu avant
il y a moins de gens qui s'investissent comme moi dans l'association
je donne du temps pour l association ce que mes enfants me reprochent

Pour les experts du domaine, cette étape est assez mécanique par le fait qu'elle est avant tout une simple traduction des données brutes en variables identifiables. Pour les personnes du domaine, cette étape est un choix difficile dans les expressions à garder ou non. En effet, pour diverses raisons (information qui n'a pas de chance d'être retrouvée ailleurs, …), certaines informations ne sont pas gardées. Dans l'exemple qui précède, les expressions « *y'a beaucoup de mouvement* » et « *tu vois on aime ça* » paraissent pour un non – expert du domaine porteur de sens, elles ne sont, cependant, pas gardées comme variables.
Pour un maximum d'objectivité, cette étape se contente d'être une étape de découverte. C'est-à-dire que les personnes, qui effectuent cette seconde étape, tentent de n'apporter aucune réflexion quant à la problématique. Elles se contentent de dévoiler un maximum de variables. Cette manière de faire évite au maximum le biais apporté par les a priori.

La troisième étape consiste à agréger les variables entre elles. L'étape précédente est une étape de découverte, cela signifie que deux variables ne sont pas agrégées si elles ne sont pas parfaitement similaires. Cette troisième étape s'intéresse à reconnaître les variables sociologiques, c'est-à-dire celles qui ont un lien avec la problématique. Les variables sociologiques peuvent donc être l'union de plusieurs variables obtenues à la deuxième étape. Cette étape permet aussi de supprimer les variables qui sont jugées comme étant en-dehors du thème abordé.
Par ses conséquences sur la table finale de concordances, cette étape peut apparaître comme décisive. L'exemple qui suit présente les variables retenues dans le thème Aménagement et concernant la cuisine. Il y a, entre autres, une variable « *la cuisine est un peu petite, trop petite* » qui est l'agrégation de deux variables, l'une indiquant que la cuisine est petite, l'autre indiquant que la cuisine est « trop petite ». Ces variables sont différentes, mais portent le même sens.

le passe-plat permet de voir les gens quand on cuisine
c'est difficile de faire rentrer les appareils ménagers actuels
j'ai renoncé au lave-vaisselle faute de place
on n'a pas besoin d'un lave-vaisselle
une petite cuisine correspond à mon style de vie
les descendants ne sont pas pratiques car on ne peut utiliser les balcons à partir de la cuisine
c'est une kitchenette car elle n'est pas aérée
la cuisine est un peu petite, trop petite
la cuisine n'est pas trop petite

La description des étapes de cette méthode et de la table de concordances obtenue en résultat confirment son caractère systématique et gérable d'un point de vue statistique. C'est d'ailleurs pour son caractère mécanique apparent et la facilité d'interprétation de ses résultats que cette méthode a été élaborée pour traiter de manière idéale les enquêtes.
Pour un traitement manuel par des experts du domaine, le schéma d'exécution paraît relativement « simple ». Comme la facilité des traitements manuels ne se retrouve pas dans les traitements informatiques, une étude sur la faisabilité a été menée afin de voir si l'état actuel de la recherche, dans en analyse informatique de données textuelles, permettrait d'utiliser réellement cette méthode comme modèle de traitement. La partie suivante étudie les possibilités de créer un traitement informatique automatique à partir de cette méthode.

2.2. Informatisation de la méthode

La méthode manuelle était un modèle de traitement. Un premier travail a, donc, consisté à comparer chacune des étapes avec les problématiques déjà existantes dans le domaine du traitement de l'information.
La première étape consiste clairement à segmenter et à classer les thèmes présents dans chacune des enquêtes. La deuxième étape consiste, dans un premier temps, à détecter et à résoudre les anaphores afin de donner plus de contenu au discours, puis, dans un second temps, à reconnaître et à extraire des informations. Enfin, la troisième étape correspond à une comparaison de requêtes. Chacune de ces problématiques est étudiée dans les quatre parties qui suivent.

2.2.1. Segmentation et classification thématique

La segmentation et la classification de thèmes sont deux problématiques fortement étudiées par la communauté traitant de manière automatique les langues naturelles, car elles sont considérées par beaucoup comme une première étape primordiale à tout traitement textuel [BOU 02-1]. La segmentation thématique consiste à trouver les bornes au-delà desquelles le texte étudié change de thème. La classification consiste à réunir les blocs bornés dans lesquels un même thème est abordé.
Ces deux problématiques trouvent leur application dans la recherche d'information [SAL 94], mais aussi dans la reconnaissance de la parole [BRU 03], [BIG 00].

La pondération la plus répandue, en traitement de l'information, est le tf*idf de Salton. Cette pondération est détaillée dans le chapitre 4 sur l'indexation. Salton utilise, à la suite de cette

pondération, la distance du cosinus [SAL 89]. Cette base peut être utilisée pour une classification ou avec une normalisation du poids des mots [SAL 94]. L'étude des liens (suppression des liens entre des blocs trop distants) [SAL 96] permet d'effectuer une segmentation thématique.
Cette méthode peut aussi être utilisée à d'autres moments du traitement comme pour la structuration du document ou pour sa description [HER 02]. Le tf*idf peut servir de pondération des termes avant l'utilisation d'une autre distance comme celle de Dice et des informations relatives aux cooccurrences pour améliorer le résultat [FER 97-1]. Le tf*idf peut aussi servir de pré-traitement à une classification hypergraphe [CLI 04]. Certains préfèrent, néanmoins, garder le cosinus mais effectuer un filtre médian et une normalisation sur le résultat final [MAT 03]. La pondération des termes peut être améliorée en ajoutant au tf*idf, le tf*isf. Le tf*isf évalue, tout d'abord, la distribution d'un terme à l'intérieur d'un document, puis la dispersion du mot dans le document [DIA 05].
La méthode [DIA 05] prend un intérêt particulier par la formation des blocs qu'elle traite. En effet, elle considère le texte comme un flux continu d'informations C'est le cas lorsque l'on traite la reconnaissance d'une source audio continue. Les blocs de texte traités sont les blocs d'une fenêtre coulissante de n phrases.
[PON 97] a aussi utilisé le principe de fenêtrage. Cependant, deux fenêtres sont utilisées. Une similarité est ainsi calculée à gauche et une autre à droite de la phrase étudiée. La meilleure segmentation est trouvée par programmation dynamique.
[FER 97-2] utilise une petite fenêtre à gauche. Le but est de conserver une mémoire épisodique, à court terme, des événements traités.

En reconnaissance de la parole, les études sont plus concernées par le modèle linguistique à adopter. Les études se portent, donc généralement, sur l'apprentissage. Dans de tels traitements, l'information mutuelle d'un mot pour un thème, et inversement, prend une place de choix [BRU 02], [BRU 03]. Les modèles les plus communs sont le modèle unigramme et le modèle cache [BIG 98]. Le premier de ces modèles traite l'attachement des mots à un thème de manière indépendante, c'est-à-dire sans fenêtrage. Le modèle cache repose sur la comparaison entre le contenu d'une mémoire cache et les distributions statistiques des mots clés des unigrammes thématiques, avec une valeur constante fixée pour chaque thème et qui est assignée à tous les autres mots qui n'entrent pas dans le cache.
Dans une étude [BIG 00], l'utilisation du modèle cache est comparée à d'autres méthodes du point de vue de la segmentation et de la classification. Au niveau de la segmentation, le modèle cache est comparé à une segmentation systématique. Une segmentation systématique consiste à segmenter de manière régulière. Au niveau de la classification, le modèle cache est comparé à une programmation dynamique basée sur l'historique. Le modèle cache utilisé aux deux niveaux apporte les plus grandes précisions (0,65 – 0,81), mais les plus petits rappels (0,29 – 0,39), dans des expériences menées sur des articles provenant du journal « Le Monde » pour lequel sept thèmes avaient été retenus. Inversement, l'utilisation des deux autres méthodes sur le même corpus offre les plus petites précisions (0,16 – 0,35), mais les plus grands rappels (0,4 – 0,6).
Deux autres études, [BIG 01-1] et [BIG 01-3], ont comparé au modèle cache, le modèle unigramme, un modèle sur l'information mutuelle, le modèle statistique de Salton et un modèle de perplexité (issu de la théorie de l'information). Ces études avaient pour but de

montrer l'effet de chacun de ces modèles sur un corpus du type journal et sur un corpus du type e-mails. Le modèle cache reste le meilleur sur le corpus du type journal, mais c'est, de loin, le modèle basé sur l'information mutuelle qui reste le meilleur pour le corpus du type e-mail.
D'autres travaux, [BIG 01-2] et [BIG 02], se sont plus particulièrement attachés au type de classification. Ces travaux ont prouvé que les résultats étaient améliorés par l'utilisation d'une classification thématique du type hiérarchique.
[KAN 01] considère l'importance des hiérarchies de thèmes. Cette hiérarchie permet de gérer des documents multi-thématiques du type dictionnaire. C'est-à-dire que chaque information du même type est répétée dans plusieurs parties (capitale, nombre d'habitants, …).Cette hiérarchie peut aussi permettre de gérer des sous-thèmes. Cette hiérarchie est obtenue par une agrégation qui autorise, entre autres, l'héritage multiple.

En dehors de ces deux grandes approches, la liste des méthodes reste très longue. De manière non-exhaustive, on peut en citer plusieurs. [KAR 94] utilise une simple analyse discriminante.
[BEE 99] représente chaque mot d'un texte par une liste binaire. La liste binaire indique la présence, ou non, d'un terme dans le voisinage du mot représenté. Une fois la représentation construite, la méthode utilise une classification par Chaînes de Markov Cachées (CMC).
[BLE 01] ajoute aux CMC classiques la notion de modèle d'aspect apportée par Hofmann.
[BOU 02-2] traite les conversations audio par des CMC. Dans cette méthode, un certain nombre d'expressions ou de mots relatifs à l'articulation d'une conversation (mots d'accueil, mots d'introduction, mots de transition, …) sont considérés comme des états.
[FOR 00] opte pour l'utilisation d'un classifieur du type réseau de neurones avec apprentissage non-supervisé (ART I). Ceci permet de tracer des liens entre les classes et de voir un même mot dans les divers contextes où il peut être employé. [LAR 02] utilise des réseaux de neurones avec des cartes de Kohonen. Cette méthode permet de segmenter les discours, mais permet aussi de déterminer les parties transitoires. Les parties transitoires sont utilisées pour classer les informations relatives au thème précédent et celles préparant au thème suivant.
[AMA 00] classe les textes oraux retranscrits par k plus proches voisins (kppv) en utilisant la distance de Kullback-Leibler. Sur le même classifieur, [ROG 03] conserve la distance euclidienne. Afin de limiter l'effet d'écrasement dû aux grandes dimensions, une sélection des mots est effectuée à partir de la loi de Zipf [ZIP 35].
[LI 03] propose une modélisation stochastique et l'utilisation de l'information de Shannon.
[FER 01] allie au traitement statistique un traitement linguistique afin d'améliorer les performances.
Enfin peuvent être cités les outils de segmentation semi-automatiques [BEU 02]. L'outil est une application interactive qui permet à son utilisateur de créer et de modifier des thèmes, c'est-à-dire des classes de mots relevant d'un même domaine sémantique.

Afin de réellement comprendre les conditions réelles d'application, il faut observer la segmentation thématique qui est faite manuellement. Dans l'exemple qui suit, les thèmes abordés sont indiqués entre crochets (l'exemple n'aborde que deux thèmes Vie Familiale **[VF]** et Aménagement **[AM])** et les questions posées par l'enquêteur sont mises en police grasse.

[VF]
Bah non le soir et le matin je garde des enfants en nounou agréée, ça c'est pas la mairie, c'est moi toute seule tu vois.
[/VF]
[AM]
D'accord. Donc ici c'est quoi comme type d'appartement ?
Alors c'est un type 4 qui fait 76m², parce que tous les types 4 font ça, et on a un descendant, on est côté parc…
Donc y'a trois chambres…
[VF]
Oui trois chambres.
Et vous vivez à combien ici ?
*Alors là en ce moment on est trois parce que y'a **** qui est là, elle a 20 ans, et **** 14* ans.
[/AM]
Et tu es arrivée y'a combien de temps ici ?
25 ans.
Ça fait un bout de temps…

Cette enquête, comme toutes les autres de la série, concerne les unités d'habitation imaginées par Le Corbusier. Le thème Aménagement trouve donc régulièrement sa place dans la discussion. On retrouve ce thème lorsqu'il est discuté de la taille de l'appartement (nombre de pièces, surface), de son orientation, du nombre de chambres et par là même du nombre de personnes logées.
Les deux dernières phrases concernent aussi le thème Vie Familiale. En effet, derrière le nombre de chambres et de personnes vivant dans le foyer, on aborde le sujet de la famille et pas seulement celui de l'aménagement. Cette superposition des thèmes rend le traitement de ce corpus très complexe.

Cette superposition n'est pas propre aux entretiens sociologiques. Chaque phrase peut concerner de multiples sujets. Son orientation dépend de la personne qui l'exprime, mais aussi de l'interprétation qui en est faite.
Lorsque quelqu'un dit : « Ce matin, je suis allé acheter du pain ». Cette personne parle-t-elle de l'acte social, économique ou gastronomique ? Peut-être justifie-t-elle son emploi du temps !
Dans l'exemple donné précédemment, deux niveaux de réflexion sont présents. Un premier niveau, l'Aménagement, est plus concret, plus physique, sont abordées des notions d'organisation matérielle. Un second niveau, la Vie Familiale, est plus humain, sont abordées des notions d'organisation humaine.
Ces deux thèmes ont une certaine proximité car ce sont les humains qui utilisent le matériel. Mais ils n'appartiennent pas au même niveau de réflexion. C'est la raison pour laquelle ils se superposent, mais restent indépendants l'un de l'autre. Cette superposition est une première difficulté de cette étape de segmentation.

La seconde difficulté réside dans le langage employé. Le discours oral est, en effet, très pauvre en vocabulaire. Ce manque se ressent dans l'identification des thèmes. Certains thèmes sont fortement caractérisés par certains sigles, termes ou mots. D'après l'exemple, on peut lister pour l'Aménagement : type 4, m².
Un thème comme Vie Familiale trouve, par contre, bien plus difficilement des sigles, termes ou mots caractéristiques. Pour le prouver, il suffit de prendre un nouvel exemple :

> *Bah non le soir et le matin je garde des enfants en nounou agréée, ça c'est pas la mairie, c'est moi toute seule tu vois.*
> *[...]*
> ***Et vous vivez à combien ici ?***
> *Alors là en ce moment on est trois parce que y'a **** qui est là, elle a 20 ans, et **** 14* ans.

Les extraits qui précèdent appartiennent, sans conteste, au thème Vie Familiale. Ils montrent l'organisation familiale, au niveau des activités de la mère (nounou agréée) et au niveau de la composition (2 enfants).
Des mots qui semblent être caractéristiques peuvent être extraits, par exemple :« garde », « enfants », « nounou » et « mairie ». Ces mots pourraient être liés aux services de l'unité d'habitation. C'est-à-dire que la personne interviewée pourrait employer ces mots pour citer des services qu'elle utilise pour ses enfants. Le thème Intérieur/Extérieur de la série d'enquêtes sur les unités d'habitation de Le Corbusier concerne les services proches ou liés aux unités d'habitation. Si l'emploi des mots est changé, les extraits passent du thème Vie Familiale au thème Intérieur/Extérieur. C'est donc par le contexte que les mots sont liés au thème Vie Familiale. Dans ce cas, les mots formant le contexte sont communs, c'est-à-dire vides de sens.

La deuxième partie de l'extrait ne contient aucun mot utile. Il s'agit d'une simple phrase en langage naturel.

L'annexe 2 présente les mots caractéristiques (exclusifs) pour chacun des thèmes liés à la série d'enquêtes sur les unités d'habitation « Le Corbusier ». Le thème Vie Familiale est caractérisé par une liste de 15 mots qui ne sont ni implicites ni utilisés de manière fréquente. Le mot le plus utilisé apparaît 4 fois, il s'agit du mot « maris ».
Cela signifie qu'aucun mot ne permet de lier de manière catégorique l'exemple au thème Vie Familiale. Dans ces phrases apparaissent pourtant des variables qui paraissent essentielles à la compréhension de la structure de cette famille. Peut-on considérer comme similaires deux familles vivant dans des types 4, l'une avec 5 enfants et l'autre avec 2 enfants ? Peut-on considérer comme similaires deux familles, l'une avec des enfants de 14 et 20 ans, l'autre avec des enfants de 10 et 24 mois ?
L'annexe 2 montre clairement que des thèmes fortement concrets et physiques, comme l'Aménagement, trouvent de nombreuses caractéristiques. Elle montre aussi que les thèmes à tendance plus sociologique, comme Vie Familiale ou Association, ne se représentent dans le texte que par des formulations simples et n'ont pas de réelles caractéristiques pouvant les identifier.

Cependant, des méthodes d'apprentissage permettent d'appréhender le contexte. De plus, les mots qui caractérisent les thèmes correspondent très rarement aux mots utilisés de manière exclusive. L'annexe 2 n'est donc pas le reflet d'une base de connaissance comme elle pourrait être formée pour un tel traitement.
Enfin, l'étape de reformulation, c'est-à-dire de détection et de résolution des anaphores, pourrait être employée avant la segmentation thématique. Ceci permettrait d'améliorer la qualité du contexte des termes.

2.2.2. Détection et Résolution d'Anaphores et d'Ellipses

En informatique, dans le traitement de la langue naturelle, la résolution des anaphores est un problème important, complexe et à part entière. La majeure partie des résolutions consiste à résoudre les liens de références entre un pronom et son référé. Les communications, tant orales qu'écrites, contiennent de nombreux liens références-référés. Ces liens permettent d'alléger la communication. Voici un exemple :

> *Lou-Ann avait acheté les chaussures avec son argent de poche.*
> *Maintenant elles lui appartenaient et elle en était fière.*

Cet exemple montre bien toute la complexité de la résolution des anaphores.
Le pronom possessif « *son* » de la première phrase réfère évidemment à « *Lou-Ann* ». Dans la seconde phrase, le pronom personnel « *elles* » référence « *les chaussures* » alors que les pronoms personnels « *lui* » et « *elle* » référencent « *Lou-Ann* ». Dans ce cas, il faut noter que les deux référés, « *les chaussures* » et « *Lou-Ann* », sont cités dans la phrase précédente. Enfin, le pronom personnel « *en* » fait référence au fait que les chaussures lui appartiennent.
Les référents (ici les pronoms personnels) sont appelés anaphores et les référés sont appelés antécédents. La résolution d'une anaphore consiste à déterminer l'antécédent qui lui correspond. Pour l'exemple précédent, une résolution des anaphores et une reformulation à partir des antécédents mèneraient donc aux phrases suivantes :

> *Lou-Ann avait acheté les chaussures avec l'argent de poche de Lou-Ann.*
> *Maintenant les chaussures appartenaient à Lou-Ann et Lou-Ann était fière que les chaussures appartiennent à Lou-Ann.*

Suite à une telle reformulation, la raison principale de l'emploi d'anaphores devient évidente : elles évitent les répétitions peu supportables.

La plupart des systèmes complets d'interprétation des pronoms, implantés à ce jour, s'occupent de manière automatique de la détection et de la résolution. Pour cela, ils effectuent une résolution en étapes :

- Détection des anaphores et de leur type.
- Application d'un filtre sélectif pour éliminer les anaphores ne trouvant pas de solution.
- Constitution d'un ensemble d'antécédents potentiels pour chaque anaphore.
- Application de contraintes pour éliminer certains antécédents.
- Détermination de l'antécédent le plus probable par application de préférences.

Ce schéma de traitement, initié par [LAP 94], a ainsi été appliqué pour traiter les anaphores de plusieurs langues. Des solutions sont proposées pour résoudre les anaphores présentes en anglais, [LAP 94] et [MIT 98], en allemand, [LAP 94] (même si, dans l'article, l'accent est mis sur la résolution dans des textes anglophones), en espagnol, [PAL 01], et en français, [TRO 02-1].
Les travaux de base, [LAP 94] et [MIT 98], ne trouvent que peu d'amateurs, [HOS 04], pour une application en l'état. Les systèmes plus récents proposent de réelles améliorations dans les règles et contraintes de détection et de résolution.
[PAL 01] liste de manière exhaustive tous les types de pronoms à traiter et propose une liste de préférences pour chaque type de pronom.
[TRO 02-1] propose, en dehors des spécificités dues à la langue française, de reconnaître les insertions, c'est-à-dire les segments de texte qui ne font qu'apporter des précisions non-essentielles par rapport au discours principal. Une contrainte supplémentaire sur les antécédents est donc énoncée en fonction de la présence d'insertions. Et [TRO 02-2] propose un système de choix de préférences où l'antécédent choisi a une pondération supérieure à la somme des pondérations des antécédents les plus proches.
[JAI 04] propose une modélisation du texte par un graphe de noms et de pronoms et base la résolution sur un système pondéré de diverses heuristiques dont les poids sont définis par apprentissage et non de manière fixe.

En dehors de ce schéma de traitement, d'autres systèmes existent. En effet, [BOU 05] essaie de résoudre les anaphores à partir de connaissances sémantiques. Pour cela, il recherche une association sémantique dans l'unité thématique étudiée.
[SAL 04] émet une critique quant aux contraintes utilisées par le système précédent et prouve l'intérêt d'utiliser un lexique de synonymes extrait d'un dictionnaire.
[MAR 03] s'attache à résoudre les anaphores nominales à partir d'un système de votes sur le lien sémantique unissant l'anaphore aux possibles antécédents. Le résultat du vote correspond au nombre de résultats obtenus dans une requête sur un moteur de recherche Internet.
Mais ces dernières méthodes tentent avant tout de résoudre l'anaphore de manière semi-automatique. La détection est effectuée de manière manuelle.

D'autres méthodes mettent l'accent sur l'annotation des textes. [TUT 00] propose une annotation du type XML qui faciliterait la résolution.
[DEP 99] propose un étiquetage par classe d'objets afin de privilégier les liens entre les mots ou les groupes de mots. Par exemple, deux types d'objets peuvent être considérés : les aliments et les meubles. Si les premiers peuvent être mangés, on ne peut pas en dire autant des seconds. Par contre, avec le verbe « *laver* », la classe aliment doit être séparée en deux sous-classes : fruit, légumes, … d'un côté et gâteaux, pizzas, … d'un autre. En effet, des meubles peuvent être lavés, des aliments de type fruit ou légume peuvent être lavés, alors que des aliments de type gâteau ou pizza ne peuvent l'être.

Enfin il existe des systèmes particuliers, liés aux langues. [CHI 03] s'attache à la résolution des anaphores en langue chinoise. C'est un cas particulier car, en chinois, le thème est généralement abordé en début d'expression. S'il vient à manquer, c'est le sujet qui prend sa place. Si le sujet lui-même vient à manquer, l'expression commence alors par le verbe qui est toujours à l'infinitif.

Dans le cadre de ce travail, la résolution d'anaphores est appliquée à des entretiens oraux retranscrits. Le fait que ce corpus soit constitué de retranscriptions de conversations orales implique une structure très particulière. Si l'utilisation d'anaphores est grande dans les textes écrits, elle devient systématique à l'oral. Le manque de répétitions est avant tout marqué par un grand nombre d'ellipses. L'ellipse est un procédé de langage par lequel on sous-entend l'un des éléments d'une phrase sans nuire à la compréhension de celle-ci, par exemple :

Céline a pris du lait au chocolat et Lou-Ann à la fraise.

La résolution des ellipses est aussi importante que celle des anaphores, car ce procédé est extrêmement utilisé dans les entretiens. Il permet, en effet, de ne pas répéter la question ou les informations présentes dans la question qui vient d'être posée.
Voici quelques exemples, tous issus du même entretien (n°203), illustrant les types d'ellipses qui sont présentes dans les entretiens. Dans les exemples, les questions sont mises en gras afin de les distinguer des réponses émises.

Il y a donc évidemment les réponses brèves :

D'accord. Et quand tu es arrivée, l'appartement était déjà un peu transformé ?
Oui.

A l'origine, c'est un appart pour quatre en fait, un couple et deux enfants. Oui, ça change tout. Et au niveau déco, les petites portes hublots, c'était déjà fait ?
C'était déjà fait, oui.

Oui et puis tu n'aurais pas dans ta cuisine des chaises en formica.
Voilà.

Ici, tu n'as pas l'impression d'avoir une vie de quartier en fait ?
Pas trop non.

Et toi, tu avais plutôt, tu m'as dit que tu avais eu des apparts mais avant, quand tu vivais chez tes parents, tu vivais plutôt en maison ou en appart ?
En appart.

Ce type d'ellipse peut facilement être résolu car la plupart des réponses sont, soit une confirmation positive (« *oui* », « *voilà* »), soit une infirmation (« *non* »). Dans de tels cas, la résolution de l'anaphore consiste à simplement reformuler la question, soit dans sa forme positive, soit dans sa forme négative.

L'autre type de réponse est la réponse à une question à choix. Le choix est marqué la plupart du temps par un « *ou* ». Dans ce cas, la réponse n'est que l'énonciation succincte

de l'un des choix. La résolution de l'anaphore consiste ici à reformuler la réponse, à partir de la question, en éliminant les choix non retenus.

Parfois, les réponses brèves peuvent apporter des informations supplémentaires. Dans l'exemple suivant, une notion de quantité de personnes est ajoutée à l'affirmation. Cette notion n'est, a priori, pas attendue de manière explicite dans la question qui demande juste confirmation. Elle apporte pourtant, d'un point de vue concret, une réelle indication quant à la fréquence et à l'habitude.

Oui, c'est étrange. Et tu disais qu'à ce moment là ça t'a permis de rencontrer des gens dans l'immeuble ?
Oui, mais c'est tout. Une personne.

Enfin, il y a les jeux de questions-réponses. Dans ce cas, le contexte n'est présent ni dans la question, ni dans la réponse.

Donc, ça comprend chauffage…
Oui, c'est tout. Et puis toutes les…

Oui. Et par rapport à la Poste, tu l'as connue ? Bah oui, tu l'as connue.
Bah oui et je trouve ça vraiment dommage.

Que ça ait fermé ?
Oui, je trouve ça vraiment nul.

Et il t'arrive de faire visiter ton appart autre qu'à des gens que tu connais ? Parce que Anne m'expliquait qu'elle avait déjà trouvé des gens dans le couloir qui voulaient visiter des apparts et…
Oui, ça m'est arrivé une fois.

Ça c'est pas commun aussi.
Bah non.

Le premier exemple fait suite à une discussion sur les charges. La personne qui répond compte donner des informations supplémentaires, mais sa réponse est coupée par une autre question. L'exemple présente la réponse au complet telle qu'elle apparaît dans la retranscription. Sorties de leur contexte, cette question et cette réponse sont vides de sens.
Le second exemple montre la réaction à la fermeture de la Poste. Or le lieu, c'est-à-dire la Poste, n'est citée que dans la première question. Et la fermeture n'est abordée que dans la

seconde question qui n'a pas, au sens littéraire du terme, la forme réelle de la question. Ici, les deux réponses vont dans le même sens. Aucune information supplémentaire n'est apportée par la seconde réponse. Mais il a fallu deux questions pour établir le contexte des réponses.
Quoique plus simple, le second exemple montre bien le fait qu'un entretien est une interaction permanente. D'une manière générale, pour les questions initiales, le contexte est amené par la situation de départ (paroles échangées avant l'enregistrement, présentation par un tiers, …). Pour les questions suivantes, le contexte est amené par les questions et les réponses qui précèdent.

Cette partie a, donc, permis de faire un tour des systèmes de résolution d'anaphores et a permis de faire un tour du problème des entretiens oraux retranscrits.
Les systèmes de résolution des anaphores s'améliorent sans cesse. Ils proposent à présent de bonnes méthodes pour la détection et la résolution des anaphores. La plupart des systèmes sont liés à des méthodes linguistiques telles que l'étiquetage, la segmentation thématique ou la recherche de liens sémantiques. Ces méthodes fonctionnent, le plus souvent, à partir d'un dictionnaire, d'un lexique ou d'un logiciel extérieur.
Dans le cas des entretiens oraux retranscrits, la résolution des anaphores pourrait être adaptée pour donner plus de corps aux réponses brèves. La difficulté de ce corpus réside dans le fait que la mise en contexte est permanente à partir des questions et réponses qui précèdent.

2.2.3. Extraction d'Information

L'extraction d'information est la mise en évidence d'une information particulière. Jusqu'à présent, les systèmes d'extraction d'informations sont liés à une application particulière et nécessitent, la plupart du temps, la création d'un modèle particulier de résolution.
Par exemple, pour extraire les différents concepts présents dans des documents médicaux, des lexiques de flexions, dérivations et synonymes propres au domaine médical sont utilisés [POU 02].

[POI 01] cherche des expressions littéraires telles que « *a dit que* », « *a affirmé que* », « *a reproché à* » afin de modéliser les interventions de personnes. [POI 99] et [POI 00] cherchent, dans un texte prétraité, des expressions du type « *attentat à le* » afin d'extraire, pour les attentats, les informations concernant la date de l'événement, l'endroit de l'événement, le nombre de personnes tuées, le nombre de personnes blessées et l'arme utilisée.

Dans un autre contexte, [BOU 05] traite les transcriptions de conversations téléphoniques portant sur des incidents survenus en mer. A partir d'associations entre les étiquettes sémantiques, un modèle markovien sur les relations prédicats-arguments cherche à extraire les informations sur la description du bateau, le type d'incident, les ressources utilisées pour la mission, le lieu et les conditions de la mission.

Enfin [ARA 00-1] améliore, à partir d'une recherche d'informations, une classification bayésienne en quatre classes de petites annonces (voitures, emplois, immobiliers et autres). Dans cette méthode, chaque classe est marquée par une utilisation caractéristique de certains termes (« *salaire* » pour emploi, « *surface* » pour immobilier …). Chaque classe est aussi

caractérisée par d'autres termes trouvant une utilisation plus commune (« couleur » a une association privilégiée avec voiture, mais peut désigner un grand nombre d'autres objets).
Un poids est, ensuite, mis sur les termes de façon manuelle à partir des connaissances du domaine. Les annonces, dont le poids total ne dépasse pas un certain seuil, ne sont pas classées.

[ARA 00-2] propose une méthode afin d'étiqueter les termes ou expressions des petites annonces une fois qu'elles sont classées. Pour cela, une liste d'expressions régulières et un lexique de mots généraux et spécifiques sont créés. Les termes les plus communs peuvent ainsi être étiquetés. Une liste de mots clés permet d'étiqueter les segments qui ne le sont pas (tag de contexte) contenant ces mots clés.

Pour le traitement d'entretiens sociologiques, la première difficulté réside dans la non exhaustivité du modèle qui pourrait être créé. Comme il a été écrit dans la description de la méthode manuelle, cette étape est une étape de recherche. Les informations extraites peuvent correspondre à celles émises dans l'hypothèse, mais l'hypothèse ne peut pas prévoir toutes les variables à extraire.
Une seconde difficulté réside dans la construction du modèle définissant les informations à extraire. En effet, le corpus est un ensemble de textes oraux retranscrits en langage naturel. Manuellement, les informations sont extraites à partir d'expressions simples.
L'exemple suivant montre la simplicité des variables extraites.

> ***il y a des gens avec qui vous avez plus d'affinité que d'autres***

Cette variable présente un état de socialisation. On peut se demander quels seraient les mots à utiliser pour former un modèle pour une telle variable ?

Le mot « *affinité* » paraît être le meilleur candidat.
On retrouve cette variable dans l'entretien 113. Le mot « affinité » n'est utilisé qu'une seule fois et il l'est dans le bon sens :

> *et lui non on l'a rencontré, c'est une **affinité** comme ça*

L'entretien 12 partage aussi la variable. Dans cet entretien, le mot est utilisé deux fois. La première utilisation est en lien avec la variable. Par contre, la seconde utilisation est plus commune et n'a rien à voir avec la variable.

> *Puis vous avez aussi des gens avec qui vous êtes en **affinité** et puis d'autres bon…*

> *Mais il y a plusieurs possibilités d'aménagement parce que bon, avec mon épouse, on n'a pas toujours les mêmes goûts, les mêmes **affinités** d'intérieur.*

L'entretien 136 partage aussi la variable et, pourtant, le mot n'est pas utilisé. On peut donc s'intéresser à un autre mot comme le mot « autres » par exemple. Ce mot montre une certaine

ouverture, mais son utilisation commune est fortement supposée. L'exemple qui suit montre des expressions issues de l'entretien 136 contenant ce mot.
Le mot « autres » est un mot typique qui se prête à une utilisation dans le sens de la variable (voir les trois expressions phrases), mais qui se prête, avant tout, à une utilisation commune (voir le deuxième groupe d'expressions).

> Par contre d'autres, on a connu en étant ici ; un peu plus maintenant aussi par le biais de l'école.
> de rencontrer d'autres parents
> rencontrer d'autres enfants
>
> d'autres appartements aménagés
> d'autres maisons
> les autres pièces
> Et il y a d'autres… il y a des nuisances
> qui sont plus jeunes que les autres
> découvrirait d'autres choses
> il y a d'autres associations de l'immeuble
> d'autres associations à l'extérieur
> ou d'autres choses comme ça
> les autres immeubles

L'exemple précédent propose, par contre, un mot de grand intérêt qui ne figure pas dans le texte initial de la variable : le mot « rencontre » (sous toutes ses formes).
Ce mot semble apporter une solution. Cependant, dans un autre entretien qui partage la variable, l'entretien 402, ce mot n'apparaît qu'une fois dans les réponses :

> *vous allez en rencontrer des gens comme ça, mais il faut pas qu'ils disent*

Cette utilisation est avant tout une interpellation, ou mise en garde, pour l'enquêteur. Dans ce dernier entretien, le mot « affinité » n'est jamais utilisé, et le mot « autres » n'est présent que dans le cadre d'une utilisation commune.

Dans ce cas, la construction du modèle pourrait être aidée par un dictionnaire de synonymes et un réseau de cooccurrences. La construction du modèle n'en reste pas moins complexe et liée à l'application. Il est évident que le modèle créé pour un corpus sur des unités d'habitation ne peut être réutilisé pour un corpus sur les religions ou sur le passage des traditions.

Certaines méthodes proposent des traitements sans création d'un modèle.
[RAJ 04] s'intéresse au traitement des requêtes dans une collection de textes indexés. Il s'intéresse aussi à l'extraction de termes dans des textes scientifiques ou ayant un vocabulaire spécialisé limité.

[TOR 02] propose, lui, un traitement encore plus général. Une lemmatisation et une segmentation sont réalisées afin de créer une liste d'expressions. Les expressions sont des

suites de termes délimitées par les marques de ponctuation « . », « ! », « ? » et « ; ». Les expressions sont représentées par des vecteurs booléens indiquant si un terme est utilisé au moins deux fois.
La liste des expressions est, alors, soumise aux votes de neuf méthodes (3 mesures statistiques, l'entropie, 3 mesures dérivées de la distance de Hamming, 2 mesures mixtes). A la fin, un algorithme de décision est utilisé afin de déterminer si le segment est conservé.

Les méthodes sont différentes, mais les difficultés restent les mêmes. Pour les deux méthodes, le manque de vocabulaire spécifique rend complexe le traitement des entretiens sociologiques.

Quel que soit le type de traitement choisi, les entretiens oraux retranscrits constituent une difficulté particulière. Cette difficulté tient dans la nature même des textes. Le vocabulaire utilisé dans le langage oral est naturellement simple. Il est dit que 1000 mots suffisent à tenir des discussions. Les entretiens sociologiques abordent des problèmes simples d'un point de vue du vocabulaire. C'est-à-dire que les problèmes abordés ne demandent pas l'utilisation de mots spécifiques comme on pourrait en trouver dans des textes scientifiques.
D'une manière générale, la plus grande difficulté de l'extraction des variables réside dans le fait qu'elle est exploratoire. L'entretien est orienté par des hypothèses qui ont été émises lorsque la problématique a été posée. Mais, de manière évidente, ces hypothèses ne peuvent pas couvrir toutes les variables susceptibles d'être trouvées lors des entretiens.

2.2.4. Comparaison de Requêtes

La comparaison de requêtes est un problème dérivé de la recherche d'information.
Le but de la recherche d'information consiste à répondre à une requête. Cela revient à fournir des données ayant un maximum de cohérence avec la requête. Cela nécessite donc une comparaison du texte de la requête avec le corpus d'application.
Une comparaison de deux requêtes revient à comparer deux phrases courtes (une dizaine de mots).

[KAR 94] aborde ce problème pour trouver des réponses aux problèmes d'utilisateurs. Chaque problème est identifié un libellé. Apporter une réponse à l'utilisateur consiste à retrouver le libellé qui correspond à la requête qu'il a posée. Par exemple : « ma machine a un problème » ramène au libellé « mon ordinateur ne démarre plus ».
Cette méthode utilise une lemmatisation. Puis elle compare les mots, en utilisant entre autres les n-grammes de caractères qui les constituent. Pour cette comparaison sont extraits l'étymologie, les phonèmes, les morphèmes, les n-grammes de caractères et les contextes et domaines d'utilisation. Puis, les mots sont remplacés par leur synonyme et les mots communs (bruit) sont supprimés. Alors seulement, la comparaison est faite.

Cet exemple est l'unique pour cette partie. Mais, ce système paraît être le plus logique pour une telle comparaison. Du moins, cet exemple laisse entrevoir les différentes étapes d'un tel traitement. Une première étape consiste à effectuer des pré-traitements linguistiques, puis une recherche synonymique pour aboutir à un maximum de similarité. Une suppression du bruit trouve évidemment son utilité et nécessite un dictionnaire des non-mots. Enfin une

comparaison des phrases qu'elle soit par n-grammes de caractères, du type Levenstein ou autre.
Il faut rappeler que le but de ce traitement, dans le cas de l'analyse d'entretiens sociologiques, consiste à comparer les variables extraites entre elles. De plus, ce traitement intervient dans un univers thématique restreint puisque chaque thème est traité séparément.

3. Autres types de méthodes

Afin de ne pas se limiter au choix d'une informatisation de la méthode manuelle, il est normal de se tourner vers d'autres méthodes de traitement d'ordre plus général. Le corpus est constitué d'un ensemble d'entretiens non directifs à questions ouvertes, une première partie s'intéresse donc au traitement des questions ouvertes.
De plus, l'objectif de ce travail consiste à retrouver les groupes sociologiques et leur constitution. Une seconde partie s'intéresse à la classification et à l'indexation des textes de façon plus générale.

3.1. Traitement des Questions Ouvertes

Certains systèmes traitant des entretiens à questions ouvertes encouragent un traitement direct des données plutôt qu'un traitement par étapes comme celui de la méthode manuelle. [ACH 91] émet l'idée que les données orales devraient être traitées telles qu'elles ont été formulées, sans passage par une pré-codification thématique qui est déjà l'aboutissement d'une analyse et d'un jugement. A quoi [LEI 95] ajoute que ceci est valable pour toutes les données orales (interview, histoires de vie, conversations), d'autant plus lorsque l'on travaille sur des retranscriptions, la « donnée » a déjà subi de nombreuses modifications : le passage de la situation de face à face, à la bande magnétique et le passage de l'enregistrement de l'oral, à la transcription.

Dans la communauté de l'analyse du discours et du traitement de données textuelles en général, le traitement des questions ouvertes apparaît comme une particularité d'application mais non comme une réelle particularité de traitement. Le traitement se borne la plupart du temps à ne traiter que quelques questions ouvertes et non tout un entretien.
[BEC 98] se sert ainsi d'une analyse factorielle sur les mots les plus fréquents utilisés dans deux questions ouvertes d'ordre politique pour mettre en évidence les pensées des différentes classes d'âge. Cependant, le but de l'étude reste, avant tout, de trouver le moyen de réussir à juxtaposer plusieurs analyses. Ceci permettrait d'obtenir un résultat global représentatif des résultats locaux. L'étude a été poussée jusqu'au traitement de trois questions [BEC 00].
Une analyse des questions indépendamment les unes des autres est parfois préférable, afin d'obtenir une meilleure extraction des caractéristiques [GAR 98].

Les questions ouvertes permettent aussi un traitement conjoint avec des réponses fermées [BEC 02]. Lors d'une analyse conjointe, l'utilisation de la valeur test issue de la statistique textuelle de Lebart et Salem [LEB 94] permet d'améliorer les résultats globaux des enquêtes par questionnaire. Pour cela, il est fait une combinaison de la valeur test sur les mots des questions ouvertes et une analyse discriminante sur les questions fermées.

Mais souvent, le traitement de questions ouvertes, ou de discours, est réalisé à l'aide de logiciels déjà existants. [GAR 98] utilise ainsi ALCESTE [LOG ALC] afin de déterminer les subtilités d'un traitement séparé d'une question et de sa relance et d'un traitement joint de deux questions. ALCESTE est aussi utilisé, plus récemment, par [MOI 02] afin d'extraire les thèmes abordés dans la réponse à la question posée. Une modélisation fréquentielle des thèmes abordés sert ainsi de base à un traitement statistique et d'analyse de diversité afin d'évaluer l'importance des thèmes et donc la richesse et la diversité du corpus.
[LEI 95] est l'un des seuls à traiter des questionnaires oraux retranscrits dans leur intégralité. Il utilise pour cela le logiciel LEXICO [LOG LEX].
Le traitement des questions ouvertes apparaît donc le plus souvent comme le traitement ponctuel d'une ou plusieurs questions, le plus souvent dirigées (c'est-à-dire que les questions sont identiques pour tous les questionnaires). Le traitement global des entretiens se limite le plus souvent à l'utilisation d'un logiciel d'analyse textuelle.

3.2. Méthodes Générales sur l'Indexation et/ou la Classification de Textes

L'indexation de textes consiste à trouver une représentation vectorielle d'un texte. Ce vecteur contient les caractéristiques propres au texte et il est, la plupart du temps, utilisé pour permettre une recherche rapide de textes ou d'informations présentes dans les textes.
La classification de textes permet de créer un système de classes, indépendantes ou imbriquées, de textes. La notion de distance peut être associée au classement afin de déterminer la proximité des classes les unes par rapport aux autres.

La classification de textes trouve de multiples applications [SEB 02] : indexation automatique pour les systèmes booléens de recherche d'information, l'organisation de documents, le filtrage documentaire, la désambiguïsation du sens des mots, les moteurs de recherche.
La majeure partie du temps, la classification de textes consiste à créer un index de type numérique et à appliquer un classifieur sur l'index. Le nombre de méthodes développées dans ce domaine est si important qu'il est impossible d'en faire une liste exhaustive. Cette partie propose donc un aperçu aussi représentatif que possible du domaine.

La représentation la plus commune des textes est le vecteur de fréquence [SAL 89]. Le passage du corpus à une telle représentation crée donc une matrice à deux dimensions. Les colonnes représentent les documents. Les lignes représentent les différents mots du corpus. Chaque valeur de la matrice indique le nombre de fois où le mot apparaît dans le texte.
[CAV 94] utilise les vecteurs de fréquence avec les n-grammes de caractères et non les mots. [LAB 01] étudie des fréquences normalisées. Il espère ainsi ôter le biais amené par les différences de longueur des textes. [MOT 01] propose, pour améliorer l'indexation, de passer à la représentation en trois dimensions. La troisième dimension sert à représenter des informations de type structurel (balises). [DEV 00] représente les textes essentiellement par leurs caractéristiques structurelles : nombre total de mots, longueur moyenne des mots, nombre de phrases, …, fréquence d'apparition des mots outils.

L'indexation fait parfois intervenir des méthodes liées davantage au traitement de la langue naturelle [RAJ 04]. Ainsi, [DAS 04] ne se contente pas uniquement des mots, mais ajoute à la représentation des termes (préfixes, suffixes, …), des paires de mots, …
Cependant d'autres choisissent une autre modélisation. [SAN 02-1] représente chaque texte sous forme d'un graphe. Ce graphe est construit à partir des relations linguistiques que les unités textuelles les plus pertinentes partagent entre elles. Pour cette représentation particulière, les fonctions nécessaires au traitement sont liées à la représentation. Ainsi la classification consiste à trouver les composantes connexes des graphes et à appliquer une Classification Ascendante Hiérarchique avec saut minimum.

Pour les représentations sous forme vectorielle, une étape optionnelle existe : la sélection des caractéristiques. Cette étape consiste à pondérer la représentation afin de mettre en valeur certains mots plus que d'autres. Le but usuel de cette étape est l'élimination partielle des mots outils par une pondération proche de 0. Pour cela, plusieurs méthodes sont disponibles.

- Le tf*idf, [SAL 94] et [REN 03], essaie de valoriser les mots spécifiques de chaque document, en relativisant les fréquences.
- Le DF (Document Frequency) [ROG 02] [LIU 03] est le nombre de documents dans lesquels apparaît chaque caractéristique.
- L'IG (Information Gain / Gain d'Information) [ROG 02] [LIU 03] est la quantité d'information apportée par chaque caractéristique aux classes.
- Le Chi² [ROG 02] [LIU 03] mesure le manque d'indépendance entre une caractéristique et une catégorie.
- Le TS (Term Strength / Force d'un Terme) [LIU 03] mesure la probabilité qu'une caractéristique apparaisse dans deux documents similaires.
- L'En (Entropy / Entropie) [LIU 03] qui est la réduction d'entropie d'un document lorsque la caractéristique est ôtée
- Le TC (Term Contribution / Contribution d'un Terme) qui correspond à la transposée de la similarité de Salton. Le TC se veut plus performant que le DF.
- La longueur normée [REN 03] résout les problèmes liés aux différences de taille entre les documents.

La sélection de caractéristiques est un sujet d'étude à part entière. Il existe aussi, dans ce domaine, une grande quantité de méthodes proposées et parfois propres à un problème. On peut encore citer, par exemple, [DAS 04] qui aborde le sujet sous un aspect plus linguistique.

Une fois que l'index est réalisé, il peut être utilisé, dans le cadre d'une recherche d'information. Il peut aussi être utilisé pour calculer les distances entre les textes ou en entrée d'un classifieur. [LAB 01] utilise la mesure de distance de Bray-Curtis. Salton [SAL 94] utilise la mesure du cosinus. [LEL 02] compare au cosinus, précédemment cité, la mesure du Chi² et le cosinus dans l'espace distributionnel.
D'un point de vue des classifieurs, [MOR 02] utilise une Analyse Factorielle des Correspondances (AFC). [TAN 05] améliore les k-plus proches voisins en pondérant les classes suivant le nombre d'individus qu'elles contiennent. [JOR 05] tente d'extraire le contexte d'utilisation des mots (différence entre le lit pour dormir et le lit de la rivière) en utilisant un réseau de neurones. [CAV 94] calcule, avec sa méthode semi-supervisée, un profil pour chaque classe puis la distance de chaque texte à chacun des profils. [SHA 00] calcule,

dans le même genre, les centroïdes des classes et les normalise. [DEV 00] opte pour les SVM. [ZHA 00] compare les SVM au perceptron et à des algorithmes Winnow (entre le perceptron et les SVM). [REN 03] préfère les Réseaux Bayésiens. [KLO 05] utilise les réseaux Bayésiens sous forme arborée. [SIN 04] utilise les K-means, et [BRO 99] les Chaîne de Markov Cachées (CMC / HMM).
La classification de textes tient sa spécificité de la représentation qui est faite des données textuelles, c'est-à-dire de l'indexation qui est faite des textes. Le reste du traitement suit un schéma classique de classification. Un schéma identique a déjà été observé, précédemment, pour la classification thématique [FOR 00].
Certaines méthodes sortent de ce schéma. Dans le cadre de la classification thématique, une indexation originale consiste à représenter chaque texte par une image [REY 94]. Une carte des mots est tracée : si un mot apparaît, dans un texte aux positions x et y alors les points (x,x), (x,y), (y,x) et (y,y) sont marqués. La segmentation thématique est réalisée à partir d'une analyse des images : étude de l'évolution de la densité.

Cette confusion des méthodes amène avant tout une question très importante : qu'est-ce qui est classé ? Ce sont évidemment les textes qui sont classés. Mais il faut y voir une question bien plus profonde : sur quelle base les textes sont-ils classés ?
Les deux dernières étapes (sélection de caractéristiques et classifieurs) sont générales et trouvent leur application dans tous les domaines de classification (images, sons, textes, …). La sélection des caractéristiques et la classification traitent des données numériques et cela peu importe ce que ces données représentent.
L'indexation des textes est donc une étape primordiale. C'est la première étape du traitement et elle consiste à passer d'un texte à une forme vectorielle numérique (ou autre). De nombreuses méthodes choisissent de représenter le contenu des textes. [DEV 00] choisit de représenter leurs informations structurelles.
Ainsi, la classification textuelle se heurte à deux sortes de difficultés. Il y a les difficultés « habituelles » dans le domaine général de la classification. Cela signifie qu'il faut trouver le meilleur système de classification pour les données à traiter. Mais, il y a, avant tout, la difficulté de la représentation des textes du corpus, car c'est sur la base de cette représentation que les textes sont classés.

4. Schéma général de la solution retenue

Ce chapitre a présenté différentes solutions à ce travail : des solutions logicielles, une solution manuelle pouvant être informatisée et des solutions d'ordre plus général.

L'analyse qui a été faite des logiciels indique qu'aucun ne peut être retenu comme une solution au problème posé par ce travail.
La solution manuelle a prouvé son efficacité. Cependant, chaque étape de son informatisation se heurte au problème de la langue orale. De plus, chacune des étapes nécessite la mise en place d'un certain nombre d'outils (outils linguistiques, dictionnaires, lexiques, …) assez coûteux pour une utilisation restreinte. En effet, des thèmes sont traités de manière récurrente, mais la liste des thèmes n'est pas fermée. Pour chaque nouveau thème étudié, il est donc nécessaire de compléter les outils existants. Pour cela, il faut définir une base d'apprentissage, ce qui est complexe à cause du problème de la langue orale et du manque de bases

d'apprentissage dans le domaine. Pour ces diverses raisons, il a été choisi de ne pas retenir cette solution.

La solution retenue est une solution en trois étapes semblable à la majeure partie des méthodes de classification.

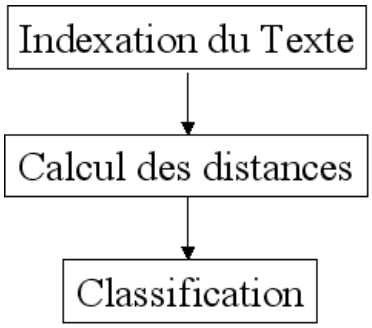

Figure 1 Schéma général du Traitement en 3 Etapes

Dans une telle chaîne de traitements, les premiers traitements ne peuvent pas être fixés sans que les derniers le soient. C'est la raison pour laquelle, il est préférable de remonter la chaîne de traitements. Ainsi, une fois la classification fixée, les distances peuvent être étudiées et fixées, etc.

En remontant la chaîne de traitements, on trouve tout d'abord la troisième étape. Elle consiste à classer les textes indexés et à offrir une visualisation aussi explicite que possible des proximités existant entre les textes. C'est pour cette raison qu'il a été choisi d'effectuer une classification par arbre. Un arbre est une structure facilement compréhensible. De plus, la classification par arbre fonctionne par voisinage. Cela signifie qu'un arbre présente à la fois le comportement global des textes les uns par rapport aux autres, mais aussi leur comportement local. Dans le cadre de ce travail, cet aspect apparaît comme un avantage. Un système de classe, comme d'autres méthodes pourraient en fournir, peut sembler trop strict lorsque l'on traite des entretiens. Il paraît moins surprenant de parler d'une forte dissimilarité plutôt que de parler de deux classes distinctes. Le chapitre 2 étudie différentes méthodes de classification arborée et propose une sélection.

La seconde étape a pour but de calculer les dissimilarités entre les entretiens. Ce calcul est une étape importante, car si l'indexation essaie de représenter les textes eux-mêmes, le calcul des dissimilarité tâche de représenter la proximité des textes entre eux. Il est donc nécessaire de trouver une mesure de dissimilarité fiable du point de vue de la classification. Le chapitre 3 propose et étudie deux sortes de normalisation des données et une liste aussi exhaustive que possible de mesures. A partir d'expérimentations, une sélection est faite parmi le nombre important des mesures proposées.

Enfin, l'étape d'indexation consiste à passer de manière automatique d'un texte brut à une représentation numérique du texte. Il a été vu dans la partie 3.2 que l'indexation des textes est l'étape primordiale et qu'elle dépend complètement des objectifs de classification. Il est donc

normal que ce soit cette étape qui ait demandé le plus de travail de réflexion. Le but est de fournir une méthode d'indexation qui offre, à la suite de la chaîne de traitements, la représentation la plus significative de chaque entretien.
Les méthodes actuelles proposent une représentation du contenu à partir de vecteurs de fréquences ou proposent une représentation de la structure à partir de caractéristiques structurelles des textes.
Dans le cas d'entretiens sociologiques oraux retranscrits, la structure est naturellement inexistante, car la forme écrite n'est pas la forme originale.
Le contenu ne semble, a priori, pas non plus une solution à l'indexation. Le problème du langage naturel oral a été soulevé dans la partie 2.2 sur l'informatisation de la méthode manuelle. Ce problème apparaît comme une difficulté dans la représentation des entretiens. En effet, les entretiens utilisent le langage oral courant, sans spécificité de vocabulaire.
De plus, tous les thèmes sont abordés par tous les entretiens. Le contenu semble donc pauvre et commun à tous les entretiens. Il semble donc difficile à une représentation par vecteurs de distinguer les textes les uns par rapport aux autres.
Une solution est proposée par ce travail. Il est, en effet, émis l'hypothèse que le sens est porté par la structure des textes plus que par leur contenu. Cela signifie que l'organisation du discours serait propre à chacun. Il est donc proposé de classer les entretiens non pas sur leur fond, mais sur leur forme.
La chapitre 4 présente les méthodes d'indexation citées précédemment et présente des méthodes qui essaient de modéliser la forme du langage. Deux nouvelles sortes de méthodes sont proposées. Tout d'abord, il est proposé trois méthodes qui étudient la structure globale des textes. Une première méthode propose de représenter la structure d'un texte par une image de taille fixe. Les deux méthodes suivantes utilisent un automate pour construire une méthode à l'aide de la structure.
Puis, il est proposé une méthode qui étudie la dynamique des textes. Cette méthode s'intéresse donc à l'évolution textuelle. C'est en essayant de représenter la structure et la dynamique que ce travail tente de modéliser le raisonnement.

Chapitre 2 - Classifications et Visualisation

Ce chapitre est consacré aux méthodes de classification et à la méthode de visualisation adoptées pour ce projet.
Comme expliqué dans la présentation générale de la méthode suivie par ce projet, le choix a été fait d'utiliser des méthodes de classification par arbre. Ces méthodes ont un double avantage. Tout d'abord, elles ne définissent pas des classes en tant que telles, mais gardent la notion de proximité des éléments à classer. En sociologie, comme certainement en littérature en général, les classes ne sont pas fermées, c'est la raison pour laquelle une présentation des distances est préférable. Puis, le second avantage est la représentation elle-même de la classification. Une représentation arborée est une représentation simple et explicite. Ces deux caractéristiques en font une représentation idéale et largement utilisée dans l'étude des textes [BAR 98], [DUB 99], [LAB 01], mais aussi et surtout en phylogénie où elle fait l'objet de recherche [GUI 03].
La phylogénie peut être considérée comme la construction de l'histoire évolutive d'un ensemble d'espèces. Les relations qui existent entre les espèces sont alors représentées sous forme d'un arbre (le plus souvent binaire).La phylogénie trouve de nombreuses applications en biologie : classification, taxonomie, tests de paternité, études des variations géographiques, mise en évidence de nouvelles espèces, analyse des comportements reproducteurs, recherche virale...
Il existe plusieurs méthodes de construction d'arbres phylogénétiques que l'on peut classer en trois catégories principales :
– Les méthodes de distances telles que UPGMA et WPGMA.
– Les méthodes probabilistes qui ont recours à un modèle de l'évolution et qui se basent sur une analyse élémentaire des caractères.
– Les méthodes cladistiques qui se basent également sur l'analyse des caractères.
Compte tenu des étapes qui précèdent dans le déroulement général de la méthode, seules les méthodes de distances trouvent un intérêt dans ce projet.
La première partie de ce chapitre présente, en détail, des méthodes issues de la phylogénie, dont certaines sont issues du domaine de la classification, et une méthode développée pour l'étude de textes. La seconde partie présente deux visualisations possibles pour la représentation d'arbre. Elle présente aussi le choix fait pour ce projet et les raisons de ce choix. La troisième partie présente des expérimentations menées sur de simples jeux de distances afin d'exposer, de manière concrète, les différences entre ces classifications. Enfin, une dernière partie fait le bilan du chapitre.

1. Arbres de Distances

Cette partie a pour but de présenter un ensemble de méthodes de classification par arbres à partir des distances entre les éléments. L'étape précédente consiste donc à calculer la matrice des distances. Cette étape faisant l'objet d'un chapitre d'étude, on considère, ici, simplement que la matrice des distances correspond à un tableau à double entrée, dans lequel on retrouve la valeur des distances entre les éléments pris deux à deux.
Le principe général des arbres de distance est de réduire cette matrice à un unique élément. Pour cela, un couple, de distance minimale, est choisi parmi l'ensemble des couples d'éléments possibles. Ce couple est réuni pour former une classe, représentée à présent

comme un unique élément. La matrice est alors réduite et les distances sont recalculées. Le regroupement hiérarchique des éléments permet d'obtenir un arbre.

Cependant, dans le détail, chaque méthode est spécifique. Le re-calcul des distances est l'une des grandes spécificités de chaque méthode. Cette partie présentera dans un premier temps les méthodes appelées en phylogénie « Unweighted Pair Group Method with Averages » (UPGMA), « Weighted Pair Group Method with Averages » (WPGMA) ainsi que les méthodes de « Lien Simple » et de « Lien Complet ». Ces méthodes sont les standards de la classification arborée.
Puis il sera étudié deux variantes : le « Neighbor Joining » (NJ) et l'ADDTREE. Ces deux méthodes résolvent les problèmes de distances ultramétriques.
Enfin, il sera vu la méthode des groupements, régulièrement utilisée pour l'étude des textes.
Toutes ces méthodes sont des méthodes de classification hiérarchique non supervisée (CAH), qui permettent une utilisation de manière complètement automatique.

1.1. UPGMA, WPGMA, Lien Simple et Lien Complet

UPGMA a été présenté par Sokal et Michener en 1958 [SOK 58] et WPGMA par McQuitty en 1966 [MCQ 66]. Ces deux techniques proposent une utilisation de la moyenne (Average) des distances lors du re-calcul des distances, contrairement aux deux autres méthodes, Lien Simple et Lien Complet, qui choisissent l'une des deux distances. Quoiqu'il en soit ces quatre méthodes suivent l'algorithme des méthodes du type PGMA énoncé précédemment.
C'est-à-dire qu'ayant la matrice des distances en entrée, une sélection dans les couples est faite afin de choisir le couple présentant la plus petite distance. Les deux membres du couple sont regroupés en une classe. La matrice est réduite et les distances sont re-calculées autour de la classe créée. Cette dernière apparaît donc, d'un point de vue des distances comme un élément parmi les autres.
C'est donc dans le re-calcul des distances que ces méthodes se différencient. Les méthodes du Lien Simple et du Complet choisissent, respectivement et de manière catégorique, la distance minimum et la distance maximum du couple formé, aux autres éléments.
Les méthodes UPGMA et WPGMA proposent, comme écrit précédemment, une représentation plus ou moins égale des membres du couple. La distance finale est de la forme :

$$d(C,k)=\alpha_i \times d(i,k)+\alpha_j \times d(j,k) \quad (1)$$

où C représente la classe formée, i et j les deux membres du couple formant la classe et k l'élément pour lequel la distance est calculée. α_i et α_j représentent les pondérations de chacune des distances. Il est évident que α_i et α_j respectent la condition : $\alpha_i + \alpha_j = 1$.

La nouvelle distance calculée par WPGMA sera une simple moyenne, c'est-à-dire que $\alpha_i = ½$ et $\alpha_j = ½$. La nouvelle distance calculée par UPGMA tient quant à elle compte de la cardinalité des deux ensembles réunis et laisse, ainsi, un poids plus important au membre qui domine le couple. Dans les premières itérations, chaque membre du couple aura donc un poids équivalent, puis, à force de regroupement, selon la classification, les poids pourront rester équivalents ou l'un pourra être plus important que l'autre. Si on note |I| et |J|, les

cardinalités des ensembles représentés par, respectivement, i et j, les poids associés à chaque membre du couple sont :

$$\alpha_i = \frac{|I|}{|I|+|J|} \text{ et } \alpha_j = \frac{|J|}{|I|+|J|} \quad (2)$$

Compte tenu de leur forme, ces deux méthodes sont souvent confondues dans la littérature. En effet, si la méthode WPGMA, associant le même poids aux deux membres du couple, apparaît, d'un point de vue visuel, comme une méthode non-pondérée, elle suit, d'un point de vue théorique, une pondération inversement proportionnelle à l'effectif de chaque sous-arbre.

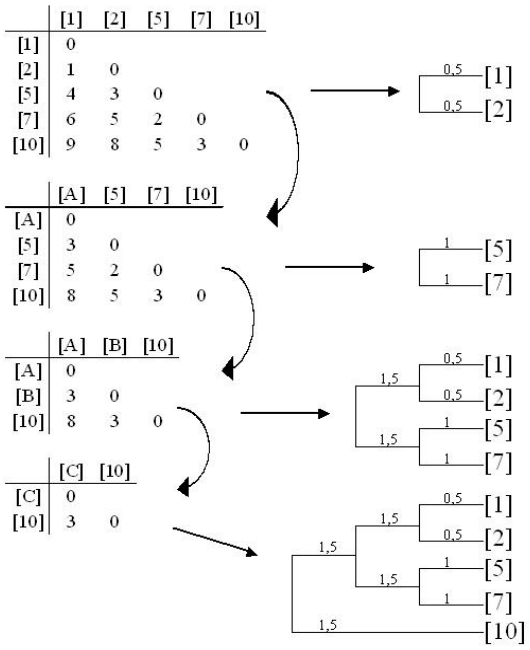

Figure 2 Déroulement de la méthode de Lien Simple

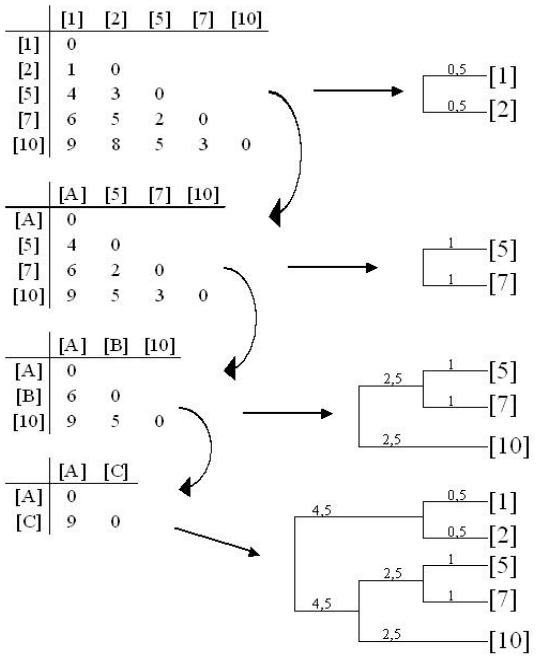

Figure 3 Déroulement de la méthode de Lien Complet

Les figures précédentes présentent le déroulement de chacune des méthodes vues précédemment, sur un même exemple. Le jeu de données utilisé pour ces exemples est fort simple. Il s'agit d'un ensemble de 5 éléments dont la distance correspond à la différence entre les chiffres qu'ils représentent. Ainsi [1] est à une distance de 1 de [2] et à une distance de 4 de [5], …

Dans la Figure 2, le déroulement de la méthode de Lien Simple laisse lors de l'avant-dernière itération le choix entre deux regroupements possibles. Le choix s'est ici porté selon l'ordre du tableau. Cet exemple en particulier présente l'un des inconvénients de ce type de classification. En effet, le choix de l'autre regroupement influe sur toute la formation de l'arbre. Le choix qui a été fait a, d'ailleurs, pour résultat un arbre dont la structure générale est différente de celles obtenues par les autres méthodes. C'est-à-dire, qu'ici l'élément 10 n'appartient à aucune autre classe si ce n'est à la classe globale.

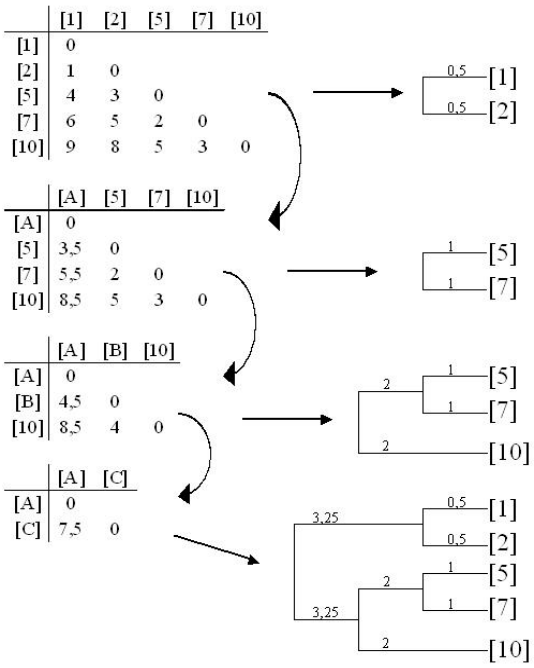

Figure 4 Déroulement de la méthode WPGMA

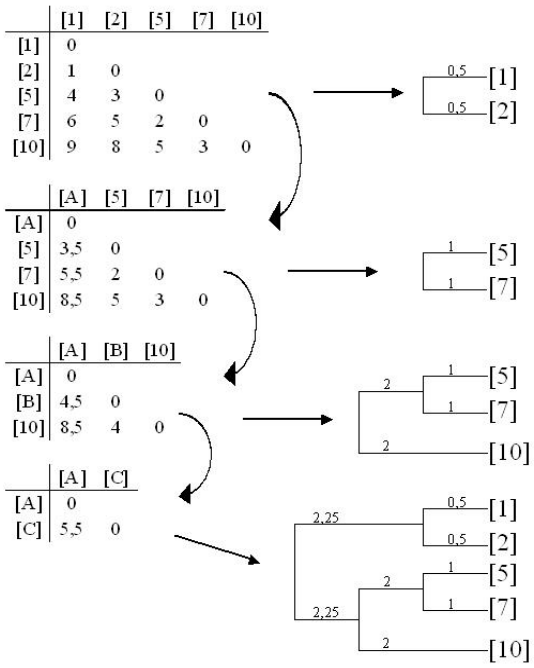

Figure 5 Déroulement de la méthode UPGMA

La Figure 3, comme les figures suivantes, ne présente, quant à elle, aucun choix. Cette méthode se démarque surtout par la taille des derniers liens. En effet, la méthode des liens complets conserve chaque fois le maximum des distances. Lors des dernières itérations, il ne reste plus que les plus grandes distances. La distance entre l'élément [1] ou [2] et l'élément [10] passe ainsi à 12. Cette même distance était de 5 avec la méthode du lien simple.
Avec respectivement 9,5 et 7, ce sont donc les méthodes WPGMA et UPGMA qui donnent la meilleure représentation de la distance initiale. La Figure 4 et la Figure 5 montrent d'ailleurs que sur un tel jeu de données, la différence entre ces deux méthodes n'est visible qu'à la fin, lorsque des regroupements avec des classes sont faits. C'est donc, pour cet exemple, sur le calcul de la dernière distance que ces deux méthodes se différencient.
Cependant, il faut signaler que les méthodes WPGMA et UPGMA ne trouvent LE bon arbre que si et seulement si il existe un arbre ultramétrique.
La racine, qui n'est pas représentée sur les figures précédentes, correspond au point le plus à gauche des arbres représentés. C'est-à-dire que la racine se trouve, sur l'arbre, à l'opposé des feuilles. Un arbre est ultramétrique si les feuilles sont équidistantes de la racine. En partant de la matrice des distances, on peut définir une matrice symétrique comme étant ultramétrique si et seulement si, pour tout i, j et k, max(d(i,j) ; d(i,k) ; d(j,k)) n'est pas unique. Enfin, une matrice symétrique admet un arbre ultramétrique si et seulement si elle est ultramétrique.
Dans l'exemple qui précède, il suffit de prendre les 3 premiers éléments, [1], [2] et [5], pour prouver que la matrice n'est pas ultramétrique. Il n'existe donc pas d'arbre ultramétrique associé à cette matrice de distances. Les méthodes présentées ici ne présentent donc pas la meilleure représentation possible.
Les méthodes NJ et ADDTREE permettent cependant d'établir l'arbre correct à partir de la matrice des distances sans que la matrice n'ait le besoin d'être ultramétriques. Les parties suivantes présentent chacune de ces méthodes.

1.2. NJ

Le Neighbor Joining a été présenté en 1987 par Saitou et Nei [SAI 87]. Cette méthode, à l'inverse des deux précédentes, autorise ce qui est appelé en phylogénie des taux de mutations différents sur les branches. Cela se traduit par des branches de longueurs différentes. Par rapport aux méthodes de type PGMA, toutes les feuilles ne sont pas situées à égale distance de la racine, c'est-à-dire que les arbres finaux ne sont pas ultramétriques.
La méthode NJ est une heuristique basée sur le minimum d'évolution. C'est un algorithme d'agglomération qui n'examine pas toutes les topologies possibles.
Cette méthode a pour but de montrer les voisinages (Neighbor) entre deux groupes d'éléments reliés par un seul nœud dans un arbre sans racine. Ainsi si on considère deux voisins [1] et [2] comme n'étant qu'un seul groupe [A], alors [A] est voisin de [B], qui regroupe [5] et [7]. Le voisin de [1] reste [2], mais c'est au niveau de la classe que se forme un nouveau lien de voisinage.
Pour construire un tel arbre, il faut partir d'un arbre en étoile où tous les éléments sont reliés à un même ancêtre commun, c'est-à-dire qu'aucune classe n'est formée. Il faut ensuite chercher le couple d'éléments i et j, qui, une fois groupés en classe, minimise la longueur totale (estimée aux moindres carrés) de l'arbre. Puis réduction de la matrice et re-calcul des distances. Ainsi de suite jusqu'à ce que l'arbre entier ne forme plus qu'une classe.
Cela revient donc à minimiser :

$$C(i,j)=(n-2)d(i,j)-(R_i+R_j) \text{ où } R_x=\sum_{k=1}^{n} d(x,k) \tag{3}$$

La longueur entre un élément i et le nœud représentant la classe A (formée à partir des éléments i et j) est calculée par l'équations suivante :

$$d(A,i)=\frac{(n-2)d(i,j)+R_i-R_j}{2(n-2)} \tag{4}$$

Enfin la distance entre un couple d'éléments et les autres éléments est calculée de la manière suivante :

$$d((i,j),k)=\frac{1}{2}(d(i,k)+d(j,k)-d(i,j)) \tag{5}$$

On arrive donc à l'algorithme suivant :

Algorithme NJ

Tant qu'il y a plus d'un élément dans la matrice

1. Pour chaque Feuille calculer $u(i)=\sum_{i\neq k;k=1}^{n}\frac{d(i,k)}{n-2}$
2. Choisir tous les couples i et j pour lesquels d(i,j)-u(i)-u(j) est minimum
3. Pour chaque couple :
   3.1. Joindre tous les couples i et j : Calculer les longueurs v(i) et v(j) de i et j au nouveau nœud avec : $v(i)=\frac{d(i,j)+(u(i)-u(j))}{2}$ et $v(j)=\frac{d(i,j)+(u(j)-u(i))}{2}$
   3.2. Calculer la distance entre le nouveau nœud et chacun des autres éléments.
4. Eliminer les colonnes et les lignes correspondant aux éléments i et j et ajouter celles correspondant au nouvel élément

Comme pour les méthodes du type PGMA, un exemple de déroulement est donné dans la figure qui suit.

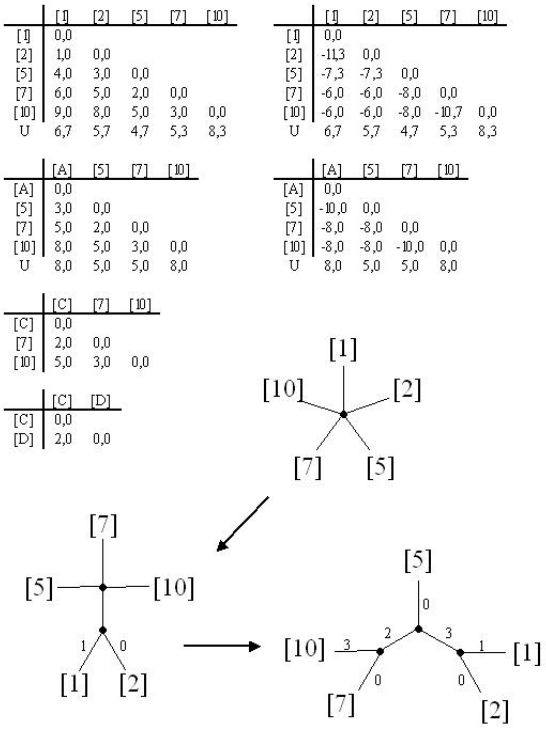

Figure 6 Déroulement de la méthode NJ

La Figure 6 ne présente pas le calcul des distances entre les éléments et le nœud qui les unit, mais présente, de manière aussi complète que possible, tout le reste de la démarche. A gauche, il y a les distances et les u calculés pour chaque élément. A droite, il y a les critères à minimiser ainsi que les u calculés pour chaque élément.

Il peut être vu que, lors de la seconde itération, deux classes peuvent être formées. Comme l'indique l'algorithme, il suffit de traiter chaque classe séparément. Les distances des éléments au nœud qui les unit ne présentent aucun problème dans le cas présent, il suffit d'utiliser le tableau des distances de gauche. Pour le calcul des distances entre les nœuds formés et les autres éléments, un tableau transitoire est affiché pour montrer que, là aussi, il suffit de calculer successivement les deux distances.
A la fin de cette seconde itération, il ne reste donc plus que deux éléments dans la matrice et la distance qui les séparent.
Le résultat final est un arbre sans racine qui respecte parfaitement les distances originales de la matrice des distances. De plus, comme le montre cet exemple, la construction d'un arbre est très rapide, même si, en détail, les calculs sont plus compliqués que ceux nécessaires aux méthodes de type PGMA.

1.3. ADDTREE

L'ADDTREE a été présenté en 1977 par Sattath et Tversky [SAT 77]. Elle apparaît dans la littérature sous le nom d'ADDTREE, Additive Tree, Preference Trees, Pretree. Cette méthode précède la méthode NJ et en est même la base. En effet, la différence significative entre ces deux méthodes tiens dans la sélection de la paire à grouper. La méthode de l'ADDTREE est basée sur la notion de voisinage.

Ainsi avec un arbre est constitué de quatre éléments (i, j, k, l), on peut former deux couples de voisins (i, j) et (k, l), si la condition des quatre points est satisfaite :

$$d(i,j)+d(k,l)\leq\big(d(i,k)+d(j,l)=d(i,l)+d(j,k)\big) \tag{6}$$

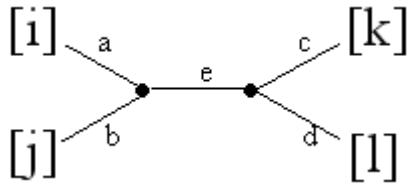

Figure 7 Relation de voisinage dans ADDTREE

Avec la Figure 7, la condition se formule de manière plus explicite :

$$(a+b)+(c+d)\leq\big((a+e+c)+(b+e+d)=(a+e+d)+(b+e+c)\big) \tag{7}$$

La force de voisinage, pour un couple d'éléments (i, j), correspond donc au nombre d'autres couples possibles de l'arbre pour lesquels le couple (i, j) satisfait la condition de voisinage. D'une manière plus générale, cela revient à calculer la distance moyenne qui sépare le couple des autres couples. C'est-à-dire que cela consiste à faire la somme des valeurs possibles de e, où e représente (Figure 7) la distance entre les nœuds associés à chaque couple.

Dans la condition de voisinage, l'égalité à droite de l'inéquation n'étant pas toujours respectée, la distance minimum est donc prise en compte. On peut donc formuler la force de voisinage de la manière suivante :

$$Fv(i,j)_{i<j} = \sum_{k<l;(k,l)\neq(i,j)} \min(d(i,k)+d(j,l);d(i,l)+d(j,k)) - d(k,l) - d(i,j) \quad (8)$$

A chaque itération, le couple sélectionné est donc celui qui montre la plus grande force de voisinage. Une fois le couple créé, comme pour la méthode NJ, la distance entre le nouveau nœud A et les autres éléments est calculée à partir de (5) et la distance d'un élément i du couple au nœud A formé correspond à la moyenne des distances aux éléments en-dehors du couple :

$$d(A,i) = \frac{1}{(n-2)} \sum_{k\neq i;k\neq j;k=1}^{n} \frac{1}{2}(d(i,j)+d(i,k)-d(j,k)) \quad (9)$$

On arrive donc à l'algorithme suivant :

Algorithme ADDTREE

A. Tant qu'il y a plus de 4 éléments dans la matrice

1. Pour chaque couple (i,j) de l'arbre, on calcule la force de voisinage $Fv(i,j)_{i<j} = \sum_{k<l;(k,l)\neq(i,j)} \min(d(i,k)+d(j,l);d(i,l)+d(j,k)) - d(k,l) - d(i,j)$
2. Choisir le couple i et j pour lequel Fv(i,j) est maximum
3. Pour chaque couple :
   - 3.1. Joindre tous les couples i et j : calculer les longueurs d(A,i) et d(A,j) de i et j au nouveau nœud
   - 3.2. Calculer la distance entre le nouveau nœud et chacun des autres éléments.
4. Eliminer les colonnes et les lignes correspondant aux éléments i et j et ajouter celles correspondant au nouvel élément

B. Former les deux derniers couples de voisinage

Comme pour les méthodes précédemment présentées, un exemple de déroulement est donné dans la Figure 8.
Cette figure présente deux séries de tableaux. Les tableaux de gauche affichent les distances. Les tableau de droite présentent les forces de voisinage de chaque couple possible. Une première observation peut être faite sur la force de voisinage. Celle-ci montre un caractère plus continu que si elle avait suivi un algorithme de comptage. Dans la première matrice des

forces de voisinage, il peut ainsi être observé plus le membre du couple associé à l'élément [1] est loin, plus la force de voisinage est faible.

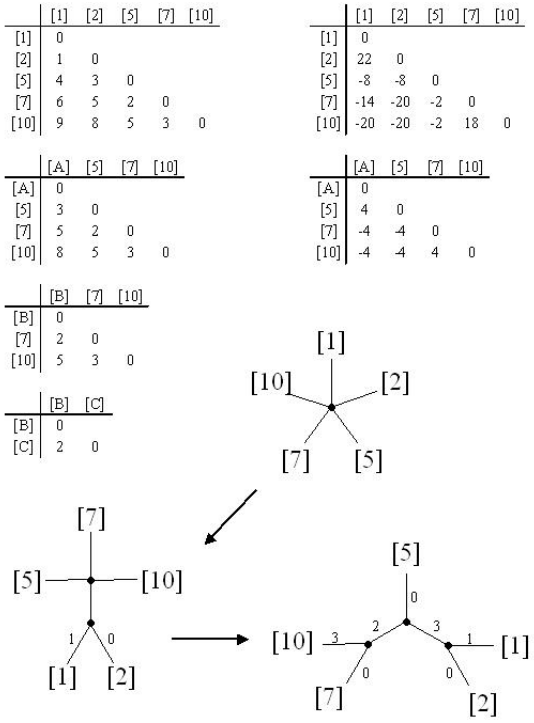

Figure 8 Déroulement de la méthode ADDTREE

Il peut, de plus, être observé que le résultat final est semblable à celui de la méthode de NJ. D'une manière générale, ces deux méthodes ne se différencieraient que dans la vitesse de convergence et dans leur complexité : $O(n^4)$ pour la méthode de l'ADDTREE et seulement $O(n^3)$ pour la méthode de NJ.

1.4. Méthode des Groupements

La méthode des groupements est issue du travail de Jean-Pierre Barthélémy, Alain Guénoche et Xuan Long [BAR 88], [BAR 98], [GUE 01], [GUE 99]. Elle a l'avantage, contrairement aux méthodes présentées précédemment, d'avoir prouvé son utilité dans des applications de classification de textes [BAR 98], [LAB 01].

Cette méthode utilise comme la méthode de l'ADDTREE la notion de voisinage. Elle se sert pour cela de la méthode du comptage de voisins ce qui permet de calculer deux scores. Le premier score, noté s*, est le « score strict » et il correspond à l'utilisation d'une inégalité stricte dans (6). Le second score, noté s, est le « score large » et il correspond au respect de l'égalité tel qu'elle est écrite dans (6). Si λ correspond à la taille minimum d'un groupement de l'arbre, le score fort maximum, noté s**, des couples d'éléments réels de l'arbre vaut :

$$s^{**}=\frac{(n-\lambda)(n-\lambda-1)}{2}\leq\frac{(n-2)(n-3)}{2} \qquad (10)$$

Deux éléments i et j sont ainsi considérés comme pré-voisins si et seulement si, avec λ=1, ils satisfont l'inéquation suivante :

$$s(i,j) \geq s^{**} - (n - \lambda - 1) \tag{11}$$

L'ensemble des pré-voisins constitue des pré-groupements où chaque élément n'apparaît qu'une seule fois. Dans le cas où l'ensemble des pré-voisins est vide, le critère est affaibli au score large maximum obtenu, et chaque pré-groupement devient automatiquement un groupement. Dans le cas où des pré-groupements ont pu être créé à partir d'un ensemble non vide de pré-voisins, chaque pré-groupement ne satisfaisant pas le critère (11) où, cette fois-ci, λ correspond à la taille du pré-groupement est éliminé. A nouveau, dans le cas d'une couverture nulle, ce critère est affaibli.

La suite de l'algorithme conserve les calculs des méthodes précédentes (5) et (10) servant respectivement à calculer la dissimilarité d'un couple à un non-classé et la distance d'un élément au nœud de la classe (formule générale). Il est à noter que ces formules sont, quand même, modifiées afin de tenir compte de la taille λ de la classe formée. Chaque distance correspond ainsi à une représentation moyenne des distances par rapport au groupe.
Enfin, si une itération conduit à une matrice des distances ayant moins de 4 éléments, et néanmoins plus d'1 (si il ne reste qu'un seul élément, plus aucun calcul n'est nécessaire), la formule (5) est utilisée 1 ou 3 fois selon qu'il reste 2 ou 3 éléments.
Cette méthode est donc fortement similaire à celle de l'ADDTREE. La seule réelle différence entre ces deux méthodes tient dans le critère de sélection des couples de voisinage, l'un prend le couple de force maximum, l'autre prend tous les couples ayant une force suffisante.

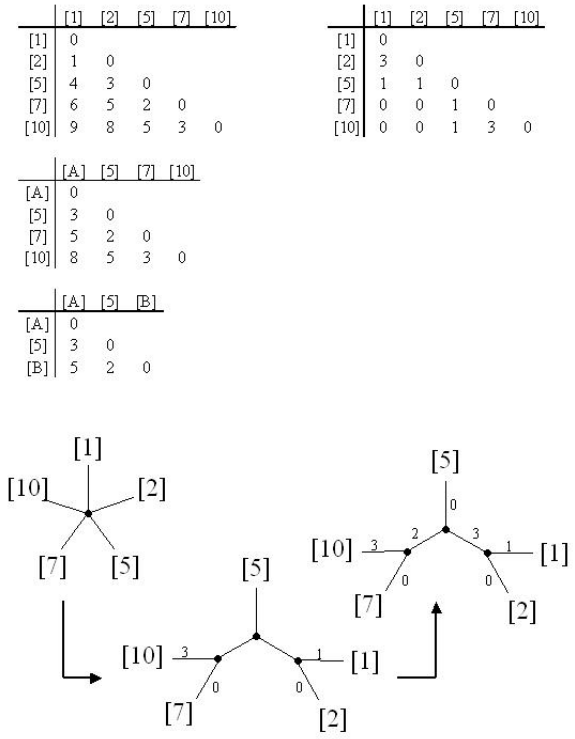

Figure 9 Déroulement de la méthode des Groupements.

La Figure 9 présente le déroulement de cette méthode sur le même exemple que précédemment. A gauche de la figue sont les matrices des distances et à droite les matrice de force de voisinage par la méthode du comptage.

2. Visualisation

Les courtes expérimentations de présentation ont permis de montrer deux types de visualisation, c'est-à-dire deux types d'arbre : l'arbre dit enraciné (UPGMA, …) et le X-arbre ou arbre en étoile (NJ, …). Pour ce projet, il a été choisi de conserver uniquement la visualisation par X-arbre. Tout d'abord, elle est nécessaire au déroulement des trois deniers algorithmes. Ensuite, un unique mode de visualisation permet d'offrir un meilleur terrain aux différentes comparaisons. Enfin, il est préférable de se détacher de toute notion de racine qui pourrait donner à l'arbre des sens incontrôlés. Tandis qu'une représentation en étoile, même si elle possède un centre, paraît montrer une continuité sans frontière d'un bout à l'autre de l'arbre.
Sans entrer dans le détail, une représentation en étoile est obtenue en divisant de manière dichotomique l'espace angulaire offert. Un tel système permet, en effet, d'éviter tout croisement. Il a pourtant, sur des jeux de données importants, l'inconvénient de « tasser » les éléments les uns aux autres et d'autant plus si les distances sont « petites ».
De plus, la notion de distance est, par choix, marquée par la longueur des segments graphiques. La présence chiffrée de ces distances, même si elle est intéressante sur des jeux de données de petite taille comme le jeu utilisé lors de la partie précédente, rend l'affichage brouillon et empêche une bonne appréciation de la représentation.

La Figure 10 présente l'affichage tel qu'il est obtenu à la suite d'une application réelle.
Au-delà des critères mentionnés précédemment, cette représentation prend bien plus de sens que celles de la partie précédente. Une représentation graphique, et non numérique, des distances laisse apercevoir que la classification NJ (il en va de même pour la méthode ADDTREE et la méthode des Groupements) reconstitue la dimension unique à la base du jeu de données.

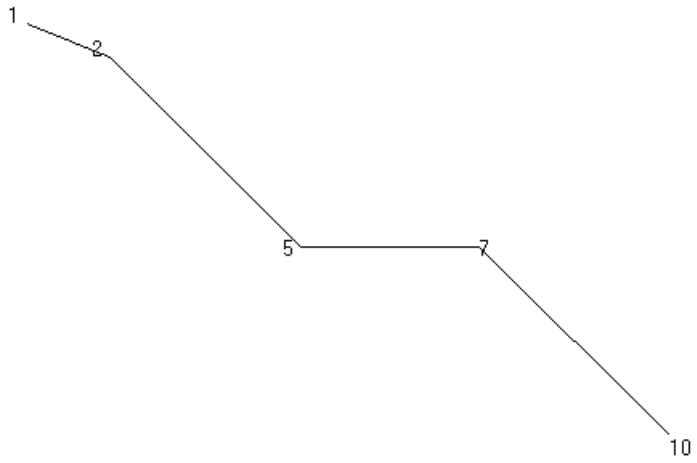

Figure 10 Résultat graphique réel pour la méthode NJ sur le jeu de données 1-2-5-7-10

Les proportions, gardées graphiquement sur les deux axes, permettent même de constater que la distance entre les éléments 2 et 5 est identique à celle entre les éléments 7 et 10. Cette

illustration montre l'efficacité d'une telle représentation ou, tout du moins, montre l'effet de visualisation cherché pour le projet. C'est-à-dire une visualisation simple à comprendre et qui respecte les distances de la matrice initiale.

3. Expérimentations

Pour clore ce chapitre cette partie présente les résultats réels des méthodes obtenus sur deux jeux de données. Le premier est dans la suite du jeu de données exemple, puisqu'il s'agit des distances entre les 101 premiers nombres (0 à 100). Le second jeu de données est issu de la phylogénie est représente les distances génétiques entre divers animaux.

3.1. Suite Numérique

Ce jeu de données est donc constitué des chiffres allant de 0 à 100 qui sont représentés dans les matrices des distances par la distance absolue entre les chiffres. Cette expérimentation n'a aucun sens en soi à part montrer les différences de réactions face à des données mono-dimensionnelles.

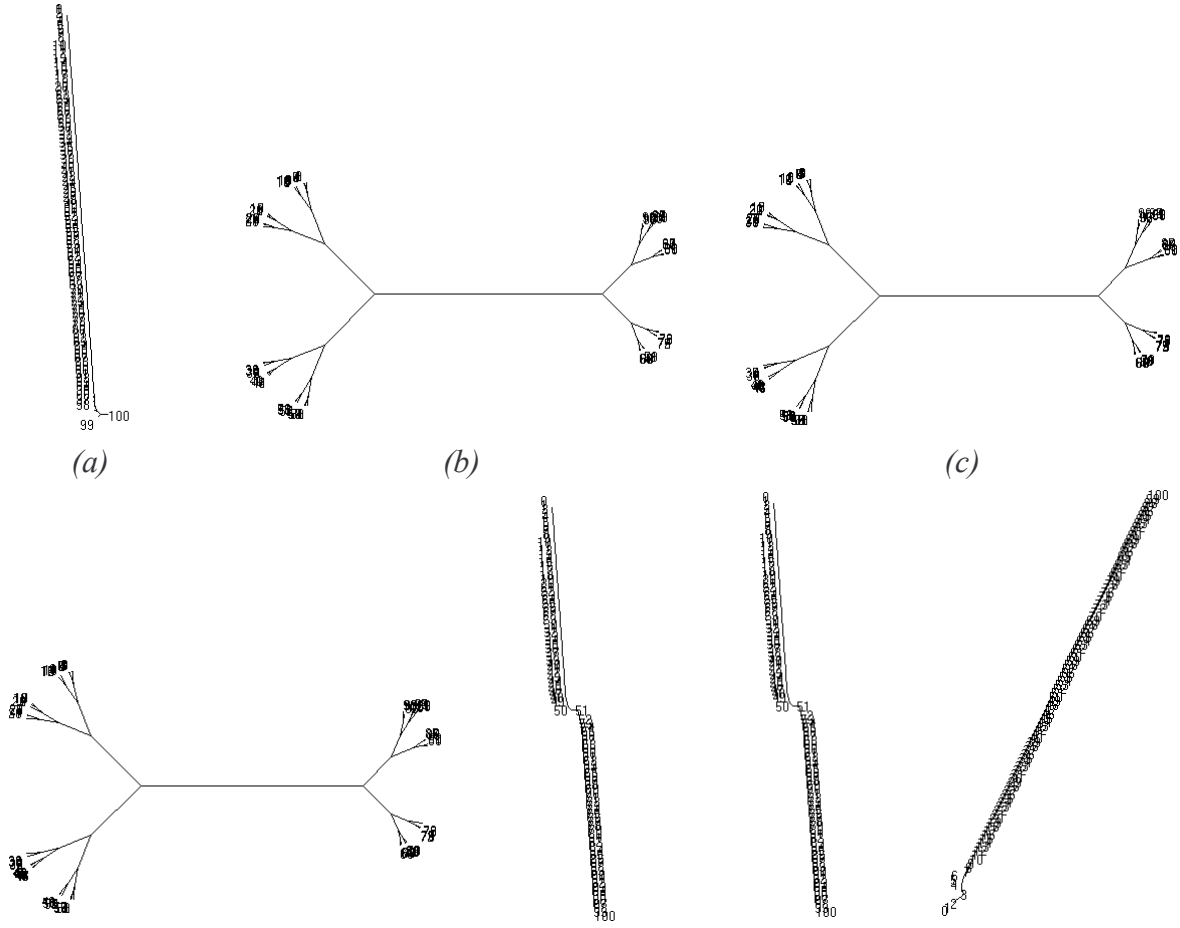

Figure 11 Résultats des différentes méthodes sur la suite numérique
((a) Méthode des liens simples, (b) Méthode des liens complets, (c) Méthode WPGMA, (d) Méthode UPGMA, (e) Méthode NJ, (f) Méthode ADDTREE, (g) Méthode des Groupements)

Face à ces figures, on peut distinguer deux types de comportements : les méthodes qui semblent respecter le caractère linéaire et celles qui ne le respectent pas visuellement.
Les méthodes présentant visuellement un respect du caractère linéaire des données sont la méthode des liens simples et les méthodes basées sur un arbre en étoile (NJ, ADDTREE, Groupements). La méthode des groupements s'affiche graphiquement de manière différente des autres, elle n'en garde pas moins le respect du comportement linéaire.
Une étude plus détaillée des arbres révèle cependant les différences cachées.

	Nb de Nœuds	Distance Feuille - Nœud	Distance Nœud - Nœud
Liens Simples	100	1 (0)	1(0)
Liens Complets	100	1,01 (0,01)	10,68 (16,92)
WPGMA	100	1 (0,05)	5,81 (8,26)
UPGMA	100	1 (0,05)	5,84 (8,46)
NJ	100	0,02 (0,14)	0,99 (0,07)
ADDTREE	100	0,02 (0,14)	0,99 (0,07)
Groupements	99	0,02 (0,14)	1 (0,06)

Tableau 1 Description des arbres (nombre de nœuds et distances)

Le Tableau 1 indique pour chaque méthode le nombre de nœuds créés (le nombre de feuille est identique pour chaque méthode et correspond au nombre de données en entrée), la distance moyenne des feuilles au nœud (et l'écart type) et la distance moyenne des nœuds fils à leur nœud père (et l'écart type).

Ainsi la méthode des liens simples qui apparaissait visuellement comme les méthodes basées sur un arbre en étoile, montre une grande différence d'un point de vue des distances. Cela signifie, entre autres, que seules les distances de l'élément [1] aux autres éléments sont respectées, toutes les autres assument une erreur de 1. Certes, cette erreur paraît faible dans l'absolu pour les grandes distances, mais elle est élevée, d'un point de vue relatif, pour les courtes distances : d(2,3)=2. De plus, il faut signaler que cette erreur dépend du jeu de données. Dans un jeu de données, moins régulier et avec des distances croissantes entre les groupes ([1], [2], [4], [7], [11], …), cette erreur serait bien plus importante et même d'un point de vue absolu.
Il faut, outre ce point, signaler que la méthode des groupements se distingue par un nombre total de nœuds de 99. Cela vient de la particularité de cette méthode à proposer des groupements de plus de deux éléments. Enfin il faut signaler, l'infime différence entre les méthodes WPGMA et UPGMA et l'absence de différence entre les méthodes NJ et ADDTREE.

A la vue de tels résultats, le premier choix se porterait donc sur les trois méthodes respectant la linéarité des données. En effet, une séparation en groupes bien distincts, comme le proposent les méthodes PGMA, apparaît déjà comme une distorsion visuelle de la réalité. Une telle représentation laisse croire à l'existence de classes alors qu'il n'en est rien. L'expérimentation suivante permet d'apporter un second avis quant aux choix des méthodes à conserver.

3.2. Données Phylogéniques

Les données, utilisées pour cette expérimentation, sont issues, comme écrit précédemment, de la phylogénie et plus précisément d'une étude de Sarich de 1969 [SAR 69]. La matrice d'entrée représente les distances génétiques entre divers animaux :

	Chien	Ours	Racoon	Belette	Phoque	Otarie	Chat	Singe
Chien	0							
Ours	32	0						
Racoon	48	26	0					
Belette	51	34	42	0				
Phoque	50	29	44	44	0			
Otarie	48	33	44	38	24	0		
Chat	98	84	92	86	89	90	0	
Singe	148	136	152	142	142	142	148	0

Tableau 2 Distances génétiques entre divers animaux

Les arbres obtenus par chacune des méthodes sont représentés par les figures suivantes :

Un bref aperçu des données (Tableau 2) permet de distinguer plusieurs groupes. Tout d'abord, il est clairement visible que le singe est différent de toutes les autres espèces animales, avec une distance moyenne proche de 140. Puis, le chat se différencie aussi des autres espèces avec une distance moyenne de 90. De plus, un premier groupe de voisins proches peut être imaginé entre le Phoque et l'Otarie qui partagent la plus petite distance commune 24. Enfin, un second couple tend à se créer entre l'Ours et le Rancoon puisqu'ils présentent une distance commune de 26 soit la seconde plus petite distance.

A nouveau, dans leur représentation, les méthodes du type PGMA se montrent radicales. Ce côté radical, dans la discrimination du singe, a pour conséquence le « tassement » des autres espèces. Aussi radicales soient elles, ces figures représentent les premiers sentiments face au jeu de données.
Les méthodes basées sur les X-arbres présentent des classifications plus « adoucies ». La distance entre chaque espèce y est représentée par la longueur de liens. Ainsi le Singe et le Chat se distinguent des autres espèces par leurs distances au centre, et même si ils semblent proches l'un de l'autre, la distance des liens pour aller de l'un à l'autre indique le contraire.

Si à nouveau rien ne peut distinguer les classifications NJ et ADDTREE, la méthode des Groupements se distingue de toutes les autres classifications par la création d'une classe commune Rancoon – Ours – Chien. Si cette classe se justifie localement par le fait que la distance du Rancoon à l'Ours est la seconde plus petite distance et celle de l'Ours au Chien est la troisième, la jointure dans une même classe du Rancoon et du Chien ne paraît pas si évidente (distance deux fois plus élevée qu'entre le Rancoon et l'Ours).

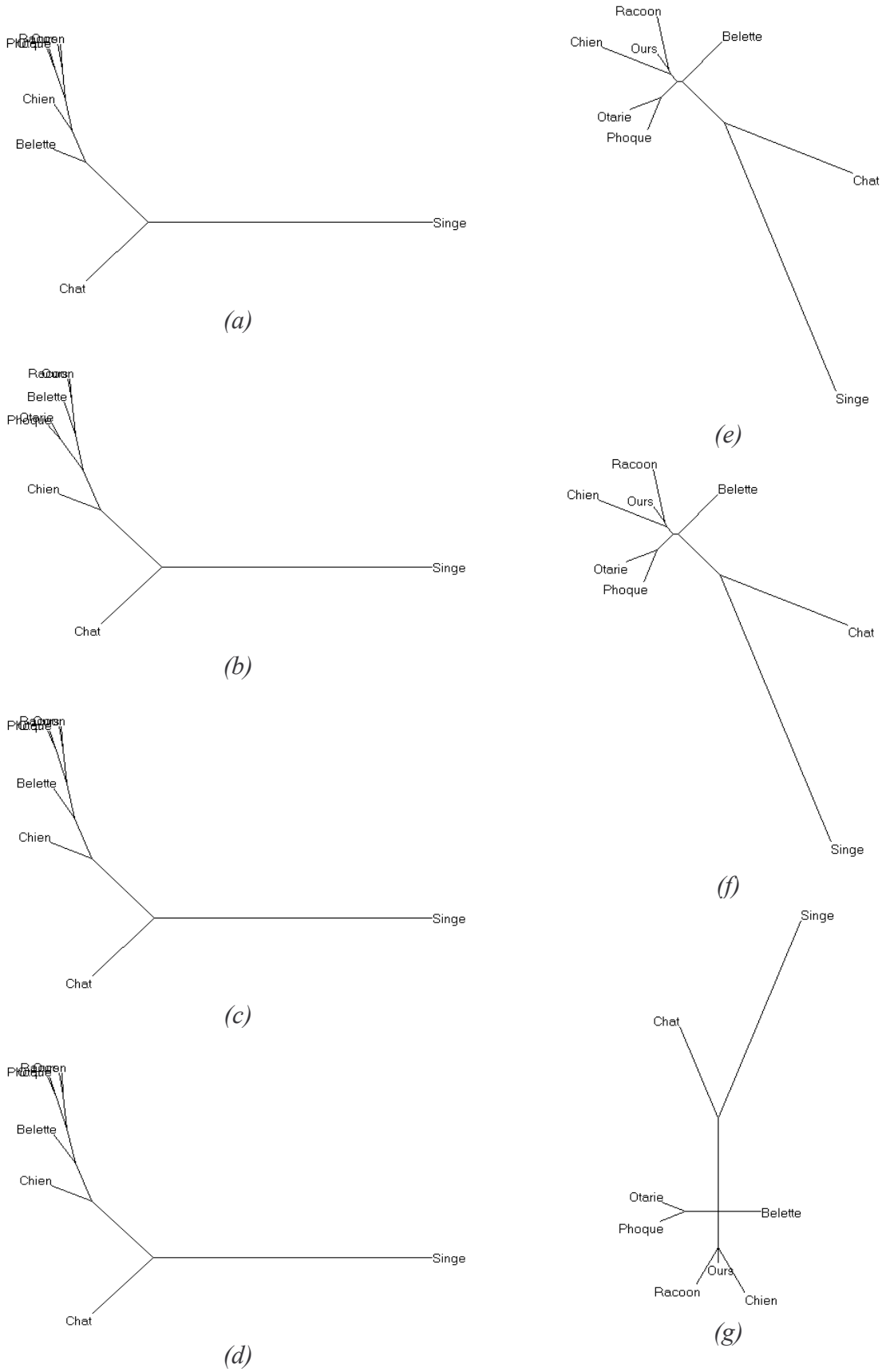

Figure 12 Résultats des différentes méthodes sur les données phylogéniques ((a) Méthode des liens simples, (b) Méthode des liens complets, (c) Méthode WPGMA, (d) Méthode UPGMA, (e) Méthode NJ, (f) Méthode ADDTREE, (g) Méthode des Groupements)

3.3. Erreur de Classification

Au-delà d'un simple contrôle visuel, on peut calculer une erreur de classification. En effet, la classification arborée permet, par sa connexité, de recalculer les valeurs de la matrice des distances. Pour un couple d'éléments, une erreur peut être calculée comme la différence absolue entre la distance originale du couple et la distance recalculée. L'erreur de classification correspond à la somme des erreurs pour chaque couple distinct.
Le Tableau 3 présente les différentes erreurs obtenues lors des deux expérimentations et par chacune des méthodes.

Nom Fichier	0A100	Phylogénie
Liens Simples	10000,00	4244,00
Liens Complets	1128496,00	5662,00
WPGMA	497740,00	5236,81
UPGMA	507293,00	5106,93
NJ	0,00	40,39
ADDTREE	0,00	40,39
Groupements	74,50	548,86

Tableau 3 Erreurs de classification sur chacune des expérimentations par chacune des méthodes

Il peut être à nouveau observé les similarités entre les méthodes WPGMA et UPGMA d'un côté et les méthodes NJ et ADDTREE de l'autre côté.
Mais cette erreur permet de montrer à quel point la méthode des liens simples s'est éloignée de la matrice des distances et surtout à quel point, malgré une représentation visuelle assez proche du résultat idéal, la méthode des groupements s'est éloignée elle aussi de la matrice des distances.
Pour un jugement moins relatif, le Tableau 4 présente les coefficients de corrélation entre les diverses matrices de distances.

Nom Fichier	0A100	Phylogénie
Liens Simples	99,99494%	97,02901%
Liens Complets	75,04291%	96,23660%
WPGMA	75,17057%	96,49613%
UPGMA	75,07031%	96,42891%
NJ	100,00000%	99,93335%
ADDTREE	100,00000%	99,93335%
Groupements	99,99968%	98,66940%

Tableau 4 Corrélations entre la matrice originale des distances et celles recalculées des méthodes

Ces résultats sont complémentaires des résultats précédents. Le coefficient de corrélation ne diffère que très peu, pour les 101 données consécutives, entre la méthode des liens simples et

la méthode des groupements. Le tableau précédent indique pourtant un rapport de plus de 100 entre les erreurs respectives. A nouveau, sur ce même jeu de données, l'incompatibilité des 3 méthodes PGMA avec des données linéaires est constatable, et le score parfait des méthodes NJ et ADDTREE.
Le jeu de données phylogéniques est à la fois plus simple et plus complexe. Il est plus simple car il possède moins de données, l'erreur totale est donc limitée. Il est plus complexe car il provient de données réelles et en cela provoque de réelles erreurs. La méthodes des liens simples se montre à nouveau la meilleure des méthodes du type PGMA. La méthode des groupements montre un score relativement bon. [DUB 99], en présentant la méthode des groupement, annonce des scores de 73% pour les méthodes PGMA et 92% pour la méthode des groupements. Les scores présentés dans le tableau ne sont donc pas à généraliser. Quoiqu'il en soit les méthodes NJ et ADDTREE montrent à nouveau leur efficacité.

4. Bilan

Ce chapitre a étudié sept méthodes de classification arborées. Les quatre premières méthodes, PGMA, appartiennent au domaine de classification en général. Les deux méthodes suivantes, NJ et ADDTREE, sont issues du traitement de données biologiques, et plus particulièrement de la phylogénie. Contrairement aux premières, elles tentent d'établir la notion de voisinage d'un point de vue général et pas seulement local. Enfin la dernière méthode, méthode des Groupements, a été développée dans le cadre général des classifications arborées mais a prouvé son utilité dans le domaine de la classification de textes.
Les premières méthodes apparaissent comme catégoriques dans leur classification, et laissent apparaître des classes là où il n'y en a pas nécessairement. Ce côté catégorique se rapproche des a priori que peut avoir l'esprit humain lors d'une analyse superficielle des données (voir l'expérimentation sur les Données Phylogéniques de la partie 3.2). Cet aspect va à l'encontre du but du projet qui est de fournir une représentation aussi objective que possible des données.
Les deux méthodes issues de la biologie ne se sont distinguées l'une de l'autre que par des différences infimes dans les distances. Si infimes qu'elles en sont imperceptibles. D'un point de vue théorique, cependant, ces deux méthodes se distinguent l'une de l'autre par leur complexité, et donc par leur temps de calcul.

Enfin, la méthode des Groupements, utilisée pour les données textuelles, a l'avantage de traiter les classes à λ éléments. Cependant, comme l'indique [BAR 98], cette méthode fournit une approximation de la dissimilarité initiale moins bonne que celle que l'on obtiendrait à l'aide d'une approche métrique (ou géométrique). De plus, sa complexité algorithmique est trop élevée. La seule obtention de la liste des paires de sommets réels ordonnés par scores décroissant est en $O(n^4)$. Elle doit donc être réservée à des corpus de taille moyenne (de l'ordre de tout ou partie de l'œuvre d'un auteur).
Les erreurs, plus faibles que celles des méthodes PGMA, ont été confirmées par des expérimentations sur des jeux de données relativement simples. La complexité s'est remarquée par un temps d'exécution supérieur à celui nécessaire pour l'ensemble des autres méthodes.

De par des résultats presque parfaits et une complexité moindre à résultat comparable, la méthode NJ constitue la meilleure méthode pour l'application souhaitée. C'est donc elle qui a été retenue pour le projet.

Chapitre 3 - Normalisation et dissimilarité

Ce chapitre s'intéresse à résoudre la seconde étape du projet. Cette étape consiste à calculer les dissimilarités entre les index obtenus à la suite de la première étape afin de fournir une matrice de dissimilarités à la méthode de classification de la troisième étape.

La mesure des dissimilarités est une étape très importante dans tout système de classification. Il est donc nécessaire d'étudier les diverses mesures disponibles à cet effet. L'utilisation d'une mesure géométrique peut, en effet, fournir un résultat complètement différent de celui d'une mesure de corrélation.
Une liste de mesures aussi exhaustive que possible a été étudiée. Ces mesures sont issues du traitement de données réelles et du traitement de données binaires. Les mesures issues du traitement de données binaires peuvent, dans le cas d'une formulation non-ensembliste, être appliquées aux données réelles à condition que celles-ci soient réelles à valeurs dans l'intervalle [0,1]. Afin de répondre à une telle condition, une normalisation est effectuée, dans tous les cas, avant le calcul des dissimilarités.

Ce chapitre s'articule en quatre parties. Dans une première partie, les deux types de normalisation étudiées pour ce projet. Puis, dans une seconde partie, les mesures de dissimilarités sont listées. La troisième partie est consacrée à l'étude expérimentale des normalisations et des mesures. Enfin, dans une dernière partie, le bilan est tracé et un premier choix est effectué.

1. Normalisation

L'étape précédente a pour résultat un index. C'est-à-dire un ensemble de vecteurs de données. Ces données sont supposées réelles et positives ou nulles.
Cependant, pour le projet, il a été choisi de n'opérer des calculs de dissimilarités que sur des données appartenant à l'intervalle [0,1]. Une normalisation des données est donc nécessaire. Comme le projet s'attache aux entretiens d'une série, c'est-à-dire à un corpus fermé, aucun texte n'est ajouté au corpus une fois que celui est constitué, la normalisation est effectuée à partir de l'ensemble des valeurs de chaque caractéristique. C'est-à-dire que cette normalisation est effectuée à partir de l'ensemble des textes du corpus et de manière indépendante sur chaque caractéristique. Cela permet de conserver le sens relatif des données comparées.

Afin d'effectuer un tel recalage, il y a le choix est principalement entre deux types de normalisations : la normalisation par la somme et la normalisation par le maximum.
La normalisation par la somme est une normalisation probabiliste. Elle permet d'assurer que la somme des valeurs d'une caractéristique soit égale à 1. Elle a cependant, le grand défaut de dépendre du nombre de données, c'est-à-dire du nombre de textes traités. En effet, à même échelle, plus il y aura de textes, plus la somme initiale sera élevée et donc plus les valeurs normalisées seront faibles.
La normalisation par le maximum remédie à ce problème. Cette normalisation assure une valeur maximum pour chaque caractéristique égale à 1. Par contre, à l'inverse de l'autre normalisation, pour chaque caractéristique, la somme n'est pas limitée et dépend du nombre et

de la distribution des valeurs de la caractéristique. Si toutes les valeurs d'une caractéristique sont proches du maximum, la somme sera proche du nombre de textes étudiés : il y a une dépendance au bruit.

Si le but recherché par ces deux méthodes est le même, le résultat sur les mesures de dissimilarité peut être très différent. Le rapport terme à terme entre les deux jeux de données normalisées est toujours proportionnel, pour chaque caractéristique, au rapport entre la somme des données et la valeur maximum de la caractéristique. Le rapport entre les dissimilarités dépend de la caractéristique et des textes comparés.

Pour bien observer cette variation dans les rapports et comprendre l'enjeu du choix de la normalisation, il suffit de prendre le simple exemple d'un ensemble de quatre textes indexés par deux caractéristiques. Le Tableau 5 présente un tel jeu de données.

	c1	c2
t1	3	3
t2	3	0
t3	3	100
t4	4	4
Somme	13	107
Max	4	100
Rapport	3,25	1,07

Tableau 5 Exemple de données avec quatre textes et deux caractéristiques

Le tableau présente aussi pour chaque caractéristique la somme des valeurs et la valeur maximum ainsi que le rapport entre ces deux valeurs. Après normalisation et calcul des dissimilarités par la distance euclidienne, le rapport pour chaque valeur des matrices de distance peut être calculé. Le Tableau 6 présente les rapports obtenus pour chaque valeur de dissimilarités pour les données du Tableau 5.

	t1	t2	t3
t2	1,07		
t3	1,07	1,07	
t4	3,23	2,96	1,10

Tableau 6 Rapport entre la distance euclidienne après normalisation par la somme et la distance euclidienne après normalisation par la valeur maximum.

Il peut être observé que le rapport entre les deux dissimilarités calculées varie du simple au triple. D'un point de vue de la classification, les conséquences sont importantes. Cela signifie que la dissimilarité entre les textes t1 et t2 et les textes t1 et t3 est quasiment similaire. Par contre, la dissimilarité entre le texte t1 et t4 est trois fois plus importante avec une méthode plutôt qu'une autre.

Sur des données réelles, peuvent provoquer deux classifications différentes. La méthode de classification choisie se base sur les distances pour créer le regroupement. Entre deux matrices de distances différentes, les regroupements sont différents.

Il n'existe pas de solution générale à un tel choix. La méthode de normalisation dépend aussi de la mesure employée. La distance de Gower, qui normalise par la différence entre la valeur maximum et la valeur minimum de la caractéristique, est insensible à la variation de normalisation. La distance euclidienne est naturellement influencée par le nombre de caractéristiques. Plus les valeurs seront faibles, plus la distinction entre deux textes sera difficile. La distance euclidienne est donc avantagée par une normalisation par la valeur maximum. Inversement, la mesure du cos-θ tient son pouvoir discriminant de sa normalisation par une somme de valeurs au carré. Donc plus les valeurs sont petites, plus le pouvoir discriminant sera fort. Cette mesure est donc avantagée par une normalisation par la somme.

Le choix de la normalisation se fait donc au moment de choisir la mesure de dissimilarité. Ce n'est donc pas le choix de deux méthodes indépendantes mais d'un couple de méthodes liées l'une à l'autre séquentiellement.

2. Calcul de dissimilarités

Les mesures de dissimilarités, qu'elles soient destinées au traitement de données binaires ou de données réelles, trouvent de nombreux domaines d'application : biologie ([GAV 03], [LOU 04], [MEY 02], [MEY 04], [SEL 05]), écologie [ROO 04], sécurité des réseaux [YE 01], traitement d'images ([KUL 01], [OMH 04]), traitement de données textuelles ([CHU 01], [LAB 01], [LEV 98], [LEE 01], [MUR 04], [SAL 96]) ou traitement de données en général ([HU 04], [SAN 02-2])
Ces mesures peuvent être, comme le fait ce chapitre, simplement présentées et étudiées dans un cadre purement mathématique [POL 05], dans le cadre de l'étude des systèmes d'analyse de données [BIS 00] ou dans le cadre particulier de l'étude de la symétrie d'une mesure [JOH 01]. Ces mesures peuvent aussi faire l'objet d'études de leurs propriétés métriques [ZHA 03].

Une distance est nommée ainsi lorsqu'elle respecte les contraintes de positivité, de réflexivité de symétrie et la contrainte de l'inégalité triangulaire. Ce projet s'est intéressé à l'ensemble des mesures dans sa globalité. La plupart des mesures utilisées lors de ce projet sont communes et issues de la littérature. Les références gardées et citées ne le sont que pour la quantité importante de mesures distinctes présentées.

Les mesures ont été répertoriées et expérimentées pour vérifier leur pertinence dans la résolution de notre problématique. Aucune étude sur les contraintes n'a été menée. C'est la raison pour laquelle c'est le terme de mesure de dissimilarité qui est employé.

Une cinquantaine de mesures ont donc été répertoriées. Pour ne pas surcharger inutilement ce chapitre par les formulations, les mesures sont listées en Annexe 3. De même, cette annexe présente la représentation graphique de la classification par la méthode NJ sur deux jeux de données différents.

La liste des mesures, et leurs différents noms, utilisées pour le traitement des données binaires est la suivante (par ordre alphabétique) :

- Braun, Blanque
- Chi-2
- Dice, Sorensen – Dice, Czekanowsky, Harmonic Mean
- Dispersion
- Equivalence
- Faith
- Hamann
- Hamann II
- Hamming
- Information Mutuelle
- Jaccard, Jaccard – Needham
- Kulzinsky I, Kulezinski, Kulczinski
- Kulzinsky II
- McConnaughy
- Michael
- Mountford
- Mozley – Margalef
- Ochiai I, Cosinus
- Ochiai II
- Pearson (Phi de), Sokal & Sneath V
- Rogers – Tanimoto
- Rogers - Tanimoto II, Anderberg, Sokal & Sneath II
- Russell – Rao
- Simpson, Inclusion
- Sokal-Michener, Simple Matching
- Sokal et Sneath I
- Sokal et Sneath IV
- Yule (Q de)
- Yule (Y de)

De même, la liste des mesures, et leurs différents noms, utilisées pour le traitement des données réelles est la suivante (par ordre alphabétique) :

- Bhattacharyya
- Bray-Curtis
- Canberra
- Chi-2
- Clark, Divergence
- Cosinus, Distance Angulaire
- Cos-θ
- Euclidienne Standardisée
- Gower
- Hellinger
- Information Statistique

- Jensen – Shannon
- Kullback – Leibler
- Mahalanobis
- Minkowski
    - Manhattan - city block
    - Euclidienne
    - Chebychev – Chessboard
- Orloci
- Pearson (Corrélation de)
- Skew
- Soergel

Il faut signaler que la mesure de Kullback – Leibler n'est pas utilisée en l'état en raison de son non-respect de la symétrie. Les formulations conservées pour le projet sont les propositions de symétries proposées par [JOH 01]. Ces propositions sont détaillées en annexe.

Les distances ne pouvant être jugées sur leur formulation, des expérimentations ont été menées pour étudier les qualités de classification de chacune. La partie suivante présente ces expérimentations.

3. Expérimentations

Les expérimentations avaient deux buts. Le premier consiste à associer à chaque mesure la méthode de normalisation qui lui convient le mieux. Le second but consiste à trouver les couples offrant les meilleurs résultats de classification.
Des expérimentations sur trois jeux de données différents ont été menées. Les deux premiers jeux de données sont des jeux habituels dans le domaine de la classification, il s'agit du jeu de données Iris et du jeu de données Soybean.
Le premier jeu de données est constitué de trois classes et de 150 individus. Chaque individu est représenté par quatre caractéristiques. Le deuxième jeu est quant à lui constitué de 376 individus répartis en 19 classes représentées par 35 caractéristiques.

Un troisième jeu de données a été conçu de manière artificielle. Deux jeux de 10 échantillons de 250 valeurs ont été générés en suivant une loi normale. L'écart-type pour les deux générations a été fixé à 2, la moyenne du premier jeu a été fixée à 5 et la moyenne du second jeu a été fixée à 8. Ces moyennes assurent une superposition des classes. La Figure 13 présente la distribution de l'ensemble des valeurs intervenant pour chacun des deux jeux de 10 échantillons.

Cette figure confirme de manière visuelle une réelle superposition des deux jeux d'échantillons. Chaque jeu sert à définir une classe et chaque échantillon définit les caractéristiques. Ce jeu aléatoire est donc constitué de deux classes et d'un total de 500 individus. Chaque individu est représenté par dix caractéristiques.

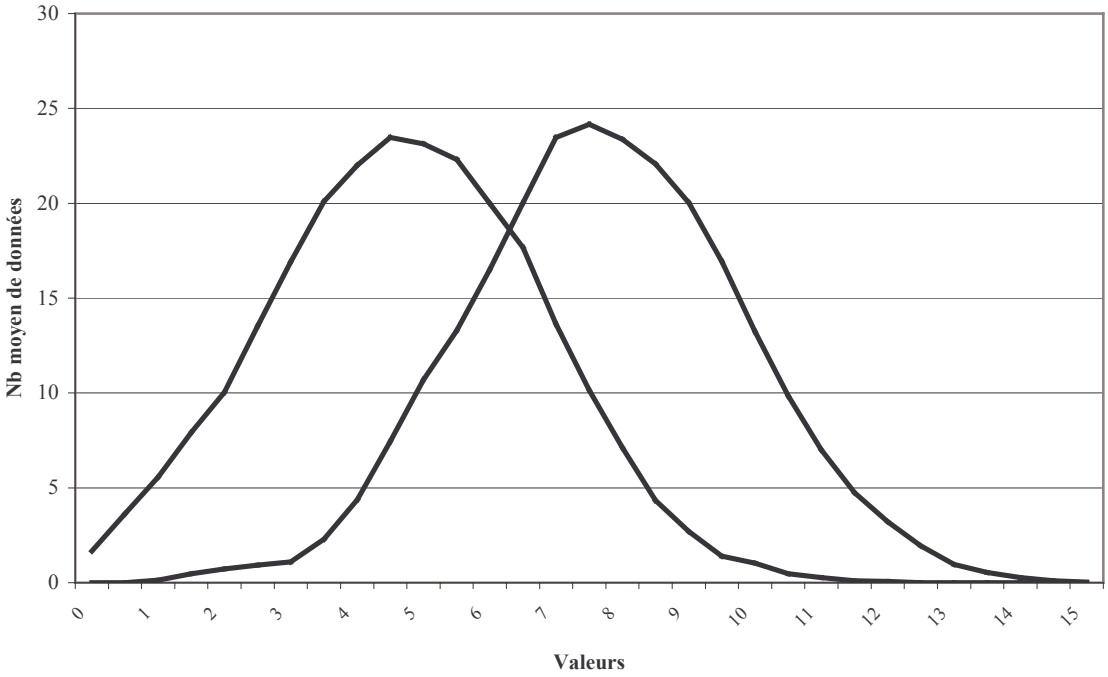

Figure 13 Distribution des deux jeux d'échantillons

Dans le cas d'une méthode de classification comme celle choisie, les comparaisons sont difficiles. Le projet s'est, en effet, orienté vers des méthodes de classification non-strictes, c'est-à-dire que les classes ne sont pas définies en tant que telles. D'un arbre, il peut n'être formé qu'une unique classe, comme il peut en être formé autant qu'il y a de textes. La formation des classes a donc été réalisée à partir du nombre de classes réelles à former. L'arbre de classification est utilisé comme un simple arbre à racine et les nœuds les plus proches du nœud racine sont supprimés jusqu'à obtention du bon nombre de classes. La méthode NJ forme un arbre binaire. Ainsi pour obtenir deux classes, il suffit de supprimer le nœud racine ; pour obtenir les 3 classes du jeu de données Iris, il suffit de supprimer le nœud le plus proche du nœud racine parmi les deux nœuds directement raccordés au nœud racine, etc. Une classe créée est comparée à la classe réelle majoritairement représentée dans cette classe créée.

C'est l'évaluation des classifications qui permettra de réaliser un choix. Cette évaluation est basée sur une combinaison de la précision et du rappel. Cette évaluation permet ainsi de vérifier qu'un maximum d'éléments des classes créés sont correctement classés (précision) et qu'un minimum d'éléments des classes réelles soient mal classés (rappel).
La précision correspond donc au rapport entre le nombre d'éléments bien classés d'une classe créée et le nombre d'éléments de cette classe. Le rappel correspond au rapport entre le nombre d'éléments bien classés d'une classe créée et le nombre d'éléments que devrait compter la classe réelle.
Une mesure notée F_1 permet de représenter en une seule valeur les deux valeurs précédentes. Sa formulation est la suivante :

$$F_1 = \frac{2 \times rappel \times précision}{rappel + précision} \quad (6)$$

Les valeurs du rappel et de la précision sont comprises dans l'intervalle [0,1] car ce sont tous les deux des rapports. L'idéal revient à avoir autant de bien classé que compte la classe créée et que compte la classe réelle. Dans ce cas, le rappel et la précision sont égaux à 1. De même la valeur de F_1 sera de 1. Inversement, plus le rappel et plus la précision sont faibles et plus F_1 sera faible.

Le Tableau 7 et le Tableau 8 présentent les résultats de F1 obtenus pour chacun des jeux de données avec l'utilisation de chacune des mesures. Le premier tableau présente plus précisément les résultats lorsque la normalisation est faite par la somme des valeurs de chaque caractéristique et le second tableau présente les résultats lors d'une normalisation par la valeur maximum de chaque caractéristique.
Dans chacun des tableaux, une colonne moyenne indique la moyenne des valeurs sur les 3 jeux de données. Les mesures, dans chacun des tableaux, sont d'ailleurs organisées par ordre de valeur moyenne décroissante. De plus pour une visualisation plus rapide des résultats, le meilleur résultat de chaque colonne est mis en gras.
Cette mise en évidence permet de voir que les meilleurs résultats sont obtenus sur le jeu de données aléatoire dont les données suivent une loi normale. Le meilleur résultat du Tableau 7 est donc celui obtenu par la symétrie IV de la mesure de Kullback-Leibler avec 0,842 et le meilleur résultat du Tableau 8 correspond à celui de la distance euclidienne standardisée avec un résultat de 0,898.

Cependant si le meilleur résultat est obtenu par une normalisation par la valeur maximum, le résultat moyen est meilleur pour la normalisation par la somme des valeurs de chaque caractéristique (0,363) que par la normalisation par maximum (0,345). De plus, pour les mesures offrant des bons résultats dans les deux tableaux, il peut être observé que le résultat est toujours meilleur pour la normalisation par la somme : Mahalanobis (0,4929/0,4915), Bhattacharyya (0,498/0,437), Soergel (0,432/0,414). Enfin, à part Gower qui par son insensibilité à la normalisation est le maximum commun aux deux tables, les résultats obtenus à la suite d'une normalisation par la valeur maximum est généralement moins bon que ceux obtenus à la suite d'une normalisation par la somme.

Du point de vue des mesures, il peut être observé avec surprise que les pires scores sont obtenus lorsque la mesure du cosinus est utilisée. Elle constitue pourtant la base de la comparaison documentaire de Salton. A la suite de telles expérimentations, elle apparaît pourtant comme la mesure la moins apte à être utilisée pour une classification quelconque.
En ce qui concerne les résultats du haut des tableaux, il peut être constaté, comme écrit précédemment que la distance de Gower offre les meilleurs résultats moyen. Cette mesure offre aussi le meilleur résultat sur le jeu Iris lors d'une normalisation par le maximum des valeurs de chaque caractéristique. Les autres maximaux locaux sont la distance de Soergel pour le jeu Iris, Mahalanobis pour le jeu Soybean et la symétrie IV de Kullback-Leibler pour le jeu aléatoire dont la distribution suit une loi normale lors d'une normalisation par la somme. Les maximaux sont la distance de Russell-Rao pour le jeu Soybean et la distance euclidienne standardisée pour le jeu aléatoire dont la distribution suit une loi normale lors d'une normalisation par le maximum.

	Iris	Soybean	Loi Normale	Moyenne
Gower	0,583	0,309	0,791	**0,561**
Bhattacharyya	0,597	0,418	0,481	0,499
Mahalanobis	0,547	**0,462**	0,470	0,493
KullBack-Leibler Symétrie IV	0,567	0,050	**0,842**	0,486
Cos-Téta	0,588	0,272	0,470	0,443
Bray Curtis	0,580	0,286	0,461	0,442
Soergel	**0,633**	0,187	0,478	0,433
Michael	0,327	0,447	0,495	0,423
Dispersion	0,327	0,447	0,494	0,423
KullBack-Leibler Symétrie I	0,607	0,157	0,489	0,418
Jensen-Shannon	0,566	0,112	0,564	0,414
Manhattan	0,310	0,198	0,719	0,409
Dice	0,328	0,374	0,499	0,400
Hamman II	0,328	0,374	0,499	0,400
Jaccard	0,328	0,374	0,499	0,400
Rogers-Tanimoto II	0,328	0,374	0,499	0,400
Hellinger	0,566	0,155	0,470	0,397
Kulczinski 2	0,566	0,127	0,498	0,397
Chi-2	0,328	0,366	0,494	0,396
Sokal-Sneath 4	0,323	0,373	0,491	0,396
Russel-Rao	0,325	0,354	0,497	0,392
McConnaughy	0,331	0,343	0,499	0,391
Equivalence	0,331	0,343	0,499	0,391
Chebychev	0,570	0,111	0,488	0,390
Pearson (Phi de)	0,331	0,332	0,499	0,388
Yule (Y de)	0,583	0,073	0,492	0,383
Yule (Q de)	0,607	0,049	0,492	0,382
Simpson	0,331	0,300	0,499	0,377
Information Statistique	0,313	0,191	0,564	0,356
Ochiai 1	0,331	0,217	0,499	0,349
Braun	0,304	0,224	0,482	0,337
Kulczinski 1	0,331	0,135	0,499	0,322
KullBack-Leibler Symétrie II	0,331	0,134	0,499	0,321
Clark	0,328	0,188	0,446	0,321
Ochiai 2	0,331	0,101	0,499	0,310
Euclidienne	0,330	0,111	0,487	0,309
Pearson	0,331	0,117	0,477	0,308
KullBack-Leibler	0,265	0,175	0,485	0,308
…	…	…	…	…
Cosinus	0,234	0,033	0,346	0,204

Tableau 7 Valeur de F1 pour les différents jeux de données et les différentes mesures avec une normalisation par la somme des valeurs

	Iris	Soybean	Loi Normale	Moyenne
Gower	**0,583**	0,309	0,791	**0,561**
Mahalanobis	0,547	0,458	0,470	0,492
Euclidienne Standardisée	0,327	0,156	**0,898**	0,460
Euclidienne	0,329	0,161	0,865	0,452
Bhattacharyya	0,572	0,284	0,458	0,438
Russel-Rao	0,325	**0,484**	0,494	0,435
Information Statistique	0,560	0,256	0,458	0,425
Soergel	0,580	0,187	0,478	0,415
Clark	0,578	0,158	0,474	0,403
Chebychev	0,331	0,081	0,773	0,395
Michael	0,331	0,327	0,499	0,386
KullBack-Leibler	0,560	0,108	0,476	0,381
Mozley	0,331	0,309	0,499	0,380
Kulczinski 2	0,327	0,314	0,499	0,380
Hellinger	0,560	0,104	0,471	0,378
Faith	0,322	0,297	0,494	0,371
Information Mutuelle	0,327	0,269	0,499	0,365
Hamman	0,322	0,271	0,498	0,364
Jaccard	0,322	0,271	0,498	0,364
Ochiai 2	0,331	0,248	0,492	0,357
Bray Curtis	0,300	0,284	0,465	0,350
Manhattan	0,508	0,071	0,468	0,349
Dispersion	0,325	0,218	0,494	0,346
Simpson	0,331	0,192	0,495	0,340
Ochiai 1	0,327	0,191	0,499	0,339
Pearson (Phi de)	0,323	0,184	0,495	0,334
McConnaughy	0,331	0,160	0,499	0,330
Dice	0,327	0,163	0,499	0,329
Kulczinski 1	0,331	0,143	0,499	0,324
Yule (Q de)	0,331	0,141	0,499	0,324
Equivalence	0,331	0,133	0,499	0,321
Chi-2	0,318	0,148	0,490	0,319
Pearson	0,312	0,129	0,495	0,312
Yule (Y de)	0,331	0,105	0,499	0,312
Sokal-Sneath 4	0,325	0,113	0,494	0,311
Rogers-Tanimoto	0,317	0,151	0,462	0,310
Canberra	0,324	0,126	0,476	0,309
KullBack-Leibler Symétrie III	0,331	0,093	0,499	0,308
…	…	…	…	…
Cosinus	0,252	0,039	0,380	0,224

Tableau 8 Valeur de F1 pour les différents jeux de données et les différentes mesures avec une normalisation par la somme des valeurs

Il faut aussi noter la double bonne performance de la distance de Bhattacharyya, la meilleure étant obtenue par une normalisation par la somme. Si les expérimentations n'avaient été menées que sur les deux premiers jeux de données, la distance de Bhattacharyya aurait présenté un résultat moyen meilleur que celui de Gower. La distance de Gower n'aurait, en effet, occupé que la 3^{e} place après Bhattacharyya et Mahalanobis. Enfin, on peut noter, la bonne performance moyenne (0,442) de la distance de Bray-Curtis qui est utilisée par Dominique Labbé dans son étude sur Corneille et Molière.

Les expérimentations tendent donc à limiter l'utilisation des mesures de dissimilarité à un ensemble de sept mesures. Cinq de ces mesures trouvent une application idéale à la suite d'une normalisation par la somme : Bhattacharyya, Gower, Kullback-Leibler Symétrie IV, Mahalanobis et Soergel. Et deux autres mesures sont plus performantes d'un point de vue d'une classification par la méthode choisie à la suite d'une normalisation par la valeur maximum : Euclidienne Standardisée et Russel-Rao.

Cette dernière mesure citée, Russel-Rao constitue d'ailleurs l'unique mesure issue du traitement de données binaires. Une critique à cette remarque tient à faire rappeler que les expérimentations n'ont été menées que sur des données de type réel. Un jeu de données binaires, comme c'est le cas des index fournis par les sociologues dans le cas de l'application réelle, peuvent, néanmoins, être considérés comme des jeux de données normalisés pour lesquels l'utilisation d'une valeur se fait de façon maximum ou ne se fait pas.

Ces expérimentations ne sont pas des preuves. Mais elles apparaissent comme de très bons exemples d'utilisation de mesures.

4. Bilan

Ce chapitre s'est attaché à présenter de manière concise deux normalisations et une cinquantaine de mesures de dissimilarités. Il a été prouvé que normalisation et mesure de dissimilarité ne pouvaient être choisies de manières indépendantes.

A partir d'expérimentations sur trois jeux de données variés, la centaine de combinaisons a été étudiée sur la qualité de classification de chaque combinaison. La qualité a été calculée à partir du rappel et de la précision. Ainsi huit couples normalisation-mesure ont été conservés. Il s'agit des couples : Somme – Bhattacharyya, Somme – Gower, Somme – Kullback-Leibler Symétrie IV, Somme – Mahalanobis, Somme – Soergel, Maximum – Euclidienne Standardisée et Maximum – Russel-Rao.

La seconde étape est donc constituée d'un ensemble de propositions. La proposition retenue sera propre à chaque mesure d'indexation choisie. En effet, certaines propositions montrent un résultat moyen bon, peu importe le jeu de données. Mais, les mesures sont, généralement, autant liées au jeu de données pour lequel elles doivent calculer la matrice de dissimilarités qu'à la normalisation à appliquer avant le calcul des distances. La chapitre suivant, qui présente le travail sur la représentation des données textuelles effectué pour ce projet, utilise lors de ses expérimentations l'ensemble de ces combinaisons afin d'obtenir les meilleurs résultats.

Chapitre 4 - L'indexation : Représentation codée des textes

*Les chapitres précédents ont permis d'étudier les deux étapes aboutissant à un système de comparaison. Ce chapitre s'intéresse donc à la première étape : l'indexation ou la manière de coder un texte pour le représenter sous forme numérique. Au-delà de l'étude de telles méthodes, il faut s'intéresser aux traitements primaires des textes bruts. De même, cette représentation peut être améliorée de manière adaptative à l'aide de méthodes qui pondèrent les caractéristiques pour une meilleure discrimination. La partie 1 présente les traitements possibles avant et après l'indexation. Cette partie aborde les différentes unités textuelles qui peuvent être extraites des textes (mots, lemmes, n-grammes) et les différents pondérations qui peuvent améliorer la représentation des textes (tf*idf, recalage). La partie 2 s'intéresse aux méthodes d'indexation. Dans un premier temps, cette partie présente des méthodes, présentées dans le chapitre 1, qui permettent de représenter le contenu ou la structure des textes de manière statistique. Dans un deuxième temps, cette partie présente les premières solutions proposées par ce travail. Il s'agit de trois méthodes, une méthode statistique et deux automates, qui essaient de représenter l'organisation du discours. Dans un troisième temps, il est présenté une autre méthode complètement nouvelle : l'évolution textuelle. Cette méthode représente l'évolution des textes à partir des indexations locales des textes.*
Un bilan est réalisé dans la partie 3.
Tout au long de ce chapitre des expérimentations sont menées en suivant le protocole de test décrit dans le chapitre précédent. C'est-à-dire que la coupure est placée de façon à obtenir le nombre de classes souhaitées et que deux classifications sont comparées à l'aide de la mesure du F1.
Il a été choisi d'effectuer les expérimentations de ce chapitre sur des corpus accessibles, reconnus et pour lesquels une classification existe. De telles bases sont très rares. C'est la raison pour laquelle, ces expérimentations sont menées sur les corpus Amaryllis et NewsGroups. Ces deux corpus sont constitués respectivement de 131 et 200 textes (pour NewsGroups, une limitation aux 10 premiers textes de chaque thème a été effectuée). Ce sont des textes courts, c'est-à-dire qu'il font généralement moins d'une page. Ces corpus sont donc l'inverse de ceux du domaine réel d'application où les textes font en moyenne une quinzaine de pages et où le corpus est le plus souvent constitué de moins de 100 textes. Les analyses des expérimentations tiennent compte de cette différence de corpus et la comparaison est faite relativement au corpus.

1. Pré-traitements et Post-traitements

Le traitement de données textuelles brutes amène toujours la question du découpage des textes en unités textuelles. Les mots, pour la rapidité et la simplicité de leur extraction, sont le plus souvent employés. Les mots correspondent, pourtant, à un niveau assez élevé du langage, au même titre que la phrase. Il est donc normal de s'intéresser à l'extraction de telles unités. La partie 1.1 présente les divers découpages possibles ainsi que leurs avantages et inconvénients.

L'indexation de textes bruts est souvent considérée comme une représentation brute. Les post-traitements permettent, par l'ajout d'une pondération aux caractéristiques de l'index, d'améliorer les représentations. La partie 1.2 présente tout d'abord la pondération tf*idf qui est utilisée pour changer l'importance des unités textuelles suivant l'utilisation qui en est faite

dans les textes et dans le corpus. Puis, la pondération par recalage des fréquences, qui peut être considérée comme une normalisation, permet de limiter les effets de longueur qui peuvent se produire entre les différents textes.

1.1. Pré-traitements : Les unités textuelles

Les unités textuelles ont un rôle prépondérant dans l'indexation. Elles constituent, en effet, l'unité de base des textes. Ces unités de base peuvent être un paragraphe ou une phrase, mais, dans le cas d'une indexation, de telles unités seraient trop unique pour avoir un réel intérêt. Pour une plus grande cohérence, il est préférable d'utiliser des unités de taille réduite. Cela implique une augmentation de la cardinalité de l'ensemble des unités et un taux d'utilisation des unités assez élevé.
C'est la raison pour laquelle l'unité textuelle généralement choisie est le mot. Cependant, chaque mot se présente sous une multitude de formes possibles : singulier-pluriel, masculin-féminin, formes d'un verbe, … Certaines formes ne changent pas l'aspect réel du mot. D'autres formes, en revanche, imposent des changement radicaux.
Face à une telle problématique, un choix « philosophique » est à faire. Ce choix revient à répondre à la question : Lorsque l'on compare deux formes d'un même mot, faut-il les considérer comme un seul et même mot ou comme deux mots séparés ?
Au-delà de ce changement de forme, d'autres traitements, plus basiques, peuvent être effectués sur les textes bruts. Ces changements réduisent la taille de l'ensemble des mots distincts (ou unités textuelles) en réduisant l'alphabet à partir duquel ils sont construits.
La partie 1.1.1 s'intéresse aux traitements qui peuvent être réalisés sur l'alphabet. La partie 1.1.2 et la partie 1.1.3 s'intéressent aux unités textuelles citées précédemment. Et la partie 1.1.4 fait un bilan et une comparaison de ces unités textuelles.

1.1.1. Alphabet

Une réduction de l'alphabet consiste dans un premier temps à mettre en minuscules tous les caractères. Cette confusion entre minuscule et majuscule n'a aucune incidence ni sur la comparaison des textes, ni sur leur compréhension.

De même, il peut être procédé à une « désaccentualisation ». Le principe est de ramener chaque caractère accentué à sa forme basique. Ce traitement ramène donc le mot « livré », participe passé du verbe « livrer » et le mot « livre », l'objet, à une même forme. Il faut cependant signaler que cet exemple fait figure d'exception et que, la majeure partie du temps, ce traitement n'a pas de réelle incidence sur la compréhension même du texte.

Il peut aussi y avoir une suppression des caractères « bruits ». L'identification du bruit, au niveau de l'alphabet, est propre à chaque méthode. Cette suppression consiste, généralement, à supprimer tous les caractères non lettre, c'est-à-dire les marques de ponctuation, les chiffres, … Cependant certaines méthodes s'appuient sur ces caractères. Par exemple, les chiffres ont forte une importance dans le domaine scientifique (« Carbone 14 »), mais n'ont qu'un intérêt réduit dans l'analyse d'œuvres littéraires.

Enfin, une suppression des mots « bruits » peut parfois être envisagée. Les méthodes ayant un tel recours se basent sur la distribution des mots [ZIP 35].

Pour ce projet, sauf exception, la mise en minuscule, la désaccentualisation et la réduction de l'alphabet ont été effectuées systématiquement. Les deux premières réductions permettent d'élargir les informations communes aux textes. La troisième réduction permet une réduction des caractères sans intérêt. Néanmoins, une méthode représentant la structure des textes, présentée partie 2.1.3, nécessite la totalité de l'alphabet et la présence de tous les mots. Pour cette méthode, les textes ne sont pas pré-traités.

1.1.2. Lemmes

Comme écrit précédemment, dans la langue française, les mots n'ont généralement pas une forme unique. Si les noms ne possèdent que deux formes, singulier et pluriel, les adjectifs sont accordés, sauf exception, en genre et en nombre et les verbes offrent autant de formes qu'il existe de temps de conjugaison et de types de sujets. D'un point de vue étymologique, les formes sont encore plus nombreuses. En prenant le mot « faveur », on obtient, par exemple, les mots « favori », « favoriser », « favorable » et toutes les formes qui en découlent.
Sans aller jusqu'au rattachement étymologique, il est parfois préféré de rattacher toutes les formes d'un mot à leur forme canonique. La lemmatisation permet ce rattachement de tout mot à son lemme. La plupart des lemmatiseurs utilisent une analyse syntaxique et un dictionnaire pour un tel résultat. Cela demande donc un traitement linguistique très lourd et conduisant à la présence de nombreux cas de désambiguïsation. [ENG 92] propose une heuristique liant la majorité des formes à une forme unique sans aucun traitement linguistique. L'heuristique consiste à remplacer chaque mot par la sous-chaîne de caractères rassemblant les premières lettres qui composent ce mot jusqu'à l'obtention de deux voyelles non-consécutives. Les doublons de lettres consécutives sont éliminés. Pour les trois formes étymologiques du mot « faveur » citées précédemment, l'heuristique n'offre qu'une unique forme : « favo ». Le Tableau 9 montre d'autres exemples de résultat.

Mot	Forme de rattachement
outrage	Outra
illicite	Ili
automatique	Auto
automobile	Auto

Tableau 9 Exemples de mots et leur forme de rattachement obtenue par [ENG 92]

Si les deux premiers exemples semblent satisfaisants, les deux derniers le sont beaucoup moins. D'une manière générale, le rattachement des formes d'un mot à une forme unique est un problème complexe.
L'absence de règles linguistiques dans une telle lemmatisation ne permet pas d'appeler réellement lemmes les formes trouvées. Comme l'heuristique se base sur la graphie des mots, il est employé, pour la suite du projet, le terme de « lemme graphique ».
En dehors des mots et des lemmes, d'autres formes peuvent être extraites des textes. Ce sont des unités d'un niveau moins élevé. Ce sont les n-grammes.

1.1.3. n-grammes

Les n-grammes sont, en effet, une autre solution d'unités textuelles. Cette solution est à mi-chemin entre l'extraction des mots et les traitements linguistiques ou graphiques de lemmatisation. Les n-grammes de caractères sont largement utilisés en classification de textes

([BIS 02-1], [CAV 96]) et en recherche et extraction d'informations ([BIS 02-2], [CRO 96], [KAR 94], [LEL 98]). Il s'agit d'utiliser une fenêtre de taille fixe et de la glisser sur le texte. Ainsi, pour une fenêtre de taille fixée à 3, le groupe de mots, « la table », sera représenté par les 3-grammes suivants : « la_ », « a_t », « _ta », « tab », « abl », « ble ». Le sigle « _ » permet d'indiquer l'espace. L'utilisation des n-grammes de caractères permet [JAL 02] de capturer automatiquement les racines des mots les plus fréquents, d'être indépendant des langues traitées, d'être tolérant aux fautes d'orthographe et aux déformations et de ne pas nécessiter d'uniformisation des données. Les n-grammes présentent aussi des intérêts de stockage [GU 00]. La taille habituellement recommandée pour le traitement de langues comme le français ou l'anglais est une taille de 5 [DAM 95].

1.1.4. Bilan

Cette partie a permis de présenter deux types d'unités textuelles autres que les mots : les lemmes et les n-grammes.

Les n-grammes ont essentiellement l'avantage de considérer un texte comme un flux d'informations. Mais ils ont aussi l'avantage d'offrir une diversité encore plus grande de motifs. Les corpus Amaryllis et NewsGroups comptent respectivement 9059 et 9338 mots différents. Ces motifs sont réduits au nombre de 3363 et 5756 à la suite d'une lemmatisation graphique. L'évolution du nombre de motifs suivant la longueur des n-grammes utilisés a une évolution exponentielle : 27 1-grammes différents, 508 et 687 2-grammes différents, … et 80034 et 102115 6-grammes différents. La Figure 14 présente, de manière graphique, cette évolution. Cette figure présente aussi la longueur, en nombre de motifs, du texte le plus long du corpus. Cela permet de représenter l'avantage des n-grammes, c'est-à-dire la quantité d'informations représentées. En effet, lorsque l'on compte la présence de moins de 10000 mots, il faut compter trois fois plus de n-grammes.

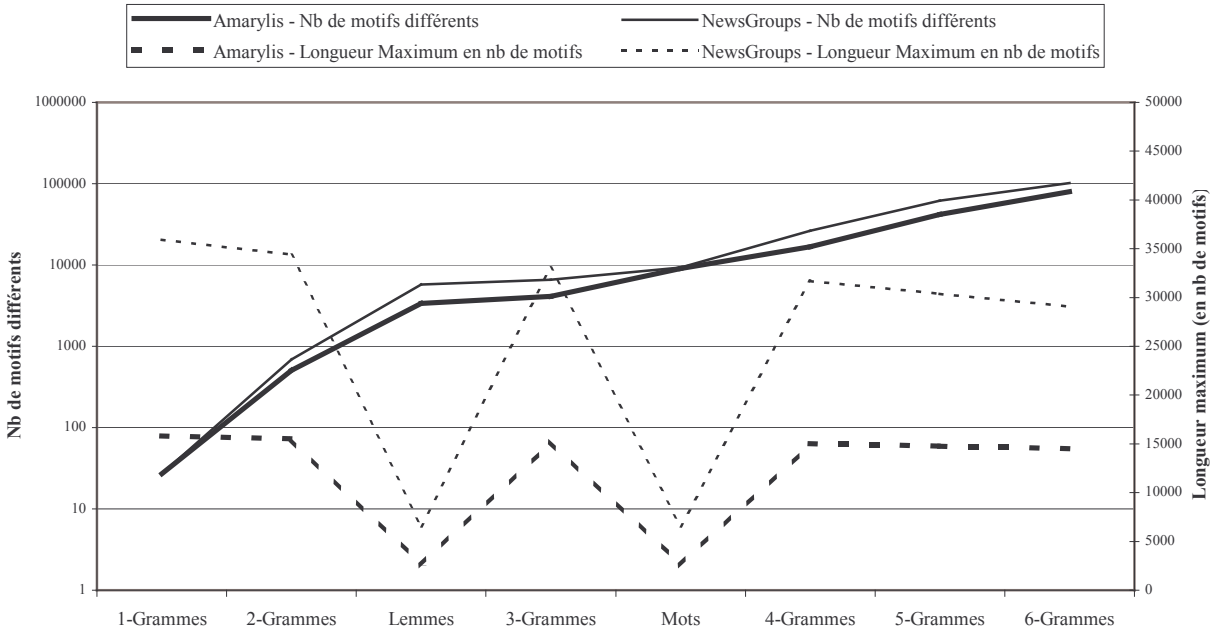

Figure 14 Evolution du nombre de motifs et de la longueur maximum suivant les motifs choisis

Les n-grammes présentent donc de sérieux avantages de découpage. Néanmoins, la quantité d'information qu'ils apportent peut devenir un désavantage. L'intérêt d'un découpage en unités textuelles peut s'évaluer par le rapport entre la longueur maximum et le nombre de motifs différents. Si ce rapport est égale à 1, cela signifie qu'en moyenne chaque texte utilise de manière unique chaque motif. Un rapport supérieur à 1 indique une forte utilisation de chaque motif et donc une base de comparaison. Inversement un rapport inférieur à 1 indique la présence de nombreux motifs utilisés de manière unique, c'est-à-dire utilisés une seule fois dans un unique texte. Un découpage avec un tel rapport n'a aucun intérêt pour une comparaison de textes. Le Tableau 10 présente les rapports pour chaque type de motif et pour chaque corpus.

	1-grammes	2-grammes	Lemmes	3-grammes	Mots	4-grammes	5-grammes	6-grammes
Amaryllis	585,70	30,57	0,75	3,71	0,28	0,90	0,35	0,18
NewsGroups	1330,85	50,06	1,14	5,02	0,70	1,20	0,49	0,28

Tableau 10 Rapport Nombre de motifs Maximum / Nombre de motifs différents pour chaque type de motif et pour chaque corpus

Les 1-grammes, par la quantité réduite de motifs différents, ont le rapport le plus élevé. Il a été écrit précédemment que les mots ne constituaient pas forcément les meilleures unités textuelles de découpage. Il peut, en effet, être constaté que leurs rapports sont assez faibles. Seuls les 5-grammes et les 6-grammes présentent des rapports plus faibles que ceux des mots. Il peut donc être conclu que l'utilisation de tels motifs est moins intéressante que l'utilisation des mots. Inversement, l'utilisation des lemmes graphiques permet de rehausser les rapports des mots, un tel traitement est donc d'un intérêt certain.
Ainsi pour le projet, les types de motifs retenus sont les mots, les lemmes graphiques et les n-grammes de longueur 1 à 4.

1.2. Post-traitements : Pondération

L'indexation est un traitement indépendant pour chaque texte. C'est-à-dire que, généralement, l'indexation d'un texte ne dépend que du texte lui-même. Les autres textes n'ont aucune influence lors de la création de l'index. L'indexation est donc générale.
Un système de pondération permet d'adapter les valeurs d'indexation au corpus traité. Une pondération est, en effet, appliquée à chaque valeur afin qu'elle tienne compte des valeurs des autres textes ou des valeurs des autres caractéristiques du texte.
La première pondération permet de mieux valoriser les informations discriminantes. La pondération par tf*idf est la plus connue dans le traitement des textes.
La seconde pondération permet, quant à elle, d'adapter les valeurs de chaque index pour qu'ils soient comparables. Les textes sont souvent d'une taille différente les uns des autres. Ces différences peuvent être compensées par un recalage des données.
Enfin, dans certains cas, les données donnent de meilleurs résultats si aucune pondération n'est appliquée.
La partie 1.2.1 présente la pondération tf*idf et la partie 1.2.2 présente le recalage des données.

1.2.1. tf*idf

Cette sélection de caractéristiques a été proposée par Salton [SAL 94]. Elle correspond à la multiplication du tf à l'idf.
Le tf (Terme Frequency) essaie d'apporter une représentation linéaire à l'évolution des fréquences observées à l'aide des courbes de Zipf.

$$tf(c_{ij}) = \log(1 + c_{ij}) \tag{7}$$

L'idf (Inverse Document Frequency) se caractérise comme l'inverse de la fréquence totale d'un terme. Cet inverse est normalisé par le nombre de documents et par un logarithme. L'idf est indépendant du texte i traité car il offre une évaluation globale de la caractéristique.

$$idf(c_{ij}) = \log\left(\frac{n}{Card(\{i \in [1,n] \mid c_{kj} > 0 \forall j \in [1,N]\})}\right) \tag{8}$$

Le tf*idf pondère donc chaque valeur selon l'utilisation qui est faite de la caractéristique dans le document mais aussi dans les autres documents. Cette pondération est fortement marquée pour une utilisation des fréquences, elle peut, cependant, être utilisée sur tous les types de données suivant une évolution exponentielle.

1.2.2. Recalage

Le recalage des données est essentiellement utilisé pour les fréquences. C'est la raison pour laquelle cette pondération porte habituellement le nom de recalage des fréquences. Ce recalage consiste à diviser les valeurs calculées sur un texte par la somme des valeurs. Du point de vue des fréquences, cela permet de comparer entre eux les textes. C'est-à-dire que cela permet de comparer entre eux des textes comparables du point de vue de la différence de taille. Le recalage consiste à passer de l'ordre des fréquences absolues à l'ordre des probabilités.
Comme pour la pondération tf*idf, cette pondération trouve son application sur tous les jeux de données.

2. Les méthodes d'indexation

Cette partie s'intéresse aux méthodes permettant de passer de textes formatés en motifs de base à des index multidimensionnels. Les méthodes considèrent, jusqu'à présent, les textes comme un unique ensemble de données. Cela signifie que les méthodes d'indexation proposent une indexation globale du texte, sans tenir compte du contexte local. En d'autres termes, cela signifie que le discours en lui-même, son organisation, n'est pas représenté. Il a été développé lors ce travail plusieurs alternatives, [MAR 04-2], [MAR 05-2], qui essaient de représenter l'organisation du discours d'une part et son évolution d'autre part. La différence d'état d'esprit entre les méthodes habituelles et les méthodes proposées par ce travail peut être résumée simplement en disant que les méthodes habituelles génèrent un index global à partir des données globales, l'alternative génère un index global à partir des données locales.
Cette partie se détaille en trois sous-parties. La première sous-partie présente les méthodes d'indexation basées sur une étude globale des textes. La deuxième sous-partie présente des

méthodes alternatives qui représentent l'organisation du discours. La troisième sous-partie présente une autre méthode alternative qui représente l'évolution des textes.

2.1. Les méthodes d'indexation basées sur la représentation globale

Cette partie présente trois méthodes différentes d'indexation de textes. La première méthode, présentée dans la partie 2.1.1, propose de représenter statistiquement le contenu des textes à partir de vecteurs. La seconde méthode, présentée dans la partie 2.1.2, est réputée pour permettre de distinguer visuellement deux textes de natures différentes. La comparaison est réalisée au niveau de la distribution des unités textuelles, c'est-à-dire au niveau de la structure des textes. La partie 2.1.3 propose, quant à elle, une représentation purement structurelle des textes.

2.1.1. Vecteurs

La représentation pas vecteur de fréquences est la plus commune [SAL 89] pour la représentation de textes sous forme d'index.

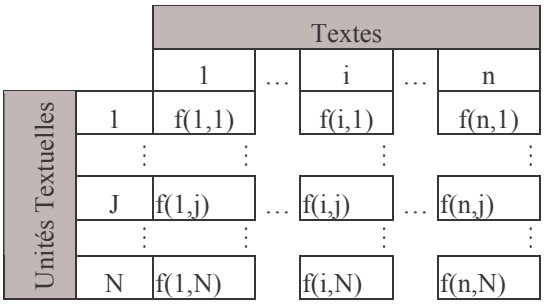

Figure 15 Représentation des textes par vecteur de fréquences

Chaque texte est représenté par un vecteur dont les caractéristiques suivent un index commun à tous les textes du corpus. Le corpus peut donc être représenté sous la forme d'une matrice (Figure 15) où chaque colonne représente un texte, chaque ligne représente une unité textuelle et où la valeur indiquée en ligne i et colonne j est celle de la fréquence de l'unité textuelle j dans le texte i.
Cette représentation reste la base d'une majorité de méthodes. Les différences entre ces méthodes se trouvent dans la sélection des caractéristiques et les méthodes de classification.

Le Tableau 11 et le Tableau 12 présentent les résultats de classification obtenus sur les corpus Amaryllis et NewsGroups, la mesure de F1, présentées dans la partie 3 du chapitre 3, est utilisée.

Pour le corpus Amaryllis, la pondération dominante est la pondération par le recalage des fréquences. Seule l'indexation sur les mots ne suit pas cette tendance. Il peut, néanmoins, être remarqué que le score obtenu par la distance euclidienne suite à un recalage apparaît comme le troisième meilleur score de la colonne.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	10,91%	15,02%	11,79%	11,60%	13,08%	11,13%
	Euclidienne Standardisée	17,18%	12,73%	12,91%	15,66%	12,13%	14,84%
	Gower	28,13%	13,31%	11,13%	12,93%	17,77%	13,95%
	KullBack-Leibler Symétrie IV	11,07%	12,26%	8,62%	8,09%	8,30%	10,11%
	Mahalanobis	12,49%	7,69%	10,98%	9,44%	12,71%	12,88%
	Russel-Rao	15,00%	18,53%	12,55%	15,77%	16,56%	20,99%
	Soergel	26,58%	19,04%	17,71%	12,58%	12,91%	25,72%
Recalage	Bhattacharyya	12,15%	18,62%	19,19%	30,07%	35,77%	21,90%
	Euclidienne Standardisée	27,45%	**43,66%**	18,72%	35,39%	32,99%	**41,86%**
	Gower	11,29%	18,62%	24,21%	**36,13%**	**37,68%**	13,79%
	KullBack-Leibler Symétrie IV	11,54%	11,36%	7,74%	14,72%	8,05%	13,22%
	Mahalanobis	9,55%	29,49%	**26,31%**	30,02%	27,39%	29,91%
	Russel-Rao	8,65%	8,73%	13,87%	12,35%	12,46%	8,73%
	Soergel	12,45%	13,60%	17,79%	26,08%	14,14%	14,54%
Tfldf	Bhattacharyya	10,61%	12,43%	8,58%	12,68%	12,56%	7,50%
	Euclidienne Standardisée	25,37%	8,46%	7,83%	11,85%	17,99%	14,84%
	Gower	19,86%	18,89%	17,35%	14,08%	12,43%	13,12%
	KullBack-Leibler Symétrie IV	11,07%	8,05%	8,65%	13,25%	8,46%	13,40%
	Mahalanobis	11,65%	13,76%	11,10%	13,27%	12,57%	12,33%
	Russel-Rao	**28,17%**	25,41%	15,06%	12,91%	19,54%	21,46%
	Soergel	25,96%	18,75%	15,85%	12,09%	12,21%	21,67%

Tableau 11 Score F1 sur le corpus Amaryllis avec une indexation par vecteur de fréquences

Cette exception ne présente que peu d'importance puisque l'indexation par les mots n'offre que l'avant-dernier score. Seule l'indexation par les 1-grammes, c'est-à-dire les 27 caractères (26 lettres + Caractère spécial) a un résultat pire.

Le meilleur score est obtenu par l'indexation par lemmes graphiques, suivie par l'indexation par les 4-grammes, 3-grammes et 2-grammes. Les lemmes graphiques présentent donc un double intérêt, puisqu'ils constituent le meilleur produit nombre de motifs différents * longueur maximum et qu'ils offrent le meilleur résultat de classification pour une indexation par vecteurs.

Du point de vue des mesures de dissimilarités, la distance euclidienne standardisée donne les meilleurs résultats globaux (moyenne de 33,35%), elle est suivie par la distance de Mahalanobis (25,45%) puis par la distance de Gower (23,62%).

De ce premier corpus, il peut donc être conclu que les deux premières étapes doivent suivre les directives suivantes : lemmatisation graphique, indexation par vecteur, pondération par recalage des fréquences, normalisation par la valeur maximum de chaque caractéristique et enfin calcul des distances par la distance euclidienne standardisée.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	10,51%	8,21%	7,73%	5,77%	14,04%	10,70%
	Euclidienne Standardisée	8,00%	9,35%	4,33%	11,20%	18,33%	8,67%
	Gower	7,17%	7,08%	5,51%	6,27%	12,17%	12,26%
	KullBack-Leibler Symétrie IV	5,94%	5,83%	6,28%	5,68%	4,30%	4,27%
	Mahalanobis	9,24%	11,55%	5,30%	15,32%	11,83%	11,20%
	Russel-Rao	7,76%	7,78%	7,18%	4,12%	4,49%	7,73%
	Soergel	16,93%	16,94%	7,69%	10,25%	7,39%	17,20%
Recalage	Bhattacharyya	12,43%	6,59%	4,65%	4,30%	**21,73%**	8,94%
	Euclidienne Standardisée	15,76%	**28,55%**	**10,67%**	18,26%	16,72%	18,10%
	Gower	13,29%	12,34%	7,56%	**25,87%**	14,23%	**23,75%**
	KullBack-Leibler Symétrie IV	4,17%	5,95%	6,27%	8,34%	5,51%	8,63%
	Mahalanobis	14,55%	20,83%	6,30%	16,84%	11,72%	24,45%
	Russel-Rao	4,71%	8,06%	4,72%	5,88%	6,91%	4,49%
	Soergel	21,50%	18,21%	5,97%	5,43%	16,37%	7,63%
Tfldf	Bhattacharyya	18,32%	14,82%	9,09%	12,93%	7,84%	17,68%
	Euclidienne Standardisée	21,26%	19,78%	7,04%	5,20%	16,81%	21,76%
	Gower	4,55%	12,87%	7,61%	10,47%	8,68%	20,12%
	KullBack-Leibler Symétrie IV	4,10%	5,82%	5,66%	5,72%	8,82%	5,90%
	Mahalanobis	9,55%	6,28%	5,01%	8,18%	13,70%	20,37%
	Russel-Rao	7,55%	7,63%	4,75%	8,30%	10,58%	10,04%
	Soergel	**23,49%**	22,40%	8,46%	5,62%	10,35%	12,77%

Tableau 12 Score F1 sur le corpus NewsGroups avec une indexation par vecteur de fréquences

Le corpus issu des NewsGroups propose des scores moins élevés que les précédents. La conclusion est, cependant, identique. La pondération après indexation offrant les meilleurs scores globaux est, à nouveau, la pondération par recalage des fréquences.
La meilleure indexation est, à nouveau celle par les lemmes graphiques. Il faut, au niveau des motifs, noter la perturbation pour le reste du classement. L'indexation par les 2-grammes offre, en effet, le second score, suivi par l'indexation par les 4-grammes. Les mots prennent une place et passent devant le score des 3-grammes.
De même, au niveau des mesures de dissimilarités, la distance euclidienne normalisée reste la meilleure des mesures (moyenne de 18,01%), les scores suivants sont inversés et la distance de Gower prend la seconde place (16,17%) et la distance de Mahalanobis la troisième (15,78%).

Ces expérimentations apportent un premier schéma d'indexation – classification. Mais au-delà de ce schéma, ces expérimentations laissent à supposer que la qualité des motifs et la qualité des mesures de dissimilarités sont variables suivant le type de corpus étudié. Cette remarque n'est pas confirmée pour les meilleurs résultats de cette méthode d'indexation. Mais elle est confirmée par les nombreuses divergences entre les deux corpus pour les résultats qui suivent.

2.1.2. Zipf

La représentation d'un texte par la méthode de Zipf [ZIP 35] est, généralement, associée aux courbes qui en découlent. Cette méthode est une variante de la précédente. La grande différence tient dans le fait que, pour cette représentation, aucun index général n'est créé. Les valeurs de chaque vecteur sont ordonnées par ordre décroissant de leur fréquence. [COH 97] a montré qu'il est ainsi graphiquement possible de distinguer des textes écrits de manière naturelle de textes artificiels.

Le Tableau 13 et le Tableau 14 présentent les scores de classifications obtenus par cette méthode d'indexation pour les corpus Amaryllis et NewsGroups.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	10,18%	8,42%	11,76%	8,58%	11,92%	8,30%
	Euclidienne Standardisée	10,21%	**14,57%**	12,16%	11,21%	11,87%	11,98%
	Gower	10,91%	8,42%	16,37%	10,41%	12,09%	11,47%
	KullBack-Leibler Symétrie IV	10,18%	8,46%	10,04%	10,55%	12,08%	8,62%
	Mahalanobis	8,42%	8,42%	13,21%	10,37%	11,87%	**12,27%**
	Russel-Rao	12,28%	12,05%	11,12%	12,86%	12,32%	11,83%
	Soergel	10,21%	8,46%	12,58%	10,33%	12,15%	11,52%
Recalage	Bhattacharyya	10,18%	8,42%	8,26%	13,15%	11,92%	8,42%
	Euclidienne Standardisée	10,18%	10,61%	11,84%	11,82%	11,89%	11,31%
	Gower	10,86%	11,34%	8,54%	12,28%	11,87%	11,73%
	KullBack-Leibler Symétrie IV	10,25%	8,46%	12,63%	10,55%	12,08%	8,69%
	Mahalanobis	10,18%	11,34%	13,95%	12,57%	12,03%	12,02%
	Russel-Rao	11,41%	11,06%	8,73%	11,84%	12,03%	10,97%
	Soergel	10,84%	11,01%	15,59%	11,48%	12,08%	8,22%
TfIdf	Bhattacharyya	10,84%	8,42%	8,13%	13,09%	11,92%	8,42%
	Euclidienne Standardisée	10,77%	11,01%	15,34%	**16,90%**	12,08%	12,09%
	Gower	14,02%	10,98%	**18,07%**	10,60%	12,08%	11,73%
	KullBack-Leibler Symétrie IV	10,94%	8,46%	12,35%	10,55%	12,08%	8,65%
	Mahalanobis	10,07%	11,01%	15,99%	11,24%	**12,16%**	12,09%
	Russel-Rao	**12,49%**	12,38%	12,13%	12,49%	12,03%	11,74%
	Soergel	10,06%	8,46%	15,10%	10,55%	12,08%	11,77%

Tableau 13 Score F1 sur le corpus Amaryllis avec une indexation par Zipf

Avec le corpus Amaryllis, cette méthode d'indexation obtient de meilleurs scores avec la pondération par tf*idf. Ces résultats sont, par contre, clairement, inférieurs aux résultats de l'indexation par vecteurs. Le meilleur résultat de l'indexation par vecteur, sur le même corpus est, en effet, plus de deux fois supérieur au meilleur résultat obtenu par l'indexation par Zipf.

Ici aussi les trois meilleures mesures de dissimilarités sont la distance euclidienne standardisée (moyenne de 13,03%), puis la distance de Gower (12,91%) et enfin la distance de Mahalanobis (12,21%). L'écart entre ces mesures est faible et celui avec les mesures suivantes l'est tout autant.

Le type de motifs offrant le meilleur score est le 1-gramme. Ce choix n'apparaît pas comme un simple hasard, puisque la colonne des 1-grammes regroupe, dans la majorité des cas, le meilleur score de la ligne. Cela aurait tendance à signifier que l'évaluation de la structure d'un texte se fait idéalement à son plus bas niveau.

Les scores baissent encore d'un cran avec le corpus NewsGroups. Le rapport entre les meilleurs scores est, à présent, proche de 3. Il faut noter, comme pour le corpus Amaryllis, une domination des 1-grammes. Compte tenu des mauvais scores, cette remarque est relative et les différences notées ne sont que très faibles.

La pondération qui offre les meilleurs scores est le recalage des fréquences. Pour cette pondération, les meilleurs scores ne sont, d'ailleurs, pas obtenus par les 1-grammes. Au-delà de cette observation, la distance de Russel-Rao a le meilleur score global (moyenne de 7,61%) et présente l'un des pires scores dans la colonne des 1-grammes.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	4,15%	5,82%	6,51%	4,01%	5,19%	4,25%
	Euclidienne Standardisée	5,81%	6,06%	4,27%	5,89%	5,63%	4,15%
	Gower	7,87%	5,69%	5,69%	5,81%	5,96%	6,09%
	KullBack-Leibler Symétrie IV	4,41%	6,21%	6,25%	6,05%	5,76%	6,03%
	Mahalanobis	5,77%	5,94%	5,61%	5,89%	5,59%	6,00%
	Russel-Rao	5,60%	5,99%	4,35%	**6,45%**	5,85%	6,06%
	Soergel	5,75%	5,81%	6,92%	5,99%	5,97%	5,95%
Recalage	Bhattacharyya	5,17%	6,09%	8,42%	3,55%	4,69%	4,15%
	Euclidienne Standardisée	5,74%	**7,32%**	9,88%	5,54%	5,19%	5,90%
	Gower	6,08%	6,16%	8,06%	5,82%	5,62%	5,90%
	KullBack-Leibler Symétrie IV	4,30%	6,05%	4,46%	6,05%	4,33%	6,03%
	Mahalanobis	**9,08%**	6,15%	7,36%	3,90%	6,58%	6,35%
	Russel-Rao	6,77%	9,79%	5,78%	4,33%	**9,25%**	**9,73%**
	Soergel	6,01%	6,05%	6,59%	6,01%	5,81%	6,00%
Tfidf	Bhattacharyya	5,41%	5,82%	**10,41%**	4,84%	5,27%	4,19%
	Euclidienne Standardisée	5,63%	5,52%	5,35%	5,93%	5,80%	5,65%
	Gower	5,87%	5,90%	10,03%	6,01%	5,83%	5,97%
	KullBack-Leibler Symétrie IV	4,41%	6,02%	4,12%	6,05%	5,76%	6,03%
	Mahalanobis	5,77%	5,96%	7,73%	6,08%	5,72%	5,63%
	Russel-Rao	5,94%	6,13%	4,58%	6,35%	5,92%	5,58%
	Soergel	6,07%	6,05%	7,83%	6,02%	6,00%	6,00%

Tableau 14 Score F1 sur le corpus NewsGroups avec une indexation par Zipf

Inversement, la distance de Bhattacharyya avec la pondération du tf*idf offre le meilleur résultat de la table et de la colonne des 1-grammes. Son résultat global est pourtant l'un des plus faibles (5,99%).
Avec le second meilleur score de la table (10,03%) et le second meilleur score global (6,60%) la distance de Gower sur les 1-grammes avec une pondération par le tf*idf apparaît être le meilleur compromis.

Cette méthode d'indexation est bien plus mauvaise que la méthode d'indexation par les vecteurs de fréquences. Cette méthode confirme l'importance du lien entre le corpus étudié, la représentation et la mesure. Elle ajoute au lien, l'importance de la pondération qui présenterait des spécificités selon le corpus. Cette remarque prend plus de valeur avec cette méthode d'indexation. Les résultats sont, en effet, si faibles que le schéma proposant les meilleurs résultats ne s'impose par de manière radicale face aux autres résultats.

2.1.3. Information structurelle

[DEV 00] propose une représentation particulière des textes. Celle-ci a pour but de modéliser la structure du texte à partir d'une liste de caractéristiques.
Ces caractéristiques sont :

- le nombre total de mots utilisés
- la longueur moyenne des mots utilisés
- le nombre de phrases
- la longueur moyenne des phrases
- le nombre de paragraphes
- le nombre de lignes blanches
- la longueur moyenne des lignes
- le nombre de caractères lettre
- le nombre de caractères en majuscule
- le nombre de caractères non-lettre
- le nombre de chiffres
- le rapport entre le nombre de mots courts (longueur <5) et le nombre de mots total
- le rapport entre le nombre de mots utilisés de manière unique et le nombre de mots total
- la fréquence des 5 premiers mots courts communs les plus utilisés

La dernière caractéristique est donc une série de caractéristiques. De plus, il faut signaler que les caractéristiques ont été légèrement modifiées de la proposition originale. Celles-ci restaient trop attachées à leur utilisation aux mails. La méthode originale utilise, pour ses 5 dernières caractéristiques, les mots outils. Dans l'adaptation qui en est faite, un mot outil est défini comme un mot court apparaissant très fréquemment. Ainsi, l'adaptation proposée se base complètement sur la description statistique.
Cette méthode est donc une pure représentation textuelle obtenue de manière statistique.

Le Tableau 15 et le Tableau 16 présentent les scores obtenus à la suite d'une indexation par la structure pour les corpus Amaryllis et NewsGroups.

Pour le corpus Amaryllis, la mesure de dissimilarité est fixée, mais la pondération semble liée au type de motifs. Ainsi, une indexation par les mots offre de meilleur résultat sans pondération et une indexation par les lemmes offre de meilleurs résultats si un recalage est effectué.
Il faut rappeler que le recalage constitue une normalisation des valeurs de chaque texte par la somme de ces mêmes valeurs. Cela signifie que l'indexation par les mots est plus performante

lors d'une utilisation de manière absolue alors que l'indexation par les lemmes graphiques est plus performante lors d'une utilisation relative.

		Mots	Lemmes
Id	Bhattacharyya	14,32%	8,09%
	Euclidienne Standardisée	22,42%	14,54%
	Gower	**23,28%**	13,79%
	KullBack-Leibler Symétrie IV	8,46%	8,17%
	Mahalanobis	19,33%	13,21%
	Russel-Rao	16,89%	8,73%
	Soergel	10,91%	14,54%
Recalage	Bhattacharyya	12,40%	12,64%
	Euclidienne Standardisée	8,65%	14,70%
	Gower	7,55%	**20,53%**
	KullBack-Leibler Symétrie IV	8,38%	8,73%
	Mahalanobis	8,65%	14,89%
	Russel-Rao	15,54%	17,51%
	Soergel	12,63%	8,17%
TfIdf	Bhattacharyya	8,00%	8,30%
	Euclidienne Standardisée	8,00%	8,30%
	Gower	11,30%	8,13%
	KullBack-Leibler Symétrie IV	11,69%	8,22%
	Mahalanobis	8,00%	8,30%
	Russel-Rao	15,99%	11,26%
	Soergel	8,00%	8,30%

Tableau 15 Score F1 sur le corpus Amaryllis avec une indexation par la Structure

Le corpus issu des NewsGroups confirme la même tendance que celle affirmée lors de l'étude du corpus Amaryllis. Dans ce cas, la distance de Gower sur les lemmes à la suite d'un recalage se distingue nettement des autres cas possibles.
Le score de 28,58% atteint par ce schéma est le meilleur score atteint pour le traitement du corpus NewsGroups. Il faut noter que ce score dépasse le meilleur score pour ce corpus de l'indexation par vecteurs de seulement 0,03%. La différence sur le corpus Amaryllis est de l'ordre de 20%, favorable pour l'indexation par vecteurs.

L'indexation par structure s'insère donc dans une suite d'opérations (Lemmes – Indexation par Structure – Recalage – Gower) qui offre pour un corpus des résultats comparables à ceux de l'indexation par vecteurs. Cette indexation présente surtout l'avantage de ne contenir que 18 caractéristiques. Lié au fait que le produit nombre de motifs différents * longueur maximum est le plus faible, cette suite d'opérations apparaît comme l'une des plus intéressantes.

		Mots	Lemmes
Id	Bhattacharyya	15,06%	5,59%
	Euclidienne Standardisée	14,01%	16,36%
	Gower	18,50%	22,46%
	KullBack-Leibler Symétrie IV	7,30%	5,78%
	Mahalanobis	10,36%	14,87%
	Russel-Rao	9,11%	4,68%
	Soergel	**19,92%**	8,95%
Recalage	Bhattacharyya	14,75%	10,21%
	Euclidienne Standardisée	8,44%	14,31%
	Gower	15,47%	**28,58%**
	KullBack-Leibler Symétrie IV	6,94%	6,17%
	Mahalanobis	12,19%	20,07%
	Russel-Rao	14,84%	9,76%
	Soergel	18,54%	18,92%
Tfldf	Bhattacharyya	6,90%	13,67%
	Euclidienne Standardisée	8,45%	13,70%
	Gower	10,11%	13,64%
	KullBack-Leibler Symétrie IV	8,33%	11,26%
	Mahalanobis	8,51%	13,48%
	Russel-Rao	8,72%	8,87%
	Soergel	8,81%	13,64%

Tableau 16 Score F1 sur le corpus NewsGroups avec une indexation par la Structure

2.1.4. Bilan des méthodes basées sur une représentation globale

Cette partie a permis d'étudier trois méthodes d'indexation basées sur une représentation générale du texte. C'est-à-dire que le lien entre les textes et leur forme indexée est direct. De ces six méthodes, deux méthodes peuvent être retenues.

La méthode d'indexation par les Vecteurs permet de représenter les textes par leur contenu. C'est la méthode la plus couramment utilisée. Les scores de classification obtenus par cette méthode restent les meilleurs pour un schéma de classification comme celui de ce projet. La taille démesurée des index est le plus gros désavantage de l'indexation par Vecteurs.

La méthode d'indexation par la Structure permet de représenter les textes par les informations structurelles des textes. Cette méthode est donc basée sur la forme plus que sur le fond. Cette méthode d'indexation obtient des résultats inférieurs à ceux de l'indexation par Vecteurs pour un index ne contenant qu'une vingtaine de caractéristiques.

2.2. Méthode représentant l'organisation du discours

La méthode d'indexation par la Structure montre que la structure pourrait être une base pour la comparaison des textes. Si une majorité de méthodes s'attachent au contenu sous toutes ses

formes, Mme Denèfle rappelle, avec sa culture de sociologue, qu'il existe 50 façons de dire : « Il fait beau ». Cette affirmation pourrait à elle seule contredire tous les protagonistes du contenu. Cependant, ce qu'elle sous-entend, c'est qu'il existe un lien fort entre le contenu et la structure. Le contenu n'existe et ne prend son sens que parce que la structure existe. Et inversement, une structure sans contenu n'a aucune cohérence. Cela est d'autant plus vrai avec des textes issus d'entretiens sociologiques. Il a été vu à maintes reprises que le contenu était « vide » d'un point de vue linguistique, c'est-à-dire que les textes sont essentiellement constitués de mots vides de sens lorsqu'ils sont pris indépendamment. Il n'y a que deux solutions. Soit, il faut passer à une couche supérieure jusqu'à ce qu'ils trouvent un sens. C'est un peu le principe des unités de contexte élémentaires d'Alceste [LOG ALC]. Soit, au contraire, il faut s'enfoncer à un niveau si fin du texte que, tel l'ADN, il dévoile sa structure et donc sa signature. Le style étant propre à chacun, comparer des personnes consisterait à comparer les textes auxquels elles sont associées. Il nous a semblé important de rechercher la structure des textes tout en conservant les statistiques. Trois méthodes ont été développées dans ce sens. Les sous-parties 2.2.1, 2.2.2 et 2.2.3 présentent ces trois méthodes originales crées et développées pour ce travail et ayant pour but d'extraire et de représenter une structure des textes.

2.2.1. Images

L'indexation par l'image est la première indexation développée pour ce projet. La version finale est la conséquence de plusieurs travaux [MAR 03-2], [MAR 04-1], [MAR 04-2]. Cette méthode propose un changement radical de la représentation. Cette méthode d'indexation peut paraître aussi originale que [BAV 02] qui représente les textes par les lois de la thermodynamique. L'indexation par l'Image présente surtout l'avantage de proposer une représentation de taille fixe.
Les méthodes d'indexation de textes analogues à celles vues précédemment se ramènent à un condensé d'informations des textes originaux : vecteurs statistiques, par exemple. Ce condensé d'informations sert de nouvelle représentation (index) pour les textes de l'étude. Le problème de tels index est qu'ils n'admettent aucune limite prévisible de taille.
Les deux premières méthodes d'indexation présentées, par Vecteurs et par Zipf, dépendent, en effet, du nombre de mots distincts du corpus ou des textes eux-mêmes. Or le nombre de mots différents est fortement variable d'un texte à l'autre et d'un corpus à l'autre. Et, avant tout, ce maximum dépend de la longueur des textes.
Pour résoudre ce problème, le nombre d'informations pourrait être limité. Mais la comparaison de deux textes, l'un court et l'autre bien plus long, conservera l'ensemble des informations pour le texte court et une quantité trop faible d'informations pour le texte long. Pour ces méthodes, il n'existe donc pas une taille fixe et universelle.

L'indexation par l'Image résout ce problème en transformant chaque texte en une image de taille fixe. C'est une idée originale car, a priori, un texte apparaît comme un signal mono-dimensionnel alors qu'une image est plane. Les voisinages usuels d'un point n'ont pas du tout la même structure dans les deux types d'espace mono et bi – dimensionnels.
Cependant en transformant chaque texte en une image de taille fixe, cette méthode propose une structure d'index quasiment universelle. Cette représentation n'est pas complètement universelle car elle dépend, comme cela sera détaillé dans la suite, d'un patron. Or ce patron^, pour être efficace, doit être organisé suivant la langue étudiée.

Par contre, cette méthode a l'avantage de pouvoir être appliquée aussi bien sur des textes assez longs comme des œuvres littéraires que sur des textes plus courts, des entretiens à questions ouvertes ou des discours politiques. Cette méthode pourrait être appliquée à des textes « abstraits » comme un texte représentant un son [DEL 02] ou un texte représentant une image [NIK 02], [LIN 00] utilisant un large éventail de caractères pouvant être ordonnés les uns par rapport aux autres. Cette méthode est inspirée de celle proposée par [DES 99] et [LES 03] pour le traitement de chaîne ADN.

Le texte, dans sa structure même, comporte un aspect temporel très marqué, et il est nécessaire de lire le texte entier pour se forger une impression globale. Au contraire, l'observation d'une image laisse instantanément une sensation globale. Evidemment les études occulométriques montrent qu'un balayage temporel contextuel a lieu mais sa rapidité n'en laisse aucune conscience à l'observateur. L'image d'un texte peut donc être attendue comme une représentation globale de la structure et du contenu du texte organisée de manière à offrir une étude rapide des informations locales. L'image qui est construite par la méthode est donc une représentation particulière des unités textuelles propres à chaque texte. La partie 2.2.1.1 présente les pré-traitements nécessaires à la méthode.
Une image peut être considérée comme un ensemble de motifs placés de manière harmonieuse dans un univers à deux dimensions.
La partie 2.2.1.2 présente l'organisation qui est suivie par la méthode et la partie 2.2.1.3 présente la démarche à suivre pour obtenir l'image finale. Enfin la partie 2.2.1.4 présente l'indexation numérique qui est faite de l'image.

2.2.1.1.Pré-traitements

Un alphabet correspond au jeu complet de caractères de l'ensemble des textes étudiés. L'alphabet français est constitué de 26 lettres en minuscules, 26 lettres en majuscules, de 10 caractères numériques, de lettres accentuées, de marques de ponctuation, …
Les alphabets suivent quelques variations avec les langues et avec le formatage des textes (certains ne sont écrits qu'en majuscules, certains sont écrits sans accents, …). Pour cette méthode, il a été donné plus d'importance à la longueur des n-grammes qu'à la quantité de caractères différents.
De plus, il a été choisi de conserver l'image la plus symétrique possible. Pour cela, il a été choisi de créer une image carrée. Ces caractéristiques permettent une plus grande ouverture dans le choix de l'indexation de l'image. Par contre, cela nécessite que l'alphabet utilisé dans le texte ait une cardinalité ayant une racine carrée entière.
Compte tenu que l'alphabet utilisé en langue française (et anglaise) est composé à la base de 26 lettres (c'est-à-dire sans accent, sans majuscule) cette méthode s'est limitée à 25 caractères. C'est, en effet, le carré le plus proche de 26. Ainsi, la limitation dans le nombre de lettres est faite mais les plus représentatives peuvent toutes être gardées.
Pour réussir une telle limitation, les textes sont mis en minuscules et sont désaccentués. De plus, tout caractère non lettre est regroupé sous la forme d'un caractère dit spécial noté « @ ». Donc, l'ensemble des lettres est restreint à 24 en regroupant sous une même forme les lettres « c » et « k » et les lettres « v » et « w ».
Le nouvel alphabet ne compte ainsi plus que 25 caractères.

Cette réduction d'alphabet donne plus de poids à la lecture en n-grammes. En effet, le nouvel alphabet réduit le formatage du texte et permet à la méthode de lecture d'être plus tolérante aux erreurs sur ce point. Cette réduction a aussi pour avantage de fixer des limites prévisibles quant au nombre de n-grammes possibles : 25^n.

2.2.1.2.Patron de création

Pour que la méthode d'indexation soit cohérente, il faut qu'elle soit appliquée sur des images ayant au départ la même structure sous-jacente, c'est-à-dire la même organisation interne. Pour que cela soit possible, il faut que le schéma d'organisation des motifs, le patron de création, soit le même pour toutes les images du corpus.
Le patron de création est très important car c'est lui qui donnera un sens local et global aux images créées. C'est, en effet, lui qui fera ressortir de l'image un ensemble structuré représentatif du texte ou un ensemble désordonné de motifs. Cette partie s'attache à expliquer ce qu'est le patron et les règles sur lesquelles il est fondé.

Le but étant de créer une image carrée et l'alphabet étant constitué de 25 caractères différents, il suffit d'organiser les motifs représentant ces caractères dans une matrice de taille 5x5. L'organisation des motifs n'a aucune raison de suivre l'ordre alphabétique, il n'a jamais été dit que cet ordre était représentatif des textes, qu'ils soient en français, en anglais, ou même en latin. D'une manière générale, chaque langue se caractérise par la fréquence de ses 1-grammes de caractères [MAR 04-2]. Cet ordre est si représentatif qu'il est utilisé pour résoudre les systèmes cryptographiques de base. Il faut ajouter le fait qu'en cryptographie, l'étude se fait sur plusieurs longueurs car les 1-grammes s'articulent autour de groupes de fréquences : E / ASITN / RULO / D / CMP / VQGFBH / JX / YZ / KW (en anglais E / TA / ONISRH / LDCU / PFMW / YBGV / KQXJZ). Sur une telle liste, il peut être observé que le regroupement de lettres précédemment effectué suivait déjà l'ordre de fréquence des lettres, puisque ce sont les deux lettres les moins fréquentes qui sont regroupées avec leur semblable. A cet ordre, il faut évidemment ajouter le caractère spécial qui tient, naturellement, la tête puisqu'il représente l'ensemble des caractères non-lettre (espace, ponctuation chiffres, …).

Pour que cet ordre soit représenté dans l'image, le système du point chaud (ou de création des dunes) a été imité. La Figure 16 montre le schéma de répartition des caractères qui a été suivi et la Figure 17 montre le patron, ainsi, obtenu pour la langue française. Avec un tel schéma, on trouve en haut à gauche de la matrice le caractère le plus fréquent, dit point chaud, et plus l'éloignement de ce point est important plus la fréquence d'apparition en langue française est faible.

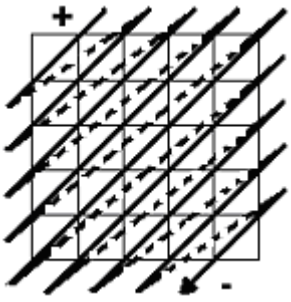

Figure 16 Organisation des classes de fréquence

@	E	S	N	O
A	I	R	D	V
T	U	C	Q	B
L	M	G	H	X
P	F	J	Y	Z

Figure 17 Patron de création pour la langue française

Le patron est donc représentatif des textes en langue française. Une fois celui-ci adopté, c'est le même patron qui sera appliqué sur l'ensemble des textes en langue française. C'est en cette précision que la méthode n'est pas complètement universelle. Il est, de toute façon, rare de comparer tels quels les contenus de textes écrits dans des langues différentes. Et une comparaison des structures n'aurait aucun sens si elle était basée sur la structure d'une langue en particulier.
Le patron donne ainsi l'organisation d'une image pour les caractères. Comme cela a été écrit précédemment, cela n'aurait un intérêt que pour différencier les langues (structure primaire des textes).

Il a été vu précédemment le patron de création pour les 1-grammes de caractères. Pour les n-grammes de caractères, la méthode est identique.
Puisque l'ordre des fréquences d'apparition des 1-grammes de caractères est considéré comme un ordre naturel pour le corpus, c'est cet ordre qui détermine aussi le patron des n-grammes de caractères (pour tout n). C'est donc le même patron qui est appliqué de manière récursive à l'intérieur de ses propres cases. On s'approche ainsi de la construction d'une image fractale. La Figure 18 montre de façon explicite cette répétition.

Chaque répétition récursive est appelée couche. Ainsi pour trouver l'information relative au motif « ab », il faut, dans la première couche, aller dans la case correspondant au caractère « a » (1ère colonne, 2ème ligne). Puis dans la seconde couche, il faut aller dans la case correspondant au caractère « b » (5ème colonne, 3ème ligne). Les coordonnées du motif « ab » dans le patron des 2-grammes de caractères (dimension : 25x25) est donc 5ème colonne, 8ème ligne.

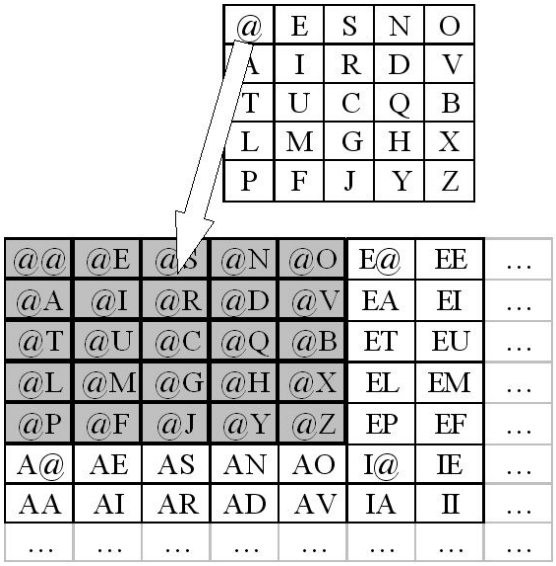

Figure 18 Construction récursive pour la représentation des n-grammes

La représentation de la structure est d'autant plus précise que le nombre n de caractères des n-grammes est important. L'apport de précision suit une évolution logarithmique alors que la taille des matrices suit une évolution exponentielle [MAR 03]. L'équilibre a donc été trouvé avec n=4. Ce nombre reste inférieur au 5 proposé par [DAM 95], mais il reste néanmoins

régulièrement utilisé dans le domaine. Avec un tel nombre de couches, la taille des matrices est donc de 625x625.

2.2.1.3.Du patron à l'image

L'image est donc un ensemble agencé de pixels où chaque motif représente un n-gramme. Chaque n-gramme étant représentatif par sa fréquence, cette information est traduite par le niveau de gris du motif correspondant. Plus un n-gramme est fréquent, plus le motif qui le représente aura un niveau de gris élevé. L'absence du motif est signifié par un niveau de gris de 0 et le motif du n-gramme le plus fréquent a une valeur de 255. Cette représentation est donc une représentation relative. Les figures suivantes montrent le résultat obtenu pour les 1 à 4 grammes de caractères sur un entretien sociologique (423). Un entretien est utilisé car il s'agit d'un texte long donc porteur de nombreuses unités textuelles. La structure de l'image est donc mieux représentée. De plus, pour une meilleure visualisation, l'ordre de niveau de gris est inversé. Ainsi, plus un point sera foncé, plus il relèvera d'une fréquence importante. Dans les figures qui suivent, la taille d'affichage a été modifiée pour que toutes les figures aient la même taille à l'affichage. De plus, il a été, exceptionnellement ajouté un cadre afin de définir les limites de chaque figure.

La Figure 22 paraît pratiquement entièrement blanche. C'est une conséquence normale de l'évolution de la longueur des n-grammes. En effet, avec l'accroissement de la longueur des n-grammes, c'est-à-dire avec l'augmentation de n, le nombre de motifs possibles devient de plus en plus important : 25^n. Ainsi pour n = 4, il y a 390625 motifs possibles. Seuls 5347 motifs différents sont présents dans le texte. Si tous les points non blancs de l'image étaient regroupés dans une zone précise, cette zone ne couvrirait que $1/73^e$ de l'image, c'est à dire juste un peu plus de 1%.

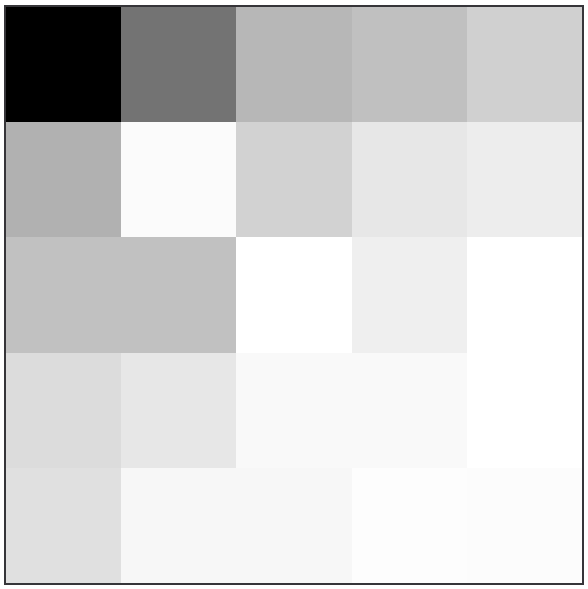

Figure 19 Image de l'entretien 423 à partir de ses 1-grammes (taille réelle 5x5 pixels)

Figure 20 Image de l'entretien 423 à partir de ses 2-grammes (taille réelle 25x25 pixels)

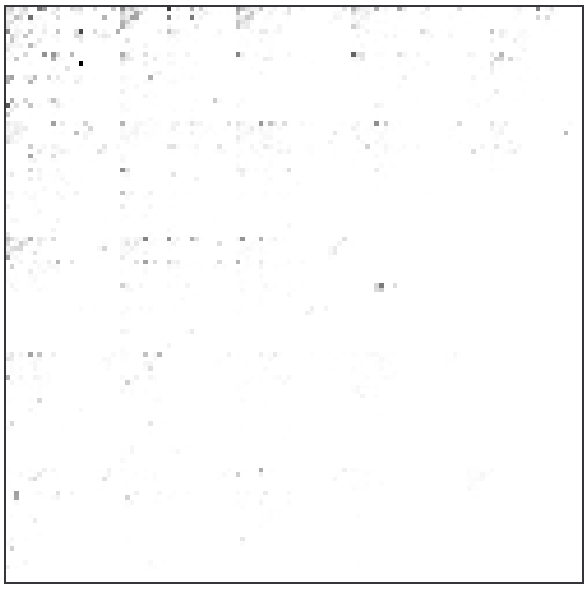

Figure 21 Image de l'entretien 423 à partir de ses 3-grammes (taille réelle 125x125 pixels)

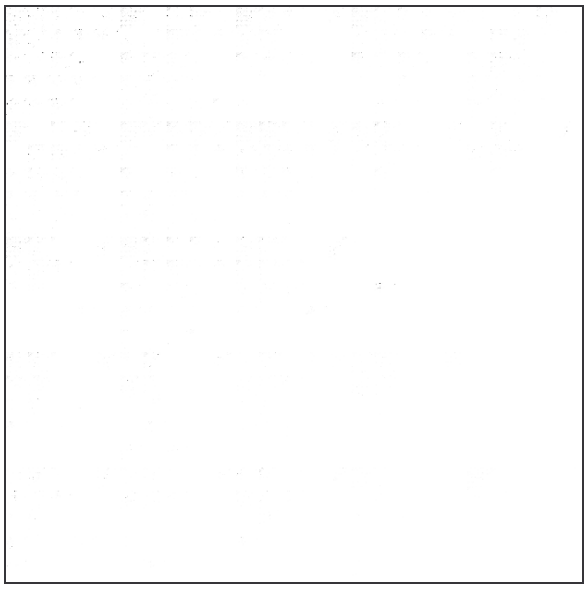

Figure 22 Image de l'entretien 423 à partir de ses 4-grammes (taille réelle 625x625 pixels)

Une autre raison réside dans la distribution des fréquences d'apparition. Il y a en effet une augmentation du rapport entre la fréquence du n-gramme le plus fréquent et celle des autres.
Le rapport entre le nombre de motifs différents et le nombre de motifs possibles ne pouvant être changé (le nombre de motifs présents dans le texte ne peut être augmenté), un recalage des fréquences est réalisé à partir d'une double échelle logarithmique. Le rapport précédemment décrit est donc diminué d'autant. La Figure 23 montre l'image créée après le recalage. Comparée à la Figure 22, la Figure 23 donne une réelle impression visuelle. Elle laisse, de plus, apparaître sa construction fractale.

Cette dernière image représente de manière explicite une grande quantité d'informations. La combinaison du nombre de motifs possibles et du nombre de valeurs pour chaque motif a pour conséquence une presque unicité de chaque représentation. Cette unicité permet d'identifier chaque texte de manière unique. Cependant, la structure de l'image incite les textes partageant une forte similarité à avoir un comportement assez proche.
La partie suivante montre comment essayer de tirer parti de cette similarité de comportement pour créer l'index.

La Figure 24 présente une image qui a été obtenue à partir d'un texte artificiel. Ce texte a été construit en suivant les statistiques d'apparition des caractères décrites précédemment. En comparant cette figure à la Figure 23, il peut être observé que la distribution des caractères dans un texte naturel suit des lois bien plus complexes qu'une simple distribution statistique de caractères. Cette indexation permet donc, réellement, de représenter visuellement la nature des textes.

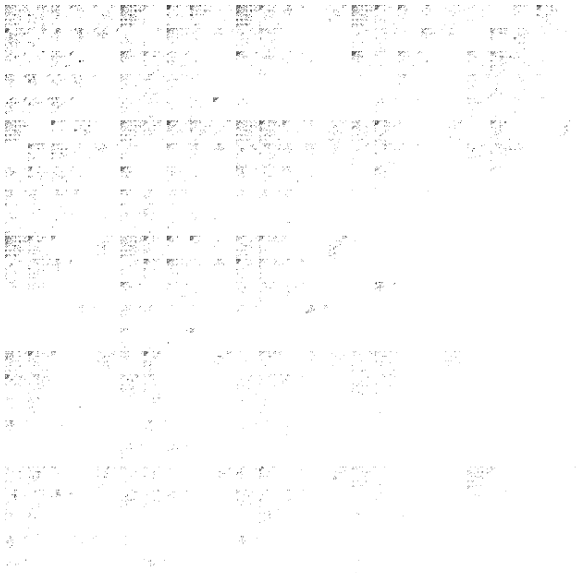

Figure 23 Image de l'entretien 423 à partir de ses 4-grammes et d'un recalage (taille réelle 625x625 pixels)

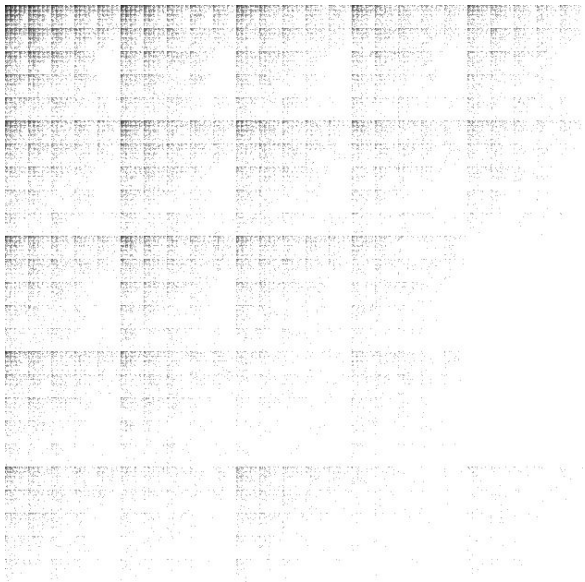

Figure 24 Image d'un texte artificiel suivant les statistiques d'apparition des 1-grammes (taille réelle 625x625 pixels)

2.2.1.4.Indexation

Les images, dans cette méthode, sont construites de manière récursive par répétition du même patron. Les images présentent, comme les figures précédentes le montrent, un caractère fractal. C'est la raison pour laquelle il a été décidé de les évaluer à partir de mesures fractales. Les mesures fractales ont pour but d'apprécier le caractère auto-structurant du motif répété. La structure d'une image, c'est-à-dire de son contenu, est une représentation de la structure du texte. Evaluer une image avec une telle mesure revient à évaluer la structure du texte qui lui est associé. Il a donc été choisi d'appliquer une méthode basée sur la dimension fractale de l'image comme méthode d'indexation.

La dimension fractale d'une image mesure la façon dont l'image occupe l'espace. La dimension sera donc propre à chaque image.
Si un segment de droite, ensemble géométrique auto-similaire, peut être considéré comme l'union de 3 sous-segments identiques entre eux et semblables au segment initial, la dimension fractale est de 1 : longueur L = 1 du segment, nombre N = 3 de sous-segments, log(N)/log(N/L) = log(3)/log(3/1) = log(3)/log(3) = 1.
Pour un carré constitué de 9 sous-carrés identiques, la dimension fractale est de 2 : taille L = 3 du carré, nombre N = 9 de sous-carrés log(N)/log(N/L) = log(9)/log(9/3) = log(9)/log(3) = 2.
…
Dans le cas d'ensembles plus complexes, la notion de dimension d'auto-similarité se généralise par exemple en dimension de masse.

A partir d'un point d'un ensemble X et en notant $m(X_k)$ la mesure de $X_k = X \cap [\ 0\ ;\ k\ ]^2$ l'expression de la dimension de masse devient :

$$D = \lim_{k \to 0} \frac{\ln(m(X_k))}{\ln(k)} \tag{9}$$

Comme il a été expliqué précédemment, la dimension fractale est une mesure d'occupation de l'espace. Pour la méthode d'indexation par Image, chaque image semble être constituée de sous-images et l'unité de base de chaque image est le point, ou pixel. La masse pour un tel ensemble se calcule donc à partir des points.
De plus, l'image est la représentation d'un texte et les points sont les représentations d'unités textuelles. La masse de l'image est donc une mesure d'utilisation d'un type d'unité textuelle et de sa structuration.
Puisque l'image apparaît visiblement de type fractal, la masse peut être calculée sur l'image entière, mais aussi sur la totalité des sous-images. L'ensemble des mesures assure ainsi une représentation de la structure globale et des structures locales propres à l'utilisation de chaque caractère. Ainsi, c'est la structure dans sa totalité qui est représentée.

Cette méthode est donc une nouvelle approche qui a pour but de changer la représentation de données textuelles en données graphiques. Cette représentation, basée sur le contenu, permet de mettre en valeur les structures globale et locales du texte représenté. Ainsi, comme pour la méthode précédente, l'indexation tente de représenter la signature du texte.

Le Tableau 17 et le Tableau 18 présentent les scores obtenus par l'indexation par Image pour les différentes mesures de dissimilarités et les différentes pondérations.

		4-grammes
Id	Bhattacharyya	10,69%
	Euclidienne Standardisée	**14,33%**
	Gower	10,79%
	KullBack-Leibler Symétrie IV	13,17%
	Mahalanobis	14,19%
	Russel-Rao	11,48%
	Soergel	10,94%
Recalage	Bhattacharyya	11,75%
	Euclidienne Standardisée	13,55%
	Gower	11,04%
	KullBack-Leibler Symétrie IV	11,97%
	Mahalanobis	13,43%
	Russel-Rao	7,36%
	Soergel	13,77%
TfIdf	Bhattacharyya	7,69%
	Euclidienne Standardisée	7,69%
	Gower	8,42%
	KullBack-Leibler Symétrie IV	7,69%
	Mahalanobis	7,69%
	Russel-Rao	8,73%
	Soergel	7,69%

Tableau 17 Score F1 sur le corpus Amaryllis avec une indexation par Image

Pour le corpus Amaryllis, les scores sont moins bons que ceux obtenus avec les méthodes d'indexation précédentes. D'une manière relative, la pondération par tf*idf présente les pires des scores. C'est sans pondération que cette méthode s'avère être la plus performante. Deux mesures de distances présentent le même type de scores : la distance euclidienne standardisée et la distance de Russel-Rao.

Pour le corpus issu de NewsGroups, la tendance est presque complètement inversée. La pondération par le tf*idf rend les meilleurs scores. Le corpus issu des NewsGroups offrait, pour les méthodes d'indexation par Vecteurs et par Zipf, des scores plus faibles que ceux obtenus pour Amaryllis. La méthode d'indexation par Image, comme la méthode d'indexation par Structure, offre des scores comparables sur les deux corpus.
Dans ce cas, une comparabilité des scores signifie que les scores, pour le corpus issu des NewsGroups, sont faibles. Ils sont deux fois plus faibles que ceux obtenus par la méthode d'indexation par Vecteurs.

L'indexation par Image ne présente donc pour seuls intérêts son côté théorique et son faible nombre de caractéristiques d'indexation. La liste des désavantages, à ajouter aux résultats assez faibles, réduit à néant l'utilisation d'une telle méthode.

		4-grammes
Id	Bhattacharyya	7,19%
	Euclidienne Standardisée	7,39%
	Gower	6,84%
	KullBack-Leibler Symétrie IV	6,07%
	Mahalanobis	7,09%
	Russel-Rao	6,04%
	Soergel	6,44%
Recalage	Bhattacharyya	5,38%
	Euclidienne Standardisée	8,13%
	Gower	6,50%
	KullBack-Leibler Symétrie IV	6,26%
	Mahalanobis	4,51%
	Russel-Rao	4,30%
	Soergel	7,17%
TfIdf	Bhattacharyya	**14,17%**
	Euclidienne Standardisée	**14,17%**
	Gower	**14,17%**
	KullBack-Leibler Symétrie IV	**14,17%**
	Mahalanobis	**14,17%**
	Russel-Rao	8,45%
	Soergel	**14,17%**

Tableau 18 Score F1 sur le corpus NewsGroups avec une indexation par Image

En effet, aux mauvais résultats, il faut ajouter le fait que cette méthode est plus performante lorsqu'elle utilise les 4-grammes. Les 4-grammes sont le type de motifs ayant le pire produit

nombre de motifs différents * longueur maximum. De plus, cette méthode d'indexation impose de suivre un patron unique pour chaque langue traitée. Cette contrainte enlève à cette méthode toute ouverture d'utilisation. Enfin, expérimentalement, l'indexation par Image est instable dans ses performances à configuration fixée. Cela signifie qu'aucune chaîne de traitements ne permet d'assurer un résultat général. Enfin, cette méthode apparaît, aux yeux de certains, comme une mauvaise dérivée de l'indexations par Vecteurs.

A ces deux dernières critiques, il faut répondre que cette méthode est basée sur le caractère auto-similaire de l'image de représentation. C'est l'auto-similarité qui renvoie la structure des textes étudiés. Cela signifie qu'une quantité minimum d'information est nécessaire pour que la structure de l'image apparaisse. Dans un cas de faible information comme ça l'est pour les textes des corpus Amaryllis et NewsGroups (la plupart des textes ont moins d'une page), l'image n'est constituée que d'une faible quantité de points. Cette faible quantité de points n'est pas suffisante pour afficher une structure qui puisse être évaluée. En d'autres termes, cela confirme que cette méthode d'indexation n'est utilisable que sur des textes suffisamment longs. Le « suffisant » n'a pas été étudié puisque c'est la longueur du texte qui permet de se détacher des méthodes linguistiques.

2.2.2. Automate 1D

Avant de détailler la méthode qui est basée sur un automate, quelques rappels sont présentés sur les automates finis déterministes. Ces rappels permettent de présenter les bases théoriques des automates utilisés dans cette méthode et dans la méthode de la partie 2.2.3 concernant l'Automate Peintre. Après ces quelques rappels, l'automate et la méthode d'indexation sont détaillés.

2.2.2.1.Rappels sur les automates

De manière intuitive, on peut voir un système de reconnaissance comme une machine permettant de lire un mot à travers différentes manipulations [NOU 92]. Cette machine, appelée automate, permet donc, par extension, de reconnaître un langage.
Il existe plusieurs types d'automates, cependant ils ont tous une structure commune. Ainsi, un automate est composé de trois parties :

- Une bande en entrée, finie ou infinie, sur laquelle va s'inscrire le mot à lire. Une bande, en entrée, est divisée en cellules ; le mot à lire étant formé d'une suite de symboles (de l'alphabet), un symbole (et un seul) est logé dans une cellule.
- Un organe de commandes qui permet de gérer un ensemble fini d'états. La gestion des états se fait à travers une fonction spécifique, dite fonction de transition.
- Eventuellement, une mémoire auxiliaire de stockage.

[AHO 86], [NOU 92] définissent un automate fini non-déterministe comme un modèle mathématique défini par le cinq-uplet A(X,E,e0,t,F), avec :

- X : un ensemble de symboles d'entrée (l'alphabet des symboles d'entrée)
- E : l'ensemble des états
- e0 : un état qui est distingué comme l'état de départ ou état initial
- t : une fonction de transition, qui fait correspondre des couples état-symbole à des ensembles d'états
- F : un ensemble d'états distingués comme états d'acceptation ou états finals.

Un automate fini non déterministe peut être représenté graphiquement comme un graphe orienté étiqueté, appelé graphe de transition, dans lequel les nœuds sont les états et les arcs étiquetés représentent la fonction de transition. Ce graphe ressemble à un diagramme de transition, mais le même caractère peut étiqueter deux transitions ou plus en sortie d'un même nœud et les arcs peuvent être étiquetés par le symbole spécial ε (neutre) au même titre que les symboles d'entrée.

Un automate fini déterministe, [AHO 86], [NOU 92], est un cas particulier d'automate fini non déterministe dans lequel :

- aucun état n'a de ε-transition, c'est-à-dire de transition sur l'entrée ε
- pour chaque état e et chaque symbole d'entrée a , il y a au plus un arc étiqueté a qui quitte e.

Un automate fini déterministe a au plus une transition à partir de chaque état sur n'importe quel symbole. Si on utilise une table de transition pour représenter la fonction de transition de l'automate fini déterministe, alors une entrée dans la table de transition est un état unique.

2.2.2.2.L'automate et l'indexation

L'indexation par automate 1D est une nouvelle méthode développée pour ce travail. Cette méthode d'indexation tente de représenter la structure du texte par l'enchaînement des informations qui le composent. C'est donc la dynamique de la structure qui est mise en évidence. C'est-à-dire que la structure n'est pas représentée de manière globale, mais à partir de ses comportements locaux, à partir de son évolution. Les textes sont une représentations mono-dimensionnelle de l'information. Cet automate tente de conserver cet aspect mono-dimensionnel.
L'appellation « automate 1D » provient donc du fait que l'automate essaie de construire un signal d'amplitude bornée à partir du texte.

Le signal évolue dans une amplitude discrète. C'est-à-dire que chaque valeur de l'amplitude est représentée par un état. De plus, chaque état n'est relié qu'aux états de va-leurs directement supérieure et inférieure. C'est-à-dire que l'automate ne propose que deux types de transitions, la transition « Aller vers le bas » qui permet d'aller dans l'état de valeur directement inférieure et la transition « Aller vers le haut » qui permet d'aller dans l'état de valeur directement supérieur. Les états aux extremums n'ont, naturellement, qu'un seul voisin.

L'ensemble des symboles d'entrée correspond à l'ensemble des unités textuelles pouvant être traitées. Afin de laisser un maximum de liberté à l'automate, il est placé, initialement sur l'état qui se situe au centre de l'amplitude. C'est-à-dire l'élément e0 correspond à l'état central.. Il a été choisi, afin de ne pas devoir effectuer d'étude préalable des textes du corpus, de ne pas affecter les transitions avant le traitement. C'est-à-dire que les fonctions de transition sont définies au fur et à mesure du traitement. L'affectation suit une liste de règles énoncées dans la partie qui suit.

L'amplitude du signal est bornée a priori, ils font donc au maximum évité les effets de bords. C'est-à-dire qu'il faut éviter au maximum de se rapprocher des limites de l'amplitude et en aucun cas les franchir.
Par contre, la signal n'a d'intérêt que s'il oscille. Il faut donc éviter son immobilité.
Les règles ont été constituées afin de répondre à ces contraintes. Elles sont, néanmoins, aidées par la distribution naturelle des unités textuelles dans les textes traités.

D'un point de vue localisation, les états sont divisés en quatre classes d'effectifs identiques. Les états périphériques ont été définis comme correspondant au regroupement des deux classes d'états aux limites de l'amplitude, c'est-à-dire la classe la plus en bas et la classe la plus en haut. Les états centraux correspondent aux deux classes centrales.

Voici la liste des différentes règles :

- La première règle concerne le premier symbole d'entrée, c'est-à-dire la première unité textuelle du texte. L'action « Aller vers le haut » lui est assigné. En dirigeant le signal vers les classes périphériques, cela permet d'éviter l'immobilité initiale du signal.
- Lorsque l'automate est dans un état périphérique, la seconde règle impose de suivre la direction du centre de l'amplitude aux symboles d'entrée auxquels aucune transition n'a été affectés. Ainsi, si il arrive un symbole d'entrée auquel une action n'a été affectée et que l'automate se situe dans le quart bas des états, alors, il est assigné au symbole la transition « Aller vers le haut ». Cette règle essaie, ainsi, de donner au signal créé une allure sinusoïdale.
- La troisième règle indique le cas général : si aucune transition n'est affectée à un symbole d'entrée et que ce symbole n'est concerné par aucune des premières règles, alors il lui est affecté la même transition que celle qui est affectée au symbole qui le précède. Cette troisième règle assure, ainsi, un minimum de continuité et de cohésion dans les déplacements effectués et donc dans la formation du signal.
- Enfin la quatrième et dernière règle, dite règle de changement d'urgence, inverse l'action d'un symbole d'entrée lorsque l'action qui lui est assignée implique la sortie de l'amplitude.

La Figure 25 montre l'action de l'automate sur deux textes différents comme s'il était enregistré par un outil du type sismographe. Les 4-grammes sont utilisés comme unités textuelles. Le premier texte un entretien sociologique. Le second texte est un texte artificiel conçu statistiquement à partir des fréquences d'apparitions des caractères en langue française. Dans la figure, le centre et les quarts sont marqués par des barres horizontales. L'amplitude a été, ici, fixée à 301.

Le texte artificiel est nettement marqué par son aspect statistique et reste enfermé dans la partie centrale. Le texte en langue naturelle se différencie par une oscillation générale beaucoup plus lente. Il se caractérise surtout par des unités textuelles formatée pour descendre et cela jusqu'aux limites ce l'automate, qu'il atteindra 13 fois au total.

La Figure 26 présente le résultat obtenu si les unités textuelles utilisées sont les 1-grammes. L'évolution est complètement différente. Dans ce cas, la règle de changement d'urgence est régulièrement sollicitée. En effet, les caractères les plus fréquents apparaissent au début du

texte, il leur est donc assigné, à tous, l'action « Aller vers le haut ». Le signal est donc guidé rapidement jusqu'à la borne supérieure de l'amplitude. La règle 4 change donc les actions assignées aux caractères les plus fréquents en « Aller vers le bas ». Le signal est donc guidé rapidement jusqu'à la borne inférieure de l'amplitude. Etc.

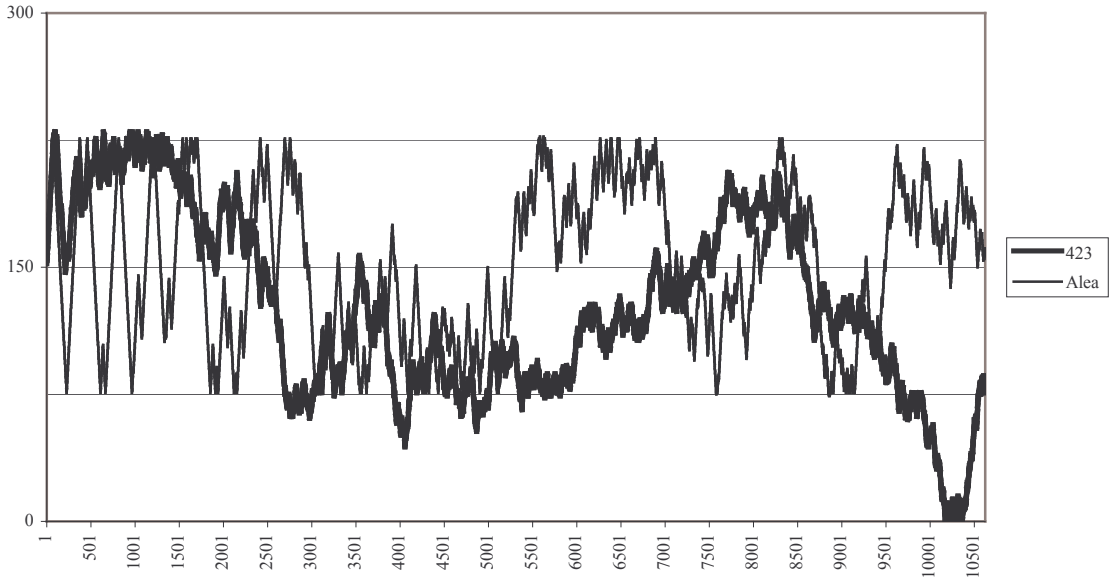

Figure 25 Evolution des automates relatifs à l'entretien 423 et à un texte artificiel avec des 4-grammes

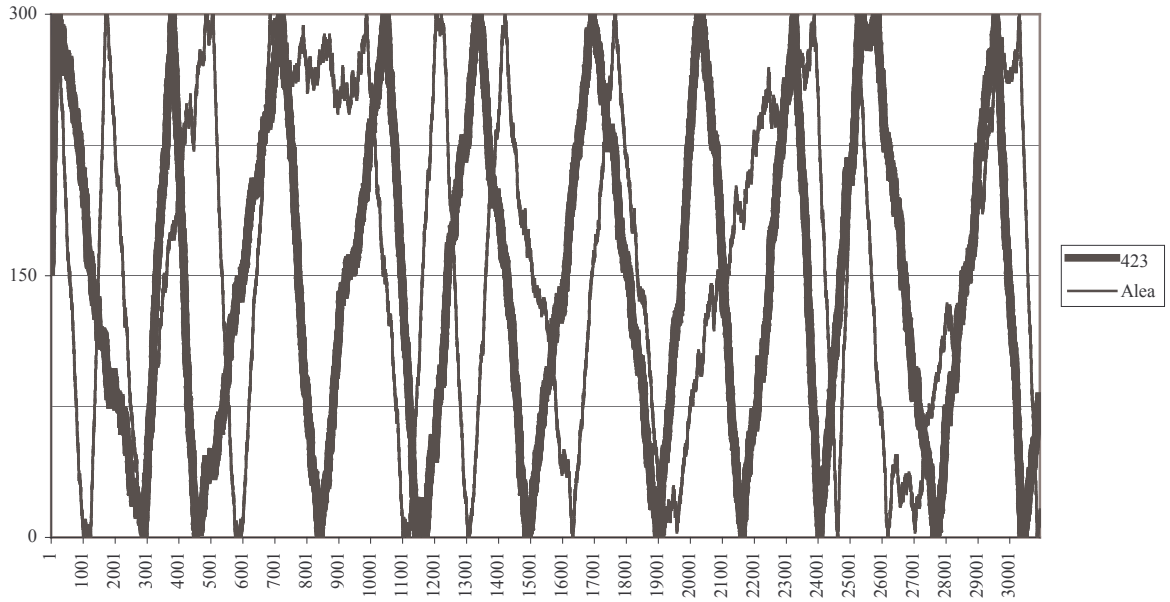

Figure 26 Evolution des automates relatifs à l'entretien 423 et à un texte artificiel avec des 4-grammes

L'automate représente donc réellement l'évolution du texte. La méthode d'indexation consiste à compter, pour chaque case, le nombre de passages effectués par l'automate. L'index représente donc l'évolution globale du texte auquel il est associé. Les textes naturels auront

une tendance à flirter avec les limites alors que les textes statistiques, c'est-à-dire suivant une structure répétitive, se remarqueront par une simple utilisation de la bande centrale.

Si cette méthode a l'avantage de fournir une représentation de taille fixée, elle a pour inconvénient la façon dont elle assigne les actions aux symboles d'entrée. L'insertion d'une unité textuelle peut, en effet, changer complètement l'allure de la courbe selon l'endroit où elle est insérée.

Le Tableau 19et le Tableau 20 présentent les résultats obtenus sur les corpus Amaryllis et NewsGroups.

Pour le corpus Amaryllis, une chaîne de traitements se distingue des autres. Avec plus de 1,20% de différence avec le second score globale la distance de Bhattacharyya à la suite d'un recalage des fréquences offre le meilleur score globale (moyenne de 13,80%) et le meilleur score locale (19,95%). Il y a donc une nette amélioration par rapport à l'indexation par Image. Le score reste, néanmoins, inférieur à ceux des méthodes d'indexation par Vecteurs ou par Structure.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	12,94%	**14,72%**	11,13%	12,36%	11,80%	10,70%
	Euclidienne Standardisée	7,55%	7,69%	12,00%	13,56%	10,23%	10,38%
	Gower	11,61%	11,93%	11,66%	7,87%	8,38%	10,42%
	KullBack-Leibler Symétrie IV	10,86%	9,48%	11,18%	14,11%	12,00%	11,76%
	Mahalanobis	7,83%	12,80%	12,46%	9,49%	8,58%	10,61%
	Russel-Rao	8,73%	8,73%	7,96%	7,60%	8,73%	8,73%
	Soergel	11,24%	12,78%	11,53%	10,95%	8,58%	10,65%
Recalage	Bhattacharyya	10,27%	**14,72%**	11,55%	**19,95%**	**15,63%**	10,70%
	Euclidienne Standardisée	7,55%	7,69%	12,00%	13,56%	11,55%	10,38%
	Gower	11,61%	11,93%	11,66%	7,87%	8,38%	9,47%
	KullBack-Leibler Symétrie IV	10,66%	9,48%	11,10%	10,38%	11,93%	11,76%
	Mahalanobis	7,83%	12,80%	12,46%	13,66%	8,58%	10,61%
	Russel-Rao	8,73%	8,73%	7,96%	7,60%	8,73%	8,73%
	Soergel	11,24%	12,78%	11,14%	11,68%	8,58%	10,44%
TfIdf	Bhattacharyya	8,65%	9,09%	**17,31%**	17,31%	13,65%	8,62%
	Euclidienne Standardisée	8,54%	6,39%	**17,31%**	17,31%	8,58%	**16,26%**
	Gower	8,73%	7,55%	**17,31%**	17,31%	8,58%	8,73%
	KullBack-Leibler Symétrie IV	**13,18%**	11,48%	**17,31%**	17,31%	8,58%	7,96%
	Mahalanobis	8,69%	7,21%	**17,31%**	17,31%	8,69%	8,65%
	Russel-Rao	8,73%	8,73%	8,69%	8,69%	8,73%	8,73%
	Soergel	8,73%	6,51%	**17,31%**	17,31%	8,58%	12,32%

Tableau 19 Score F1 sur le corpus Amaryllis avec une indexation par l'Automate 1D

Au tableau des meilleurs scores globaux, le recalage et la distance de Bhattacharyya sont suivis par trois distances nécessitant la pondération par tf*idf. Ces trois distances offrent des scores globaux proches les uns des autres. L'amplitude entre ces trois distances, après une

normalisation par le tf*idf, est de 0,24%. Ces trois distances sont, dans l'ordre décroissant des scores globaux, la symétrie IV de Kullback-Leibler, la distance de Bhattacharyya et la distance euclidienne standardisée. D'une manière générale, les scores moyens obtenus par pondération par tf*idf sont meilleurs que les scores moyens par recalage. Le recalage est donc ponctuellement la meilleure pondération.

Du point de vue des types de motifs, les 1-grammes et les 2-grammes offrent les meilleures performances moyennes (respectivement 12,78% et 13,01%). Cette inégalité est plus fortement marquée par le couple recalage – Bhattacharyya.

Pour le corpus issu des NewsGroups, la tendance au niveau des motifs est inversée. C'est, en effet, les 1-grammes avec une moyenne de 9,48% qui offrent le meilleur résultat, suivis par les 3-grammes avec 8,69% et seulement les 2-grammes avec 8,15%.
Avec ce corpus, il faut aussi remarquer que la majorité des bons résultats sont obtenus avec une normalisation par tf*idf. Cette normalisation obtient le meilleur résultat moyen avec 9,05%, la normalisation par recalage est dernière avec 6,31%.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	6,09%	6,21%	12,22%	4,77%	8,89%	7,52%
	Euclidienne Standardisée	9,48%	6,65%	6,61%	4,89%	4,35%	7,89%
	Gower	5,34%	3,13%	6,53%	7,74%	7,94%	7,01%
	KullBack-Leibler Symétrie IV	9,33%	4,94%	5,65%	3,94%	3,63%	5,23%
	Mahalanobis	**10,75%**	**8,81%**	**14,49%**	6,75%	7,41%	4,89%
	Russel-Rao	5,03%	4,75%	6,14%	5,80%	4,75%	4,75%
	Soergel	7,95%	5,30%	7,59%	6,91%	7,15%	8,21%
Recalage	Bhattacharyya	6,09%	6,54%	12,22%	4,77%	8,89%	8,29%
	Euclidienne Standardisée	9,48%	6,65%	6,08%	4,89%	4,35%	**8,91%**
	Gower	5,23%	3,13%	7,50%	6,85%	7,94%	7,01%
	KullBack-Leibler Symétrie IV	8,19%	4,94%	4,74%	3,67%	5,68%	3,92%
	Mahalanobis	6,07%	**8,81%**	5,10%	6,52%	7,41%	8,06%
	Russel-Rao	5,03%	4,75%	4,20%	5,80%	4,75%	4,75%
	Soergel	7,95%	5,30%	6,45%	4,46%	5,80%	7,73%
TfIdf	Bhattacharyya	3,73%	4,58%	14,17%	**14,17%**	**14,17%**	4,54%
	Euclidienne Standardisée	4,07%	4,67%	14,17%	**14,17%**	**14,17%**	4,54%
	Gower	4,74%	5,58%	14,17%	**14,17%**	**14,17%**	4,62%
	KullBack-Leibler Symétrie IV	4,74%	5,93%	14,17%	**14,17%**	**14,17%**	6,58%
	Mahalanobis	4,74%	4,74%	14,17%	**14,17%**	**14,17%**	4,25%
	Russel-Rao	4,20%	4,75%	8,45%	8,45%	8,45%	4,75%
	Soergel	4,74%	4,74%	14,17%	**14,17%**	**14,17%**	4,25%

Tableau 20 Score F1 sur le corpus NewsGroups avec une indexation par l'Automate 1D

Ainsi la chaîne de traitements Recalage – Bhattacharyya, qui offrait le meilleur résultat observé avec le corpus Amaryllis, semble n'être qu'une performance ponctuelle. Pour les mêmes conditions, c'est-à-dire avec les 2-grammes, ce couple n'offre qu'un score de 4,77%, soit l'un des plus faibles de l'expérimentation.

La normalisation par tf*idf semble donc plus robuste aux changements de corpus. Les 1-grammes et la distance de Gower ont été conservés pour leurs bonnes performances générales.

Il a donc été proposé dans cette partie, une nouvelle méthode qui tente de représenter un texte par sa structure, ou plutôt par son évolution. C'est en effet la séquence particulière des mots, leur ordre, … qui permet de représenter le texte. Les expérimentations montrent que cette méthode offre de meilleurs résultats qu'une représentation globale par image.
Cependant, le manque de robustesse de la méthode l'empêche de fournir des résultats meilleurs. En effet, le fait d'ajouter un simple mot, unique ou non, dans le texte ou d'en inverser deux peut former, selon l'endroit de la modification, des changements très importants au niveau de la représentation. De plus, la représentation conservée, c'est-à-dire l'histogramme de comptage des passages, n'est vraisemblablement pas la meilleure représentation qui puisse être retirée d'une telle méthode. Cette représentation oublie toute notion de passage. Deux représentations pourraient être identiques pour des courbes complètement différentes, ne partageant même pas une certaine symétrie.
Quoiqu'il en soit, cette méthode est, comme la méthode de Zipf, détachée du contenu, c'est-à-dire que les textes ne sont pas comparés par rapport à leur contenu. Cette méthode est composée d'un nombre fixé auparavant de caractéristiques et ce nombre est bien inférieur à ceux de la méthode vectorielle et de la méthode de Zipf. Et cette méthode assure de meilleurs résultats que la méthode de Zipf.
Cela indique donc que l'évolution, si elle est correctement capturée pourrait servir à représenter les textes.

2.2.3. Automate Peintre

Comme la méthode d'indexation précédente, cette méthode est basée sur le fonctionnement d'un automate. Et comme la méthode d'indexation par Image, elle a pour but de transformer un texte en image. La méthode d'indexation par Automate Peintre tente donc de s'approprier les qualités des différentes méthodes présentées précédemment : représentation du contenu en l'utilisant comme signature, représentation de l'évolution en lui donnant une part active dans la création, représentation de la structure en utilisant des règles de construction.
Contrairement à l'automate linéaire où les actions étaient fixées selon le type de symbole d'entrée et selon le type d'états, cette méthode base l'affectation de ses transitions sur les statistiques d'apparition de chaque symbole d'entrée.
Cette méthode est décrite en deux parties : la partie 2.2.3.1 décrit, non sans quelques définitions, l'automate de création et la partie 2.2.3.2 décrit, plus particulièrement, la méthode d'indexation.

2.2.3.1.L'automate

L'automate crée une image à partir d'un texte. Afin de mieux présenter les états possibles que peut prendre l'automate, il faut définir la notion d'image. Une fois ces définitions émises, les caractéristiques peuvent être étudiées et les états en découler. A partir des états, la table de transition peut être créée.

Soit $I_{m,n}$ l'ensemble des images en niveaux de gris de taille m*n. L'ensemble $I_{m,n}$ a une cardinalité de 256^{m*n} et peut être défini comme :

$$I_{m,n}=\left\{I^{k}_{m,n}\right\}_{k\in[1,256^{m\times n}]} \tag{10}$$

De même, chaque image peut être considérée comme un ensemble connexe de points dans un espace à deux dimensions :

$$I^{k}_{m,n}=\left\{I^{k}_{m,n}(i,j)\right\}i\in[1,m]\times j\in[1,n] \tag{11}$$

Enfin, chaque point peut prendre l'une des valeurs de niveaux de gris possibles :

$$I^{k}_{m,n}(i,j)\in[0,255] \tag{12}$$

Pour créer un signal, il suffit de modifier les valeurs d'un signal silence (valeurs initialisées à 0) sur toute la durée du signal. Une image est un ensemble de points affectés d'une valeur de niveau de gris. Sur le même principe que le son, créer une image consiste à partir d'une image dite « silence », à modifier positivement ou négativement les valeurs de tous les points de cette image.
La formule (6) définit l'ensemble des valeurs possibles des points d'une image en niveaux de gris. Compte tenu de cette remarque et de la volonté d'offrir un degré de liberté le plus large et le plus symétrique possible à la modification, l'image dite « silence » correspond à l'image dont tous les points sont initialisés au niveau de gris moyen.
La caractérisation des modifications est donc assez simple, leur ordre l'est beaucoup moins.

Pour un signal, le temps fournit cet ordre de modification. Or, aucun chemin réellement naturel ne permet de parcourir une image. Il existe, certes, des chemins de type fractal, par exemple la courbe de Peano [NIK 02] et [LIN 00]. Mais ils ne peuvent pas être considérés comme des parcours naturels de l'image.
L'objectif étant de représenter un texte et son évolution, il paraît normal d'opter pour un parcours lié au contenu du texte. Un point dit de référence est donc défini comme le point à partir duquel se fait la modification de l'image. De façon à conserver un maximum de degrés de liberté aux déplacements de ce point de référence, la position « silence » correspond au centre de l'image.

L'image finale est donc construite par modifications successives. Toute modification effectuée constitue la modification de la valeur du point de référence ou la modification de la position du point de référence. Ainsi, chaque état de l'automate est un 2-uplet contenant une image de dimension m*n en niveaux de gris et un vecteur de position (i,j) avec $1\leq i\leq m$ et $1\leq j\leq n$ donnant la position du point de référence. L'automate possède donc $m*n*256^{m*n}$ états possibles. Si m=n=1, cela représente 256 états ; si m=n=2, cela en représente $1,7*10^{10}$ …

Cet ensemble est fini, l'automate l'est donc aussi. L'état initial est défini par le 2-uplet (image « silence », position « silence »), c'est-à-dire le 2-uplet constitué d'une image dont tous les points ont un niveau de gris moyen et d'une position de référence au centre de l'image. L'ensemble des états finals possibles est l'ensemble des états.

Les transitions sont des ponts d'un état à un autre et ne dépendent que du symbole en entrée. Du fait de la cardinalité importante de l'ensemble des unités textuelles, il est difficile de trouver un jeu absolu de transitions pour chaque symbole. L'alphabet a donc été, initialement, limité à 25 symboles. Pour cela, un alphabet relatif est construit pour chaque texte.
Pour chaque texte, les unités textuelles sont rangées par ordre de fréquence décroissante. Et 25 classes d'effectifs égaux sont créées. Cela correspond donc à un découpage en 25 classes de même effectif de l'index de Zipf. Pour plus de robustesse, à fréquence égale, les unités textuelles sont classées par ordre alphabétique. La première classe compte donc les unités textuelles les plus fréquentes et la 25^{e} classe compte les unités textuelles les moins fréquentes. Ces 25 classes forment les 25 symboles de l'alphabet d'entrée de l'automate.

Le choix des transitions repose sur les analyses du langage qui ont pu être faites par Zipf [ZIP 35] et qui sont confirmées par ses expérimentations. Cela est traduit par des actions qui sont, avec la décroissance des fréquences, de plus en plus catégoriques. Inversement, l'action la plus souvent réalisée consiste à ne rien faire (loi du moindre effort). La construction de la table s'est faite par blocs, on en compte six au total. Le premier bloc est constitué des 7 premières actions, ce sont les actions primaires : ne rien faire, changements simples de valeur, déplacements horizontaux et verticaux. Le deuxième bloc est constitué, en plus, des 4 actions suivantes. L'automate est, ainsi, autorisé aux déplacements verticaux. Le troisième bloc est constitué des 17 premières actions. L'automate est autorisé à effectuer des changements de valeurs plus importants. Le quatrième bloc va jusqu'à la 21^{e} action et permet des déplacements horizontaux et verticaux plus importants. Le cinquième bloc ajoute deux changements radicaux de valeur. Enfin le dernier bloc contient toutes les actions et permet des changements radicaux de positions.

Le nombre d'états, même s'il est fini, est très important. Cela signifie qu'aucun graphique ne pourrait représenter la totalité de l'automate et qu'aucune table des transitions ne peut être présentée dans son ensemble. Avec les 25 actions choisies, une telle table serait composée de $25*m*n*256^{m*n}$ lignes. En prenant m=n=1, c'est-à-dire les images constituées d'un seul point, la table aurait 6400 lignes.
La table des transitions est donc adaptée au problème.

Les transitions agissent tant au niveau de la position du point de modification que du niveau de gris du point de l'image sur lequel est le point de modification. La table de transitions présente donc les changements effectués sur ces paramètres. La position de référence est représentée par ses coordonnées X et Y. X représente la position horizontale, X=0 correspond à la gauche de l'image et X=m correspond à la droite de l'image. De même, Y représente la position verticale, Y=0 correspond au bas de l'image et Y=n correspond au haut. Le niveau de gris est, quant à lui, représenté par NdG. NdG=0 correspond à la valeur de niveau de gris la plus basse, le noir, et NdG=255 correspond à la valeur de niveau de gris la plus haute, le blanc.

Le Tableau 21 présente la table de transition.

Symbole d'entrée	Action réalisée
1	Rien
2	NdG=NdG +1
3	NdG=NdG –1
4	Y=Y+1
5	Y=Y-1
6	X=X+1
7	X=X-1
8	Y=Y+1 X=X+1
9	Y=Y+1 X=X-1
10	Y=Y-1 X=X+1
11	Y=Y-1 X=X-1
12	NdG=NdD+5
13	NdG=NdG-5
14	NdG=NdG+7
15	NdG=NdG-7
16	NdG=NdG+10
17	NdG=NdG-10
18	Y=Y+2
19	Y=Y-2
20	X=X+2
21	X=X-2
22	NdG=255
23	NdG=0
24	Y=n/2 X=m/2
25	Y=n/2-Y X=m/2-X

Tableau 21 Table des transitions réalisées selon le symbole d'entrée

La Figure 27 présente l'évolution du score pour les corpus Amaryllis (Ama) et NewsGroups (NG) suivant le nombre d'actions conservées. Les valeurs maximum, moyennes et minimum des scores obtenus pour chaque corpus sont représentées par une courbe.

Cette évolution permet de déterminer la taille des blocs à conserver. Comme pour les méthodes précédentes, le score obtenu pour les NewsGroups est toujours inférieur à celui obtenu pour le corpus Amaryllis. Les évolutions sont, néanmoins, similaires.

Les évolutions des scores minimum et moyens ne varient que très peu. Par contre, les scores maximum présentent des variations suivant le nombre d'actions conservées. Il peut donc être conclu qu'un nombre trop faible d'actions, comme la limitation aux actions primaires, ou un

nombre trop élevé, comme l'autorisation des divers actions radicales, constitue une liste de règles respectivement trop restrictives et trop permissives pour l'automate. De plus, l'automate se contente d'actions simples puisque le meilleur score est atteint par le bloc de 11 actions.
Cela signifie donc que l'automate offre les meilleurs résultats lorsqu'il est autorisé à effectuer des changements simples de valeur, des déplacements simples dans les huit directions.

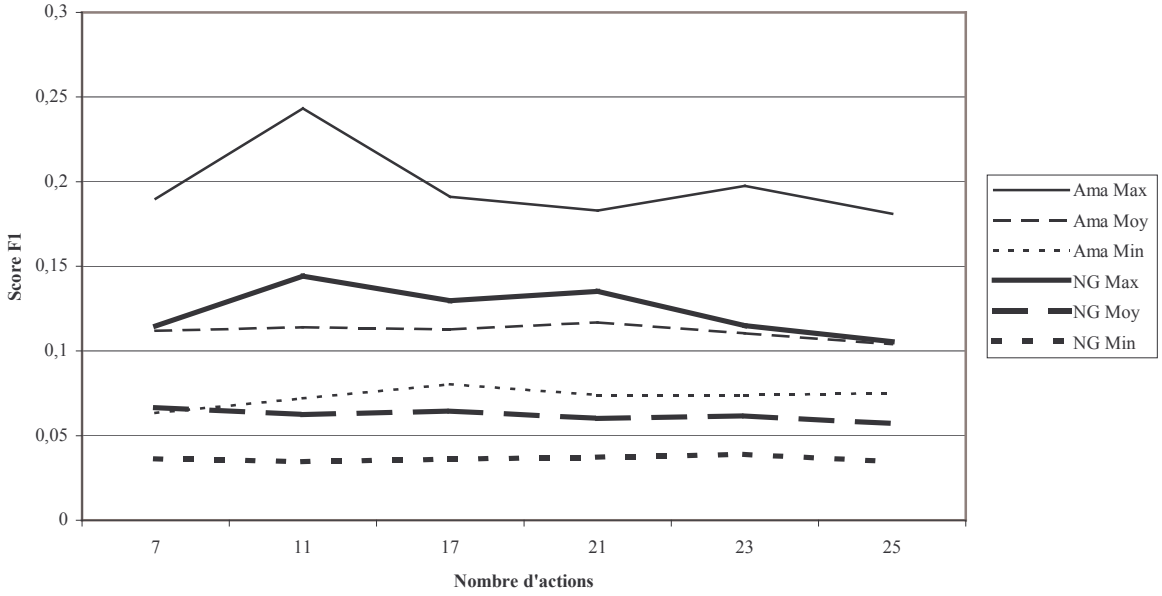

Figure 27 Evolution du Score suivant le nombre d'actions conservées

La Figure 28 présente, pour exemple, le déroulement dans la création d'une image à partir d'un texte artificiel avec les quatre premiers états obtenus et l'état final. La position du point de référence est indiquée par un cadre rouge.
Partant de l'état 0, la première transition effectue un assombrissement. Les deuxième et troisième transitions effectuent des déplacements, tout d'abord vers la gauche, puis vers le bas. La quatrième effectue un éclaircissement. L'image finale est composée de multiples points, dont les niveaux de gris ne sont, parfois, que très légèrement modifiés. Cependant, il peut être noté que certains points peuvent changer plusieurs fois de valeur. La valeur du point à la position initiale du point de référence est différente dans l'image finale de celle de l'image de l'état 1.

De plus, il faut signaler que cette illustration ne présente qu'une partie de l'image réelle. Pour éviter tout effet de bord, aucune taille n'est fixée, a priori, pour les images. La méthode d'indexation, présentée dans la partie qui suit, dépend de la taille de l'image. Cette représentation n'est donc pas, dans l'absolu, de taille fixe. Cependant, les expérimentations montrent qu'il existe des limites naturelles.

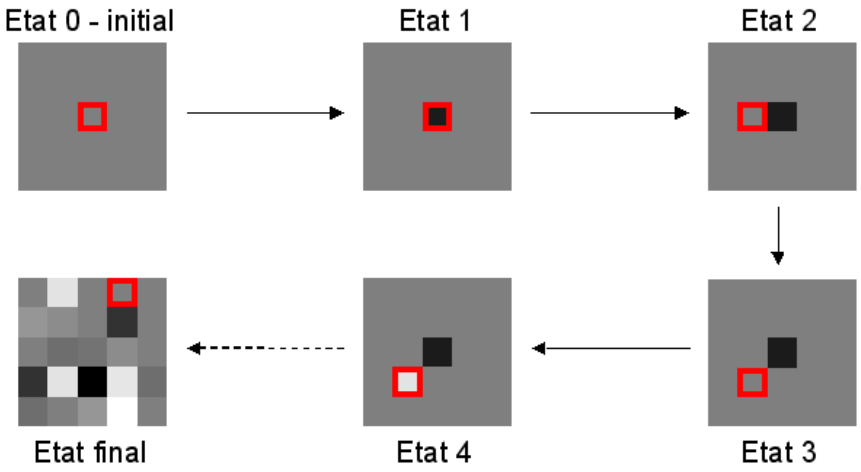

Figure 28 Exemple de création d'image

2.2.3.2.L'indexation

Les images étant en niveau de gris basées sur un niveau de gris moyen, l'effet visuel est difficilement appréciable comme le montre la Figure 29. On arrive juste à distinguer les points les plus clairs et les points les plus sombres.

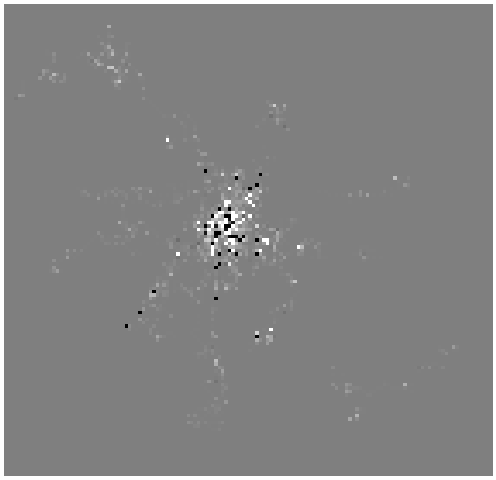

Figure 29 Image obtenue par l'automate (25 actions) pour l'entretien 423 utilisations des 4-grammes

Ainsi, généralement, on utilise la trace de l'automate, c'est-à-dire le chemin emprunté (le sillon du crayon), qui donne une meilleure impression visuelle. La Figure 30 et la Figure 31 présentent deux types de traces laissées par l'automate sur, respectivement, un entretien sociologique et un texte artificiel respectant les fréquences d'apparition des 1-grammes avec les 25 actions.

Figure 30 Trace de l'automate (25 actions) pour l'entretien 423 à partir de ses 4-grammes (taille réelle 149x125 pixels)

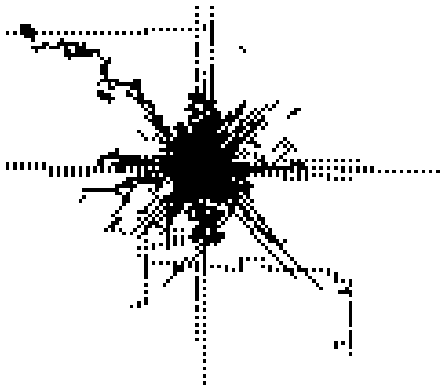

Figure 31 Trace de l'automate (25 actions) pour un texte artificiel suivant les statistiques d'apparition des 1-grammes (taille réelle 125x105)

La première impression laissée, et d'autant plus lorsque l'on voit le processus de création, est la ressemblance avec une agglomération. C'est la raison pour laquelle, dans la suite des commentaires et explications, le vocabulaire tiré de l'urbanisme sera utilisé à titre d'image.

Dans chacune des agglomérations, il peut être distingué le cœur et les différentes boucles périphériques constituées de constructions ou de fondations. De plus, il peut être distingué un système de voirie et de canalisation. Les différences entre constructions et fondations et entre voiries et canalisations correspondent à des différences de profondeur, c'est-à-dire que les unes ont une valeur supérieure au niveau normal et les autres une valeur inférieure.
Ainsi la Figure 30 montre l'équivalent d'un petit bourg à la française avec ses coins et ses recoins. Il peut être observé des quartiers qui se forment à la périphérie, « une extension à échelle humaine ». La Figure 31 montre, quant à elle, des allures de cité industriello-commerciale récente et en voie de développement rapide. Tout est fortement marqué autour des grands axes. Il faut de plus signaler que, contrairement à l'autre image, cette figure ne présente que le cœur de la cité, les 4 axes partant en suivant les 4 axes cardinaux sortent, en réalité, du champ comme pour rejoindre des cités voisines. En arrêtant la carte aux limites de ces voies, la carte de la Figure 31 serait plus de cent fois plus grande que la carte Figure 30 qui regroupe la totalité des informations.

A nouveau le texte semble présenter sa structure. Juger cette structure revient à juger l'évolution des diverses couronnes qui la compose. Le principe est donc identique à celui de la méthode précédente à la différence près que, dans le cas présent, le point chaud est un point central. Une série de dimensions de masse va donc aussi être calculée. Cependant, dans le cas des agglomérations, une attention particulière doit être portée aux périphéries, car c'est de là que proviendront les divergences. Sans détail, il est difficile de distinguer un hyper-centre d'un autre. Le problème des mesures additives pour les espaces à n dimensions tient dans la réduction d'importance des données des dernières dimensions. Dans notre cas, cela revient à dire que si un quartier est construit assez loin de la ville, il ne participera qu'à moindre effet dans l'indexation à cause de la surface de la couronne à laquelle il appartient. En notant i le numéro de couronne en partant du centre, la surface d'une couronne i peut être calculée par :

$$SC_i = 8i \tag{13}$$

et la surface englobée par une couronne i peut être rappelée avec :

$$SE_i = (2i+1)^2 \tag{14}$$

En considérant que x_0 et y_0 représentent la position initiale du point de référence, la suite permettant le calcul d'une masse pour la couronne i est donc définie par les caractéristiques suivantes :

$$U_0 = I_{m,n}(x_0, y_0) \tag{15}$$

$$U_i = U_{i-1} + (i+1)^2 \times \left(\sum_{y=y_0-i}^{y_0+i} \left(|127 - I_{m,n}(x_0 - i, y)| + |127 - I_{m,n}(x_0 + i, y)| \right) + \sum_{x=x_0-i+1}^{x_0+i-1} \left(|127 - I_{m,n}(x, y_0 - i)| + |127 - I_{m,n}(x, y_0 + i)| \right) \right) \tag{16}$$

$$V_i = \frac{U_i}{SE_i} \tag{17}$$

La formule précédente explicite la pondération linéaire (i+1) qui est attribuée aux données issues de la dernière couronne. Cette équation somme, de manière absolue, les différences au niveau normal. Cela revient à compter et évaluer l'ensemble des déformations visibles et invisibles du paysage.
L'index est donc constitué de la série de valeurs pour les différentes couronnes de la carte, c'est-à-dire de i = 0 à min(m/2,n/2).

De manière plus concrète, la Figure 32 présente l'évolution des dimensions pour les deux textes présentés précédemment.

Cette figure présente, avant tout, l'effet de la pondération aux couronnes. Cette pondération permet d'assurer une évolution constante aux représentations ayant une structure en-dehors des couronnes centrales. Inversement, comme c'est le cas avec le texte de l'entretien 423, tout texte ne possédant une structure sur les couronnes périphériques voit sa masse chuter progressivement.

Cette nouvelle méthode, basée sur un automate, permet donc de représenter la structure des données textuelles, c'est-à-dire l'organisation à la fois locale et générale des informations contenues dans le texte.

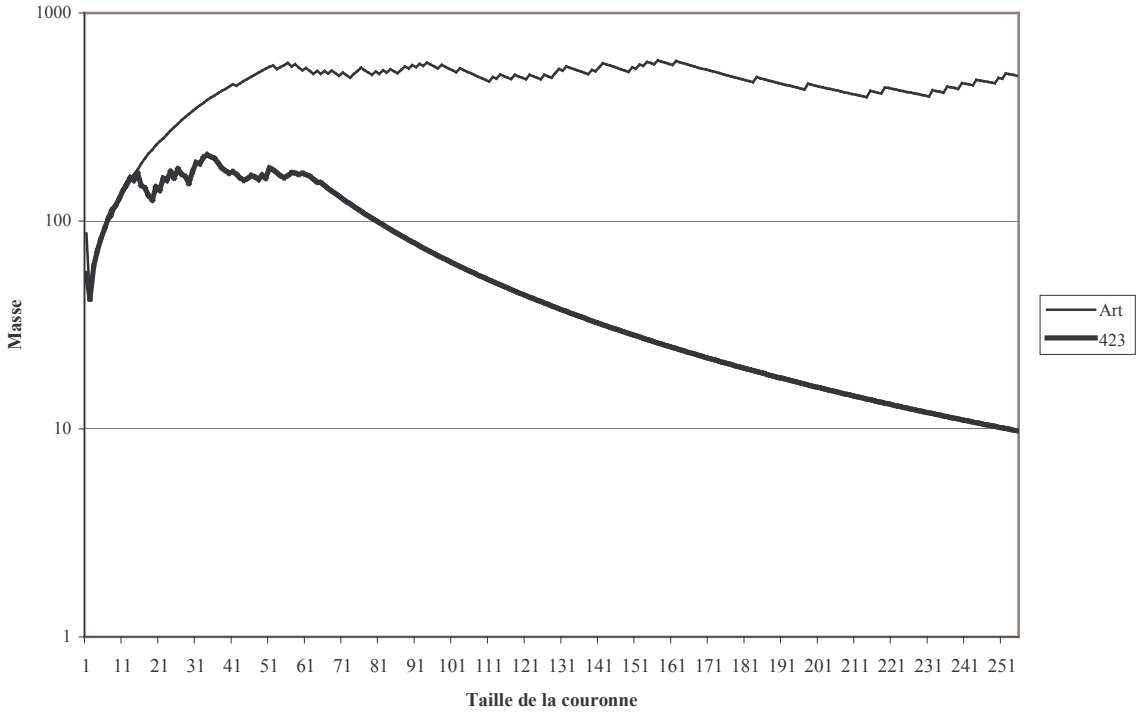

Figure 32 Evolution des masses d'un texte artificiel et de l'entretien 423

Le Tableau 22 et le Tableau 23 présente les scores obtenus par la méthode d'indexation par l'automate peintre sur le corpus Amaryllis et sur le corpus issu de NewsGroups.

Pour le corpus Amaryllis, il peut tout d'abord être constaté une augmentation du score moyen (11,41%) général par rapport à celui de l'automate linéaire (10,97%). Cela reste, néanmoins inférieur au score moyen obtenu par l'indexation par vecteur (16,43%). Pour ce corpus, il y a une nette domination de la distance de Russel-Rao et de la normalisation par recalage.
Les mots sont le type de motifs qui offrent, ici, le meilleur résultat. Les lemmes graphiques et les 1-grammes offrent, néanmoins, des scores assez proches.

Pour le corpus des NewsGroups, aucune combinaison ne se distingue des autres. C'est la distance de Gower appliquée sur des données non-normalisées qui présente le meilleur score moyen. Du point de vue des motifs, la meilleure moyenne est obtenue par les 4-grammes. Ces motifs donnent sur les deux corpus, une bonne moyenne, mais aucun résultat ponctuel intéressant.

Les résultats de l'automate peintre se rapprochent de ceux de la normalisation par vecteurs. Par contre ces expérimentations ne permettent pas de fixer les paramètres. Selon le corpus, la mesure de dissimilarité peut être la distance de Gower ou celle de Russel-Rao. De même, la normalisation peut être non-existante ou un simple recalage des fréquences. Enfin, les motifs peuvent être soit des Mots, des lemmes, des 1-grammes ou des 2-grammes.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	10,23%	12,04%	8,13%	11,66%	10,99%	10,67%
	Euclidienne Standardisée	10,67%	11,54%	12,06%	11,94%	13,19%	12,18%
	Gower	10,22%	12,11%	7,96%	12,03%	7,78%	12,34%
	KullBack-Leibler Symétrie IV	10,45%	12,27%	11,42%	12,25%	10,58%	11,97%
	Mahalanobis	11,21%	8,58%	12,40%	12,13%	10,83%	10,76%
	Russel-Rao	15,47%	**19,70%**	10,16%	9,41%	12,14%	13,46%
	Soergel	9,99%	12,17%	8,17%	12,90%	7,64%	11,60%
Recalage	Bhattacharyya	10,64%	11,89%	11,60%	12,16%	10,58%	11,53%
	Euclidienne Standardisée	7,21%	8,26%	9,66%	9,46%	10,16%	11,47%
	Gower	7,64%	11,43%	8,69%	7,83%	10,01%	12,19%
	KullBack-Leibler Symétrie IV	11,68%	11,84%	11,42%	12,24%	10,53%	11,97%
	Mahalanobis	8,50%	11,33%	8,58%	11,42%	10,47%	10,52%
	Russel-Rao	**24,33%**	18,39%	**20,37%**	**17,55%**	**16,23%**	**19,42%**
	Soergel	10,36%	12,32%	10,44%	13,09%	9,96%	8,30%
Tfldf	Bhattacharyya	10,33%	11,69%	8,17%	12,23%	10,54%	11,47%
	Euclidienne Standardisée	15,73%	12,22%	10,55%	11,35%	10,23%	12,05%
	Gower	15,48%	11,37%	8,22%	12,79%	10,06%	11,63%
	KullBack-Leibler Symétrie IV	10,03%	8,38%	11,42%	12,93%	10,51%	11,85%
	Mahalanobis	11,17%	11,63%	8,17%	12,18%	10,52%	10,59%
	Russel-Rao	11,87%	11,56%	11,49%	12,20%	11,96%	11,94%
	Soergel	8,05%	11,47%	8,22%	12,93%	10,27%	11,62%

Tableau 22 Score F1 sur le corpus Amaryllis avec une indexation par l'Automate Peintre

Quoiqu'il en soit, cette méthode montre, à nouveau, que l'évolution peut servir de base à la comparaison de textes. Le problème de cette méthode est qu'elle représente de manière « tassée » l'évolution. Cela signifie que les traces de l'évolution ne sont pas gardées. Il n'est gardé que le résultat final de cette évolution. On ne peut donc pas dire que l'évolution est correctement représentée.

Cette méthode assure, néanmoins des scores meilleurs que ceux des méthodes Images et Automate Linéaire, c'est la raison pour laquelle elle a été conservée pour la suite des expérimentations.

		Mots	Lemmes	1-grammes	2-grammes	3-grammes	4-grammes
Id	Bhattacharyya	5,44%	3,96%	3,96%	7,16%	3,46%	6,45%
	Euclidienne Standardisée	3,77%	7,33%	5,79%	5,42%	6,66%	6,96%
	Gower	7,87%	**10,53%**	5,73%	10,28%	7,51%	7,52%
	KullBack-Leibler Symétrie IV	4,74%	4,64%	4,35%	4,43%	4,22%	6,54%
	Mahalanobis	4,54%	8,83%	5,69%	5,82%	3,73%	**7,69%**
	Russel-Rao	4,62%	4,03%	5,86%	5,85%	5,47%	6,41%
	Soergel	6,85%	5,61%	4,10%	3,98%	4,44%	6,58%
Recalage	Bhattacharyya	3,94%	6,57%	5,76%	4,03%	**8,36%**	6,32%
	Euclidienne Standardisée	6,27%	3,92%	6,28%	**10,55%**	5,73%	6,12%
	Gower	8,10%	6,94%	5,64%	6,85%	5,14%	7,28%
	KullBack-Leibler Symétrie IV	4,68%	4,32%	4,37%	4,43%	4,22%	6,61%
	Mahalanobis	7,20%	7,32%	**7,02%**	5,53%	5,73%	6,22%
	Russel-Rao	**8,33%**	7,59%	6,25%	3,49%	7,50%	5,75%
	Soergel	5,62%	5,70%	4,03%	4,41%	4,38%	6,47%
Tfldf	Bhattacharyya	4,65%	4,32%	6,78%	3,94%	3,46%	4,75%
	Euclidienne Standardisée	6,80%	5,57%	6,33%	5,58%	7,59%	4,75%
	Gower	4,13%	5,67%	6,11%	3,94%	5,72%	6,45%
	KullBack-Leibler Symétrie IV	4,74%	4,75%	4,37%	4,03%	4,22%	6,54%
	Mahalanobis	5,95%	5,90%	5,98%	4,82%	5,01%	4,75%
	Russel-Rao	6,13%	5,44%	5,91%	5,20%	5,46%	5,75%
	Soergel	5,65%	6,28%	4,33%	5,56%	4,44%	6,45%

Tableau 23 Score F1 sur le corpus NewsGroups avec une indexation par l'Automate Peintre

2.2.4. Bilan des méthodes basées sur l'organisation du discours

Cette partie a permis d'étudier trois méthodes d'indexation basées sur l'organisation du texte. De ces trois méthodes et compte tenu des résultats obtenus, une seule méthode peut être retenue.

La méthode d'indexation par Automate Peintre permet une indexation des textes par une représentation globale de leur évolution. Cette méthode d'indexation permet d'obtenir des résultats similaires à la méthode d'indexation par les informations structurelles. Cette méthode a, surtout, l'avantage de prouver que les textes peuvent être modélisés par leur évolution. De plus cette indexation a un nombre réduit de caractéristiques.

L'évolution des textes est basée sur l'organisation du contenu des textes. Elle est donc propre aux auteurs, aux humeurs, aux pensées … Dans le cadre d'une indexation sociologique des textes, l'évolution constitue donc une intéressante base d'indexation. C'est suite à un tel bilan que la méthode basée sur l'évolution présentée dans la prochaine partie a été créée.

2.3. Des méthodes d'indexation basées sur l'évolution des textes

Les représentations précédentes, peu importent leurs qualités, ont le gros défaut d'offrir une représentation globale du texte. C'est-à-dire que tout le contenu ou toute la structure du texte

sont représentés de manière générale. Ce concept tue littéralement l'idée de raisonnement et d'évolution d'un texte.
Or un texte dépend d'un raisonnement que ce raisonnement soit scientifique, littéraire ou philosophique. Toute communication fait l'objet d'une réflexion sur l'ordre, l'organisation des informations. Un texte peut donc être vu comme un ensemble de textes. Chaque partie du texte a une certaine indépendance qui fait sa spécificité. Et chaque partie de texte est liée de manière logique à la partie qui la précède et à celle qui la suit. Dans le cas d'une absence de raisonnement, c'est-à-dire une suite d'informations sans lien, seule l'indépendance des textes sera représentée. Qu'il y ait un raisonnement ou non, un texte peut être segmenté en diverses parties.
L'évolution d'un texte peut être définie comme l'ensemble des liens logiques présents ou absents entre les parties successives d'un texte. L'évolution est donc une représentation de logique textuelle. Plus précisément, elle représente la logique textuelle de l'auteur. La logique est la base de toute structure, or la structure et le contenu sont liés. La comparaison de deux auteurs ou de deux textes revient donc à comparer leurs évolutions. De cette proposition, il faut étudier la méthode de segmentation d'un texte en parties et « l'effet mémoire » qui accompagne la segmentation. Chacun de ces paramètres est détaillé dans une partie. Il s'agit respectivement des parties 2.3.1 et 2.3.2. La partie 2.3.3 présente, quant à elle, une expérimentation permettant de faire un choix pour chacun des paramètres. La partie 2.3.4 détaille comme l'indexation est réalisée à partir d'une telle représentation et la partie 2.3.5 présente des expérimentations pour évaluer cette méthode d'indexation. Enfin la partie 2.3.6 fait un bilan sur cette méthode d'indexation.

2.3.1. Segmentation des textes

L'évolution textuelle évalue les liens logiques entre les différents segments de texte. La première contrainte réside dans le fait que si un texte est articulé autour de liens logiques quelconques, les pré-traitements de préparation de texte doivent conserver ces liens ou, tout du moins, ne pas les détruire. Pour cela, il faudrait pouvoir repérer les articulations des textes. Une telle procédure semble trop coûteuse pour être mise en œuvre. De plus, le nombre de parties ne correspondrait pas au nombre de parties désirées. Enfin, la taille de chaque partie serait trop variable pour qu'une réelle comparaison puisse être faite.
Chaque partie devant être indépendante et préférablement de taille égale aux autres parties, la solution consiste donc à découper le texte en parts égales. Plus il y aura de parties et plus les parties seront courtes et plus les segmentations seront alignées sur les articulations du texte. Inversement moins il y aura de parties et plus elles seront longues et plus il y aura matière à comparaison. La taille des parties est donc un choix important.

2.3.2. La mémoire

L'effet mémoire représente, lui aussi, un choix. Ce choix peut se résumer ainsi : la comparaison doit-elle se situer entre deux parties segmentées, tel que c'est fait en segmentation thématique, ou doit-elle se situer entre une partie et les parties qui la précèdent ? Une segmentation sans mémoire correspond donc à une segmentation en parties distinctes de chaque texte. Une segmentation avec mémoire correspond à une représentation du texte à différents stades d'évolution : en plus du présent, il y a tout le passé.
Le premier choix permet de comparer indépendamment deux structures. Le second permet d'évaluer l'intégration d'une partie de texte au texte qui la précède. Chacun de ces choix

paraît refléter une partie de la notion d'évolution. Cependant, il a été supposé, avant les expérimentations, que c'est l'intégration d'une partie de texte au texte qui la précède qui reflète plus l'idée d'évolution. En effet, cela signifierait que chaque partie est ajoutée au reste de façon à former un tout cohérent.

2.3.3. Choix des paramètres

La Figure 33 présente l'évolution des scores moyens obtenus sur les deux corpus Amaryllis et NewsGroups. Ces scores moyens reflètent les scores ponctuels.

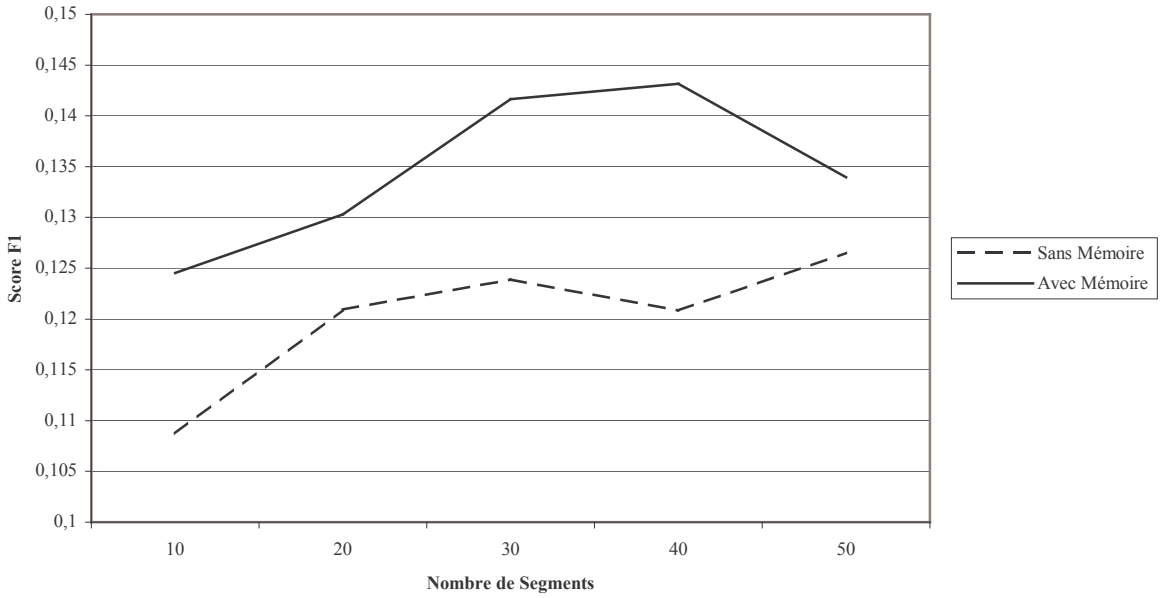

Figure 33 Evolution des scores suivant le nombre des segments et l'utilisation de mémoire

De manière évidente, il peut être confirmer que l'utilisation d'une mémoire dans la segmentation apporte de bien meilleurs scores que si celle-ci n'est pas utilisée. Cela signifie que l'évolution textuelle correspond à l'intégration de nouvelles données textuelles au texte en cours.

Du point de vue des segments, les expérimentations montrent que les meilleurs scores sont obtenus avec 40 segments. Dans toutes les expérimentations menées, ce sont toujours les nombres de 30 et 40 segments qui s'avèrent être les meilleurs découpages. Un découpage plus important semble ne pas apporter suffisamment de matière à la comparaison. Un découpage moins important ne permet pas de représenter les diverses articulations des textes. Pour le projet, il a été conservé le nombre de 40 segments.

La première étape de l'indexation est donc terminée. Chaque texte est transformé en un corpus de 40 textes. Le premier texte du corpus correspond au premier $1/40^{e}$ du texte original. Le dernier texte du corpus correspond au texte original. L'ensemble des textes intermédiaires correspond au texte précédent auquel il a été ajouté $1/40^{e}$ de texte. L'évolution du texte est donc représentée par une série de données textuelles.

2.3.4. L'indexation

L'évolution est donc représentée au niveau des unités textuelles par une explosion du corpus et une multiplication des informations. Chaque texte est représenté par un corpus d'évolution dont les textes d'évolution expriment l'évolution du texte représenté. Evaluer l'évolution revient donc à calculer les n-1 distances successives entre les textes d'évolution.
D'une manière plus pratique, chaque texte d'évolution est donc indexé par une méthode de représentation globale, puis les distances successives sont calculées.

De la partie précédente, il a été retenu trois méthodes d'indexation sur la représentation globale des textes. Ces méthodes sont utilisées pour indexation des textes d'évolution. Contrairement aux autres méthodes, la méthode d'indexation par l'automate peintre a l'avantage de proposer une représentation de l'évolution. Cette méthode est donc pressentie comme la meilleure méthode pour représenter l'évolution locale des textes.
De même les expérimentations ont permis d'associer à chaque méthode d'indexation les pré-traitements, les pondérations, les normalisations et les mesures de dissimilarités les mieux adaptés.
L'indexation et les dissimilarités successives peuvent donc être calculées. Chaque texte original est donc indexé par son évolution, c'est-à-dire par la série de dissimilarités.

Cette indexation étant nouvelle, aucune mesure ne lui est encore associée. L'index est constitué de valeurs réelles, les sept mesures fixées dans le chapitre précédent peuvent donc être ré-appliquées.
La partie qui suit étudie de manière expérimentale cette méthode par les différentes méthodes d'indexation locale qui peuvent être utilisées et par les mesures de dissimilarités globales les mieux adaptées.

2.3.5. Expérimentations

Les scores entre les corpus Amaryllis et NewsGroups sont, d'une manière générale, comparables. Cela signifie que les méthodes choisies adoptent un comportement similaire sur les différents corpus étudiés. Une étude des deux corpus paraît donc redondante. Les résultats présentés sont donc limités à ceux du corpus Amaryllis, mais les conclusions sur le corpus NewsGroups sont identiques.

Le Tableau 24 présente les scores obtenus si l'indexation utilisée localement est l'indexation par Vecteurs. Il faut rappeler que les lemmes graphiques et la distance euclidienne standardisée suite à un recalage des fréquences sont les mieux adaptés à une indexation par Vecteurs.

Les meilleurs scores de la méthode sur la représentation globale étaient de l'ordre de 40%, le score moyen était par contre de 16,43%. L'indexation par l'évolution, avec 24,44%, assure donc un score maximum plus faible. Cependant, le score moyen de l'indexation par l'évolution est de 18,35%. L'évolution assure un score moyen plus élevé que la représentation globale. Du point de vue des mesures de dissimilarités, c'est la distance de Mahalanobis qui paraît être la mieux adaptée.

Mesure	Score F1
Bhattacharyya	14,46%
Euclidienne Standardisée	20,90%
Gower	21,83%
KullBack-Leibler Symétrie IV	16,60%
Mahalanobis	**24,44%**
Russel-Rao	11,92%
Soergel	18,31%

Tableau 24 Scores de l'indexation par l'évolution pour une indexation locale par les Vecteurs

Le Tableau 25 présente les scores de l'indexation par l'évolution avec une indexation locale par la méthode des informations structurelles. A cette méthode, les lemmes graphiques, la distance de Gower et le recalage des fréquences sont les mieux adaptés.

Mesure	Score F1
Bhattacharyya	14,89%
Euclidienne Standardisée	10,15%
Gower	13,37%
KullBack-Leibler Symétrie IV	12,82%
Mahalanobis	**15,63%**
Russel-Rao	7,21%
Soergel	15,54%

Tableau 25 Scores de l'indexation par l'évolution pour une indexation locale par les Informations Structurelles

Les conclusions sont identiques à celles pour une utilisation de l'indexation par Vecteurs. C'est-à-dire que le score maximum diminue de plus de 20% pour une représentation globale à 15,63% pour une indexation par l'évolution. De plus , là aussi le score moyen augmente. Il passe de 11,89% à 12,80%. L'augmentation est donc moins grande mais toujours significative. Enfin, il faut noter que, là aussi, c'est la distance de Mahalanobis qui apparaît être la mesure de dissimilarité la mieux adaptée.

Enfin, le Tableau 26 présente les scores obtenus lorsque l'indexation locale est l'automate peintre. Il faut rappeler que pour l'automate peintre, quatre types de motifs (mots, lemmes, 1-grammes et 2-grammes), deux types de normalisation (sans normalisation et recalage des fréquences) et deux types de mesures de dissimilarités (Gower et Russel-Rao) ont été conservés.

Il peut, tout d'abord, être observé l'augmentation du score maximum. Le score maximum pour une représentation globale était de 24,33%, avec l'évolution le maximum atteint est de 30,24%. De plus, comme pour les deux méthodes précédentes, le score moyen augmente, il passe de 11,41% à 14,05%. L'augmentation tant locale que moyenne indiquerait donc que c'est la méthode d'indexation la mieux adaptée à la mesure de l'évolution.

			Mots	Lemmes	1-grammes	2-grammes
Id	Gower	Bhattacharyya	14,33%	10,94%	13,46%	**29,39%**
		Euclidienne Standardisée	**21,98%**	16,86%	7,83%	18,74%
		Gower	18,46%	13,35%	12,50%	18,95%
		KullBack-Leibler Symétrie IV	7,87%	8,09%	11,85%	12,72%
		Mahalanobis	16,69%	13,38%	11,87%	13,33%
		Russel-Rao	16,12%	**19,03%**	18,15%	19,52%
		Soergel	11,81%	15,09%	14,91%	19,29%
	Russel-Rao	Bhattacharyya	10,37%	16,09%	19,46%	17,81%
		Euclidienne Standardisée	10,37%	16,09%	18,95%	17,81%
		Gower	11,83%	11,72%	19,58%	15,94%
		KullBack-Leibler Symétrie IV	13,43%	12,05%	8,73%	12,48%
		Mahalanobis	10,37%	16,09%	8,73%	17,81%
		Russel-Rao	8,73%	8,73%	8,73%	8,73%
		Soergel	10,46%	14,64%	8,73%	16,39%
Recalage	Gower	Bhattacharyya	16,34%	12,22%	14,81%	20,16%
		Euclidienne Standardisée	21,80%	8,17%	10,89%	18,53%
		Gower	14,92%	13,98%	11,83%	16,26%
		KullBack-Leibler Symétrie IV	12,58%	8,00%	11,54%	11,24%
		Mahalanobis	11,28%	17,83%	11,88%	22,61%
		Russel-Rao	8,50%	11,20%	11,28%	15,82%
		Soergel	19,67%	21,92%	12,24%	15,52%
	Russel-Rao	Bhattacharyya	16,72%	7,92%	14,80%	13,61%
		Euclidienne Standardisée	16,72%	7,92%	14,80%	13,61%
		Gower	11,32%	17,20%	**30,24%**	12,71%
		KullBack-Leibler Symétrie IV	11,63%	12,83%	14,17%	12,59%
		Mahalanobis	17,43%	11,52%	14,96%	13,61%
		Russel-Rao	8,73%	8,73%	8,73%	8,73%
		Soergel	11,99%	12,89%	14,62%	12,40%

Tableau 26 Scores de l'indexation par l'évolution pour une indexation locale par l'Automate Peintre

Par les scores, les paramètres peuvent être fixés. C'est donc sans pondération et avec la distance de Gower que la méthode d'indexation locale assure les meilleurs résultats. Il est intéressant de voir que cela correspond à la chaîne de traitement offrant le meilleur score pour le corpus des NewsGroups. Dans cette configuration, c'est la distance de Russel-Rao qui offre le score le plus élevé et plus précisément lorsque les motifs utilisés sont les 2-grammes.

2.3.6. Bilan sur l'indexation par l'évolution textuelle

Cette partie a présenté une nouvelle méthode d'indexation basée sur des méthodes d'indexation appliquée localement de manière à évaluer l'évolution des textes.
L'évolution a pu être définie comme l'évaluation de l'intégration de nouvelles données textuelles dans un texte. Cela indique que la création du corpus d'évolution nécessite l'utilisation d'une mémoire lors de la segmentation. C'est-à-dire que chaque texte d'évolution,

du corpus d'évolution, est constitué des données du texte qui le précède et de données qui lui sont propres.
De même, cette partie a étudié la quantité de segments à former, c'est-à-dire la quantité de caractéristiques que doit contenir l'index. Il a pu être observé qu'une trop faible quantité d'information nuit à l'évaluation de l'évolution, et qu'une quantité trop grande empêche de distinguer l'évolution.
Enfin, d'une manière générale, il peut être conclu que l'indexation par l'évolution permet d'augmenter les scores moyens des méthodes d'indexation basées sur une représentation globale des textes. Cependant, certaines méthodes d'indexation semblent mieux adaptées que les autres pour être utilisées localement dans une indexation sur l'évolution.

L'évolution permet donc une indexation performante pour un minimum de caractéristiques. De plus, l'évolution permet, d'une part, de se détacher du contenu et, d'autre part, d'évaluer le raisonnement ou, du moins, les liens logiques existant entre les parties successives. Or, une indexation sur le raisonnement, détachée du contenu, paraît être la meilleure réponse à fournir dans le cadre d'une indexation sociologique.

3. Bilan

Ce chapitre a donc permis de réaliser le processus de classification en proposant diverses méthodes d'indexation. Dans un premier temps, le découpage en unités textuelles et trois sortes de pondérations ont été étudiés. Des expérimentations montrent que les paramètres sont liés aux méthodes d'indexation et ne peuvent pas être fixés auparavant.

Au niveau des méthodes d'indexation, certaines méthodes existantes ont permis d'indexer les textes sur leur contenu et sur leur structure. Les expérimentations prouvent le bon fonctionnement de ces méthodes sur des textes journalistiques (Amaryllis) et spécifiques (NewsGroups). Dans de tels corpus, il existe des informations structurelles (passage de lignes, majuscules, …) et du contenu même si tous deux sont réduits à un condensé.

Dans le cas de textes oraux retranscrits, il n'existe pas réellement de structure. De plus, le langage oral réduit le vocabulaire aux mots les plus communs. Les méthodes existantes ne semblent donc guère adaptées au traitement des corpus sociologiques. C'est la raison pour laquelle des méthodes d'indexation qui cherchent à représenter l'évolution, c'est-à-dire le raisonnement du texte ont été développées pour ce projet. A force de méthodes, il est apparu qu'une représentation globale du résultat de l'évolution n'est pas suffisant. L'évolution a donc été étudiée de manière plus locale.

La dernière méthode constitue donc une méta-méthode, qui crée un corpus d'évolution et calcule les index et les dissimilarités d'évolution. Cette méthode a montré qu'elle améliore les résultats moyens sur les deux corpus cités précédemment. Cependant, elle semble nécessiter une méthode d'indexation locale adaptée à la mesure de l'évolution. Une méthode développée pour ce projet semble la mieux adaptée. Cette méthode est basée sur un automate peintre et a pour but de représenter de manière globale l'évolution.

Au-delà des méthodes, les expérimentations de ce chapitre montrent que l'évolution des textes, si elle est correctement représentée, permet leur indexation.

Chapitre 5 – Application aux Données : Le corpus Le Corbusier

Ce chapitre s'attache à présenter et à étudier un exemple de corpus réel d'application de ce travail. Il s'agit d'une série d'enquêtes sur les unités d'habitation imaginées par Le Corbusier. Une première partie présente le corpus en détail car c'est l'occasion de présenter l'ampleur du travail manuel effectué par les sociologues. Une seconde partie étudie les résultats obtenus par les méthodes d'indexation sur ce corpus.

1. Le corpus Le Corbusier

En 2003, commencent deux séries d'enquêtes, dirigées par Sylvette Denèfle, ayant pour but d'analyser les différentes façons d'habiter dans les unités d'habitation « Le Corbusier ». Une série est menée auprès des habitants de la cité de Firminy et l'autre auprès des habitants de la cité de Rezé.
Une courte présentation de l'architecte et du contexte de ces deux cités est donnée dans l'Annexe 1.
Cette partie va, quant à elle, présenter, tout d'abord, de manière concise l'équipe de sociologues qui a entourée Sylvette Denèfle pour cette étude. Puis, les enquêtes seront brièvement décrites. Enfin les variables sociologiques issues de l'analyse seront, en partie, détaillées.

1.1. L'équipe de sociologues

DENEFLE Sylvette – professeure de sociologie à l'Université François-Rabelais de Tours – Domaine de recherche : Genre, Sociologie urbaine, Idéologie.

DUSSUET Annie - maître de conférences en sociologie à l'université de Nantes - Domaines de recherche : Sociologie des rapports sociaux de sexe : travail domestique, emplois de proximité.

ROUX Nicole – maître de conférences en sociologie à l'université de Bretagne Occidentale - Domaines de recherche : Monde ouvrier, Politique, Femmes.

BISSON Sabrina – doctorante en sociologie (dirigée par Sylvette Denèfle) à l'Université François-Rabelais de Tours – Domaine de recherche : Sociologie urbaine.

1.2. Les enquêtes

L'analyse a débuté par les enquêtes faites dans la cité de Rezé. Les travaux sur les entretiens y sont donc plus aboutis que ceux sur la cité de Firminy. C'est la raison pour laquelle, il n'a été retenu pour unique corpus que la série d'enquêtes menées à Rezé. Les informations générales sont, cependant, vraies pour les deux séries d'enquêtes.

La série d'enquêtes menée à Rezé a abouti à un corpus de 32 entretiens répartis en 30 entretiens individuels et 2 entretiens mixtes, lors desquels plusieurs individus interviennent. Une fois les 2 entretiens mixtes séparés en entretiens individuels, le corpus est donc composé de 34 entretiens individuels.

Les thèmes abordés durant les enquêtes du corpus Le Corbusier sont :

- L'aménagement : Ensemble d'informations liées à la structure même de l'habitation et l'organisation qui en est issue.
- Les associations : Ensemble d'informations liées à la vie associative à l'investissement dans les unités d'habitation.
- Intérieur/Extérieur : Ensemble d'informations liées aux services et à l'environnement tant intérieur aux unités d'habitation que dans un voisinage proche.
- La sociabilité : Ensemble d'informations liées aux rapports humains et aux possibilités offertes en ce sens par les unités d'habitation.
- La théorie : Ensemble d'informations liées à la conception architecturale et urbanistique de Le Corbusier.
- La vie familiale : Ensemble d'informations liées à la structure familiale et à son organisation au travers de l'unité d'habitation et de ses services.

1.3. Les thèmes et les variables sociologiques

Il faut rappeler que les variables sociologiques sont extraites des entretiens et représentent du sens pour un thème abordé. Elles sont, comme expliqué plus en détail dans le chapitre suivant, des reformulations sociologiques de phrases extraites de l'entretien. Voici, pour chaque thème, les deux variables sociologiques les plus partagées :

- Aménagement
  - « le parc est un espace approprié par les habitants »
  - « l'école elle-même est un espace privé »
- Associations
  - « je ne participe pas aux activités (de l'immeuble) »
  - « j'ai d'autres responsabilités dans d'autres associations (d'autres investissements) »
- Intérieur/Extérieur
  - « utilise les supermarchés de Rezé »
  - « circule en voiture »
- Sociabilité
  - « j'ai connu la poste »
  - « au Corbusier, on se dit bonjour de manière assez systématique »
- Théorie
  - « depuis que je suis là, je me suis intéressée à Le Corbu »
  - « le fait d'habiter au Corbusier crée un intérêt, une certaine curiosité »
- Vie Familiale
  - « le parc, c'est bien pour les enfants »
  - « l'école est pratique pour emmener les enfants »

De la série d'enquêtes menées dans la cité de Rezé, il a été extrait un total de 7982 variables sociologiques réparties d'une manière globale comme le présente le Tableau 27 qui suit.

Il peut être observé qu'il existe, d'un point de vue global, trois groupes de thèmes. Tout d'abord un groupe contenant les deux thèmes les plus fournis en variables sociologiques : Sociabilité et Théorie. Ensuite, le thème Aménagement, seul dans sa catégorie. Enfin, un groupe contenant les thèmes Association, Intérieur/Extérieur et Vie Familiale qui sont les trois thèmes les moins fournis en variables sociologiques.

Thème	Nombre de variables sociologiques
Aménagement	762
Associations	250
Intérieur/Extérieur	253
Sociabilité	2604
Théorie	3820
Vie Familiale	287
Total	7982

Tableau 27 Nombres de variables sociologiques par thème abordé

D'une manière plus précise, les parties qui suivent vont détailler la distribution des variables sociologiques pour chaque thème abordé.

1.3.1. Aménagement

Le thème Aménagement, le 3ème plus fourni en variables sociologiques, n'a pas moins de 762 variables distinctes. La Figure 34 détaille la distribution de ces variables selon le nombre d'entretiens dans lesquels elles apparaissent. Il peut donc être observé que 362 variables, soit 47% d'entres elles, sont des variables uniques, c'est-à-dire abordées par un unique entretien.

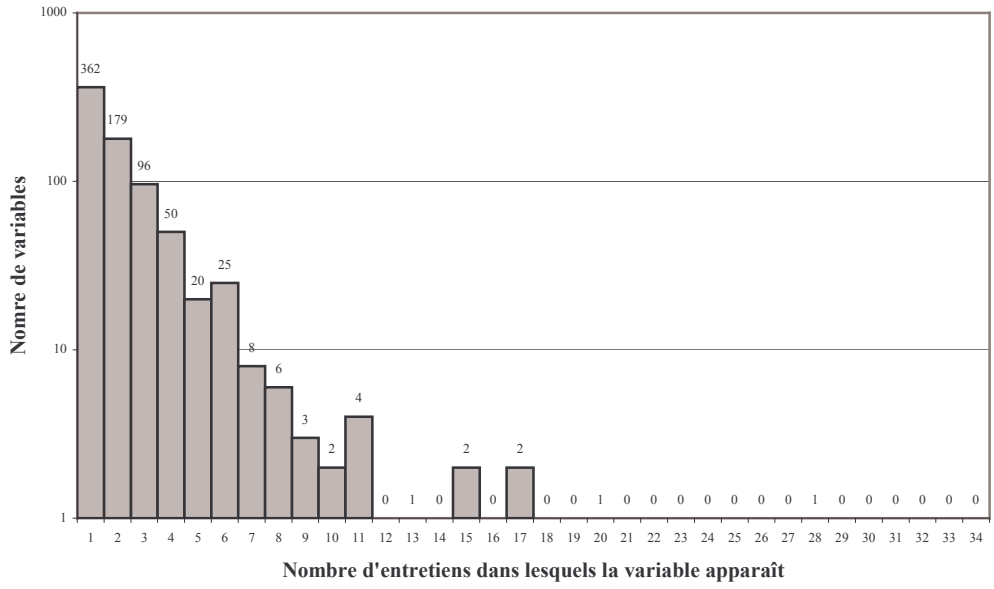

Figure 34 Distribution des variables pour le thème Aménagement

De plus, il peut être observé que deux variables sont partagées par plus de la moitié des entretiens avec respectivement 20 et 28 entretiens. A partir de telles variables, une hypothèse peut être émise sur l'existence de deux groupes, l'un formé de l'ensemble des entretiens abordant les deux variables, l'autre constitué des entretiens qui n'abordent ni l'une ni l'autre.
La liste des entretiens formant le premier est, à titre informatif, la suivante : 12, 113, 309, 402, 423, 425, 444, 53, 509, 515, 521, 523, 526, 605, 629, 630 et 633.
De même, la liste des entretiens formant le seconde groupe est la suivante : 115, 625 et 634.

1.3.2. Association

Le thème Association est le thème le moins fourni en variables sociologiques avec seulement 250 variables distinctes. Sa distribution, Figure 35, s'en ressent très nettement, puisque seulement 31 variables, c'est-à-dire à peine 12.4%, sont partagées par plus de 3 entretiens. Les 5 variables les plus partagées le sont par respectivement 7, 7, 7, 8, 13 entretiens.
Parmi l'ensemble des entretiens, un groupe de 12 entretiens peut, d'ores et déjà, être identifié comme ne partageant aucune de ces variables. Ce groupe est formé des entretiens : 50, 53, 530, 534, 61, 605, 625, 629, 630, 632, 633 et 634.
La liste des entretiens partageant 4 de ces 5 variables n'est constituée que des entretiens 12 et 117, alors que la liste des entretiens partageant les deux variables les plus fréquentes est constituée des entretiens 113, 117 et 136.

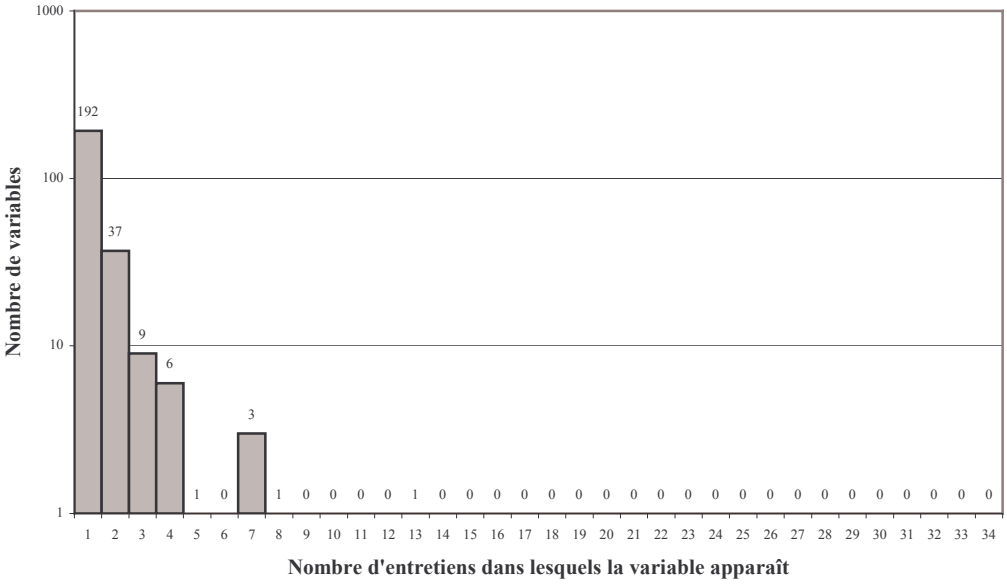

Figure 35 Distribution des variables pour le thème Association

Si les groupes formés dans ces deux premières variables sociologiques sont représentatifs des classes sociologiques formées, l'hypothèse d'un classement unique des entretiens est infirmé par la présence, par exemple, des entretiens 633 et 634 dans deux groupes différents pour le thème Aménagement et dans le même groupe pour le thème Association. Cependant, ces groupes ne sont eux-mêmes qu'hypothétiques. L'affirmation ou l'infirmation finale de cette hypothèse sera faite lorsque les différentes classifications seront présentées.

1.3.3. Intérieur/Extérieur

Intérieur/Extérieur est, certes, le deuxième thème le moins fourni en variables sociologiques différentes, mais la distribution de celles-ci est plus régulière comme le montre la Figure 36.
Comme pour les variables précédentes, on peut isoler deux ensembles d'entretiens. Un premier groupe pour lequel l'ensemble des entretiens partagent les 3 variables les plus communes (respectivement 25, 26 et 28 entretiens en commun) : 112, 113, 136, 203, 335, 402, 423, 425, 444, 53, 508, 509, 521, 523, 526, 625, 629, 630, 632, 634. Un second groupe peut être identifié comme le groupe pour lequel les entretiens ne partagent aucune des 3 variables citées précédemment : 117, 307, 414.

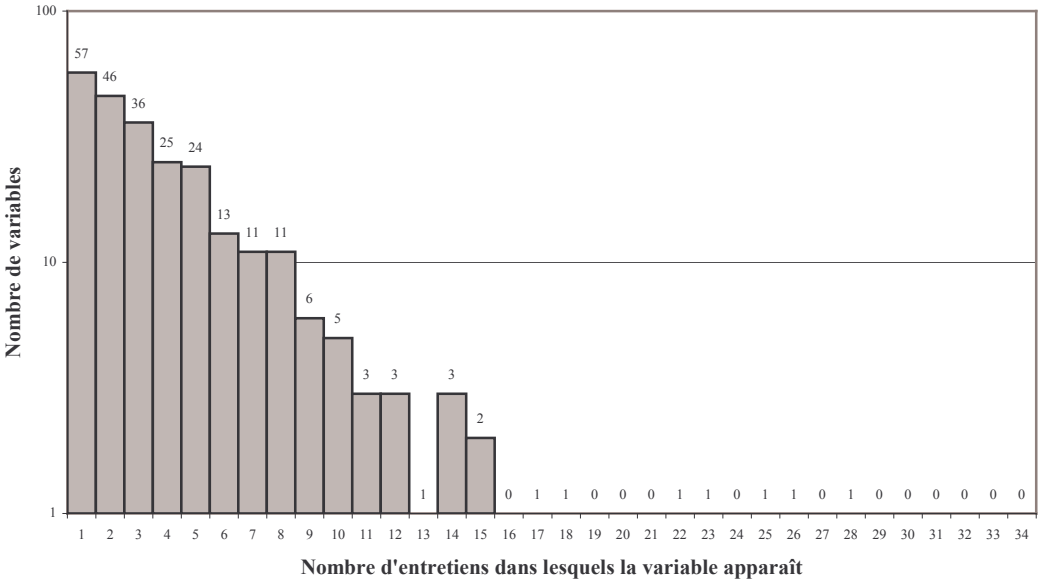

Figure 36 Distribution des variables pour le thème Intérieur/Extérieur

1.3.4. Sociabilité

Avec 2604 variables, le thème Sociabilité est le second thème le plus pourvu de variables sociologiques. Sur l'échelle logarithmique du graphique, la distribution paraît aussi régulière que pour le thème Intérieur/Extérieur qui comptait pourtant plus de 10 fois moins de variables.
La particularité de cette distribution tient au fait qu'elle paraît décroître de manière régulière. Les groupes formés à partir des deux variables les plus communes sont : 12, 115, 117, 203, 309, 423, 61, 605, 634 en ce qui concerne ceux qui partagent ces deux variables et 307, 444, 50 en ce qui concerne ceux qui ne partagent ni l'une ni l'autre.

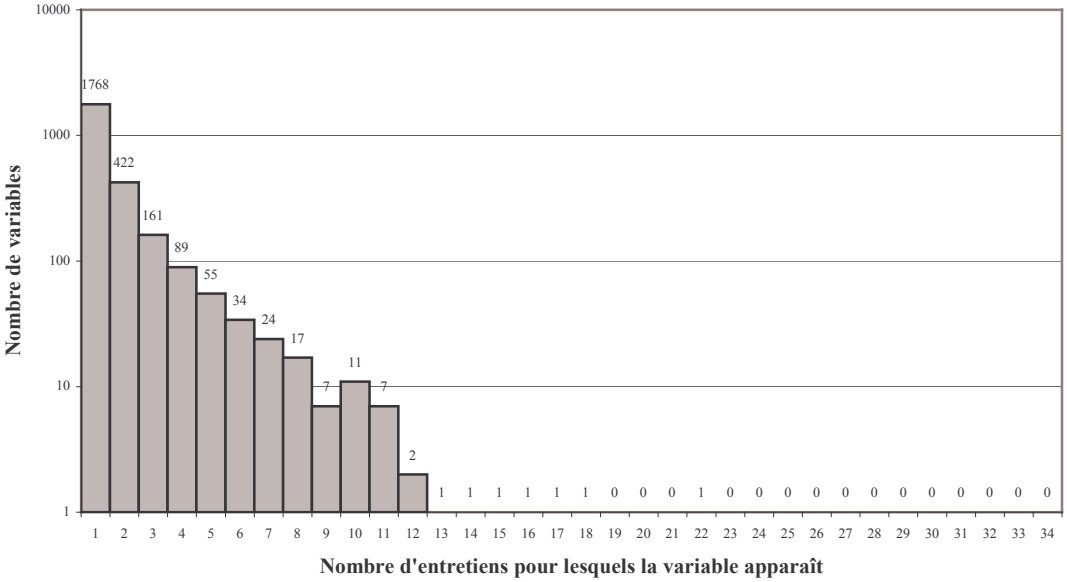

Figure 37 *Distribution des variables pour le thème Sociabilité*

1.3.5. Théorie

Le thème Théorie est celui pour lequel il y a le plus de variables sociologiques qui ont été extraites : 3820. Cependant, une fois les 3390 variables uniques enlevées, il n'en reste plus que 430 soit plus de 400 de moins que pour le thème Sociabilité pour lequel il y a 834 variables qui sont partagées par au moins 2 entretiens. Cela montre bien, qu'il ne faut pas se fier à la quantité totale de variables.
Il faut, de plus, remarquer que la distribution est très courte, puisque presque toutes les variables sont partagées par au plus 11 entretiens.

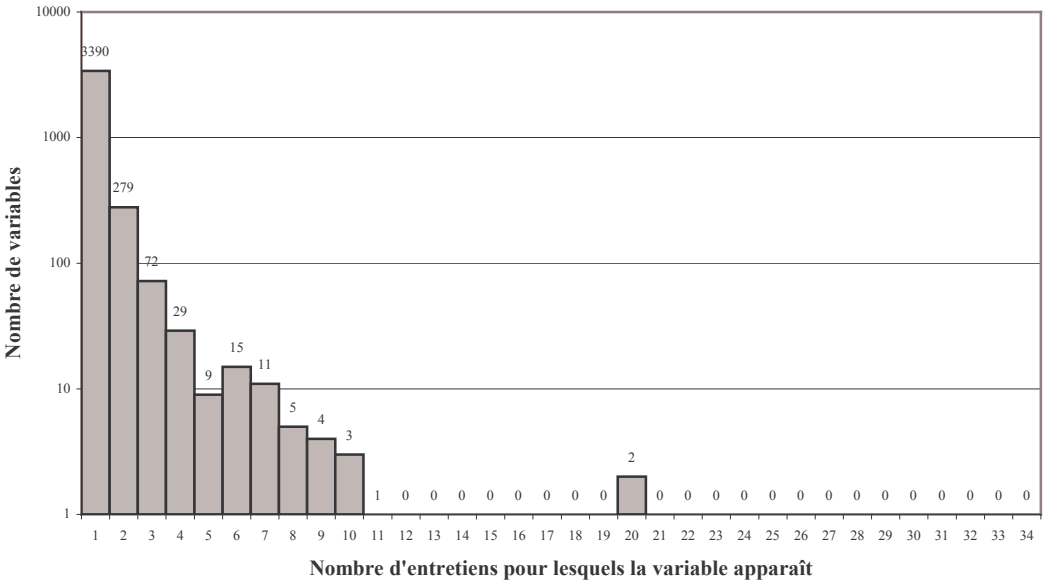

Figure 38 Distribution des variables pour le thème Théorie

Seules deux variables, citées en exemple précédemment, sont partagées par 20 entretiens :

- depuis que je suis là, je me suis intéressée à Le Corbu

- le fait d'habiter au Corbusier crée un intérêt, une certaine curiosité

Les deux ensembles d'entretiens représentés par ces deux variables sont égaux, c'est-à-dire que tout entretien d'un groupe appartient à l'autre et inversement.
Même si ces deux groupes peuvent être listés : 112, 203, 245, 307, 309, 335, 414, 423, 425, 444, 53, 508, 509, 530, 61, 605, 625, 629, 630, 634, pour ceux qui partagent les deux variables et 12, 113, 115, 117, 136, 402, 50, 515, 521, 523, 526, 534, 632, 633 pour ceux qui ne les partagent pas ; il peut être intéressant de se poser la question de la fusion de telles variables.

1.3.6. Vie Familiale

Avec 151 variables uniques sur 287, soit 52.6% d'entre elles, le thème Vie Familiale est proche, dans sa distribution, du thème Association. Cependant, dans Association la dispersion est plus longue. En effet, alors que pour Association, le groupe principal de variables est borné aux variables apparaissant dans au maximum 5 entretiens ; dans Vie Familiale, ce même groupe est borné à 7. De même, alors que les variables les plus communes d'Association apparaissent dans 13 entretiens, les plus communes de Vie Familiale apparaissent dans 17 entretiens. La *Figure 39* montre cette distribution en détail.

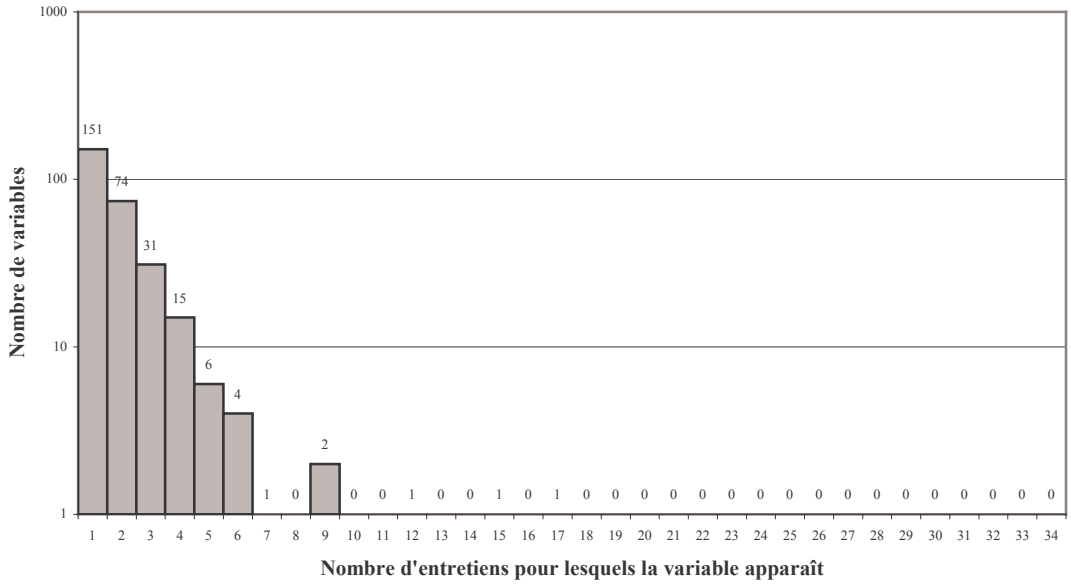

Figure 39 Distribution des variables pour le thème Vie Familiale

Comme précédemment, les entretiens formant l'intersection entre les deux variables les plus communes peuvent être listés : 309, 402, 444, 515, 521, 523, 605, 630, 633. De même, les entretiens ne partageant aucune des deux variables les plus communes sont : 112, 115, 117, 136, 203, 307, 335, 414, 50, 625, 634.

2. Quelles méthodes pour le corpus Le Corbusier ?

Cette partie étudie dans un premier temps les variables afin de créer des références de comparaison. Puis, dans un second temps, cette méthode présentent les expérimentations

menées sur le corpus Le Corbusier et les résultats obtenus par une partie des méthodes d'indexation présentées dans le chapitre précédent.

2.1. Les variables

L'indexation sociologique, faite manuellement par les sociologues, donne un index binaire de variables. Chaque valeur binaire indique la présence, ou non, de la variable dans l'entretien indexé. Cet index sert de référence à l'index obtenu par les méthodes d'indexation des entretiens proposées dans le chapitre précédent. La méthode de comparaison est la même que dans les chapitres précédents, c'est-à-dire que des classes sont formées à partir de l'arbre, puis un score F1 est calculé pour évaluer la qualité de l'indexation.
L'index constitué de l'ensemble des variables sert de référence. Il paraît donc normal de l'étudier. Les variables peuvent être étudiées afin de voir si elles ont toutes une raison d'influencer la classification. La partie 2.1.1 étudie la sélection de ces variables. Puis, le chapitre 3 a montré l'importance du choix des mesures de dissimilarités. La partie 2.1.2 étudie les mesures qui avaient été retenues dans le cadre de la création de la référence.

2.1.1. Le type de variables

Cette partie s'intéresse à la sélection des variables. Il a été vu dans la partie 1.3 les différentes distributions des variables selon le nombre d'entretiens dans lesquels elles apparaissaient. Ces distributions rappellent les courbes de Zipf. Il a été vu que la majorité des variables n'apparaissent que dans un unique entretien et que plus on cherche la diversité au niveau des entretiens et moins il y a de variables.
On peut donc émettre une critique sur l'utilité de chaque variable. On pourrait, tout d'abord, s'intéresser aux variables présentes dans de nombreux entretiens. Ces variables sont-elles du bruit ? Mais elles sont en trop faible nombre pour qu'une réflexion soit lancée dans ce sens.
Par contre, il y a de nombreuses variables qui n'apparaissent que dans un seul entretien. Or, une variable n'apparaissant que dans un unique entretien n'a aucun intérêt. Cela aurait un intérêt en psychologie où on s'intéresse plus à l'individu. Mais la sociologie étudie les hommes à travers leurs relations. Donc, le caractère spécifique des variables n'a pas d'intérêt. On peut alors se poser la question pour les variables qui ne sont abordées par uniquement deux entretiens, …
Il s'agit, en effet, de faire une réelle sélection des variables afin de limiter l'index sociologique aux variables les plus pertinentes. C'est-à-dire les variables les plus à même d'être utilisée pour une classification.
La figure suivante présente le nombre total de variables retenues selon la spécificité minimum demandée.

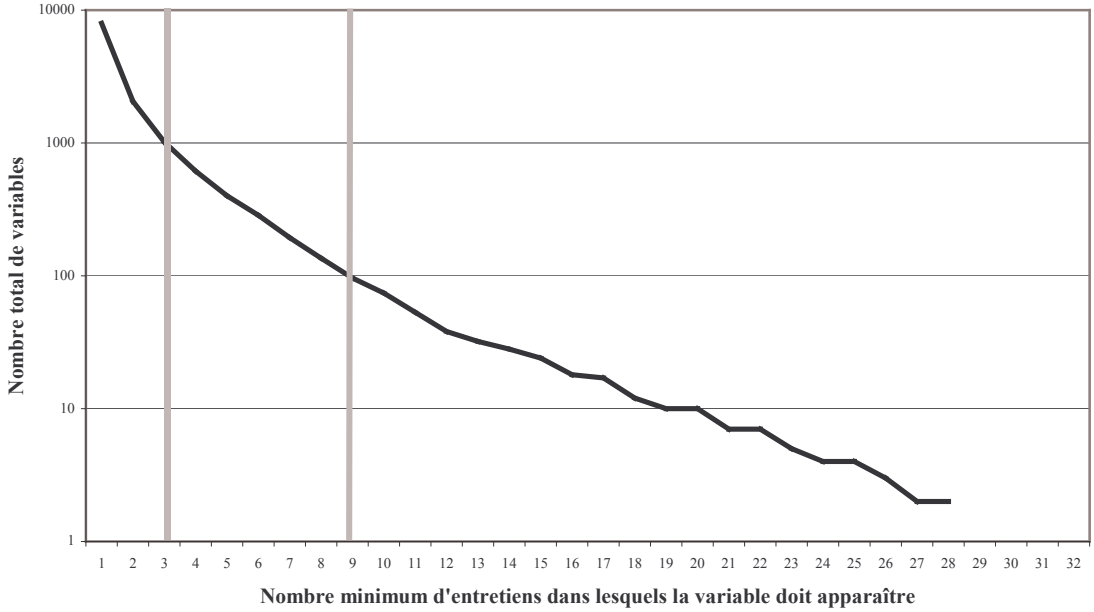

Figure 40 Evolution du nombre total de variables selon la spécificité demandée

Il faut rappeler qu'il y a au total presque 8000 variables. Une limitation aux variables apparaissant dans au moins 3 entretiens réduit le nombre total de variables à 1000. Ce nombre descend à 100 lorsque la limitation est fixée à 9 entretiens. Si on s'intéresse au limitation moyenne, c'est-à-dire 15, 16, 17 et 18 entretiens (sur 34), le nombre total de variables est, respectivement, réduit à 24, 18, 17 et 12. On se retrouve donc dans un cas, où il y a moins de variables permettant de classer que de textes à classer.
La figure suivante s'intéresse à la sélection des variables du point de vue des scores obtenus par les méthodes d'indexation de textes.

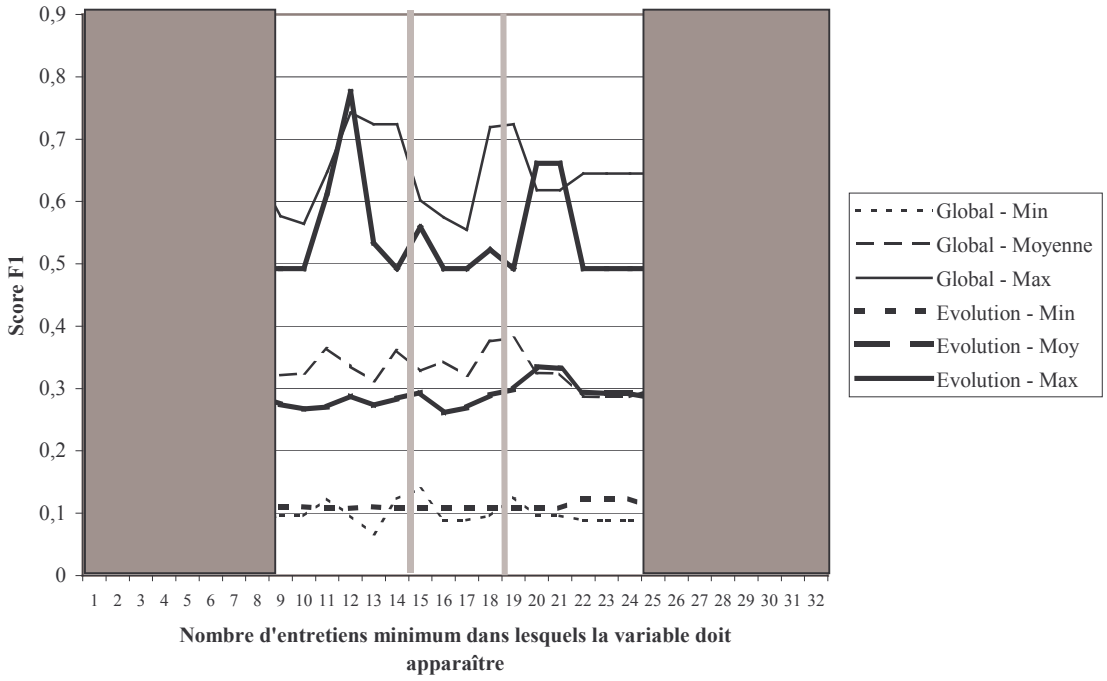

Figure 41 Evolution des scores de classification selon la spécificité demandée

Pour calculer les scores, plusieurs classifications ont été effectuées : de 2 à 15 classes. Pour chaque classification, et pour chaque méthode, un score F1 a été calculé. On ne s'intéresse ici qu'aux spécificités moyennes au sens large, c'est-à-dire aux variables qui ont une spécificité de 10 à 23. Dans la Figure 41, les quarts extrêmes des spécificités ont été grisés.
On peut, néanmoins, observer que les méthodes d'indexation globale proposent le meilleur score pour une spécificité de 1. De même, le meilleur score des méthodes d'indexation basées sur l'évolution est atteint pour une spécificité de 4. Mais, ces classifications sont trop liées aux spécificités de chaque entretien pour permettre d'évaluer les différentes relations qui existent.
Sur les spécificités moyennes, encadrées par deux barres grises, les scores chutent. Seul le score des méthodes globales remonte avec la spécificité de 18. Mais le nombre de variables (12) semble trop faible pour que la classification paraisse cohérente.
La spécificité a donc été fixée à 12. Cette spécificité est proche de la moyenne, elle permet d'effectuer une classification sur 38 variables et, sur la partie non grisée, elle offre le meilleur score pour chacune des deux méthodes.
Le Tableau 28 présente, en détail, le nombre de variables conservées pour chaque thème avec la spécificité fixée à 12. C'est donc le thème Associations qui est le moins représenté, c'est le thème qui avait le moins de variables initialement (voir Tableau 27). Il est suivi par le thème Théorie avec deux variables. Il s'agit pourtant là du thème ayant initialement le plus de variables. Il avait été remarqué dans la partie 1.3.5 que la distribution était assez courte.

Thèmes	Nb de variables représentées
Aménagement	7
Associations	1
Intérieur/Extérieur	16
Sociabilité	9
Théorie	2
Vie Familiale	3
Total	38

Tableau 28 Nombre de variables pour chaque thème avec une spécificité de 12

C'est le thème Intérieur/Extérieur qui est le plus représenté avec 16 variables, soit plus d'un tiers des variables sélectionnées.

2.1.2. La mesure de dissimilarité

Maintenant que les variables servant à la comparaison sont fixées, il faut s'intéresser aux mesures qui vont permettre le calcul des dissimilarités. Le Tableau 29 et le Tableau 30 présentent les scores obtenus selon la mesure utilisée pour calculer les dissimilarités entre les jeux de variables sociologiques. Ces scores sont, comme précédemment, calculés à partir des classification en 2 à 15 classes des entretiens.

Il peut, tout d'abord, être observé que selon l'indexation utilisée, la mesure qui obtient le meilleur score n'est pas la même. Cela montre que les méthodes d'indexation ne représentent pas les mêmes informations.

En ce qui concerne le choix de la mesure, il est évident pour l'indexation globale qui regroupe en une même mesure (Soergel) le plus grand score maximum et la plus grande moyenne.
Par contre, pour l'indexation par l'évolution, la distance de Gower propose le plus grand score maximum, mais son score moyen n'obtient que le 3e rang. La symétrie IV de Kullback-Leibler propose le plus grand score moyen, mais son score maximum n'obtient que le 3e rang.

Mesures	Min	Max	Moy	Ecart-Type
Bhattacharyya	20,69%	59,79%	35,95%	10,13%
Euclidienne Standardisée	17,56%	67,43%	35,14%	8,60%
Gower	12,32%	49,23%	32,36%	9,01%
KullBack-Leibler Symétrie IV	12,32%	56,65%	32,67%	11,45%
Mahalanobis	9,58%	49,23%	28,67%	10,28%
Russel-Rao	14,33%	54,56%	31,86%	9,78%
Soergel	12,32%	**74,23%**	**38,40%**	12,06%

Tableau 29 Détail des scores obtenus pour l'indexation globale

Mesures	Min	Max	Moy	Ecart-Type
Bhattacharyya	11,02%	49,23%	27,90%	10,95%
Euclidienne Standardisée	10,78%	49,23%	23,52%	12,03%
Gower	15,30%	**77,62%**	28,21%	13,17%
KullBack-Leibler Symétrie IV	23,77%	50,21%	**38,41%**	7,61%
Mahalanobis	12,86%	56,42%	33,32%	12,07%
Russel-Rao	11,02%	49,23%	23,57%	12,03%
Soergel	11,02%	49,23%	27,03%	11,06%

Tableau 30 Détail des scores obtenus pour l'indexation par l'évolution

Une solution est peut-être apportée par la distance de Mahalanobis qui occupe le 2e rang à la fois au niveau des scores maximum et à la fois au niveau des scores moyens. Avec ces mesures de dissimilarités ce sont donc quatre classifications qui sont proposées auxquelles vont être comparées les classifications obtenues par les méthodes d'indexation.

2.2. Les entretiens

La partie précédente a permis de fixer quatre références. Une référence a été trouvée pour la comparaison avec les méthodes d'indexation globale. Trois références différentes ont été trouvées pour la comparaison avec les méthodes d'indexation basées sur l'évolution.
Dans les figures qui suivent, la Figure 42 présente l'évolution des scores des méthodes d'indexation globale lorsqu'elles sont comparées à la référence obtenue avec la distance de Soergel. La Figure 43, la Figure 44 et la Figure 45 présentent l'évolution des scores des méthodes d'indexation basées sur l'évolution lorsqu'elles sont comparées à la référence obtenue avec, respectivement, le distance de Gower, la symétrie IV de Kullback-Leibler et la distance de Mahalanobis.
Le protocole de test est similaire à celui appliqué précédemment. De la classification par arbre, plusieurs classification sont obtenues. La comparaison des indexations est donc faite selon qu'il soit formé 2 à 15 classes.

Pour l'indexation globale, il a été gardé les trois méthodes : Structure, Vecteurs et Villes. Le libellé Villes désigne la méthode d'indexation par Automate Peintre, en référence à la formation des images. Il faut rappeler qu'il a été choisi de faire fonctionner les méthodes Structure et Vecteurs sur des lemmes graphiques, avec une pondération par recalage des valeurs et que les distances qui leur sont associées sont, respectivement, la distance de Gower et la distance Euclidienne Standardisée. Pour l'indexation Villes, il a été choisi une application sur les 2-grammes de caractères, aucune pondération et la distance de Gower.
Pour l'indexation sur l'évolution, les méthodes d'indexation précédentes sont appliquées sur 30 et 40 parties de textes.

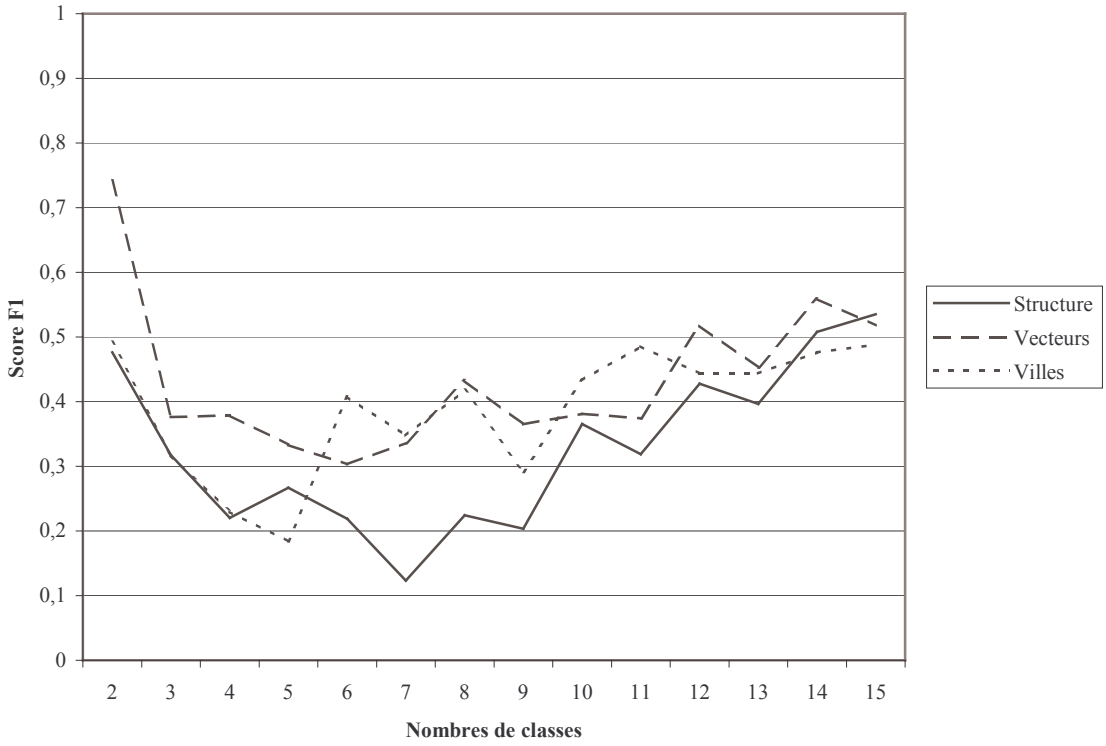

Figure 42 Evolution des scores selon le nombre de classes des méthodes d'indexation globale (référence calculée avec la distance de Soergel)

La méthode d'indexation globale présente d'assez bons résultats pour une classification en deux classes. Mais jusqu'à 12 classes, les scores restent sous la barre des 0,5 et même régulièrement sous la barre des 0,4.
Il faut noter que les meilleurs scores sont, généralement, obtenus par la méthodes des vecteurs qui reste donc, incontestablement, la meilleure représentation globale d'un texte. La méthode sur les informations structurelles proposent, sans conteste, les pires scores. Cela confirme les suppositions qui avaient été faite sur le manque complet de structure des entretiens qui sont des discours oraux retranscrits. Enfin, il peut être noté, les scores intéressants obtenus par la méthode Villes qui propose, parfois, les meilleurs scores, mais qui reste en-dessous de la barre de 0,5.

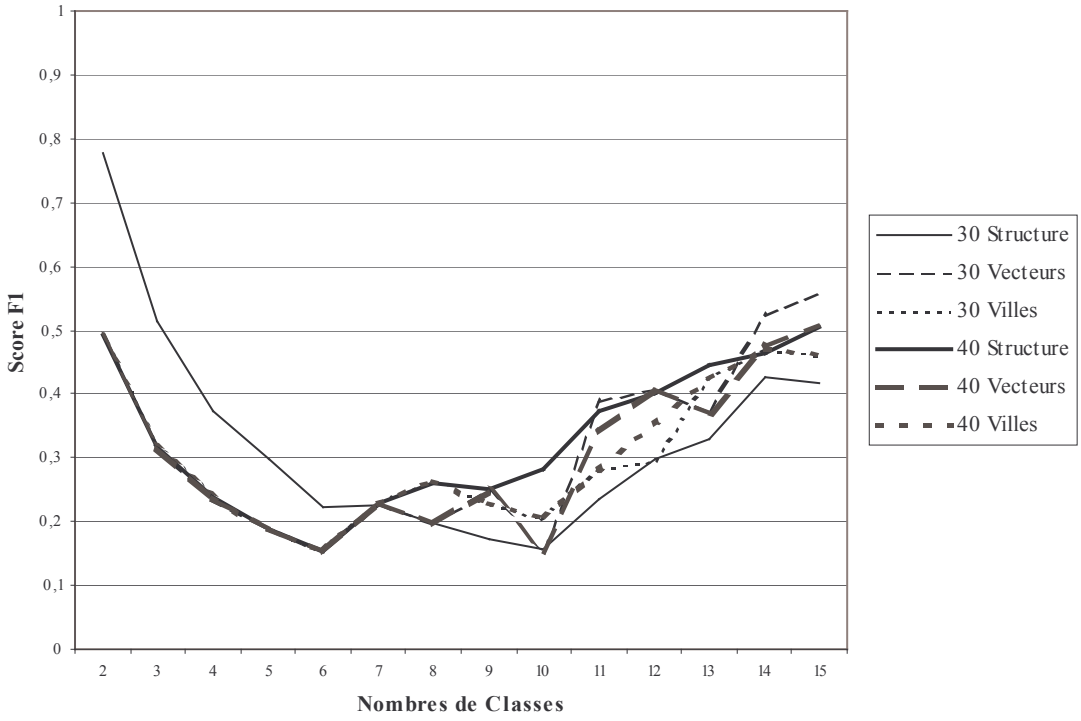

Figure 43 Evolution des scores selon le nombre de classes des méthodes d'indexation sur l'évolution (référence calculée avec la distance de Gower)

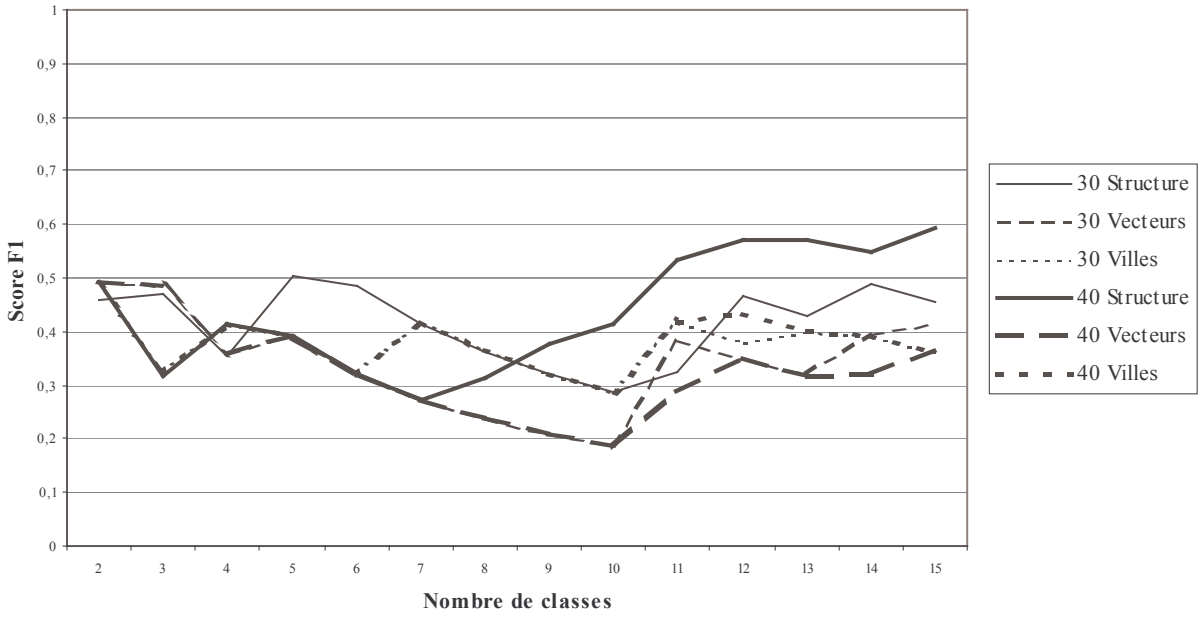

Figure 44 Evolution des scores selon le nombre de classes des méthodes d'indexation sur l'évolution (référence calculée avec la symétrie IV de Kullback-Leibler)

La cuvette entre les scores des classifications en 2 classes et les classifications en 15 classes est, aussi, observable dans la comparaison des méthodes d'indexation basées sur l'évolution et de la référence calculée par la distance de Gower. Cette cuvette est, dans ce cas, bien plus

creusée, puisque, pour 5 à 10 classes, les scores ne dépassent pas les 0,3. Cet effet cuvette était marqué dans le Tableau 30 par le fait que cette comparaison obtenait le plus grand écart-type.
Du point de vue du nombre de mesures d'évolution, c'est le 30 qui offre les meilleurs résultats avec presque 0,8 en utilisant la méthode Structure pour une classification en 2 classes et 0,55 en utilisant la méthode Vecteurs pour une classification en 15 classes.

Dans la comparaison des méthodes d'indexation basées sur l'évolution et de la référence calculée par la symétrie IV de Kullback-Leibler, l'effet cuvette est très peu visible. Dans le Tableau 30, l'écart-type était le plus faible. Cependant, seule l'indexation utilisant la méthode Structure dépasse les 0,5 lorsqu'elle utilise 40 mesures dévolution. Ce sont, d'ailleurs, les deux utilisations de la méthode Structure, pour 30 et 40 mesures d'évolution, qui ont le meilleur score moyen avec, respectivement, 0,41 et 0,43.

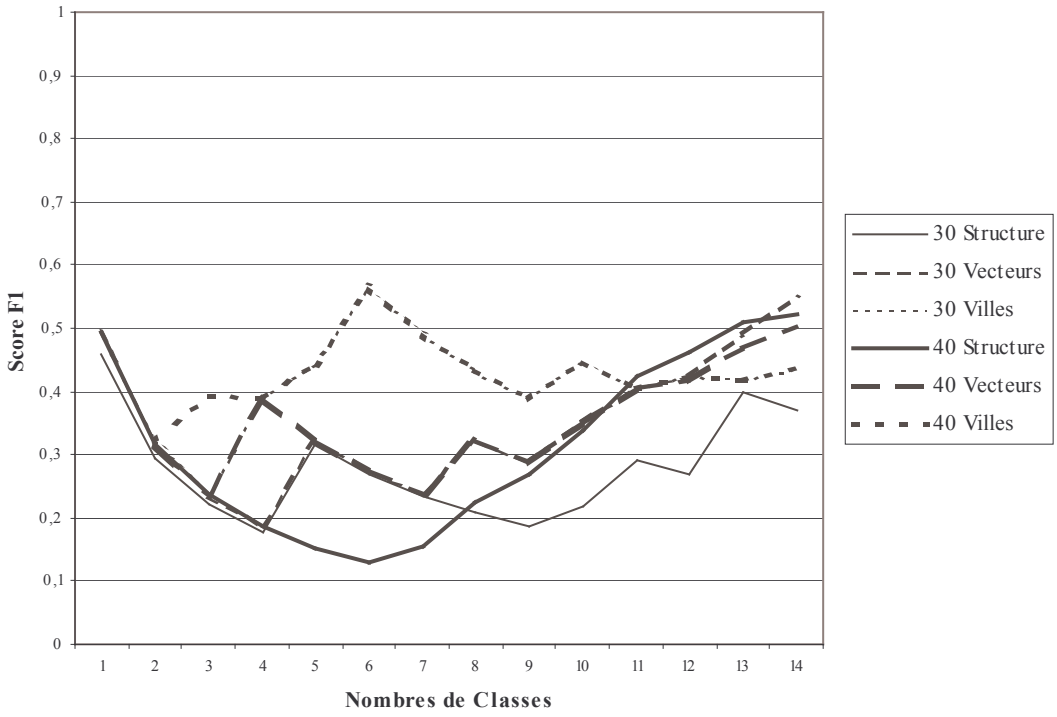

Figure 45 Evolution des scores selon le nombre de classes des méthodes d'indexation sur l'évolution (référence calculée avec la distance de Mahalanobis)

La Figure 45 montre une évolution complètement différente. Presque toutes les méthodes utilisées forment à nouveau la cuvette. Mais l'utilisation de la méthode Villes, pour un découpage de 30 ou de 40, est la seule, de toutes les méthodes de l'expérimentation, qui obtient un score supérieur à 0,4 sur pour les formations de classe de 4 à 10.
De plus, il faut signaler que les formations en 5 à 7 proposent des résultats supérieurs à 0,5.

2.3. Bilan des Expérimentations

Les expérimentations sur le corpus réel d'application ont permis de poser plusieurs problèmes, tant sur la sélection des variables que sur la distance à utiliser pour la classification des index créée manuellement par les sociologues.
La comparaison s'est faite à partir de plusieurs classifications contenant 2 à 15 classes. Ce découpage paraît excessif pour un ensemble initial de 34 entretiens, mais il permet de mieux évaluer l'évolution des erreurs. Cela permet surtout d'évaluer les différentes indexations sur leur comportement global (2 classes) et sur leur comportement au niveau des proximités (15 classes).

Les expérimentations ont montré que les résultats sont très variables selon la distance utilisée sur l'index sociologique. La plupart des méthodes ont montré une évolution en cuvette. Cela implique de mauvais scores de classification pour les formations ayant 4 à 10 classes.
Les méthodes d'indexation sur l'évolution comparées à la référence calculée avec la distance de Gower montrent, dans ce sens, le caractère le plus extrême. C'est, en effet, dans cette configuration que se trouve le meilleur score obtenu lors des expérimentations. Mais c'est aussi dans cette configuration que les scores maximum sont les plus faibles. Ces méthodes se caractérisent donc par un écart-type très important au niveau des scores.

On peut ainsi définir une bonne méthode comme étant une méthode proposant les meilleurs scores moyens, mais le plus faible écart-type. Ces deux caractéristiques sont, en effet, signe de stabilité. Trois méthodes suivent ces prérogatives. Il s'agit pour les méthodes d'indexation globale de la méthode des Vecteurs (moyenne 0,433 / écart-type 0,118) et pour les méthodes d'indexation sur l'évolution des méthodes utilisant la méthodes Structure (moyenne 0,438 / écart-type 0,107) et celle utilisant la méthode Villes (moyenne 0,431 / écart-type 0,057).
La méthode d'indexation globale est la seule des trois méthodes suivant une évolution en cuvette. La méthode d'indexation basée sur l'évolution utilisant la méthode Structure montre l'essentiel de ses performances sur les formations ayant plus de 10 classes. Enfin, la méthode d'indexation basée sur l'évolution utilisant la méthodes Villes est la méthode la plus stable. Ces scores sont majoritairement supérieurs ou égaux à 0,4. C'est surtout la seule méthode qui propose des scores supérieurs à 0,5 pour les formations en 5 à 7 classes qui est le nombre de classes le plus vraisemblable pour 34 individus.

Il faut, aussi, signaler que la référence utilisée pour la comparaison avec cette dernière méthode est calculée avec la distance de Mahalanobis. Cette distance avait montré le caractère le plus régulier en prenant, chaque fois, la seconde place dans les scores du Tableau 30.

Ainsi, l'évolution trouve un réel intérêt dans le traitement des entretiens sociologiques. En utilisant, localement, une méthode qui permet de représenter globalement l'évolution d'un texte, les résultats obtenus sont, d'un point de vue relatif, les plus stables que l'on peut obtenir.

Conclusion et perspectives

1. Conclusion

Ce travail de thèse avait pour objectif de fournir un outil d'analyse de données textuelles pour les sciences sociales. Ce travail s'est principalement intéressé à la classification des entretiens oraux retranscrits. Ainsi que cela a été montré à de multiples reprises, cet objectif se heurte fortement aux contraintes du langage oral. Ce travail propose donc de représenter les textes par leur structure. Pour y parvenir, plusieurs méthodes ont été créées. Ces méthodes tentent d'extraire et de représenter l'organisation des discours et l'évolution des textes.

Après avoir décrit quelques logiciels présents sur le marché, le chapitre 1 présente une méthode d'analyse manuelle créée par Sylvette Denèfle, la sociologue qui co-dirige ce travail. Une analyse montre que cette méthode peut être informatisée et que cela consiste à effectuer plusieurs traitements successifs.
Le premier traitement consiste à effectuer une résolution des anaphores et des ellipses. Dans le cas d'entretiens sociologiques oraux retranscrits, le contexte est permanent. Ce traitement permet de combler les lacunes en information de certaines parties. En d'autres termes, ce traitement permet de donner plus de corps aux textes.
Le deuxième traitement est une segmentation et une classification thématique. Les thèmes ont été fixés par les sociologues avant que la série d'entretiens ne soit effectuée. Ce traitement permet de distinguer les thèmes les uns par rapport aux autres. La difficulté de ce traitement est le discours employé. En effet, même si une résolution des anaphores et des ellipses permet de ramener les textes à la qualité du langage écrit, les thèmes abordés par les sociologues ne demandent pas l'utilisation d'un vocabulaire spécifique. Au contraire, la majorité des mots utilisés sont des termes communs et qui peuvent être utilisés dans plusieurs situations.
Le troisième traitement est une extraction d'informations. Cette extraction est exploratoire, c'est-à-dire qu'avant ce traitement l'ensemble des informations à extraire n'est pas défini. Des hypothèses permettent d'énoncer des informations susceptibles d'apporter une solution au problème sociologique, mais ces hypothèses peuvent ne pas se retrouver dans les entretiens et sont, naturellement, incomplètes. Il paraît peu vraisemblable que toutes les hypothèses puissent être formulées avant le traitement des entretiens.
Le quatrième traitement consiste à regrouper les informations extraites qui portent le même sens. Du point de vue traitement informatique, ce traitement s'apparente à la comparaison de requêtes.
Cette méthode peut donc être informatisée mais nécessite la mise en place de nombreux outils (dictionnaires, lexiques, …) et la résolution de nombreux problèmes linguistiques. Une étude a donc été menée sur les méthodes informatisées existantes.
Le traitement des questions ouvertes consiste le plus souvent à ne traiter qu'une unique question à l'aide d'un logiciel d'analyse de textes. Cette problématique diffère donc de celle du travail.
Du point de vue de la classification et de l'indexation de textes, il apparaît que l'étape primordiale consiste à indexer les textes, c'est-à-dire à transformer les données textuelles brutes en données le plus souvent numériques. Les méthodes existantes travaillent, en grande

majorité, sur des vecteurs de fréquences. Une méthode [DEV 00] propose une représentation de la structure, car la mise en page serait propre à chacun. Ces représentations ne semblent pas convenir, dans le cas du traitement d'entretiens oraux retranscrits. En effet, une méthode basée sur la mise en page de la retranscription de discours semble complètement improbable. Et, comme les discours se caractérisent par une forte utilisation des mots communs (les mots les plus fréquents) et des mots uniques (qui n'apparaissent qu'une fois), la représentation par vecteurs ne semble pas la plus appropriée pour ce genre de traitement.
Ce travail propose donc de représenter les textes par leur organisation et par leur évolution. Cette représentation est la première étape d'une classification en trois étapes. La deuxième étape consiste à calculer les distances et la troisième étape consiste à classer les textes à l'aide d'une classification par arbre. Les étapes sont étudiées dans l'ordre inverse du traitement, c'est-à-dire de la dernière à la première étape. Un chapitre est consacré à chaque étape.

Le chapitre 2 est donc consacré à la troisième étape, c'est-à-dire la classification par arbre. Ce chapitre a étudié sept méthodes différentes de classification. Les quatre premières méthodes, Liens Simples, Liens Complets, WPGMA et UPGMA, rassemblent les individus ayant la distance la plus faible. Chacune de ces distances se différencie des autres par le re-calcul des distances qui est effectué. Ces méthodes nécessitent, pour donner une bonne classification, que la matrice des distances donnée en entrée soit ultramétrique. Or les matrices de distances ne vérifient pas toujours cette propriété, ce qui implique des erreurs dans la représentation des distances.
Les trois dernières méthodes de classification, NJ, ADDTREE et la méthode des Groupements ne nécessitent pas que la matrice soit ultramétrique. Les méthodes NJ et ADDTREE sont utilisées en biologie pour l'analyse de données phylogéniques. La méthode des Groupements a été utilisée dans le cadre d'analyse de textes [BAR 98]. Ces trois méthodes fonctionnent avec la notion de voisinage et se différencient, essentiellement, dans la méthode de choix des individus à rassembler. La notion de voisinage permet de réduire les erreurs de représentation des distances. Des expérimentations montrent, cependant, que la méthode des Groupements fournit une approximation de la dissimilarité initiale moins bonne que celle obtenue à l'aide des deux autres approches. De plus, sa complexité empêche un passage à l'échelle. Les méthodes NJ et ADDTREE donnent des résultats similaires. Le choix s'est donc fait au niveau de la complexité des algorithmes. Comme ADDTREE a une complexité de $O(n^4)$, c'est la méthode NJ, avec une complexité de $O(n^3)$, qui a été choisie.

La chapitre 3 est consacré à l'étude de normalisation des données et à l'étude de mesures de dissimilarités. La deuxième étape consiste, en effet, à calculer la dissimilarité entre des représentations numériques afin de fournir une matrice à la méthode de classification vue précédemment. Il a été listé une cinquantaine de mesures utilisées habituellement pour le traitement de données binaires et le traitement de données réelles. Les mesures utilisées pour le traitement de données binaires peuvent être formulées de façon à traiter les jeux de données dont les valeurs sont comprises entre 0 et 1. Ce chapitre présente donc deux méthodes de normalisation des données : la normalisation par la somme et la normalisation par la valeur maximum. Plusieurs expérimentations confirment l'importance du lien entre une méthode de normalisation et une mesure de dissimilarité. Ces expérimentations montrent surtout que les résultats de classification varient selon la mesure de dissimilarité et la normalisation utilisées. Les résultats varient autant que les formulations. De manière expérimentale, la centaine de

combinaisons a été étudiée sur la qualité de classification de chaque combinaison. La qualité a été calculée à partir du rappel et de la précision. Ainsi huit couples normalisation-mesure ont été conservés. Il s'agit des couples : Somme – Bhattacharyya, Somme – Gower, Somme – Kullback-Leibler Symétrie IV, Somme – Mahalanobis, Somme – Soergel, Maximum – Euclidienne Standardisée et Maximum – Russel-Rao. Chacun de ces couples présentent des spécificités et obtient des résultats qui dépendent du jeu de donnes. C'est la raison pour laquelle le choix final est laissé au niveau de la méthode d'indexation.

Le chapitre 4 s'intéresse à la première étape : l'indexation des textes. Tout d'abord, les pré-traitements et les post-traitements sont étudiés, puis les méthodes d'indexation sont présentées.
Du point de vue des pré-traitements, il est, dans un premier temps, proposé de normaliser les données textuelles brutes. Cette normalisation consiste à mettre en minuscule tous les caractères, à les désaccentuer et à réduire l'alphabet à 27 caractères (26 lettres et 1 caractère spécial qui regroupe tous les caractères non-lettre). Cette normalisation est effectuée, par défaut, pour une majorité de méthodes d'indexation. Seule une méthode, qui tente de représenter la structure des textes, nécessite des informations particulières portées par les données brutes.
Dans un second temps, il est étudié les diverses unités textuelles pouvant être extraites des textes. Ce chapitre présente trois sortes d'unités textuelles : les mots, les lemmes et les n-grammes. Une étude est faite sur le rapport entre le nombre d'unités textuelles maximum dans le corpus (texte le plus long) et le nombre d'unités textuelles différentes pour chaque type d'unité. Cette étude montre que l'utilisation de n-grammes d'une longueur supérieure ou égale à 5 n'a aucun intérêt du point de vue de la diversité de l'information. Les unités textuelles conservées sont donc les mots, les lemmes et les n-grammes ayant une longueur comprise entre 1 et 4.
Du point de vue des post-traitements, deux types de pondération sont étudiées : le tf*idf et le recalage. Les expérimentations montrent que, dans la plupart des cas, une non-pondération ou une pondération par recalage est préférable à une pondération par tf*idf.

En ce qui concerne l'indexation, trois groupes de méthodes sont étudiés. Le premier groupe est constitué de méthodes, présentées précédemment, qui consistent à représenter les textes par des vecteurs de fréquences, par la distribution de leurs unités textuelles avec une représentation vectorielle des courbes de Zipf et par la structure des textes (de l'ordre de la mise en page).
Le second groupe est constitué de méthodes d'indexation développées lors de cette thèse. Ces méthodes proposent de représenter les textes par l'organisation du discours. Une telle représentation permet en effet de représenter la fréquence d'utilisation des unités textuelles, leur distribution et une structure des textes bien plus complète que la mise en forme.
La première de ces méthodes, [MAR 04-1], transforme chaque texte en une image de taille fixe. Les images générées sont de type fractal. L'indexation consiste donc à évaluer l'auto-similarité de l'image, c'est-à-dire si la structure globale de l'image se retrouve localement. Pour qu'une comparaison puisse être effectuée entre les images, un patron de création est fixé pour chaque langue traitée. Cela n'a pas d'importance dans le cas d'une application sur un corpus d'entretiens sociologiques, mais cela devient une contrainte pour un traitement plus ouvert.

La deuxième méthode développée lors de ce travail est un automate qui crée un signal. Cet automate a l'avantage de ne pas nécessiter d'étude des textes au préalable. Les transitions sont affectées, si nécessaire, aux unités textuelles selon la position de l'automate. Dans ce cas, chaque texte est représenté par l'histogramme vertical du signal. Cette méthode est donc très proche du texte. Elle est si proche du texte qu'elle en manque de robustesse. L'insertion d'une unité textuelle peut, en effet, changer radicalement la représentation.
La troisième méthode du groupe, [MAR 05-2], est un automate qui crée une image. L'affectation des transitions nécessite une étude statistique au préalable des textes. Plusieurs propositions de transitions ont été faites en suivant la loi du moindre effort. Les expérimentations montrent que les transitions les plus simples permettent de créer la meilleure représentation. Chaque représentation est construite à la manière d'une ville. L'indexation suit cette idée en proposant d'évaluer l'extension de la ville. Sur des textes courts, cette méthode montre des résultats comparables aux méthodes traditionnelles.
Enfin, le dernier groupe de méthode est constitué d'une « méta-méthode ». Cette méthode propose de représenter l'évolution des textes. Pour chaque texte, un corpus est constitué à partir des divers stades d'avancement du texte. C'est-à-dire que chaque texte est segmenté en plusieurs parties de taille égale et que le premier texte d'évolution est constitué de la première partie du texte, le second texte d'évolution est constitué des deux premières parties, etc. Chaque texte d'évolution est indexé et les distances successives entre les index d'évolution constituent l'index du texte original. Les expérimentations montrent que cette méthode fonctionne mieux lorsque la méthode d'indexation utilisée localement est adaptée à la représentation de l'évolution. C'est, en effet, l'automate peintre qui obtient les meilleurs résultats de classification sur des textes courts.
Ce chapitre montre donc que l'indexation par l'évolution est une solution au problème de la représentation des textes.

Le chapitre 5 est consacré à l'expérimentation sur un corpus du domaine réel d'application. Il s'agit, ici, d'une série d'entretiens menés auprès d'habitants des unités d'habitation imaginées par Le Corbusier. Dans cette série d'entretiens, six thèmes sont abordés : Aménagement, Association, Intérieur/Extérieur, Sociabilité, Théorie et Vie Familiale. Les sociologues en ont extrait presque 8000 variables sous la forme d'un index binaire où la valeur indique si la variable est présente dans l'entretien. Afin de comparer l'index obtenu par les sociologues et l'index obtenu par les méthodes d'indexation présentées précédemment, une sélection des variables a été effectuée et une étude des mesures de dissimilarités a été menée sur l'index sociologique. Il en ressort quatre classifications de référence : une classification permet de comparer avec les méthodes globales (méthodes traditionnelles et méthodes sur l'organisation du discours) et trois références différentes permettent de comparer avec les méthodes d'évolution. Sur un tel corpus, l'évolution prend toute son importance et propose la résultat le plus stable.

En conclusion, ce travail de thèse a donc permis d'aboutir à une solution particulièrement adaptée à la classification d'entretiens oraux retranscrits. Il s'agit d'une méthode qui mesure l'évolution textuelle. Si elle est associée localement à une méthode d'indexation adaptée, elle offre un résultat stable et meilleur que les autres pour les nombres de classes les plus vraisemblables. La méthode d'indexation la mieux adaptée, à l'heure actuelle, est une

méthode développée pour ce projet. Cette méthode capture et représente l'organisation du discours à l'aide d'un automate peintre. Au-delà de ces résultats, il semblerait que le raisonnement d'un texte puisse être représenté. Il s'agit d'une idée nouvelle qui va à l'encontre de toutes les méthodes qui étudient le contenu.
Cependant, le manque de corpus d'application empêche de généraliser. Les expérimentations ont montré que la longueur des textes était une limite d'application. En effet, l'évolution de textes trop court ne peut pas être réellement représentée. En-dehors de cette limite, l'évolution textuelle peut être appliquée pour l'étude d'œuvres littéraires et pour une distinction, par la structure, de textes de provenances multiples [MAR 05-2].

2. Perspectives

Comme tout travail de thèse, celui-ci n'est pas une fin en soit mais plutôt la base de travaux futurs. Cette partie énumère donc un certain nombre de travaux qui pourraient être conduits sur chacune des parties de cette thèse pour aboutir à de nouveaux résultats.

En ce qui concerne la classification, elle est pour l'instant effectuée en deux étapes. Tout d'abord un calcul des distances, puis une classification par arbre. Pour cette thèse, on s'est contenté d'étudier l'existant à partir de choix faits a priori. De plus amples études pourraient être menées sur les mesures de dissimilarité. L'influence de chaque paramètre des mesures pourrait ainsi être identifié et une mesure propre à l'analyse de données textuelles pourrait être créée.
De plus, d'autres classifieurs pourraient être étudiés. La classification pourrait être interactive [MON 02]. Dans ce cas, l'utilisateur ne se fait pas une idée à partir des distances mais à partir du comportement d'un texte par rapport à un autre. Il pourrait être étudié des méthodes nécessitant un apprentissage. Lors de cette thèse, il a été choisi de proposer une classification non supervisée, c'est-à-dire sans l'intervention à quelques moments que ce soit de l'utilisateur. Et dans le chapitre 5, il a été vu la quantité d'entretiens à apprendre pour couvrir l'ensemble des variables trouvées manuellement. Or dans ce même chapitre 5, une étude montre qu'il est préférable de ne garder que les 38 variables les plus fréquentes. Une étude au préalable des entretiens pourrait sélectionner les textes à apprendre, puis une sélection de caractéristiques pourrait réduire le nombre de variables trouvées manuellement. Enfin les textes et les variables conservées pourraient constituer une base d'apprentissage pour un classifieur du type réseau de neurones, Chaînes de Markov Cachées, SVM, …

En ce qui concerne l'indexation, on peut, tout d'abord, s'intéresser aux approfondissements possibles des méthodes développées. En effet, la méthode qui transforme un texte en image semble apporter une solution du point de vue de la représentation. Mais la dimension de masse ne semble pas assez précise pour mettre en valeur l'ensemble des informations portées par cette représentation. De plus, il pourrait être imaginé, pour une ouverture à d'autres corpus, une méthode permettant de comparer deux images de tailles différentes. La taille de l'image serait donc adaptée à la taille du texte.
Pour l'automate 1D, une nouvelle méthode d'évaluation pourrait être créée afin de rendre la représentation plus robuste. Une étude pourrait, par exemple, être faite sur l'amplitude du signal. Aucune méthode issue du traitement du signal n'a été testée alors que de nombreuses

méthodes de classification d'échantillons sonores existent. De même, des études pourraient être menées sur une transformée du signal ou sur une version filtrée du signal.
Pour l'automate peintre, la méthode d'évaluation ne semble pas mettre en valeur toutes les informations fournies par la représentation. En effet, tout d'abord, chaque couronne est prise dans sa globalité alors que dans le cas d'une ville on distingue les différents quartiers. Il pourrait donc être considéré différemment les quartiers calmes pour lesquels les constructions ne changent pas et les quartiers en mouvement pour lesquels constructions évoluent en permanence. Il faudrait pour cela définir la notion de quartier, de calme et de mouvement. Puis, l'automate est initialisé et il est créé de façon à pouvoir éclaircir autant qu'il peut obscurcir. C'est-à-dire que la ville évolue tant en profondeur (fondations, canalisations, …) qu'en hauteur (habitations, buildings, …). Or l'évaluation ne tient uniquement compte de la valeur absolue de la différence de hauteur. Une meilleure évaluation des points de la carte pourrait être imaginée.
Enfin, pour la méthode basée sur l'évolution, le découpage pourrait être étudié de façon plus précise. Il pourrait par exemple être étudié un découpage en 2^n. Le coefficient n serait propre à chaque texte et dépendrait de sa longueur. Deux textes de longueurs différentes pourraient ainsi être comparés en ramenant le plus grand index à la taille du plus petit. Il pourrait aussi être étudié un découpage effectué manuellement. Les sociologues pourraient ainsi porter des marques dans le texte qui serviraient de repère au découpage. Une autre solution consisterait à découper en fonction de l'apparition des mots les plus fréquents. Enfin cette méthode pourrait être testée sur des segments non continu de texte, par exemple uniquement sur les parties de textes abordant un thème précis.
Il pourrait aussi être créé de nouvelles méthodes d'indexation. Il pourrait donc être étudié la façon dont les modèles de Markov ou les réseaux de neurones pourraient être adaptés à la problématique posée. D'une manière plus générale, les textes pourraient être représentés à l'aide de Graphical Models.

D'un point de vue purement utilisateur, la classification pourrait devenir plus explicite si à chaque nœud de l'arbre il était spécifié les caractéristiques textuelles propres à l'ensemble des éléments raccordés à ce nœud. Par exemple, dans le cas d'une indexation sur l'évolution, il faudrait, pour chaque texte, repérer les parties caractéristiques de son évolution. Puis, il faudrait repérer les unités textuelles caractéristiques. Enfin, il faudrait extraire les expressions comprenant ces unités textuelles. Le fruit de ce travail pourrait ainsi être utilisé pour développer une solution logicielle complète d'analyse de données textuelles pour les sciences sociales et pour les études de textes en général.

Références

Les références sont organisées en trois parties.
Les textes bibliographiques sont identifiés par les trois premières lettres du nom du premier auteur, suivi de l'année, suivi occasionnellement d'un code.
Les logiciels sont identifiés par les lettres LOG et par les trois premières lettres du logiciel.
Les sites Internet sont identifiés par les lettres WWW et par une chaîne de caractères.

1. Bibliographie

[ACH 91] Achard P., Une Approche Discursive des Questionnaires : l'Exemple d'une Enquête pendant la Guerre d'Algérie, Langage et Société, vol. 55, pp. 5-40, 1991.

[AHO 86] Aho A., Sethi R. et Ullman J., Compilers, Addison Wesley, 1986.

[AMA 00] Amaral Rui et Trancoso Isabel, Topic Detection in Read Documents, 4th European Conference on Research and Advanced Technology for Digital Librairies, pp. 315-318, ISBN 3-540-41023-6, 2000.

[ARA 00-1] Aragüés Peleato Ramón, Chappelier Jean-Cédric et Rajman Martin, Using Information Extraction to Classify Newspapes Advertisments, 5èmes Journées internationales d'Analyse statistique des Données Textuelles, Lausanne - Suisse, 9 – 11 Mars 2000.

[ARA 00-2] Aragüés Peleato Ramón, Chappelier Jean-Cédric et Rajman Martin, Automated Information Extraction out of Classified Advertisements, Computer Science, vol. 1959, pp. 203 – 214, ISBN 3-540-41943-8, 2000.

[BAR 88] Barthélémy Jean-Pierre et Guénoche Alain, Les Arbres et les Représentations des Proximités, Masson, 1988.

[BAR 98] Barthélémy Jean-Pierre et Luong Xuan, Représenter les données textuelles par les arbres, 4èmes Journées internationales d'Analyse statistique des Données Textuelles, Nice - France, 1998.

[BAV 02] Bavaud François et Xanthos Aris, Thermodynamique et Statistique Textuelle : Concepts et Illustrations, , 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[BEE 99] Beeferman Doug, Berger Adam et Lafferty John, Statistical Models for Text Segmentation, Machine Learning, Special Issue on Natural Language Learning, vol. 34/1-3, pp. 177-210, ISBN 0885-6125, 1999.

[BEC 98] Becue-Bertaut Monica, Analyse Simultanée de Plusieurs Question Ouvertes, 4èmes Journées internationales d'Analyse statistique des Données Textuelles, Nice - France, 1998.

[BEC 00] Becue-Bertaut Monica et Pagès Jérôme, Analyse Factorielle Multiple Intra-Tableaux. Application à l'Analyse Simultanée de Plusieurs Questions Ouvertes, 5èmes Journées internationales d'Analyse statistique des Données Textuelles, Lausanne - Suisse, 9 – 11 Mars 2000.

[BEC 02] Becue-Bertaut Monica et Pagès Jérôme, Analyse Conjointe de Questions Ouvertes et de Questions Fermées : Méthodologie, Exemple, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[BEN 03] Beney J.G et Koster C. H. A., Classification Supervisée de Brevet : d'un Jeu d'Essai au Cas Réel, INFORSID 03, Workshop sur la Recherche d'Information, Nancy – France, 3-6 Juin 2003.

[BEU 02] Beust Pierre, Un Outil de Coloriage de Corpus pour la Représentation des Thèmes, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[BIG 98] Bigi Brigitte, De Mori Renato, El-Bèze Marc et Spriet Thierry, Combinaison de Modèles de Langages pour l'Identification de Thèmes, 22èmes Journées d'Etude sur la Parole, Martigny – Suisse, 15-19 Juin 1998.

[BIG 00] Bigi Brigitte, De Mori Renato et Spriet Thierry, Reconnaissance Thématique à partir de Textes Dictés et Adaptation Dynamique de Modèles de Langage Thématiques, 23èmes Journées d'Etude sur la Parole , Grenoble – France, pp. 301-304, Mai 2000.

[BIG 01-1] Bigi Brigitte, Brun Armelle, Smaïli Kamel, Haton Jean-Paul et Zitouni Imed, Dynamic Topic Identification : Towards Combination of Methods, Recent Advances in Natural Language Processing, Tzigov Chark – Bulgarie, pp. 255-257, 5-7 Septembre 2001.

[BIG 01-2] Bigi Brigitte, Brun Armelle, Smaïli Kamel et Haton Jean-Paul, A Hierarchical Approach for Topic Identification, International Workshop Speech and Computer, SPECOM'01, Moscou – Russie, 29-31 Octobre 2001.

[BIG 01-3] Bigi Brigitte, Brun Armelle, Smaïli Kamel, Haton Jean-Paul et Zitouni Imed, A Comparative Study of Topic Identification on newspaper and e-mail, 8th International Symposium on String Processing and Information Retrieval, Laguna de San Rafael – Chili, pp. 238-241, 13-15 Novembre 2001.

[BIG 02] Bigi Brigitte, Smaïli Kamel, Identification Thématique Hiérarchique : Application aux forums de discussions, TALN 2002, Nancy – France, 24-27 Juin 2002.

[BIS 02-1] Biskri Ismaïl et Meunier Jean-Guy, L'analyse de l'Information Multidimensionnelle au Moyen des N-Grams de Caractères, 9e Journées Francophones d'Informatique Médicale, Québec – Canada, 6 – 7 mai 2002.

[BIS 02-2] Biskri Ismaïl, Amar Bensaber Boucif et Hajji Wadii, Les N-Grams de Caractères au Service de la Recherche Documentaire Médicale dans les Bases de Données Textuelles Multilingues, 9e Journées Francophones d'Informatique Médicale, Québec – Canada, 6 – 7 mai 2002.

[BIS 00] Bisson Gilles, Chapitre XX – La Similarité : Une Notion Symbolique/Numérique, Apprentissage Symbolique – Numérique, Tome 2, pp. 169 – 201, Moulet M. et Brito P., CEPADUES, 2000.

[BLE 01] Blei David M. et Moreno Pedro J., Topic Segmentation in an Aspect Hidden Markov Model, SIGIR'01, New Orleans - USA, 9-12 Septembre 2001.

[BOU 02-1] Boufaden Narjès, Lapalme Guy et Bengio Yoshua, Découpage Thématique des Conversations : un Outil d'Aide à l'Extraction, TALN 2002, Nancy - France, 24-27 Juin 2002.

[BOU 02-2] Boufaden Narjès, Lapalme Guy et Bengio Yoshua, Segmentation en Thèmes de Conversations Téléphoniques : Traitement en Amont pour l'Extraction d'Information, TALN 2002, Nancy – France, 24-27 Juin 2002.

[BOU 05] Boufaden Narjès et Lapalme Guy, Apprentissage de relations prédicats-arguments pour l'extraction d'informations à partir de textes conversationnels, TALN 2005, Dourdan - France, 6 – 10 Juin 2005.

[BOU 05-2] Bourdan S., Cahier d'Accompagnement au Logiciel d'Analyse de Données Qualitatives QSR NVivo, Université de Sherbrooke, 2005.

[BOU 62] Bourdieu Pierre, Célibat et condition paysanne, Etudes rurales, vol. 5 - 6, pp. 32 – 136, Avril – Septembre 1962.

[BOU 80] Bourdieu Pierre, Questions de sociologie, Minuit, ISBN : 2-7073-1825-6, 277 pages, 1980.

[BOU 84] Bourdieu Pierre, La distinction : Critique sociale du jugement, Minuit, ISBN : 2-7073-0275-9, 672 pages, 1984.

[BOU 92] Bourdieu Pierre et Wacquant Loïc, Réponses pour une anthropologie réflexive, Seuil, ISBN : 2-02-014675-4, 267 pages, 1992.

[BRO 99] Brouard Thierry, Algorithmes Hybrides d'Apprentissage de Chaînes de Markov Cachées : Conception et Application à la Reconnaissance des Formes, Rapport de Thèse de Doctorat, Université François-Rabelais de Tours, 219 pages, 1999.

[BRU 02] Brun Armelle, Smaili Kamel et Haton Jean-Paul, WSIM : une Méthode de Détection de Thème fondée sur la Similarité entre Mots, TALN 2002, Nancy – France, 24-27 Juin 2002.

[BRU 03] Brun Armelle, Détection de Thème et Adaptation des Modèles de Langage pour la Reconnaissance Automatique de la Parole, Thèse, Université Henri Point-Caré, Nancy I – France, 2003.

[CAV 94] Cavnar William B. et Trenkle John M., N-Gram-Based Text Categorization, Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas - USA, pp. 161-175, 11-13 April 1994.

[CHI 03] Ching-Long Yeh et Yi-Chung Chen, Using Zero Resolution to Improve Text Categorization, Pacific Asia Conference on Language, Information and Computation 17, Singapour – Chine, 1 – 3 Octobre 2003.

[CHU 01] Chung Young Mee et Lee Jae Yun, A Corpus-Based Approach to Comparative Evaluation of Statistical Term Association Measures, Journal of the American Society for Information Science and Technology, vol. 52/4, pp. 283 – 296, 2001.

[CIB 84] Cibois Philippe, L'Analyse des Données en Sociologie, Presses Universitaires de France, 218 p., ISNB : 2 13 038359 9,1984.

[CLI 04] Clifton Chris, Cooley Robert et Rennie Jason , TopCat : Data Mining for Topic Identification in Text Corpus, Transaction on Knowledge and Data Enginerring, vol. 16/8, pp.949-964, Août 2004.

[COH 97] Cohen A., Mantegna R. N. et Halvin S., Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts, Fractals, vol. 5/1, pp.95-104, 1997.

[COH 03] Cohen William W., Ravikumar Pradeep et Fienberg Stephen E., A Comparison of String Distance Metrics for Name-Matching Tasks, KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington DC – USA, 24 – 27 Août 2003.

[CRO 96] Crowder Grace et Nicholas Charles, Using Statistical Properties of Text to Create Metadata, First IEEE Metadata Conference, 16 – 18 Avril 1996.

[DAM 95] Damashek M., Gauging Similarity with n-Grams : Language-Independent Categorization of Text, Science, vol. 267, pp. 843 – 848, 1995.

[DAO 90] Daoust François, L'informaticien, le Lecteur et le Texte, L'approche SATO, ICO : Intelligence Artificielle et Sciences Cognitives au Québec, vol. 2/3, pp. 55 – 60, 1990.

[DAS 04] Da Sylva Lyne, Indexation Automatique de Documents par Contribution d'Analyses Statistiques et Terminologiques Structurées, RIAO 2004, Avignon – France, 26-28 Avril 2004.

[DEL 00] Della Ratta Francesca et Morrone Adolfo, Du Texte aux Variables : les Contributions de l'Analyse Textuelle des Questions Ouvertes à l'Analyse Traditionnelle des Données, 5èmes Journées internationales d'Analyse statistique des Données Textuelles, Lausanne - Suisse, 9 – 11 Mars 2000.

[DEL 02] Dellandrea Emmanuel, Makris Pascal, Boiron M. et Vincent Nicole, A Medical Acoustic Signal Analysis Method Based on Zipf Law, International Conference on Digital Signal Processing, Santorini - Grèce, vol. 2., p. 615-618, Juillet 2002.

[DEP 99] Depain-Delmotte Frédéric, La Sélection de l'Antécédent du pronom dans les systèmes de traitement automatique des langues naturelles, VEXTAL'99, Venise - Italie, 20 – 22 Septembre 1999.

[DES 99] Deschavanne Patrick J., Guiron Alain, Vilain Joseph, Fagot Guillaume et Fertil Bernard, Genomic Signature : Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences, Mol Biol Evol, vol. 16/10, pp. 1391 – 1400, 1999.

[DEV 00] De Vel Olivier, Mining E-mail Authorship, Workshop on Text Mining, KDD-2000, Boston – USA, 20 Août 2000.

[DIA 05] Dias Gaël et Alves Elsa, A Language-Independent Unsupervised Topic Segmentation System based on Word Cooccurrence, 9th International Symposium on Social Communication, Santiago de Cuba – Cuba, ISBN 959-7174-05-7, pp. 588-592, 24-28 Janvier 2005.

[DUB 99] Dubrocard M. et Luong Xuan, Problèmes d'Attribution : Application de Quelques Tests Statistiques à Différents Historiens Latins, Analyse Arborée, VEXTAL'99, , 22 – 24 Novembre 1999.

[ENG 92] Enguehard C., Malvache P. et Trigano P., Indexation de Textes : l'Apprentissage des Concepts, Quinzième colloque International en Linguistique Informatique, Nantes – France, 23 – 28 Août 1992, ICCL, vol. 4, pp. 1197 – 1202, 1992.

[FER 97-1] Ferret Olivier, Grau Brigitte et Masson Nicolas, Utilisation d'un Réseau de Cooccurrences pour améliorer une Analyse Thématique fondée sur la Distribution des Mots, ISKO France 97, Lille – France, 16 – 17 Octobre 1997.

[FER 97-2] Ferret Olivier et Grau Brigitte, Une Analyse Thématique s'appuyant sur une Mémoire Episodique, 1ères journées Scientifiques et Techniques FRANCIL, vol. 1/1, pp. 161-168, Avignon – France, 1997.

[FER 01] Ferret Olivier, Grau Brigitte, Minel Jean-Luc et Porhiel Sylvie, Repérage de Structures Thématiques dans des Textes, TALN 2001, Tours – France, 2-5 Juillet 2001.

[FOR 00] Forest Dominic et Meunier Jean-Guy, La Classification Thématique des Textes : un Outil d'Assistance à la Lecture et à l'Analyse de Textes Philosophiques, 5èmes Journées internationales d'Analyse statistique des Données Textuelles, Lausanne - Suisse, 9 – 11 Mars 2000.

[GAR 98] Garnier Bénédicte et Guérin-Pace, La Statistique Textuelle pour traiter une Question Ouverte Suivie d'une Relance, 4èmes Journées internationales d'Analyse statistique des Données Textuelles, Nice - France, 1998.

[GAV 03] Gavin Daniel G., Oswald W. Wyatt, Wahl Eugene R. et Williams John W., A Statistical Approach to Evaluating Distance Metrics and Analog Assignments for Pollen Records, Quaternary Research, vol. 60, pp. 356 – 367, 2003.

[GU 00] Gu Zhong et Beleant Daniel, Hash Table Sizes for Storing N-Grams for Text Processing, Technical Report 10-00a, Department of Electrical and Computer Engineering, Iowa State University, Iowa – USA, Octobre 2000.

[GUE 01] Guénoche Alain et Leclerc Bruno, The Triangles Method to Build X-Trees from Incomplete Distances Matrices, RAIRO Operations Research, vol. 35, pp. 283 – 300, 2001.

[GUE 99] Guénoche Alain et Grandcolas S., Approximation par arbre d'une distance partielle, Mathématiques, Informatiques et Sciences Humaines, vol. 146, pp. 51 – 64, 1999.

[GUI 03] Guidon Stéphane, Méthodes et Algorithmes pour l'Approche Statistique en Phylogénie, Thèse de Biologie dirigée par Gascuel Olivier, Université Montepellier II – France, 7 Juillet 2003.

[HER 02] Hernandez Nicolas et Grau Brigitte, Analyse Thématique du Discours : Segmentation, Structuration, Description et Représentation, CIDE 5, pp. 277-288, Hammamet – Tunisie, 20-23 Octobre 2002.

[HOS 04] Hossain Aleem et Lee Mark, Weight Derivation for Salience Algorithms in Pronominal Anaphora Resolution, CLUK Research Colloquium, 6 – 7 Janvier 2004.

[HU 04] Hu Yu et Schopf Jennifer M., IBL for Replica Selection in Data-Intensive Grid Application, Rapport Technique #TR-2004-03, Université de Chicago, Computer Science Department, Chicago – USA, Avril 2004.

[JAI 04] Jain Prateek, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee et Achla M. Raina, Anaphora Resolution in Multi-Person Dialogues, Workshop on Discourse and Dialogue, SIGdial 2004, Cambridge – USA, 30 Avril – 1 Mai 2004.

[JAL 02] Jalam Radwan et Chauchat Jean-Hugues, Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clés pertinents à l'aide des n-grammes caractéristiques, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo France, 13-15 Mars 2002.

[JOH 01] Johnson Don H. et Sinanović Sinan, Symmetrizing the Kullback-Leibler Distance, IEEE Transaction on Information Theory, Mars 2001.

[JOR 05] Jörgensen Peter, Incorporating Context in Text Analysis by Interactive Activation with Competition Artificial Neural Networks, Information Processing and Management, vol. 41/5, pp ; 1081-1099, 2005.

[KAN 01] Kan Min-Yen, Klavans Judith L. et McKeown Kathleen R., Synthesizing Composite Topic Structure Trees for Multiple Domain Specific Documents, Technical Report CUCS-003-01, Columbia University, 2001.

[KAR 94] Karlgren Jussi et Cutting Douglass, Recognizing Text Genres with Simple Metrics using Discriminant Analysis, 15th International Conference Computational Linguistice, COLING'94, Kyoto - Japon, 5-9 Août 1994.

[KLO 05] Klopotek Mieczyslaw, Very Large Bayesian Multinets for Text Classification, Future Genration Computer Systems, vol. 21/7, pp. 1068-1082, Juillet 2005.

[KUL 01] Kulyukin Vladimir A. et Bookstein Abraham, Integrated Object Recognition with Extended Hamming Distance, Rapport Technique, Université DePaul, School of Computer Science, Chicago – USA, 2001

[LAB 01] Labbé Cyril et Labbé Dominique, Inter-Textual Distance and Autorship Attribution Corneille and Molière, Journal of Quantitative Linguistic, pp. 213-231, 8-3 Décembre 2001.

[LAP 94] Lappin Shalom et Leass Herbet J., An Algorithm for Pronominal Anaphora Resolution, Computational Linguistics, vol. 20/4, pp. 535-561, 1994.

[LAR 02] Largus Krista et Kuusisto Jukka, Topic Identification in Natural Language Dialogues using Neural Networks, SIGDIAL'02, Philadelphia – USA, 11-12 Juillet 2002.

[LEB 94] Lebart L. et Salem A., Statistique Textuelle, Dunod, ISBN 2-10-002239-3, Paris, 1994.

[LEB 04] Lebouc Marie-France, La Construction de l'Altérité en Contexte Marchand : le Cas de l'Animal, Thèse de Doctorat en Sciences de l'Administration, Université de Laval – Canada, 2004.

[LEE 01] Lee Lillian, On The Effectiveness of the Skew Divergence for Statistical Language Analysis, Artificical Intelligence and Statistics, pp. 65 – 72, 2001.

[LEI 95] Leimdorfer François et Salem André, Usages de la Lexicométrie en Analyse de Discours, Cahier des Sciences Humaines, vol. 31/1, pp. 131-143, 1995.

[LEL 98] Lelu Alain, Halleb Mohamed et Delpart Bruno, Recherche d'Information et Cartographie dans des Corpus Textuels à partir des Fréquences de N-Grammes, 4èmes Journées internationales d'Analyse statistique des Données Textuelles, Nice - France, 1998

[LEL 02] Lelu Alain, Comparaison de Trois Mesures de Similarités utilisées en Documentation Automatique et Analyse Textuelle, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[LES 03] Lespinats Sylvain, Deschavanne Patrick, Guiron Alain et Fertil Bernard, L'ADN en tant que Texte : Style et Syntaxe, Revue des Nouvelles Tecnologies de l'Information, vol. 1, pp. 193 – 202, 2003.

[LEV 98] Levy Joseph P., Bullinaria John A. et Patel Malti, Exploration in the Derivation of Semantic Representations from Word Co-occurrence Statistics, South Pacific Journal of Psychology, vol. 10, pp. 99-111, 1998.

[LI 03] Li Hang et Yamanishi Kenji, Topic Analysis using a Finite Mixture Model, Information Processing and Management, vol. 39/4, pp. 521-541, 2003.

[LIN 00] Lindsey Clark S. et Strömberg, Image Classification using the Frequencies of Simple Features, Pattern Recognition Letters, vol. 21, pp. 265 – 268, 2000.

[LIU 03] Liu Tao, Liu Shengping, Chen Zheng et Ma Wei-Ying, An Evaluation on Feature Selection for Text Clustering, ICML 2003, Washington DC-USA, 21-24 Août 2003.

[LOU 04] Lourenço Fernando, Lobo Victor et Bação Fernando, Binary-based Similarity Measures for Categorical Data and their Application in Self-Organizing Maps, JOCLAD 2004 - XI Jornadas de Classificação e Análise de Dados, Lisbonne – Portugal, 1-3 Avril 2004.

[MAR 03] Market Katja, Nissim Malvina et Modjeska Natalia N., Using the Web for Nominal Anaphora Resolution, Computational Treatment of Anaphora, EACL 2003, Budapest – Hongrie, 12- 17 Avril 2003.

[MAR 03-2] Marteau Hubert, Lefèvre Alexandre et Vincent Nicole, Comparaison de Textes par Mesure Fractale, Majestic'03, Marseille – France, Octobre 2003.

[MAR 04-1] Marteau Hubert, Lefèvre Alexandre et Vincent Nicole, Etude de Textes par leur Image, EGC 04, Clermont-Ferrand – France, 20 – 23 Janvier 2004.

[MAR 04-2] Marteau Hubert, Lefèvre Alexandre et Vincent Nicole, Du Texte à l'Image, RFIA'04, Toulouse – France, 28 – 30 Janvier.

[MAR 05-1] Marteau Hubert et Vincent Nicole, L'automate Textuel pour la prise en compte de l'Evolution du Texte, EGC 05, Paris – France, 19 – 21 Janvier 2005.

[MAR 05-2] Marteau Hubert et Vincent Nicole, Indexation Based on the Textual Evolution, ICMLC'O5, Guangzhou - Chine, 18 – 21 Août 2005.

[MAT 03] Mather Laura A. et Note Jarrod, Discovering Encyclopedic Structure and Topics in Text, Proceedings ok KDD'2000, Sixth ACM SIGKDD Internation Conference on Knowledge Discovery and Data Mining, Boston – USA, 20-23 Août 2003.

[MCQ 66] McQuitty L. L., Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data, Educational and Psychological Measurement, vol. 26, pp. 825 - 831, 1966.

[MEY 02] Meyer Andréia da Silva, Comparação de Coeficientes de Similaridade Usados em Análises de Agrupamenti com Dados de Marcadores Moleculares Dominantes, Rapport de Licence en Mathématique, São Paulo – Brésil, Janvier 2002.

[MEY 04] Meyer Andréia da Silva, Garcia Antonio Augusto Franco et Pereira de Souza Anete, Comparison of similarity coefficients used for cluster analysis with domiant markers in maize (Zea mays L), Genetics and Molecular Biology, vol. 21/1, pp. 83 – 91, 2004.

[MIT 98] Mitkov Ruslan, Robust Pronoun Resolution with Limited Knowledge, COLING-ACL'98, pp. 869-875, Montréal – Canada, 1998.

[MOI 02] Moine Michèle, Indicateurs de Diversité et Exploitation Statistique d'une Question Ouverte, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[MON 02] Monmarché Nicolas, Marteau Hubert, Gérard Jean-Pierre, Guinot Christiane et Venturini Gilles, Interactive mining of multimedia databases with virtual reality,

Proceedings of the Third International Conference on Virtual Reality, pp. 478-484, Hangzhou, Chine, 9-12 Avril 2002.

[MOR 02] Morin Annie, Deux Exemples d'Analyse de Données Textuelles, Colloque sur la statistique et l'analyse des données dans les sciences appliquées et économiques, Beyrouth - Liban, 16-18 septembre 2002.

[MOT 01] Mothe J., Chrisment C., Dkaki T., Dousset B. et Egret D., Information Mining : Use of the Document Dimensions to Analyse Interactively a Document Set, European Colloquium on Information Retrieval Research, Darmstadt – Allemagne, pp. 66-77, 4-6 Avril 2001.

[MUR 04] Murgue Thierry et De La Higuera Colin, Distances entre Distributions de Probabilité : Comparaison de modèles de Langages Stochastiques, Cap'04, Montpellier – France, 14 – 16 Juin 2004.

[NIK 02] Nikolaou N. et Papamarkos N., Color Image Retrieval using a Fractal Signature Extraction Technique, Engineering Applications of Artificial Intelligence, vol. 15, p. 81-96 2002.

[NOU 92] Nourredine M., Théorie des Langages, Office des Publications Universitaires, 1992.

[OMH 04] Omhover J.F., Detyniecki M. et Bouchon-Meunier B., A Region-Similarity-Based Image Retrieval System, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems - IPMU'2004, Perugia - Italie, 4 – 9 Juillet 2004.

[PAL 01] Palomar Manuel, Moreno Lidia, Peral Jesús, Muñoz Rafael, Ferrández Antonio, Martínez-Barco Patricio et Saiz-Noeda Maximiliano, An Algorithm for Anaphora Resolution in Spanish Texts, Computational Linguistics, vol. 27/4, pp. 545-567, 2001.

[POI 99] Poibeau Thierry, Mixing Technologies for Intelligent Information Extraction, Intelligent Information Integration , IJCAI'99, Stockholm - Suède, 31 Juillet – 6 Août1999.

[POI 00] Poibeau Thierry, A corpus-based approach to Information Extraction, Journal of Applied System Studies, vol. 2/2, 2000.

[POI 01] Poibeau Thierry et Balvet Antonio, Corpus-based lexical acquisition for Information Extraction, Adaptive Text Extraction and Mining, IJCAI'2001, Seattle – USA, 5 Août 2001.

[POL 05] Pollard David, Chapter 3 – Total Variance Distance Between Measures, Asymptotia, En Cours d'Ecriture.

[PON 97] Ponte Jay M. et Croft W. Bruce, Text Segmentation by Topic, Proceedings of the first European Conference on Research and Advanced Technology for Digital Libraries, ISBN 3-540-63554-8, 1997.

[POU 02] Pouliquen Bruno, Indexation de textes médicaux par extraction de concepts, et ses utilisation, Manuscrit de Thèse, Université de Rennes I, 7 Juin 2002.

[RAJ 97] Rajman Martin et Besançon Romaric, Text Mining : Natural Language Techniques and Text Mining Applications, Seventh IFIP 2.6 Working Conference on Database Semantics (DS-7), Leysin - Suisse, 7-10 Octobre 1997.

[RAJ 04] Rajman Martin et Besançon Romaric, Text Mining – Knowledge Extraction from Unstructured Textual Data, International Federation of Classification Societies, Chicago – USA, 15 – 18 Juillet 2004.

[REI 86] Reinert Max, Un Logiciel d'Analyse Lexicale : Alceste, Les Cahiers de l'Analyse de Données, vol. 4, pp. 471-484, 1986.

[REN 03] Rennie Jason D. M., Shih Lawrence, Teevan Jaime et Karger David R., Tackling the Poor Assumptions of Naive Bayes Text Classifiers, ICML 2003, Washington DC-USA, 21-24 Août 2003.

[REY 94] Reynar Jeffery C., An Automatic Method of Finding Topic Boundaries, 32nd Annual Meeting of the Association for Computational Linguistics, Student Session, pp. 331-333, Las Cruces - USA, 1994.

[ROG 03] Rogaski Mark, Dynamic Topic Analysis : Classification without Established Classes using Distance Threshold, iCML'03, Piscataway – USA, 3-8 Décembre 2003.

[ROG 02] Rogati Monica et Yang Yiming, High-Performing Feature Selection for Text Classification, CIKM International Conference on Information and Knowledge Management, McLean - USA, 4-9 Novembre 2002.

[ROO 04] Rooney Thomas P., Wiegmann Shannon M., Rogers David A. et Waller D. M., Biotic Impoverishment and Homogenization in Unfragmented Forest Understory Communities, Conservation Biology, vol. 18/3, pp. 787 – 798, Juin 2004.

[SAI 97] Saitou N. et Nei M., The Neighbor-Joining Method : A New Method for Reconstructing Phylogenetic Trees, Mol. Biol. Evol., vol. 4, pp. 406 - 425, 1987.

[SAL 04] Salmon-Alt Susanne, Résolution automatique d'anaphores infidèles en français : Quelles ressources pour quels apports ?, Session Poster, TALN 2004, Fès - Maroc, 19 – 21 Avril 2004.

[SAL 89] Salton Gerard, Automatic Text Processing – The Transformation Analysis and Retrieval of Information by Computer, Addison Wesley Publishing Company, Reading, MA, 1989.

[SAL 94] Salton Gerard et Allan James, Automatic Text Decomposition and Structuring, RIAO'94, Paris – France, 6-20 Ocrobre 1994.

[SAL 96] Salton Gerard, Singhal Amit, Buckley Chris et Mitra Mandar, Automatic Text Decomposition using Text Segments and Text Themes, HyperText'96, Washington DC-USA, 16-20 Mars 1996.

[SAN 02-1] SanJuan Eric et Ibekwe-SanJuan Fidelia, Terminologie et Classification Automatique des Textes, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002.

[SAN 02] Santos-Pereira Carla M. et Pires Ana M., Detection of Outliers in Multivariate Data : A Method Based on Clustering and Robust Estimators, CompStat 2002, ISBN 3-7908-1517-9, Berlin – Allemagne, 24 – 28 Août 2002.

[SAR 69] SARICH V. M., Pinniped Origins and the Rate of the Evolution of Carnivore Albumins, Systematic Zoology, vol. 18, pp. 286 – 295, 1969.

[SAT 77] Sattah S. et Tversky A., Additive Similarity Trees, Psychometrika, vol. 42, pp. 319 – 345, 1977.

[SEB 02] Sebastiani Fabrizio, Machine Learning in Automated Text Categorization, ACM Computing Surveys, vol. 34/1, pp. 1-47, 2002.

[SEL 05] Selinski Silvia et Ickstadt Katja, Similarity Measures for Clustering SNP Data, Rapport Technique, Statistics Department, Université de Dortmund – Allemagne, 2005.

[SHA 00] Shankar Shrikanth et Karypis George, A Feature Weight Adjustment Algorithm for Document Categorization, Workshop on Text Mining, KDD 2000, Boston – USA, 20-23 Août 2000.

[SIN 04] Sinka Mark P. et Corne David W., The BankSearch Web Document DataSet : Investigating Unsupervised Clustering and Category Similarity, Journal of Network and Computer Applications, vol. 28, pp. 129-146, 2004.

[SOK 58] Sokal R. R. et Michener C. D., A Statistical Method for Evaluating Systematic Relationships, University of Kansas Science Bulletin, vol. 38, pp.1409 – 1438, 1958.

[TAN 05] Tan Sangbo, Neighbor-Weighted K-Nearest Neighbor for Unbalanced Text Corpus, Expert System With Application, Elsevier, vol. 28/4, pp. 667-671, 2005.

[TIM 97] Timini Ismaïl, Analyse du discours assistée par ordinateur, Version 3AD95, ACH-ALLC'97, Association for Computer and the Humanities and the Association for Literary & Linguistic Computing, Kingston – Canada, 3 – 7 Juin 1997.

[TIM 98] Timini Ismaïl, SYTEM 3AD : Un Outil de Classification à Caractère Linguistico-Mathématique, Colloque International de Veille Stratégique Scientifique et Technologique, Toulouse – France, 19 – 23 Octobre 1998.

[TOR 02] Torres Juan-Manuel, Velázquez-Morales Patricia et Meunier Jean-Guy, Condensés de Textes par des Méthodes Numériques, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo - France, 13 – 15 Mars 2002

[TRO 02-1] Trouilleux François, Insertion et Interprétation des Expressions Pronominales, Atelier « Chaînes de Références et Résolveurs d'Anaphores », TALN 2002, Nancy – France, 24-27 Juin 2002.

[TRO 02-2] Trouilleux François, A Rule-based Pronoun Resolution System for French, DAARC'02, Lisbonne – Portugal, 18 – 20 Septembre 2002.

[TUT 00] Tutain Agnès, Trouilleux François, Clouzot Catherine, Gaussier Eric, Zaenen Annie, Rayot Stéphanie et Antoniadis Georges, Annoting Large Corpus with Anaphoric Links, DAARC'00, Lancaster – Grande Bretagne, 16 – 18 Novembre 2000.

[YE 01] Ye Nong et Li Xiangyang, A Scalable Clustering Technique for Intrusion Signature Recognition, IEEE Workshop on Information Assurance and Security, ISBN 0-7803-9814-9, West Point – USA, 5 – 6 Juin 2001.

[ZHA 00] Zhang Tong, Large Margin Winnow Methods for Text Categorization, Workshop on Text Mining, KDD 2000, Boston – USA, 20-23 Août 2000.

[ZHA 03] Zhang Bin et Srihari Sargur N., Properties of Binary Vector Dissimilarity Measures, JCIS CVPRIP 2003, Cary - USA, 26-30 Septembre 2003.

[ZIP 35] Zipf G. K., The Psychology of Language, an Introduction to Dynamic Philology, M.I.T. Press, Cambridge, Massachussetts, 1935.

2. Logiciels

[LOG 3AD] 3AD95 (Approximation de l'Analyse Automatique du Discours), développé par I. Timimi, avec connexion à l'analyseur morphologique CRISTAL, version PC-Windows en cours.

[LOG ALC] ALCESTE (Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte), développé sur Mac par Max Reinert (Laboratoire de Psychologie de Toulouse) et commercialisé par la Société IMAGE (Informatique Mathématique et Gestion), Version 6.4.

[LOG ALT] ATLAS.ti, Société Scientific Software Development GmbH, Version 5.0

[LOG HYP] HYPERBASE (Logiciel Hypertexte pour le Traitement Documentaire et Statistique des Corpus Textuels), développé par Etienne Brunet, chercheur à l'INaLF.

[LOG LEX] LEXICO, développé par André Salem et autres chercheurs du Laboratoire de Lexicologie et Textes politiques (ENS de Saint-Cloud)et l'équipe SYLED-CLA2T Université Sorbonne Nouvelle, Paris 3, Version 3

[LOG MOD] MODALISA, Module Interviews du Logiciel Modalisa, développé et commercialisé par la société KYNOS, Version 5.

[LOG NVI] NVivo, Société QSR, Version 2.0 (version suivante : NVivo 7)

[LOG SAS] SAS (Statistical Analysis System)

[LOG SAT] SATO (Système d'Analyse de Textes par Ordinateur), développé par François Daoust, au Centre A.T.O. de l'Université de Québec à Montréal. Il est couplé avec le logiciel Deredec (Dépistage de Relation de Dépendance en Contexte), qui comporte un module d'analyse de la GDSF (Grammaire De Surface du Français) et qui a été développé par Pierre Plante à l'Université du Québec, Montréal.

[LOG SPA] SPAD-T, (Système Portable pour l'Analyse des Données Textuelles), module de Spad, développé par Ludovic Lebart et A. Morineau, et commercialisé par le CISIA, Version 6.0.

3. Sites Internet

[WWW Firminy]
http://www.ville-firminy.fr/lecorbusier/index.htm

[WWW Jenny]
http://pagesperso.aol.fr/jacquesjenny/

[WWW LeCorbusier01]
http://www.objectifreussir.ch/fr/cadre_repertoire/Metier/Architecte/le_corbusier.html

[WWW LeCorbusier02]
http://fr.wikipedia.org/wiki/Le_Corbusier

[WWW Maisons Radieuses]
http://www.maisonradieuse.org

[WWW WIK]
http://fr.wikipedia.org

Annexe 1 - Le Corbusier

Cette annexe présente de manière succincte, dans un premier temps, l'homme qui se cache derrière le pseudonyme « Le Corbusier », puis, les deux cités qui ont fait l'objet d'entretiens, Firminy dont les entretiens n'ont pas été utilisés et Rezé, pour laquelle l'avancée de l'analyse sociologique étant plus aboutie, les entretiens ont été utilisés.

1. L'homme

(WWW LeCorbusier01), (WWW LeCorbusier02) Né le 6 octobre 1887 à La Chaux-de-Fonds, Le Corbusier était le pseudonyme de Charles-Édouard Jeanneret-Gris. De nationalité suisse puis française, cet architecte célèbre était le porte-drapeau de ce que l'on appelle de nos jours le mouvement moderne (également nommé mouvement international ou style international), avec Ludwig Mies van der Rohe, Walter Gropius, et Theo van Doesburg. Il œuvra également dans les domaines de l'urbanisme et du design.

En 1900, il entame une formation de graveur-ciseleur à l'École d'Art de La Chaux-de-Fonds (son père était horloger). Son professeur, Charles L'Eplattenier a créé à l'École d'Art, le Cours supérieur qui fera référence et où se formera Jeanneret. C'est lui qui dirigera son jeune élève vers l'architecture en 1904. Il apprend la technique du béton auprès d'Auguste Perret à Paris. Il construit peu mais produit de nombreux articles manifestes sur l'homme moderne à travers la revue "Esprit Nouveau".
En 1917, l'architecte quitte La Chaux-de-Fonds pour s'installer à Paris : les montagnes neuchâteloises sont, en effet, bien trop exiguës pour qu'un tel arbre puisse y déployer toutes ses racines. Il devient alors Le Corbusier (le nom de son grand-père maternel); sa carrière internationale prend son envol.

En 1925, à l'Exposition Internationale des Arts Décoratifs (qui donne naissance à l'art nouveau), il construit un pavillon manifeste de son art. En 1928, il organise le Congrès International d'Architecture Moderne (CIAM).
Il est naturalisé français en 1930.
Architecte, urbaniste, essayiste, peintre, sculpteur, Le Corbusier fournit une oeuvre gigantesque. Il peut être cité de manière concis : la Villa Savoye à Poissy (1928), la chapelle Notre-Dame du Haut à Ronchamp (1950), le couvent Sainte Marie de la Tourette à Eveux (1953) et le Modulor.
Le Corbusier est l'auteur de plus de 40 livres et essais.

Bardé de médailles et d'honneurs, il meurt à Cap-Martin (Alpes-Maritimes) au cours d'une baignade dans la Méditerranée, en 1965.
Il portait d'habitude des lunettes aux grands verres ronds cerclés de noir. Il était lié d'amitié avec l'artiste breton Joseph Savina. Il figure sur le billet de 10 francs suisse mis en circulation le 8 avril 1997.

Figure 46 Le Corbusier

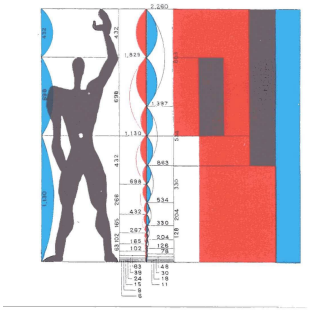

Figure 47 Le Modulor

Figure 48 La Villa Savoye

Figure 49 La Chapelle Notre Dame du Haut

Figure 50 Le Couvent Sainte Marie de la Tourette

Figure 51 Le billet de 10 francs Suisse

2. Le Corbusier à Firminy : Unité d'Habitation 1965/1967

(WWW Firminy) L'Unité d'Habitation de Firminy est le dernier exemple de "cité-jardin verticale" cher à Le Corbusier. Elle est également le seul et unique témoin d'une extension du quartier de Firminy- vert qui prévoyait à l'origine trois unités et leurs équipements connexes (commerces, passerelles et parkings).
C'est en effet la dernière des cinq unités après Marseille, Nantes-Rezé, Berlin et Briey en Forêt et également la plus grande par le nombre de logements (414 appartements). Elle a été réalisée entre 1965 et 1967 sous la conduite d'André Wogensky à qui fut confié la responsabilité du projet suite à la mort de Le Corbusier survenue en août1965.

Le bâtiment de 130 m de long par 21 m de largeur et 55 m de hauteur est orienté selon un axe nord-sud qui offre aux appartements la double exposition grâce à une conception traversante. L'ensoleillement est donc maximal et seul le "brise-soleil" vient réguler la quantité de lumière.
Les "rues intérieures" sont au nombre de sept, elles desservent 17 niveaux d'habitations dimensionnés au "Modulor" soit 2.26 m de hauteur par 1.78 m de largeur pour le module de base.
L'école sur le toit-terrasse est la plus grande réalisée par Le Corbusier pour les Unités d'Habitation. Elle occupe toute la longueur de l'édifice sur deux niveaux. Huit classes sont

prévues à l'origine pour un statut d'école maternelle et primaire. Actuellement ces espaces sont occupés par 32 élèves répartis en une classe maternelle. C'est une partie de l'édifice relativement méconnue, mais dont la conception s'avère fort intéressante.

Le classement de l'unité de Firminy au titre des monuments historiques intervenu en septembre 1993, n'autorise que des modifications internes, puisque les façades, les pilotis ainsi que le toit terrasse et l'école ont été intégralement classés.
Il est à noter que la " cité radieuse " de Firminy fonctionne toujours avec un statut de logement social, cette spécificité en fait de ce point de vue, la dernière à se situer dans la lignée des ambitions de son concepteur.
Elle connaît également un renouveau certain depuis le début des années 90 puisque actuellement l'ensemble de la partie sud est occupée, soit environ 184 logements.

Figure 52 Firminy vue d'ensemble

Figure 53 Firminy vue profilée

3. Le Corbusier à Rezé : les Maisons Radieuses 1953/1955

(WWW Maisons Radieuses) Au lendemain de la seconde guerre mondiale, des quartiers entiers de Nantes sont à reconstruire. « La Maison familiale », société coopérative d'HLM, souhaite renouveler l'image traditionnelle des logements sociaux par une opération expérimentale exemplaire. Les travaux débutèrent en Juin 1953 et durent 18 mois ; les premiers habitants s'installent le 16 mars 1955 et le bâtiment est inauguré en Juillet 1955.
Grand paquebot visible de très loin (108 m de long, 52 m de haut, 19 m de large), la Maison Radieuse surprend par son aspect massif. Montée sur pilotis, sa structure en béton armé abrite six « rues intérieures » qui desservent les 294 logements en duplex (montants et descendants) sur 17 niveaux, avec double orientation Est Ouest ou exposés au Sud. Les appartements donnent sur des loggias aux couleurs très vives. Le parc de 6 hectares accentue l'effet paysager de l'ensemble.

Le bâtiment, inscrit aux Monuments Historiques depuis 2001, a été rénové plusieurs fois depuis les années 80.
En 1988, Loire Atlantique Habitations, entreprise sociale pour l'habitat, copropriétaire principale qui gère 55% des logements, a effectué une mise aux normes de son parc locatif.
Entre 1996 et 1999, après trois années de diagnostics très poussés engagés par le Syndicat de copropriété et Loire Atlantique Habitations, les façades de l'immeuble sont entièrement restaurées.
En 2004, une dernière tranche de travaux concerne le traitement des menuiseries extérieures et le changement des doubles vitrages.