N° d'ordre : 2007telb0062

THÈSE

Présentée à

l'ECOLE NATIONALE SUPERIEURE DES TELECOMMNUNICATIONS DE BRETAGNE

en habilitation conjointe avec l'Université de Bretagne Sud

pour obtenir le grade de

DOCTEUR de l'ENST Bretagne

Mention **« Les Mathématiques et leurs Interactions »**

par

Christian MAUCERI

« *Indexation et isotopie* : *vers une analyse interprétative des données textuelles*»

Soutenue le 14 Décembre 2007 devant la Commission d'Examen :

Composition du Jury

- *Rapporteurs* : François RASTIER, Directeur de recherche, CNRS
  Monique SLODZIAN, Professeur, INALCO

- *Examinateurs* : Diem HO, IBM Academy of Technology, IBM Europe
  Ioannis KANELLOS, Professeur, ENST Bretagne
  Philippe LENCA, Maître de conférence, ENST Bretagne
  Pierre-François MARTEAU, Directeur du VALORIA, Université de Bretagne Sud

Remerciements

La première personne à qui je voudrais exprimer ma gratitude est mon épouse Anne qui m'a supporté sans jamais faillir lors de ces quatre dernières années consacrées à cette entreprise exigeante qu'est la rédaction d'une thèse pour une personne salariée. Je voudrais aussi, pour les mêmes raisons remercier mes filles Laure et Camille qui ont eu la gentillesse de comprendre que leur père ne soit pas aussi disponible qu'il aurait du l'être durant cette même période. Cette thèse leur est dédiée.

Je voudrais remercier Ioannis Kanellos, mon directeur de thèse, pour ses conseils avisés, sa patience, son ouverture d'esprit et sa grande générosité.

Je suis conscient de l'honneur que Monsieur Rastier et Madame Slodzian m'ont fait en acceptant d'être mes rapporteurs. De façon générale je souhaite adresser mes remerciements aux membres du jury – donc outre les personnes déjà citées, Messieurs Ho, Lenca et Marteau.

Je suis particulièrement redevable à Diem Ho pour les innombrables heures passées ensemble sur des projets difficiles, loin de chez nous et pour les longues conversations que nous avons si souvent tenues sur la classification automatique et l'interprétation des données. Une grande partie des idées de cette thèse sont nées là.

Je voudrais enfin, remercier la compagnie IBM qui m'a offert des moyens modernes de travail permettant de gérer mon temps au mieux pour concilier deux activités souvent antagoniques.

Table des matières

1 Introduction

Dans une lettre à Atticus Cicéron demande à son ami de lui envoyer deux copistes[1] afin qu'ils collent sur ses livres des sillybi : de fines bandes de parchemin portant le titre et parfois les auteurs des ouvrages sur lesquels ils étaient apposés. L'index, le *sillybi*, est né de la nécessité pratique de retrouver un rouleau de parchemin, un *volumen*, dans une bibliothèque. Ce sont les Grecs qui systématisent l'index et le catalogue. Callimaque décrit le contenu de la bibliothèque d'Alexandrie dans des tables, les *pinakes*, qui suivent un classement par catégorie et par genre. Très tôt, donc, le contenu des bibliothèques est organisé afin d'en faciliter l'accès : le premier livre de l' « Histoire Naturelle » de Pline l'Ancien est une immense table des matières décrivant de façon détaillée les trente six autres volumes. Le volumen est un livre qui se déroule, il ne se prête pas au repérage même si certains signes typographiques arrivés jusqu'à nous en facilite la lecture : la manicule ☞[2] désigne les parties importantes du texte, le pied de mouche ¶ sépare les parties du texte. Le codex s'impose à partir du premier siècle[3], il facilite le feuilletage, la comparaison de différents passages du texte, il introduit la notion de page nécessaire aux progrès ultérieurs du livre.

La prédication et l'exégèse de la Bible aux XIIe et XIIIe siècles motivent la création de nouveaux outils destinés à trouver rapidement « statim invenire » les textes que l'on cherche : les recueils d'exempla, les concordances et les recueils de distinctiones en sont les représentants les plus marquants. Les exempla sont des anecdotes édifiantes destinées à aider les prédicateurs à rédiger leurs sermons, elles sont organisées en rubriques classées par ordre alphabétique. Les rubriques peuvent apparaître dans le corps du texte, des manicules en indiquent le début. Les concordances sont des index alphabétiques dont chaque mot est accompagné des phrases dans lesquelles il apparaît. Les recueils de distinctiones sont des dictionnaires d'interprétations spirituelles donnant pour chaque mot important des exemples d'emploi. Le Moyen Age apparaît donc comme une période féconde qui révolutionne l'indexation et dont les préoccupations herméneutiques ont un écho dans le monde documentaire contemporain.

[1] Esclaves de bibliothèque.

[2] Si la main sort d'une manche on l'appelle manchette et par extension la note qu'elle désigne

[3] Dés le IIe siècle tous les manuscrits de la Bible sont des codex

Paradoxalement, les débuts de l'imprimerie n'apportent pas d'innovation majeure dans la pratique de l'indexation. C'est le livre en tant que support qui est considéré plus que son contenu, il acquiert dés le XVe siècle sa forme actuelle. L'uniformisation des exemplaires d'un même ouvrage permet à des communautés distantes de lecteurs de se référer aux mêmes parties d'un livre et de généraliser l'usage de la référence bibliographique. A la fin du XVIIIe siècle la fiche de bibliothèque s'impose dans la constitution des catalogues, elle permet des renvois multiples et démultiplie les accès. Entre 1874 et 1876 Melvil Dewey invente une méthode de classification qui entend organiser la totalité du savoir en dix classes, chacune d'elles divisées en dix sous-classes elles-mêmes divisées en dix et ainsi de suite. Au développement des sciences et techniques à la fin du XIXe siècle et au début du XXe correspond un accroissement considérable du nombre des bibliothèques et plus particulièrement des bibliothèques scientifiques. Les pratiques se standardisent, les supports se modernisent et la profession de documentaliste se structure. Sous l'impulsion d'Eugen Wüster la terminologie s'affirme comme une discipline autonome qui veut s'affranchir de la linguistique et en tout cas de la langue[4].

Il est difficile de dire quelles sont les inventions qui resteront attachées au XXe siècle tant il fut fécond, il n'est cependant pas douteux que l'ordinateur vienne en tête : en un demi-siècle il s'est imposé partout, il n'est pour ainsi dire plus d'activités qui n'en dépendent. Le texte ne peut échapper à cette révolution. La digitalisation, sans supprimer les livres et les journaux[5], permet d'affranchir le texte du support papier : des collections que n'aurait pu accueillir une bibliothèque municipale autrefois peuvent être aujourd'hui contenues dans un ordinateur individuel. Au delà des capacités de stockage d'une machine isolée, c'est l'accès par le Web à un espace textuel virtuellement illimité qui est désormais possible. C'est un lieu commun de comparer la bibliothèque de Babel de Jorge Luis Borges au Web, mais le parallèle s'impose jusque dans l'absurde : on y trouve tout et n'importe quoi[6]. Quels sont les outils et les méthodes d'indexation disponibles pour relever le défi que pose cette démesure ?

Traditionnellement, indexer un document consiste à le décrire à l'aide de mots-clés significatifs, les termes. La terminologie est traversée par un fort courant positiviste issu du

[4] Eugen Wüster, espérantiste convaincu, aspire à une plus grande compréhension entre les peuples, par une approche qui privilégie le concept sur le terme il pense œuvrer pour une internationalisation de la science.

[5] Qui restent les supports privilégiés pour la lecture.

[6] Cf. « la nature informe et chaotique de presque tous les livres. » de la bibliothèque de Babel. [Borges 1944] « La biblioteca de Babel »

cercle de Vienne (Eugen Wüster), influant profondément sur la pratique de l'indexation. Ce courant fige le terme dans une théorie de la signification supposant que le sens est indépendant du contexte. Il est pourtant patent que le terme ne s'interprète que contextuellement. Il convient donc de le redéfinir dans un cadre interprétatif en le situant dans une syntagmatique qui étudie son rapport au texte (Rastier, « Le terme : entre ontologie et linguistique » [Rastier 2001a]). Par exemple, le terme peut être vu comme un thème c'est-à-dire une isotopie générique, la récurrence syntagmatique d'un sème générique. L'isotopie se manifeste par la cooccurrence, dans un texte, d'unités lexicales partageant un même sème. Le lieu de cette cooccurrence peut être toute partie du texte, voire le texte entier. Quel est donc le cadre d'une indexation par des termes ainsi définis et en particulier quel peut être son support informatique ? La détection d'une isotopie est le résultat d'une interprétation. Contrairement à l'approche componentielle, les isotopies précèdent et définissent le sème qui les caractérise. Elles sont attendues par le lecteur et assurent la cohérence de l'énoncé. En ce sens, le sème ne relève que de l'esprit, de l'humain, la machine ne manipule que des chaînes de caractères, des signifiants. Les processus d'identification des isotopies s'en trouvent allégés et clarifiés. Allégés en ce sens qu'il n'est pas nécessaire de maintenir des dictionnaires de sèmes : seuls les ensembles d'unités lexicales cooccurrentes sont nécessaires. Clarifiés en ce sens que la machine ne peut que proposer des cooccurrences à son utilisateur, en aucun cas elle ne produit du sens.

La principale raison d'être de l'indexation est de retrouver efficacement un texte parmi d'autres. Lors d'une indexation on ne considère pas un texte isolé mais un ensemble de textes. Cela pose la question de savoir comment collectivement les termes opposent ou rapprochent les textes qu'ils indexent. La qualification de cooccurrences au rang de corrélats supportant un terme suppose que ce dernier est l'expression d'une isotopie. Les cooccurrences susceptibles de participer au rapprochement de plusieurs textes sont donc d'un intérêt particulier.

Nous voyons donc se préciser une méthode d'indexation par isotopie :

- Recherche de cooccurrences,
- Qualification de ces cooccurrences au rang de corrélats sur la base de leur capacité à rapprocher des textes.

Ainsi, nous partons du postulat que le sens est l'apanage de l'esprit : la machine ne permet qu'une accélération de certaines tâches reproductibles, mécaniques. Parmi ces tâches le

repérage de chaînes de caractères dans un texte électronique est notable, à tel point qu'une forme d'indexation consiste à indexer un texte par les mots qu'il contient. Une telle indexation manque tout à la fois de précision et de robustesse. Manque de précision car certains index sont trop fréquents, manque de robustesse car seuls les textes contenant les termes d'une requête peuvent être retrouvés. A l'opposé l'indexation documentaire traditionnelle est précise et robuste mais elle est lente et peut difficilement être automatisée. Elle souffre en outre d'un problème de consistance, en effet, deux indexations différentes d'un même document ne comptent, en moyenne, que 30% à 40% de termes communs (voir par exemple « *Indexing consistency in MEDLINE* » [Funk et al. 1983]). Ceci nous rappelle que la lecture et la compréhension d'un texte sont un acte hautement subjectif qui varie d'un individu à l'autre ou même d'un moment à l'autre de la vie d'une personne. Aussi est-il important de permettre à une seule personne d'indexer un grand nombre de textes dans un laps de temps aussi court que possible. Les directives de la norme ISO 5963 [ISO 1985] spécifient que l'indexation comprend trois étapes :

1. l'examen du document et l'établissement de son sujet,
2. l'identification des principaux concepts présents dans le sujet,
3. l'expression de ces concepts dans les termes du langage d'indexation.

La première phase est lente. Les rédacteurs de la norme précisent d'ailleurs les parties du document qui doivent être lues : le titre, le résumé, la table des matières, les premiers paragraphes, les illustrations ainsi que les mots ou groupes de mots soulignés ou écrits dans une typographie inhabituelle.

Lors de la seconde phase, l'indexeur identifie les concepts, il ne doit utiliser que les notes prises lors de la première phase sans revenir au document. Il est par ailleurs indiqué que l'indexeur doit être aidé par des directives et listes de contrôles éditées par l'organisme en charge du fonds documentaire.

Dans la troisième phase l'indexeur doit vérifier que les concepts précédemment identifiés existent dans le langage d'indexation et les traduire en termes préférés. Si certains concepts n'existent pas il doit chercher de nouveaux termes dans des dictionnaires, des encyclopédies ou d'autres listes d'autorités.

Plusieurs critiques viennent à l'esprit à propos de ce modèle. Il ne tient pas compte de ce qui distingue les textes les uns des autres et de ce qui les rapproche : le texte est isolé du fonds auquel il appartient. Il se réfère à une notion de concept qui aurait un sens en dehors

des textes qu'il est supposé décrire : il est en particulier significatif que l'identification des concepts se fasse à partir de notes de lectures. D'une façon générale, le texte comme les termes qui le décrivent sont extraits de leurs contextes. Le présupposé à l'origine de cette approche ontologique est que le terme se définit au travers du triangle sémiotique. Dans la tradition Aristotélicienne le concept met en relation le mot et la chose (le concept, le mot et la chose sont les trois sommets du triangle), et bien sûr, les concepts pas plus que les choses ne varient selon les langues : il est ainsi précisé dans la norme que les notions ne sont pas liées aux langues individuelles. Cette approche est en totale opposition avec la linguistique de Saussure qui cherche précisément à s'affranchir de la référence. La question est de savoir comment une approche linguistique de l'indexation permettrait d'atteindre des niveaux de rapidité et de consistance satisfaisants tout en préservant un niveau de précision et de robustesse acceptable.

La recherche d'une indexation consistante n'est pas la recherche d'une indexation objective mais celle de l'interprétation consistante d'un ensemble cohérent de textes. Ceci amène naturellement à la notion de corpus initialement définie par la philologie et l'herméneutique. Un corpus est structuré en fonction d'une typologie des textes qui le composent, par l'identification des relations qu'ils entretiennent et par les utilisations qui en sont faites. L'indexation d'un corpus ou tout du moins de certaines de ses parties remarquables est une alternative à l'indexation traditionnelle. Par indexation d'un corpus il faut entendre une indexation qui prend en compte la détermination du local par le global, une indexation où la description d'un texte dépend du corpus auquel il appartient (un même texte pouvant bien sûr appartenir à différents corpus et être indexé différemment selon le corpus que l'on considère). En accord avec une démarche structuraliste cette détermination se fonde sur l'établissement d'identités et de différences entre les textes du corpus en fonction des termes qui les décrivent. Les termes indexant les documents induisent naturellement des fonctions de similarités basées sur le nombre de leurs termes communs. Elles permettent de regrouper automatiquement les textes, l'analyse de ces regroupements permettant d'évaluer la pertinence et la cohérence de l'indexation comme un tout.

La recherche des termes d'indexation doit aussi être globale, tenir compte de l'ensemble du corpus. Il est pour ainsi dire impossible de lire et d'acquérir une vision globale d'un corpus un tant soit peu conséquent dans des temps acceptables. Il est par contre possible d'en lire rapidement le vocabulaire. Une personne reconnaît facilement parmi les mots d'un corpus

ceux qui pourraient caractériser un thème. Ces mots correspondent à ce que l'indexeur s'attend à rencontrer, ils font écho aux présomptions d'isotopie et sont susceptibles de s'inscrire dans des systèmes de cooccurrences. Ces mots peuvent être collectés et servir de point d'ancrage pour une recherche automatique de cooccurrences à différents pallier du texte. Elles sont alors proposées au jugement humain pour être qualifiées au rang de corrélats et servir de support à l'indexation. Les corrélats retenus sont alors affectés aux isotopies qu'ils supportent par la médiation d'un jeu d'étiquettes.

En résumé la méthode d'indexation qui vient d'être esquissée se décompose en six phases :

1. construction d'un corpus,
2. lecture de son vocabulaire et identification de ses éléments caractéristiques,
3. recherche de cooccurrences autour de ces éléments,
4. qualification de certaines de ces cooccurrences au rang de corrélats et affectation d'étiquettes terminologiques,
5. regroupement des textes du corpus en classes sur la base de leur indexation,
6. vérification de la cohérence des regroupements obtenus.

Les phases 3 et 5 sont purement automatiques, les phases 2, 4 et 6 sont des activités exclusivement intellectuelles recourant à des outils de concordance pour les nécessaires retour au texte. Le temps consacré à ces dernières est fonction d'un compromis entre la rapidité et la qualité de l'indexation désirée. Les bénéfices attendus de ce type d'indexation sont sa cohérence, sa rapidité, sa robustesse, son indépendance de référentiels externes, son ancrage dans le texte.

2 Agrégation de similarités et interprétation

La science moderne prend racine dans le doute et la méfiance de la philosophie cartésienne. Les théories scientifiques sont des hypothèses de travail dont la validité dépend des résultats qu'elles produisent [Arendt 1968b]. Même si l'opposition entre les sciences de la nature et les sciences humaines s'estompe. Même s'il est admis depuis la première moitié du XX siècle que dans toute expérience l'observateur introduit un facteur subjectif dans les processus objectifs de la nature, la question de l'objectivité prête encore trop souvent à confusion. « Toute expérimentation est une question posée à la nature, une question à laquelle elle est contrainte de répondre. Or, chaque question contient, dissimulé, un jugement *a priori*: chaque expérimentation, en tant qu'expérimentation, est une prophétisation. » [Schelling 1799] Ou, comme le souligne Hannah Arendt [Arendt 1968a], la confusion est de supposer qu'il pourrait y avoir des « réponses sans questions et des résultats indépendants d'un être qui questionne ».

L'analyse des données se situe aux confluents des statistiques et des sciences humaines. Elle est utilisée dans toutes sortes d'activités allant de la gestion de production dans les entreprises au marketing en passant par la lexicographie ou l'anthropologie. Les modèles mathématiques qu'elle utilise avec leurs cohortes de graphes, de tables et de pourcentages, lui confèrent, aux yeux du profane, une autorité qui, bien que caricaturée depuis fort longtemps[7], reste indubitable. Pire, l'analyse des données utilisée par des esprits mal préparés voire même malveillants peut amener à tromper pour servir des intérêts économiques, politiques ou sociaux. En d'autres termes, les méthodes statistiques utilisées en analyse des données ne sont pas une garantie « d'objectivité ».

Tout ceci est évidemment connu, les panels d'échantillonnage sont une obsession des sondeurs[8], les prédictions sont confrontées aux résultats réels et un modèle qui ne marche pas sera forcément rejeté à la longue. Il existe cependant des domaines où la confrontation à la réalité est plus problématique voire complètement biaisée. Le cas du taux de criminalité plus élevé chez certaines minorités ethniques dans les pays riches en est un excellent exemple : la donnée en elle-même ne permet pas de répondre à la question de

[7] Mark Twain attribuait à Disraeli le fameux : « Lies - damned lies - and statistics »

[8] On connaît depuis fort longtemps le problèmes de représentativité dans les sondages en 1936 le magasine « Literary Digest », qui devint le « Reader Digest », prédit le succès du gouverneur du Kansas Alf Landon malgré un sondage auprès de plus de 10 millions de personnes : les abonnés du magazine…

savoir si la raison est d'abord culturelle ou économique. Les femmes comptent-elles moins de tueurs en série que les hommes pour des raisons culturelles[9] ou biologiques ? La question de la vérité dans les sciences humaines reste entière et malgré quelques succès notoires[10] les statistiques n'y apportent pas de réponses. Aussi, si le *Discours de la Méthode* a indubitablement fertilisé la pensée scientifique et opéré une rupture nécessaire avec la tradition scholastique, n'est-il pas temps de repenser l'analyse des données[11] en fonction de ce qui rend les sciences humaines spécifiques ?

« Ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle; c'est-à-dire, d'éviter soigneusement la précipitation et la prévention, et de ne comprendre rien de plus en mes jugements que ce qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute. » [Descartes 1999] Ainsi s'expriment le doute et la méfiance chez Descartes, qui mènent naturellement au rejet des préjugés qui marque le Siècle des Lumières, au « recours à la raison comme juge unique de tout ce qui existait » [Engel 1880]. Or le rejet des préjugés est illusoire. Il est des préjugés nécessaires à la compréhension, l'esprit vierge de toute connaissance préalable ne peut appréhender quoique ce soit. « Même si cette concrétion particulière de l'explicitation qu'est l'interprétation exacte des textes invoque volontiers ce qu'elle a « sous les yeux », la véritable « donnée première » n'est en réalité rien d'autre que l'opinion préconçue « évidente » et non discutée de l'interprète, opinion nécessairement présente au point de départ de toute interprétation comme ce qui est préalablement « posé », autrement dit prédonné dans une pré-acquisition, une pré-vision et une anti-cipation, dès lors qu'on entreprend en général d'interpréter. » [Heidegger 1927] La question au cœur de l'herméneutique est de rendre compte des préjugés qui guident la compréhension. En effet, il n'est pas toujours facile d'expliquer ce qui est évident : « l'immédiateté du sens littéral procéderait du préjugé aveuglant de la doxa » [Rastier 2001a].

[9] Par exemple, parce que l'on a moins tendance à les soupçonner.

[10] Citons pour mémoire la série de procès aux Etats-Unis entre les années 60 et 80 dans lesquelles des experts prouvèrent sans l'ombre d'un doute que la sélection des jurés était biaisée à tel point en défaveur des personnes de couleur que compte tenu de la composition de la population éligible la probabilité d'obtenir les compositions de jury constatées étaient de l'ordre de $0{,}14\ 10^{-17}$ soit moins que celle d'obtenir trois flush royales consécutives au poker.

[11] Le parallèle entre l'analyse des données et le *Discours de la Méthode* peut sembler osé mais gardons à l'esprit le primat que Descartes apportait aux mathématiques « Je me plaisais surtout aux mathématiques, à cause de la certitude et de l'évidence de leurs raisons : mais je ne remarquais point encore leur vrai usage; et, pensant qu'elles ne servaient qu'aux arts mécaniques, je m'étonnais de ce que leurs fondements étant si fermes et si solides, on n'avait rien bâti dessus de plus relevé : comme au contraire je comparais les écrits des anciens païens qui traitent des mœurs, à des palais fort superbes et fort magnifiques qui n'étaient bâtis que sur du sable et sur de la boue » [Descartes 1999] § 1-10

La question qui se pose à nous est de savoir comment intégrer cette dimension herméneutique nécessaire aux sciences humaines dans la pratique de l'analyse des données. Bien sûr, l'entreprise peut sembler insurmontable. Nous nous attacherons donc dans un premier temps au problème des données textuelles : comment dresse-t-on une conscience interprétative face à un univers de textes, circonscrit certes, mais inconnu dans ses détails ? Comment réintroduit-on une autorité sémantique en coopération avec l'illusion de l'objectivité portée depuis si longtemps par le calcul — ses méthodes, ses modèles et ses techniques ? Certes, une telle conscience aspire à l'action, dont l'application informatique constitue une étape de légitimation inévitable ; mais elle convoque autant l'homme, part oubliée dans une volonté sémantique pérenne — de créer du sens, de recevoir du sens.

2.1 Agrégation des similarités

Notre pratique de l'analyse des données s'inscrit dans une tradition d'agrégation de relations binaires que l'on peut faire remonter à l'analyse des votes de Borda et Condorcet au XVIII et dont le point d'orgue fut le célèbre article du prix Nobel d'économie Kenneth Arrow en 1951 «*Social Choice and Individual Values* » [Arrow 1951]. Quoique que Kenneth Arrow ait conclu à l'impossibilité de trouver une fonction d'agrégation des préférences individuelles respectant un certain nombre de conditions garantissant son équité[12], dans les années 80 Pierre Michaud mit en évidence que l'affaiblissement de l'une de ces conditions, la condition d'indépendance[13], permettait de trouver une telle fonction ouvrant ainsi avec Jean François Marcotorchino la voie à une approche originale de l'analyse des données basée sur l'agrégation des relations binaires sous contraintes. C'est plus particulièrement l'agrégation des relations de similarités comme outil interprétatif qui nous intéressera dans la suite de cet exposé.

[12] Comme le font remarquer Benoît Lengaigne et Nicolas Postel dans [Lengaigne et al. 2004], ce ne sont pas « les hypothèses de l'ordinalisme strict associé à Pareto qui se trouvent être au coeur des interrogations d'Arrow, mais la définition d'une rationalité fondée sur la cohérence et la transitivité des préférences. » Une démonstration rationnelle de la théorie de la main invisible d'Adam Smith affirmant que les actions d'individus rationnels et égoïstes participent plus certainement au bien être de tous que des actions apparemment plus vertueuses et désintéressées.

[13] Pierre Michaud, dans son article « Condorcet, a man of the avant-garde » [Michaud 1987], affirme que Condorcet avait déjà proposé en son temps une approche équivalente qui ne fut malheureusement jamais comprise, analyse confirmée par Pierre Crépel en 1990 « Le dernier mot de Condorcet sur les élections » [Crépel 1990]. Bien que l'on ne puisse savoir aujourd'hui si Arrow aurait considéré que cet affaiblissement de la condition d'indépendance répondait pleinement à ses préoccupations, c'est un point important toujours négligé dans les multiples exégèses du théorème d'impossibilité.

2 Agrégation de similarités et interprétation

Intuitivement une fonction de similarité sur un ensemble d'objets est une fonction qui à deux objets associe une valeur représentant l'intensité de leur ressemblance. Par exemple, on peut présenter deux objets à une personne et lui demander s'ils se ressemblent et noter sa réponse : oui, non ou « sans opinion ». Cette expérience peut se répéter pour chaque paire d'objets et plusieurs personnes. Une fonction de similarité, directement héritée de la théorie des votes, est alors fournie pour chaque paire d'objets par la différence entre le nombre de personnes (de juges) considérant que ces deux objets se ressemblent et le nombre de ceux qui pensent le contraire. De cette fonction de similarité, ou encore d'agrégation des similarités peut être dérivée une nouvelle relation de similarité entre les objets en considérant que deux objets se ressemblent si la fonction d'agrégation des similarités individuelles est positive, qu'ils ne se ressemblent pas si elle est négative et que l'on n'a pu se faire une opinion si elle est nulle.

Cependant l'on attend généralement de cette relation qu'elle soit transitive, c'est-à-dire que si un premier objet ressemble à un second qui lui-même ressemble à un troisième : que ce dernier doive ressembler au premier[14]. Pourquoi ? Parce que dans ce cas la relation collective déduite des relations individuelles est une relation d'équivalence qui opère naturellement une partition de l'ensemble considéré en classes d'équivalence. Hélas cette relation n'est en général pas transitive et le problème auquel se heurtait Condorcet en tentant agréger les préférences individuelles en préférences collectives à savoir la non transitivité de la relation collective[15] se pose pour l'agrégation des similarités, heureusement aux mêmes maux on peut appliquer les mêmes remèdes ou à tout le moins accepter le même compromis : utiliser une relation d'équivalence collective respectant au mieux les relations de similarité individuelles, c'est-à-dire, par exemple, celle qui est soutenue par le plus grand nombre de voix, celle qui maximise les accords ou ce qui revient au même celle qui minimise les désaccords.

On ramène ainsi la recherche d'une opinion collective transitive à la recherche de la solution d'un problème d'optimisation sous contrainte, élégamment formulé par P. Michaud et J.F. Marcotorchino, comme un programme linéaire [Marcotorchino, Michaud 1979]. Malheureusement, la résolution du programme linéaire devient vite impraticable quand le nombre d'objets à classer devient trop important. Or il s'avère que le problème

[14] Notons au passage que l'on n'a pas exigé de nos juges une telle rigueur.

[15] Dans ce cas un premier candidat peut être préféré à un second lui-même préféré à un troisième qui à son tour sera préféré au premier.

d'agrégation des similarités est très proche du problème de recherche d'une partition centrale étudié par Simon Régnier [Régnier 1983] dans les années 60, on consultera à ce propos l'article de J.L. Petit sur la généralisation de la méthode des partitions centrales [Petit 1983]. Simon Régnier, donc, a conçu dans les années 60 un algorithme d'une remarquable simplicité pour trouver un optimum local au problème de la recherche d'une partition centrale et donc au problème d'optimisation sous contrainte de recherche d'une relation de similarité collective qui soit transitive, ouvrant ainsi la voie, moyennant certaines astuces de programmation, à une utilisation sur une très grande échelle de ce type d'approche[16].

En général, l'agrégation des similarités est utilisée pour regrouper les enregistrements d'une base de données en classes homogènes. Pour cela chaque variable est considérée comme un critère donnant un avis sur un sujet particulier pour chaque enregistrement. Ainsi peut-on décrire des félins[17] par certaines caractéristiques morphologiques et éthologiques ; le lion est grand, lourd, chasse de grosses proies et ne monte pas aux arbres ; la panthère est moins grande, moins lourde ; le guépard n'a pas de griffes rétractiles et ainsi de suite. De cette manière le juge « poids » considèrera que le lion et le tigre se ressemblent alors que le critère « type de griffes », pour sa part, isolera le guépard des autres félins. On trouvera dans l'article de J.F. Marcotorchino et P. Michaud précédemment cité le détail des regroupements obtenus.

Malgré son élégance, à l'usage, il faut cependant constater que la méthode d'agrégation des opinions à la majorité des voix ainsi décrite souffre d'une certaine rigidité. Elle ne prend en compte que les variables qualitatives. Elle n'utilise qu'un seul type de similarité, mais surtout, elle tend à favoriser de larges classes peu homogènes.

Diverses solutions ont été proposées pour surmonter les premières difficultés mais par la nature même de sa forme d'agrégation le problème des larges classes peu homogènes semblait être chronique, inhérent à la méthode.

Dans un registre apparemment différent les années 90 ont vu l'explosion de l'utilisation des méthodes à noyaux[18] et des séparateurs à vaste marge[19] à la suite des travaux de

[16] Par exemple la segmentation de bases de données clientèles de plusieurs millions d'enregistrements prenant en compte plusieurs dizaines de variables.

[17] Exemple fétiche de J.F. Marcotorchino et P. Michaud [Marcotorchino, Michaud 1983]

[18] Kernel methods

Vladimir Vapnik sur l'apprentissage automatique. Les fonctions noyau[20] utilisées en analyse des motifs[21] sont en fait des fonctions de similarités d'une grande versatilité. Il n'est, pour s'en convaincre, que de consulter les nombreux ouvrages sur le sujet[22]. Elles permettent de traiter des données qui ne sont pas nécessairement vectorielles ou qualitatives, de les projeter dans des espaces de grande dimension sans avoir à y faire de calculs explicites et enfin de traiter des motifs non linéaires par des algorithmes destinés à des motifs linéaires. Si on utilise un critère de regroupement basé sur la densité des fonctions noyau plutôt que sur leur somme, une légère modification de l'algorithme des transferts permet d'obtenir des partitions dont les densités interclasses sont supérieures à un certain seuil et les densités intraclasses inférieures à ce même seuil. Nous disons *des* partitions car il peut en exister plusieurs (comme il peut d'ailleurs exister plusieurs partitions centrales). Cette approche que nous avons décrit dans [Ho 2007] garantissant une meilleure homogénéité des regroupements, évite les débats byzantins sur les optimums locaux[23]. Il est en effet toujours possible de trouver de telles partitions lorsqu'elles existent. Nous en présenterons une adaptation au problème de la classification de données textuelles dans la section 4.

2.2 Le problème de l'interprétation en analyse des données

Trop souvent les problèmes techniques masquent la démarche interprétative propre aux méthodes statistiques et, particulièrement, à l'analyse des données. Mais qu'entend-on au juste par *analyse des données* ? Laissons la parole à Jean Marie Bouroche et Gilbert Saporta [Bouroche, Saporta 1980]: « Les méthodes d'analyse des données permettent une étude globale des individus et des variables en utilisant généralement des représentations graphiques suggestives. » Ou, un peu plus loin : « La recherche des ressemblances ou des différences entre individus peut être un des objets de l'analyse.»

[19] Support Vector Machines, voir l'excellente présentation d'Antoine Cornuéjols [Cornuéjols 2005] pour une description en Français de qualité

[20] Kernel functions

[21] Pattern Analysis

[22] Par exemple, [Cristianini et al 2004].

[23] On trouvera toujours par la méthode des transferts une partition vérifiant les conditions de densité inter et intra classes même si stricto sensu il existe parmi les partitions considérées certaines qui réalisent une meilleure valeur de la fonction économique.

On voit apparaître au travers des termes employés : « représentations graphiques suggestives », « ressemblance » et « différence » l'objet de notre interrogation, à savoir la dimension interprétative de ce type d'analyse.

Cependant à part quelques définitions formelles concernant la ressemblance et la différence ramenées en général à la notion de distance euclidienne[24] dans des espaces multidimensionnels cette dimension interprétative est peu abordée apparaissant aux auteurs comme « allant de soi ». Bien sûr comme nous l'avons déjà souligné les problèmes de corrélation ou d'échantillonnage sont traités sérieusement mais la démarche interprétative proprement dite se réduit bien souvent à des constatations comme « semble caractériser des téléspectateurs jeunes d'un milieu peu cultivé » ou « on constate une sous-représentation des bacs classiques par rapport à la moyenne nationale ».

Comprenons nous bien, notre propos n'est pas de discréditer de telles analyses qui sont parfaitement justifiables. Il est bien plutôt de tenter d'approfondir les raisons de la méfiance du public à l'égard de méthodes statistiques qu'exprime si bien le constat de H.G. Gadamer sur la statistique qui selon lui « n'est un si remarquable instrument de propagande que parce qu'elle fait parler la langue des faits, simulant ainsi une objectivité qui dépend en réalité de la légitimité de sa manière de poser les problèmes » [Gadamer 1960] § « Le principe de l'histoire de l'efficience » p.141. De notre point de vue la question qui est posée aux données est tout aussi importante que l'interprétation de la réponse rendue. La question englobe tout autant la sélection des données que la façon dont on les code, autrement dit l'interprétation de la réponse est déjà contenue dans la question qui est posée. Dans l'exemple de la classification des félins déjà cité, les données retenues contiennent la réponse que l'on attend. Ainsi, le fait que le guépard se retrouve isolé dans une classe à part tient en particulier à ce que l'on ait retenu la rétractilité des griffes comme un critère pertinent : on s'attend justement à ce que le guépard apparaisse comme un félin singulier. La classification vient renforcer ce « préjugé », elle ne met pas en lumière un fait qui nous serait inconnu, elle le confirme ou dans le meilleur des cas, le met en évidence. L'interprétation de la réponse dans la *langue des faits* sous forme d'une classification dépend bien évidemment de la sélection des données et de leur codage. Une représentation graphique n'est suggestive que si l'on sait au préalable ce qu'elle est censée représenter.

[24] Ou du χ^2 dans le cas des variables qualitatives.

Considérons par exemple la figure 1 ci-dessous. Elle représente la matrice d'une fonction noyau, d'une mesure de similarité donc, entre cinq cents objets. Cette fonction vaut 1 (point noir) lorsqu'elle est appliquée à deux objets identiques et 0 (point blanc) lorsqu'ils n'ont rien en commun ; les valeurs intermédiaires correspondant à des degrés plus ou moins importants de similarité (niveaux de gris). Ainsi la diagonale forme une ligne noire. Ces objets ont été arrangés de telle sorte que la figure « suggère » l'existence de trois classes distinctes :

Une première, très dense regroupant les objets numérotés de 1 à 50,

Une seconde, moins dense regroupant les objets numérotés de 51 à 200,

Une troisième, encore moins dense regroupant le reste.

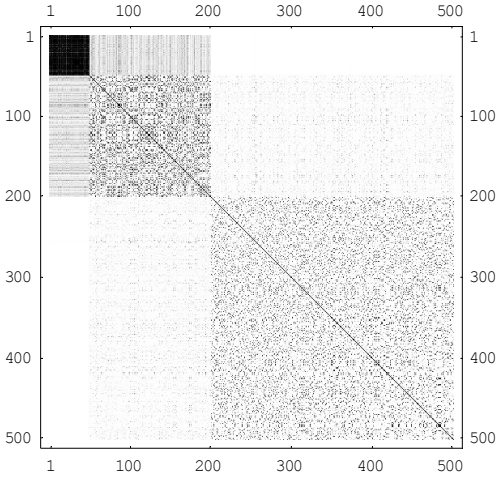

Figure 1: Matrice de Gram d'une fonction noyau

Cette même figure suggère que les objets de la seconde classe entretiennent des rapports avec ceux des deux autres classes (rectangles gris reliant les carrés sur la diagonale). Cependant, sans autre information sur la nature des données (sans connaître la question qui a été posée), il est impossible d'interpréter plus avant cette figure (la réponse à la question). Considérons maintenant cette autre figure représentant des points sur un plan formant grossièrement trois anneaux concentriques. On peut définir une fonction noyau basée sur leur distance euclidienne[25]. On comprend alors que les classes suggérées par la première figure correspondent aux anneaux de la seconde.

[25] On considère ici la fonction qui, à deux points x et y, fait correspondre la valeur $e^{-\|x-y\|}$

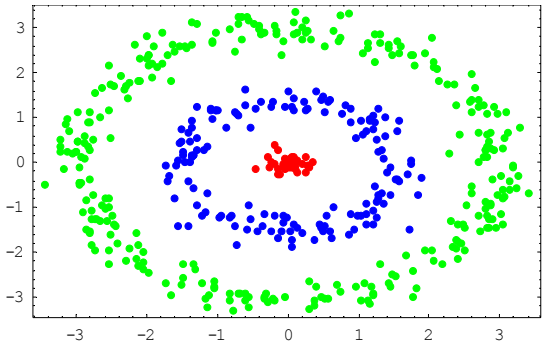

Figure : Anneaux concentriques générés aléatoirement

Nous aurions peut-être pu faire cette interprétation sans voir la seconde figure mais en connaissant la nature des données (des points sur un plan) et la formule de la fonction noyau. Mais l'interprétation de la première figure n'est possible qu'à condition de connaître la nature de ce qu'elle représente.

Cet exemple extrêmement simple va nous servir à illustrer ce que nous entendons par *approche interprétative de l'analyse des données*[26]. Il nous faut tout d'abord préciser les aspects de l'herméneutique philosophique qui nous apparaissent fondamentaux dans tout processus d'analyse des données. Tout d'abord, comme nous l'avons rappelé en introduction, le préjugé[27] contre les préjugés amène à une impasse. Il masque en fait une incapacité à reconnaître les préjugés qui guident notre jugement premier. Il occulte notre nécessaire appartenance à une tradition, à un ensemble de conventions et de règles sur lesquelles s'accorde une communauté scientifique. Ce qui est à redouter n'est pas tant le préjugé qu'ignorer son influence sur nos jugement et nous soumettre ainsi à l'arbitraire d'habitudes de pensées inconscientes. Il est donc essentiel dans tout analyse de tenir compte des « a priori » qui conditionnent notre entendement. L'herméneutique traditionnelle retenait trois éléments de compréhension, trois subtilités : la compréhension, l'interprétation et l'application. L'entreprise de Gadamer a été, dans un premier temps, de montrer l'impossibilité de séparer la compréhension de l'interprétation. Pour lui la compréhension est la compréhension du sens, or le sens n'est pas immanent aux œuvres ou aux choses, il naît de la rencontre du texte et du lecteur, de l'œuvre et du public, des données et de l'analyste dans notre cas. Le sens naissant de l'interprétation, il y a alors danger de *subjectivisation* totale. C'est le second temps de son exposé : le recours au thème

[26] Une herméneutique normative et spécialisée donc.

[27] C'est d'ailleurs lui qui donne à ce terme sa connotation péjorative, nous l'emploierons dans cette dissertation dans son sens herméneutique.

de l'application va permettre de montrer que l'interprète se met au contraire au service du sens.

L'herméneutique juridique où l'application de la loi nécessite son interprétation pour s'adapter aux cas particuliers soumis à l'appréciation du juge en est une excellente illustration. La subjectivité du juge doit être contenue dans un cadre qui assure l'application équitable de la loi pour tous, c'est cet effort qui fonde l'*autorité* de la justice. Le problème de l'application en analyse des données obéit à la même exigence de contenir la subjectivité de l'analyste lors de son interprétation des données. Enfin, comme nous le savons déjà, l'interprétation procède de la structure d'anticipation de la compréhension qui se concrétise dans le cercle herméneutique : la compréhension du particulier découle de celle du tout et celle du tout du particulier dans un mouvement constant de projection et de réajustement. On ne peut comprendre le sens du tout sans comprendre le sens des ses parties mais le sens des parties dépendent du sens du tout qu'elles forment. Ce qui est présupposé de l'objet analysé détermine la compréhension initiale de ses parties mais en retour l'affinement de cette compréhension voire même sa remise en cause modifie ce qui était présupposé au départ et ainsi de suite. Voici donc les trois règles essentielles qui, selon nous, doivent guider toute pratique de l'analyse des données.

H.G. Gadamer débute son ouvrage « Vérité et méthode » [Gadamer 1960] par une longue dissertation sur le concept de jeu, son lien étroit avec l'interprétation artistique et la richesse de ses emplois métaphoriques. Il n'échappera à personne qu'il existe un lien étroit entre l'ordinateur et le jeu. Sans même se référer aux véritables phénomènes d'addiction aux jeux vidéo ou aux jeux en ligne, la nature ludique de l'ordinateur est indubitable. L'analyse des données est une pratique essentiellement liée à l'ordinateur et contrairement à ce que semble penser V.N. Vapnik, elle n'est pas un pis-aller devant la faillite des méthodes statistiques classiques à résoudre le problème de l'inférence inductive [Vapnik 1998]. Au contraire, l'analyse des données, est une redécouverte de la dimension ludique de l'ordinateur par les statisticiens. Elle participe du même principe que la simulation sur ordinateur : le jeu. Au cœur de la notion de jeu se trouve les idées de va-et-vient, de répétition, de simulation, de virtualité et de norme. Le jeu possède un ordre propre qui détermine la conduite des joueurs sans pour autant supprimer leur liberté. Ainsi, lors de la transfiguration en art de leur jeu, les acteurs d'une pièce de théâtre, tout en se devant de respecter le texte de l'œuvre et tenir compte de la direction du metteur en scène, conservent une certaine liberté d'interprétation.

Toutes ces caractéristiques du jeu se retrouvent dans la pratique de l'analyse des données. L'analyse des données possède ses règles propres au travers de ces algorithmes et des paramètres qui les contrôlent. Une même analyse peut se répéter à l'image d'une *représentation théâtrale*. Elle se situe dans le domaine virtuel où le principe d'essais/erreurs sans conséquences réelles permet de *jouer* d'hypothèses différentes jusqu'à l'obtention d'un résultat, d'une interprétation satisfaisante. Cette dimension *ludique* de l'analyse des données ne l'exclut pas des démarches scientifiques sérieuses. Elle met seulement en évidence sa dimension herméneutique. Elle offre un élément de réponse au problème de l'application. En effet, sans totalement brider la liberté de l'analyste, les règles du jeu contraignent son interprétation et offrent surtout les moyens de sa critique : l'analyse sans être objective n'est pas le résultat d'un caprice individuel : elle est une interprétation *guidée par des règles reconnues*.

Si le jeu et ses règles, dans un certain sens, offrent une réponse au problème de l'application, nous voulons maintenant expliciter, de façon plus concrète, l'idée du cercle herméneutique dans le cas de l'agrégation des similarités appliqué à notre exemple des anneaux emboîtés. Sans perte de généralité, nous pouvons supposer qu'au début de l'analyse notre connaissance a priori des données, notre « préjugé », est que l'on analyse un ensemble de points représentant une figure que nous ne connaissons pas : notre objectif est de mettre en évidence les caractéristiques de cette figure. Connaissant les coordonnées de chaque point on peut utiliser la fonction noyau définie précédemment pour rendre compte de la proximité de deux points. L'application de l'algorithme de classification par densité de fonction noyau décrit précédemment permet de calculer des regroupements de points en fonction d'un paramètre de densité. On peut jouer sur ce paramètre pour obtenir des regroupements plus ou moins denses. Par exemple pour une valeur de 0.5 on obtient la classification représentée à la figure 2.

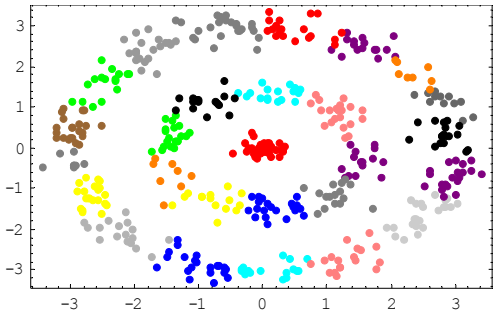

Figure 2: Regroupements obtenus pour une densité de 0.5 de la fonction noyau

Il est difficile de rendre compte de la structure particulière de la figure avec une telle classification. Cependant, notre connaissance a priori suggère que ces classes entretiennent des rapports de proximité quantifiables (détermination des parties par le tout).

La matrice suivante représente ces relations de proximité entre classes : les niveaux de gris correspondant à la densité interclasse de la fonction noyau, c'est-à-dire à la proximité des classes. Les cases de la diagonale représentent les classes, celles en dehors l'intensité des densités interclasses. Plus une case en dehors de la diagonale est sombre, plus la densité interclasse est élevée. Ainsi, par exemple, les classes 2 et 10 sont proches car la case de coordonnée (2,10) est très sombre. On peut utiliser un seuil pour décider de la proximité de deux classes, et obtenir ainsi un graphe de proximité entre classes. Il est alors possible de calculer les composantes connexes de ce graphe : une composante connexe étant un ensemble dont chaque paire de classes peut être joint par un chemin de proximité dans le graphe (détermination du tout par les parties).

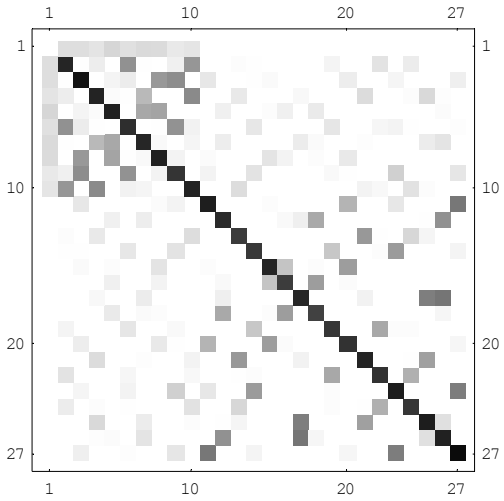

Figure 3: Matrice des densités interclasses de la fonction noyau

Les composantes connexes du graphe de proximité obtenu pour un seuil de 0.2 sont représentées à la figure 4.

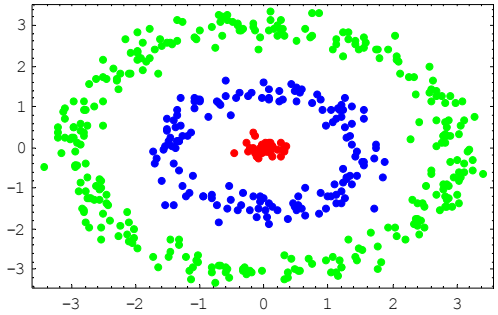

Figure 4: Composantes connexes du graphe de proximité obtenu pour un seuil de densité interclasse de 0.2

En numérotant les points en fonction des classes auxquelles ils appartiennent et de leurs composantes connexes, on obtient la matrice de la fonction noyau présentée à la figure 5 où l'on reconnaît les trois composantes connexes correspondant aux trois anneaux emboîtés de la figure initiale.

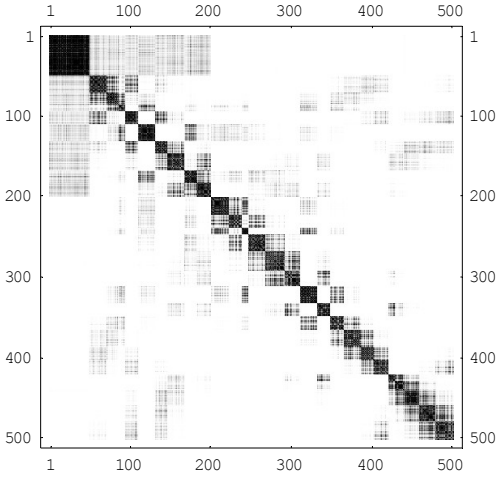

Figure 5: Composantes connexes du graphe de proximité obtenu pour un seuil de densité interclasse de 0.2

Une analyse plus approfondie de la matrice de densité met en évidence la proximité des composantes connexes : la plus centrale étant proche de la composante intermédiaire étant elle-même, quoiqu'avec moins d'intensité, proche de la plus externe.

Dans une certaine mesure notre interprétation des données, guidée par notre connaissance a priori de leur structure et d'une méthode d'agrégation des similarités, rend compte des caractéristiques de la figure telle que l'esprit l'appréhende instantanément.

En dehors des applications que l'on peut imaginer en reconnaissance des formes, notre propos est avant tout de mettre en lumière la démarche herméneutique qui, selon nous, doit guider l'analyse des données. Partant de la connaissance initiale de ce qu'une image est composée de points formant une figure reconnaissable par le jeu de leurs proximités, nous avons identifié des parties homogènes de l'image grâce à un algorithme de classification : la présupposition de la structure globale de l'image a guidé notre identification de ses parties. Nous avons alors identifié les composantes caractéristiques de l'image en considérant les relations de proximité entre ces parties : la connaissance de la nature des relations des parties entre elles nous a permis d'identifier le tout.

Ce qui peut apparaître comme une méthode découle en fait de l'essence même de la compréhension. On ne choisit pas d'entrer dans le cercle herméneutique nous y sommes déjà dans toutes nos pensées. L'interprétation est compulsive on ne décide pas d'interpréter ce qui nous entoure on le fait c'est tout : « Ce qui est décisif, ce n'est pas de sortir du cercle, c'est de s'y engager convenablement. » [Heidegger 1927]. L'appréciation de toute

interprétation est encore une interprétation, rien n'est jamais donné pour certain que ce soit dans les sciences de la nature ou dans les sciences humaines. « Le cercle n'est donc pas de nature formelle, il n'est ni objectif, ni subjectif ; il ne fait que décrire le comprendre comme le jeu complémentaire du mouvement de la tradition et du mouvement de l'interprète » [Gadamer 1960]. Le sens change en fonction des situations et des époques, ce n'est pas un objet dont on dispose, nous sommes englués dans des structures complexes d'interprétations. La nature du sens est historique, la rencontre de l'objet et de l'interprète se répète indéfiniment repoussant toujours plus loin son horizon.

Comme cela a déjà été souligné cette définition mouvante du sens pose le problème de la subjectivité de l'interprétation. Elle semble aboutir à l'impossibilité d'évaluer la pertinence de toute interprétation. La réponse de H.G. Gadamer comme on le sait est de recentrer l'herméneutique sur ce qu'il considère comme son problème fondamental « l'application » afin de rétablir *l'autorité de la tradition*. Comme nous venons de l'esquisser au travers d'un exemple simple notre propos est de nous appuyer sur l'étude du phénomène de compréhension initiée par Heidegger et poursuivie par Gadamer dans son œuvre majeure « Vérité et méthode » pour élaborer une approche herméneutique de l'analyse des données.

Nous préciserons tout d'abord notre projet d'analyse des données et son adéquation au phénomène herméneutique. Nous inspirant de la double médiation entre le réel et l'œuvre d'art d'une part et l'œuvre d'art et sa réception par le public chez Gadamer, nous étudierons la médiation entre le réel et les données lors de leur codage puis entre les données codées et leur réception par l'analyste. Nous montrerons enfin comment se pose le problème de l'application lors de ces deux moments de l'analyse.

2.3 Un projet d'analyse interprétative des données

« Le comprendre en général, n'est pas une connaissance acquise, née d'un acte cognitif, mais un mode d'être *originairement existential* qui rend tout d'abord possible l'acte de connaître et la connaissance. » [Heidegger 1927]. La compréhension précède la connaissance, la compréhension est le mode d'être de l'homme[28] c'est pourquoi la relation entre les parties et le tout et leur détermination réciproque est fondamentale. Alors même qu'elles ne peuvent échapper à cette structure fondamentale de la compréhension les sciences « objectives » en nient la réalité ou au mieux l'ignorent.

[28] Du Dasein

Les statistiques n'échappent pas à cet état de choses : elles considèrent les données qu'elles traitent comme des données objectives auxquelles sont appliqués des traitements fondés sur les mathématiques afin d'inférer des conclusions qui prennent valeur de vérité de par leurs indéniables succès dans la quantification des choses : contrôle de qualité, actuariat, traitement du signal (quoique même le traitement statistique des choses puisse poser des problèmes insolubles, les systèmes chaotiques en étant un exemple. Mais les statistiques sont utilisées pour quantifier d'autres choses que le calibrage des tomates ou la durée de vie des ampoules).

Elles sont utilisées dans d'innombrables activités humaines et c'est particulièrement là que leur déficit herméneutique se fait le plus criant. Les sondages d'opinion sont un exemple édifiant. Il est notoire que les instituts de sondages attachent un soin particulier à constituer des échantillons représentatifs des populations sachant que les réponses aux questions posées dépendent de multiples facteurs où l'âge, la condition sociale, le niveau d'études, la zone de résidence entrent en compte. Mais plus important encore, certaines réponses doivent être corrigées en fonction de considérations empiriques car l'on sait que, par exemple, certaines personnes répugneront à dire la vérité sur des sujets polémiques. On voit ainsi que des considérations et des méthodes qu'il serait téméraire de qualifier d'objectives interfèrent avec la rigueur mathématique[29] dont se prévalent les statistiques. C'est de là que naît la suspicion qui entoure les méthodes statistiques si abruptement formulée par Gadamer (Cf. infra page 17). Cette suspicion est alimentée par le flou méthodologique qui entoure l'application de méthodes réputées rigoureuses à des problèmes dont les données dépendent d'interprétations qui, pour toute aussi rigoureuses qu'elles puissent être, ne peuvent être enfermées dans le formalisme étroit des mathématiques. Le même problème s'est posé à l'intelligence artificielle. En 1986 des chercheurs du laboratoire d'intelligence artificielle du MIT en arrivent à la conclusion que :

> "On one hand, the substantial contributions of logic, mathematics, engineering and the natural sciences like physics, to AI seem to make their strategies for inquiry uncontested. On the other hand, when the subject matter is clearly linked to the concern of human sciences -- particularly linguistics, anthropology, and psychology --scientific methods devised for those areas might be more appropriate."[Mallery, Hurwitz 1987]

Le titre du chapitre était : « L'herméneutique comme une Méta Science ». Si une telle déclaration dans un laboratoire d'intelligence artificielle dans les années 80 rend compte

[29] Notons avec Heidegger que « la mathématique n'est pas plus rigoureuse que l'histoire, elle est seulement plus étroite quant à la sphère des fondements existentiaux dont elle relève. », Être et Temps

des apories de l'approche elle-même, elle propose au travers de ce titre un programme auquel nous souscrivons et que nous voudrions appliquer à l'analyse des données. L'analyse des données, comme souvent d'ailleurs les statistiques dont elle est historiquement issue, est utilisée par l'homme pour mettre en évidence et représenter des structures ou des relations entre des objets qu'il ne peut embrasser par ses seuls sens tant ils sont nombreux et complexes. La généralisation de l'ordinateur et l'accroissement de sa puissance permet aujourd'hui de construire des outils rendant possible une telle entreprise pour des objets toujours plus complexes en nombre toujours plus grand. L'ordinateur, c'est un truisme, a opéré une rupture épistémologique dans le rapport de l'homme au monde, d'une amplitude difficilement mesurable. Il apparaît comme une sorte de miroir de l'esprit et a fait naître des espoirs insensés dont l'IA est l'exemple le plus frappant. Nous savons ce qu'il est advenu des espoirs de l'IA. Malgré tout, l'ordinateur reste un formidable outil de l'esprit.

Heidegger définit l'outil par son maniement : en tant qu'outil de l'esprit, un programme d'analyse de données se définit par son maniement en vue de comprendre un tout représenté par des données que l'homme ne peut immédiatement embrasser. En d'autres termes, le maniement de cet outil, va permettre d'identifier des parties de ce tout en fonction des « a priori » de l'analyste. Le maniement de l'outil exprime ces « a priori » par le codage des données. L'identification de ces parties et des relations qu'elles entretiennent modifie les préjugés de l'analyste qui, par un nouveau maniement de l'outil, peut changer le codage initial des données et modifier par là même les l'identification des parties. Ainsi le maniement de l'outil d'analyse des données va par un jeu de projection et réajustement matérialiser la compréhension des données, matérialiser leur interprétation, donner corps à un cercle herméneutique.

Notre projet d'analyse interprétative des données est de construire un outil *en vue* de donner corps, par son maniement, au cercle herméneutique de l'interprétation des données en fonction de leur codage.

Bien sûr, il peut sembler brutal et réducteur de vouloir contraindre le cercle de la compréhension à se cantonner dans le cadre étroit de l'utilisation d'un programme d'ordinateur. Tout d'abord il ne s'agit pas là du cercle herméneutique en général mais plutôt de celui qui entre en jeu précisément dans l'utilisation d'un ordinateur pour interpréter des données digitales. La réduction a donc eu lieu en amont, elle est celle qui

advient nécessairement lors de toute représentation du réel par l'homme, que ce soit par la digitalisation sur ordinateur ou par des dessins d'animaux sur les parois d'une grotte. En second lieu, il ne s'agit pas et il ne peut s'agir de traduire sous forme de listes d'octets le processus complexe du phénomène herméneutique. Il s'agit plutôt d'en fixer les étapes importantes et d'en garder la trace. Enfin, comme cela a déjà été évoqué, nous avons l'ambition de remettre au cœur du processus d'analyse des données le problème de l'application herméneutique, i.e le problème de la validité de l'interprétation. Dans le cas qui nous intéresse, une interprétation est valide si le codage des données s'accorde avec une tradition et si les parties identifiées correspondent à une attente ou mettent en valeur des relations significatives. Bien que l'exemple que nous avons proposé permette de s'en faire une idée, la définition à ce point de l'exposé reste vague mais elle prendra plus de relief lorsque nous préciserons ce que l'on entend par codage et comment s'effectue l'identification des parties.

2.4 Le codage comme médiation avec le réel

L'analyse de l'expérience de l'art dans l'herméneutique de Gadamer a été souvent soulignée [Gadamer 1960]. Une pièce de théâtre par exemple s'articule autour d'une double médiation : une première médiation entre le réel et l'œuvre et une seconde entre l'œuvre et son exécution. La première médiation n'est pas une simple imitation du réel mais une mise en valeur de son essence, de ce que l'artiste veut que l'on reconnaisse : « dans la reconnaissance, ce que nous reconnaissons échappe pour ainsi dire, comme par illumination, à tout hasard et à toute variation des circonstances qui le conditionnent et se laisse saisir dans son essence. Il vient à être connu en tant que telle chose. » [Zarader 2005] Représenter l'essence, c'est d'une part soustraire ce qui est représenté au réel et pour le rendre à sa vérité propre mais c'est, d'autre part, accentuer certains de ses traits et en abandonner d'autres. Il y a là plus que de l'imitation du réel mais ce que Gadamer appelle une transmutation en art du réel donnant à l'œuvre son autonomie en vue d'apporter une véritable connaissance. Ce qui nous amène à la seconde médiation entre l'œuvre et son exécution. L'œuvre doit être exécutée, interprétée pour pouvoir exister et de ce fait l'interprétation de l'œuvre fait partie intégrante de l'œuvre elle-même : l'œuvre s'assimile à l'ensemble de ces représentations. Chaque nouvelle mise en scène est une réalisation de l'oeuvre, elle n'atteint son plein accomplissement que chaque fois qu'elle est jouée : « ce qui est imité dans l'imitation, créé par l'écrivain, représenté par l'exécutant, reconnu par le

spectateur, c'est tellement la chose visée, celle sur laquelle repose la signification de la représentation » [Gadamer 1960].

Mais si l'œuvre s'assimile à l'ensemble de ces représentations toute représentation n'est pas légitime, la liberté du metteur en scène et du jeu des acteurs est au service de l'œuvre, elle n'est pas l'expression d'une simple fantaisie même si il n'est pas toujours facile de rendre compte des critères d'appréciation de la justesse d'une interprétation.

On voit là se profiler la nécessité de remettre le problème de l'application au centre de l'herméneutique. Mais ce n'est pas ce qui nous préoccupe ici. Ce qui nous intéresse plus particulièrement dans l'immédiat c'est le parallèle qui nous semble possible, voire nécessaire, entre la double médiation décrite dans l'expérience de l'art et l'analyse des données. Il n'est pas question d'assimiler une analyse de données à une expérience artistique mais de façon plus prosaïque, d'affirmer qu'il existe bien une première médiation entre le réel et le codage des données et une seconde médiation entre les données codées et leur analyse, leur interprétation.

Qu'entend-on par données dans l'expression analyse des données ?

La donnée est un terme vague recouvrant diverses structures digitales, allant de l'image numérisée aux chaînes de caractères unicode représentant des textes, en passant par les enregistrements d'une base de données commerciale. Les données forment donc une réalité virtuelle, le codage de ces données est une première médiation entre cette réalité et sa représentation en vue de son analyse : ce peut être dans le cas des images digitales les résultats d'applications de filtres en vue de formaliser des informations comme la luminance, le contraste, la couleur ou le taux d'acuité. En ce qui concerne les textes numérisés, ce peut être des fréquences de trigrammes ou de mots, des collocations, des catégories grammaticales, des sèmes ou des structures plus complexes comme des molécules sémiques et des isotopies comme nous le verrons plus loin. Les enregistrements d'une base de données commerciale pourront quant à eux être discrétisés en variables qualitatives mettant en relief des informations considérées comme d'un intérêt plus particulier. Ce qu'il faut retenir c'est que cette première médiation du codage des données vise, à l'image de l'action de l'artiste, à accentuer certains traits de la réalité ou à en abandonner d'autres en fonction de ce qui veut être mis en valeur, en fonction des préjugés initiaux de l'analyste.

Pour être plus prosaïque encore imaginons le cas d'un analyste cherchant à identifier à partir de la base clientèle d'une banque les clients plus susceptibles d'acheter une carte de crédit. La tradition qui le guide, l'ensemble de ses préjugés sur ce problème donc, est celle de la profession de banquier. Cette tradition lui permet d'identifier dans le fichier client les variables qui comptent pour identifier parmi ses clients ceux à qui il veut s'adresser : ce peut être l'âge, le sexe, l'ancienneté en tant que client, le taux d'endettement, le revenu, le patrimoine, le nombre de produits bancaires déjà possédés, la domiciliation du salaire, le nombre de mouvements sur le compte courant, les hypothèques, la profession, le statut marital, le nombre d'enfants à charge, le niveau d'épargne, les agios payés et d'autres paramètres encore. Il est souvent important de tenir compte de l'historique de certaines de ces variables par exemple la tendance à la hausse ou à la baisse du taux d'endettement, l'arrivée à échéance des prêts à la consommation ou des hypothèques. D'autre part, cette même tradition considère que l'on est jeune lorsque l'on a moins de trente ans quelque soit la distribution réelle des ages des clients de la banque, de même l'importance de l'endettement n'est pas un fonction de l'écart type à l'endettement moyen mais la fonction d'un seuil fixé dépendant d'un compromis entre la politique commerciale de la banque, des lois en vigueur dans le pays où se situe l'agence du client et de décisions stratégiques de la direction générale.

Non seulement la base clientèle peut être volumineuse mais un très grand nombre de variables entrent en jeu ainsi que de multiples paramètres. Sans le recours à la tradition il serait à peu près impossible de faire un choix correct des variables à prendre en compte et de leur codification. À titre personnel nous avons eu souvent l'occasion de voir de jeunes statisticiens échouer complètement dans l'analyse de bases clientèles pour ne pas avoir su se mettre à l'écoute de la voix de la tradition. L'analyse interprétative des données accorde-t-elle une importance fondamentale à cette première médiation entre la réalité virtuelle et la codification des données au travers de laquelle s'exprime la tradition qui permet alors une précompréhension de cette réalité dans le cadre du projet justifiant l'analyse.

2.5 La similarité comme critère d'identification des parties

Dans l'expérience artistique la seconde médiation lie l'œuvre à son exécution à son interprétation. La seconde médiation en analyse des données correspond à l'identification des parties en fonction du codage des données. Cette identification se fait sur la base d'une

similarité. Comme cela est décrit dans le paragraphe 2.1, l'outil d'analyse des données permet, sur la base de cette similarité, d'agréger les données en groupes homogènes, d'identifier les parties du tout. L'analyse de ces parties et des relations qu'elles entretiennent permet d'identifier leur adéquation aux hypothèses originales ayant présidées au codage des données ou, au contraire, en cas d'inadéquation, de dégager ce qu'il faut remettre en question dans ce codage ou dans la façon dont sont définies les similarités individuelles et donc de procéder à des réajustements voire à la remise en cause de préjugés ayant présidé au codages des données. La particularité des parties identifiées dans l'outil que nous décrivons est qu'elles forment une partition c'est-à-dire que tout élément du tout appartient à une et seule de ces parties. L'analyse d'une partition se fait principalement selon deux axes : l'analyse intra classe et l'analyse inter classe.

L'analyse intra classe permet de comprendre ce qui, dans la codification des données, a provoqué le regroupement et que l'on désigne sous le terme générique de « thème ». Deux indicateurs caractérisent les thèmes : la spécificité et l'intensité. La spécificité indique le pouvoir de caractérisation de la classe par le thème :

- Un thème spécifique à une classe caractérise cette classe par rapport au tout. Il est un élément de contraste. Il affecte peu d'éléments en dehors de la classe.
- L'intensité d'un thème marque son importance dans la classe. Il affecte la plupart des éléments de la classe.

Un thème intense et spécifique est caractéristique de la classe, un thème intense et peu spécifique affecte à peu près tous les éléments considérés (et est souvent comme tel d'une trop grande généralité), un thème spécifique et peu intense peut suggérer une particularité intéressante de la classe.

L'analyse interclasse reprend les notions de centralité et de densité propres à la méthode des mots associés[30] :

- La densité reflète la cohérence du regroupement considéré. Une classe est dense si les éléments regroupés en son sein sont fortement similaires.
- La centralité reflète le pouvoir de liaison du regroupement considéré. Une classe est centrale si les éléments regroupés en son sein ressemblent à des éléments qui lui sont extérieurs.

[30] Leximappe du CSI Ecole des Mines et CDST du CNRS.

Les classes denses et centrales sont fédératrices, les éléments qu'elles contiennent reflètent plus particulièrement le tout, le point de vue ayant guidé la codification des données. Elles mettent en évidence des thèmes diffus dans le tout (centralité) et stables (densité). Les classes denses mais peu centrales isolent les éléments qu'elles contiennent, les thèmes qui les caractérisent sont stables mais peu diffus, elles regroupent généralement des éléments de peu d'intérêt. Les classes centrales mais peu denses font ressortir des thèmes diffus mais peu fédérateurs. Enfin les classes ni denses, ni centrales correspondent à des thèmes émergents.

À partir de ces considérations simples il est possible de jouer de la codification des données et du seuil de densité de similarité pour identifier des parties en accord avec le tout. Tout d'abord l'analyse interclasse permet de rapidement se faire une idée de la qualité de la classification obtenue : en effet une classification trop centrale et peu dense suggère que le seuil de densité n'est pas suffisamment élevé, à l'opposé une classification trop dense et peu centrale suggère le contraire. L'impossibilité de trouver un seuil adéquat suggère que la codification des données est trop fine ou trop grossière ne produisant que des thèmes trop ou trop peu spécifiques. D'une façon générale les thèmes intenses sont à l'origine des regroupements. Lorsqu'ils ne sont pas suffisamment spécifiques ils indiquent que le codage des données leur correspondant est trop général ou peu précis. Les thèmes manquant systématiquement d'intensité correspondent à des données au codage trop peu discriminant ou à des données anecdotiques quant au thème général. L'analyse des densités interclasses permet de mettre en évidence des graphes de relations entre parties dont les composantes connexes peuvent mettre en évidence des parties plus vastes de moindre densité mais correspondant à une structuration particulière des données. Il est enfin nécessaire de toujours vérifier que les parties identifiées correspondent à une structuration, une interprétation authentique du tout. Ce qu'il faut entendre par là est que l'interprétation du tout doit toujours s'inscrire dans la tradition de l'interprète, doit toujours répondre à une attente. Il est, en effet, possible d'obtenir une structure formelle intéressante, par exemple un ensemble de classes denses décrites par des thèmes spécifiques et intenses sans que l'on puisse en faire une interprétation cohérente. Par exemple, une classification faite il y a quelques années déjà d'articles du journal « Le Monde » basée sur une sélection de mots clefs discriminants[31] lemmatisés mettait en évidence une classe dense et peu centrale dont les thèmes intenses et spécifiques étaient « film », « chanson », « musique », etc. donc

[31] Sélection manuelle d'une liste de mots triés sur la base de leurs fréquences de document inverse

visiblement une classe consacrée aux arts et spectacles. Elle contenait des articles consacrés à Robert Badinter. Après examen il est apparu que Robert Badinter ne s'était pas reconverti dans le music-hall mais qu'il était fait référence dans une série d'articles à un discours où il parlait de la « chanson de la droite » et de la « petite musique qu'on nous sert ».

Quand les données sont très complexes comme le sont les textes, les analyses foisonnent de ce type de contre sens et dans une certaine mesure sont inévitables. Il se peut même que le type de codage ne permette pas d'éviter ce genre de problèmes, le cas cité en exemple en est d'ailleurs une parfaite illustration. Même la prise en compte de contexte locaux ne permet pas d'éliminer les termes « chanson » et « musique » des articles incriminés et seule une intervention manuelle pourra le faire, ce qui par ailleurs renforce notre conviction que l'interprétation est l'apanage de l'esprit et que l'efficacité de l'outil réside en grande part dans l'habileté de celui qui le manie. Dans d'autres cas c'est la nature des données elle-même qui peut masquer certains phénomènes, ainsi en général, l'étude du fichier client d'une banque met en évidence une vaste classe de clients inactifs : compte courant vide ou presque, peu ou pas de transactions, peu ou pas de produits bancaires, pas de domiciliation de salaire, etc. or généralement les nouveaux clients de la banque qui sont d'une grande importance commerciale se retrouvent dans cette classe dont l'étude est délaissée. Il se peut aussi que certains phénomènes soient mal interprétés, l'acquisition de nouveaux établissements peut ainsi perturber l'étude des phénomènes d'attrition de clientèle. L'analyse de données est principalement un exercice interprétatif qui ne peut que s'appuyer sur un ensemble de préjugés, la tradition. L'analyse procède par un constant mouvement de projection et de réajustement à l'image même du processus d'interprétation dans la tradition herméneutique. Ce retour sans cesse renouvelé aux « a priori » qui ont déterminé le sens de l'analyse est la condition de la justesse de l'analyse, l'interprétation des données ne fait autorité que si elle s'inscrit dans la tradition qui la sous-tend. C'est là le sens de l'application herméneutique.

2.6 L'analyse des données comme système de programmation

La prise en considération du cercle herméneutique dans l'analyse des données pourra sembler triviale et superfétatoire à certains, pourtant il est à la base d'une vaste réflexion dans les sciences humaines, animée d'un souci de rigueur de la pensée. Il n'est cependant

jamais évoqué dans les manuels traitant des statistiques et de l'analyse des données qui laissent dès lors à penser que l'interprétation va de soi et que tout scientifique digne de ce nom aborde les données en parfaite objectivité, l'esprit vierge de tout préjugé. Elle pourra sembler maladroite à d'autres voire totalement réductrice. Pourtant les statistiques et l'analyse des données sont d'un apport capital dans toute entreprise de quantification de multiples phénomènes n'obéissant pas aux lois de la nature. Ces disciplines ne sont pourtant que très rarement associées[32] à la réflexion dans les sciences humaines ou bien l'ont justement été dans des démarches douteuses.

La difficulté de penser les statistiques dans les sciences provient de ce que les statistiques, en tant que telles, sont d'abord un formalisme largement basé sur la théorie des probabilités. Trouver comment la matière des sciences prend forme dans les statistiques est le réel problème. Formaliser une tradition dans un cadre statistique, c'est devoir surmonter de multiples difficultés : quels sont les phénomènes que l'on veut ou que l'on peut mesurer ? Que signifient ces mesures ? Que peut-on en inférer ? Quelles hypothèses peut-on formuler ? Dans quelle mesure peut-on s'y fier ? Autant de questions auxquelles il n'existe pas de réponses simples. Boltzmann a eu les plus grandes difficultés à faire reconnaître ses travaux statistiques sur la théorie cinétique des gaz car il basait sa théorie sur l'existence des atomes qui n'était pas reconnue par la communauté scientifique de son temps.

Ce sont ces considérations qui nous ont conforté dans le choix de la similarité comme le concept de base au cœur d'un outil d'analyse des données. En effet, la similarité est une notion simple comprise par tout un chacun, elle s'applique à une quantité de phénomènes. Comme nous l'avons vu elle permet de calculer des regroupements d'objets sur lesquels elle opère et de ce fait renvoie naturellement à la dialectique entre le tout et ces parties soulignée par l'herméneutique. Mais au-delà de ce thème déjà amplement développé il existe un autre aspect que nous avons peu abordé qui fait de la similarité, et plus précisément des fonctions noyau, un concept extrêmement important complétant l'analyse des données proprement dite. Les fonctions noyau sont en fait des produits scalaires dans des espaces dont on ne connaît pas nécessairement la structure où sont définis les objets analysés. Cette propriété remarquable permet d'appliquer à des objets dont la nature pourrait ne pas sembler s'y prêter de prime abord des notions empruntées à la géométrie

[32] A l'exception notable du *« Centre d'analyse et de mathématiques sociales de l'Ecole des hautes études en sciences sociales »*

comme celles de distance, d'hyperplan séparateur, d'orthogonalité. C'est d'ailleurs cette propriété qui les a rendues si populaires dans les années 90 car elle permet d'utiliser les techniques de séparateur à vaste marge ou de régression à des entités aussi inattendues que des textes ou des images. Cette propriété vient donc en complément de la capacité de regroupement en parties homogènes en ce sens qu'une fois un choix de parties représentatives du tout arrêté, il est possible d'instruire la machine à ranger de nouveaux objets de nature similaire dans les parties qui leur correspondent au mieux : c'est ce que l'on appelle l'apprentissage statistique [Vapnik 1998].

On voit ainsi se dessiner l'image complète de notre idée d'un outil d'analyse des données interprétatif : dans un premier temps, en fonction d'un projet s'exprimant au travers d'une fonction de similarité et d'un codage approprié, un ensemble représentatif de données est analysé selon la méthode suggérée au travers de ces lignes afin d'en identifier les parties représentatives et les relations qu'elles entretiennent. Dans un second temps on instruit l'outil à reproduire cette analyse sur des objets de même nature. La première phase peut ainsi être considérée comme une phase d'amorçage[33], la seconde comme une phase d'apprentissage. Ce n'est pas, loin s'en faut, de l'intelligence artificielle mais une forme de programmation informelle basée sur l'analyse interprétative des données et sur les techniques statistiques d'inférence inductive. L'avantage d'une telle approche est donc de pouvoir en quelque sorte enregistrer une interprétation et de pouvoir la reproduire. Ce type d'interprétation bien qu'enfermé dans le cadre étroit de l'identification des parties d'un tout sur la base d'une fonction de similarité ne peut être l'œuvre que de l'esprit. Il ne peut donc être mécanisable.

L'indexation de textes nous semble particulièrement adaptée à cette approche. Les documents textuels sont à l'heure actuelle soit indexés par les mots qu'ils contiennent, soit interprétés par des indexeurs humains qui leur apposent des mots-clefs extraits de listes d'autorités. Dans un cas l'indexation est rapide, peu coûteuse mais souffre de limites inhérentes à l'extrême simplification des textes qui se réduisent à des sacs de mots. Dans l'autre, l'indexation est lente, coûteuse, peu régulière et l'interrogation du fonds nécessite de connaître la liste d'autorité dont sont issus les mots-clefs.

Nous proposons dans la section suivante une approche de l'indexation de textes utilisant une analyse des données textuelles se fondant sur la sémantique interprétative [Rastier

[33] bootstrap en Anglais.

1987]. Le codage des données textuelles se basera sur la notion d'isotopie qui sera utilisée pour définir des fonctions de similarités entre textes. Nous utiliserons d'autres fonctions noyau entre textes principalement pour accélérer la recherche du vocabulaire isotopant ou pour rendre la phase d'apprentissage plus effective, nous nous intéresserons en particulier aux possibilités offertes par les isotopies pour contrôler les axes principaux de systèmes d'indexation sémantique latente dans les phases d'interrogation afin de tenir compte du vocabulaire peu fréquent.

3 Indexation et sémantique interprétative

La sémantique interprétative de François Rastier est la tradition dans laquelle nous nous inscrivons pour l'étude des textes. Après une longue section consacrée à l'herméneutique philosophique une telle proclamation peut sembler étrange voire provocante quand on sait la méfiance de François Rastier à l'endroit de l'herméneutique philosophique : « En outre, l'herméneutique philosophique contemporaine s'est constituée par une dénégation des sciences du langage, dont témoigne l'oubli de Humbolt par Dilthey et le mépris des sciences en général par les heideggériens. » [Rastier 2001a] Il convient donc que nous nous expliquions rapidement sur ce point. Notre propos dans la précédente section était de mettre en évidence le déficit herméneutique dans les sciences statistiques et de proposer une démarche interprétative dans la pratique de l'analyse des données en général. La sémantique interprétative s'intéresse aux textes en particulier et participe plutôt d'une herméneutique matérielle dont l'ambition est de rapprocher linguistique et herméneutique. Il est d'ailleurs important de noter la remarque de F. Rastier : « d'autre part l'herméneutique philosophique, émancipée de certaines formes d'irrationalisme, pourrait devenir une philosophie des sciences du langage » [Abeillé et al 1994]. Nous pensons donc simplement que l'indexation des textes peut se comprendre comme une pratique de l'analyse des données textuelles, une articulation entre analyse des données et informatique linguistique. Comme nous avons fait le choix de l'agrégation des similarités pour l'analyse des données nous faisons le choix de la sémantique interprétative pour l'informatique linguistique. Ainsi, autant l'oeuvre de H.G. Gadamer nous aura inspiré dans l'étude du phénomène de la compréhension dans les sciences statistiques autant l'œuvre de F. Rastier nous inspire dans l'étude du phénomène de la compréhension des textes. Nous ne voyons aucune contradiction dans cette articulation mais au contraire une complémentarité naturelle.

3.1 L'indexation sujet aujourd'hui

L'objet de l'indexation est le document plutôt que le texte ; un document peut être un texte bien sûr mais aussi une image, un film et bien d'autres choses encore. Pour citer Suzanne Briet, bibliographe Française de grande renommée : « un document est une preuve à l'appui d'un fait » [Briet 1951]. Les documents sont cependant généralement référencés par

des textes : le titre d'une image, le synopsis d'un film, le catalogue d'une exposition, la critique d'une œuvre musicale ou bien encore les discussions au sujet d'un reportage dans un forum Internet. Le texte apparaît donc comme un objet central du monde documentaire, les problèmes que pose son indexation sont d'actualité. Les textes restent, du point de vue de l'automatisation de leur traitement, d'un abord plus accessible que la plupart des autres documents tout en conservant la puissance expressive inégalée de la langue. Les textes peuvent se représenter électroniquement sous forme de chaînes de caractères et sans vouloir revenir aux prédictions enflammées des débuts du traitement automatique de la langue (TAL), c'est un avantage indéniable, un filtrage considérable de l'information qui peut être traitée automatiquement. Cette caractéristique du texte électronique est d'ailleurs abondamment employée par les moteurs de recherche sur Internet elle n'est cependant peu ou pas utilisée par l'indexation traditionnelle pour de multiples raisons. Sans prétendre l'exhaustivité en voici quelques unes :

- L'indexation documentaire moderne est apparue et s'est structurée bien avant la venue de l'ordinateur,
- La défiance à l'endroit de la langue de la terminologie, préoccupée comme on le sait, d'universalisme dès ses origines,
- La volonté de pouvoir indexer selon une même méthodologie toutes sortes de documents.

Mais quelle est cette méthodologie ? Dans sa thèse Jens-Erik Mai [Mai 2000] résume en terme d'objets sémiotiques et d'activités d'interprétations le processus de l'indexation sujet tel qu'il est décrit dans différent documents : recommandations ISO 5963 et Codification Dewey. Il recense quatre objets :

1. Le document,
2. Le sujet,
3. La description du sujet,
4. L'indexation du sujet.

Ces objets sont d'une complexité décroissante : le document étant évidemment l'objet le plus complexe. Ils sont reliés par trois activités :

1. L'analyse du document qui, à partir du document, produit le sujet,
2. La description du sujet qui, à partir du sujet, en produit la description,

3. Enfin l'analyse du sujet produit son indexation à partir de sa description.

Or le document est un objet culturel que l'indexeur ne peut consulter en faisant abstraction de son éducation, de ses goûts, de sa sensibilité, bref de sa condition humaine. La prétention à l'objectivité et à la neutralité des recommandations ISO et CDD apparaît donc pour ce qu'elle est : une mystification destinée à dissimuler les préjugés positivistes qui prédominent dans les sciences de l'information. J.E. Mai plaide donc pour une approche interprétative de l'indexation qui permettrait de recentrer les problématiques documentaires sur la personne[34]. Il se fonde sur l'idée d' « image du monde » du Tractatus de Wittgenstein [Wittgenstein 1921] pour assurer la stabilité de l'indexation considérant que les individus d'une même pratique sociale partagent la même image du monde et par la même interprètent similairement les documents relatifs à cette pratique. En résumé, J.E. Mai ne remet pas en cause le processus de l'indexation décrit dans les recommandations ISO et CDD, mais conscient de l'inanité du principe de la séparation du sujet de l'objet dans les sciences humaines, il le resitue dans une perspective interprétative et en minimise la subjectivité par la prise en compte de normes sociales. A ce niveau de généralité peu a été dit et ce qui a été dit vaut pour tout type de documents.

En 1968, Patrick Wilson de l'Université de Californie à Berkeley, posa les principes utilisés aujourd'hui encore pour la détection du sujet d'un texte ; ces principes sont guidés par quatre questions :

1. Quel à été le but de l'auteur au moment de la rédaction de l'ouvrage ?
2. Quels sont les aspects saillants du texte ?
3. Quels sont les termes les plus utilisés dans le texte ?
4. Quel est le principe unificateur du texte ?

Chacune de ces questions permet, en un certain sens, de définir le sujet d'un texte[35] ; rien ne garantit cependant qu'elles mènent au même résultat : Wilson en conclut que le sujet d'un texte ne peut être unique et déjà se profile une *herméneutique de l'indexation*. Transposé dans le cadre de l'indexation, l'approche classique de l'herméneutique suppose de la part de l'indexeur une compréhension littérale du texte permettant son interprétation

[34] Il constate qu'il existe une abondante littérature sur les index, les catalogues et les documents mais que peu est dit sur comment l'on doit indexer.

[35] Comme c'est généralement le cas à son époque Wilson considère le texte comme un système fermé.

en vue de son indexation. Il est donc fondamental de bien définir les buts de l'indexation pour asseoir l'interprétation des textes. Hanne Albrechtsen a conçu un cadre général de discussion de l'analyse du sujet qui est résumé dans le tableau ci-après[36] :

Type d'analyse du sujet	**Type d'information**	**Méthode d'indexation**
Simple	Explicite	Extraction
Dépendant du contenu	Implicite	Affectation
Dépendant des besoins de l'utilisateur	Pragmatique	Affectation

1. Quand le type d'analyse est simple, le type d'information est explicitement contenu dans le texte et le processus d'indexation consiste à l'extraire du texte : le cas extrême étant l'indexation automatique basée sur des méthodes statistiques ne nécessitant aucune interprétation.
2. Quand le type d'analyse dépend du contenu l'analyse du document suppose une identification des sujets qui sont implicitement contenus dans le texte qui ne peut être interprété que par un indexeur humain dans le but de le décrire.
3. Quand les besoins des utilisateurs doivent en sus être pris en compte l'indexeur doit prendre en compte l'information pragmatique contenue dans le document, son interprétation doit prendre en compte les besoins des utilisateurs potentiels.

Le coût et la vitesse de l'indexation dépendent lourdement de la phase d'interprétation car elle nécessite une intervention intelligente. Cette même intervention est aussi source d'une grande variabilité comme l'on sait. Dès lors les enjeux sont clairs : quels sont les outils et méthodes permettant d'une part d'accélérer et de faciliter l'interprétation des textes et d'autre part réduire la variabilité de l'indexation ? La thèse défendue ici soutient que la sémantique interprétative des textes fournit le cadre théorique du développement d'outils facilitant et accélérant l'interprétation des textes et que l'analyse et l'indexation de corpus plutôt que de textes isolés tout en permettant le traitement simultané d'une grande quantité de documents réduit considérablement la variabilité de l'indexation.

[36]Cf. J,E. Mai [Mai 2000]

3.2 L'hégémonie ontologique vouée aux gémonies

Ce que l'on appelle positivisme ne correspond pas toujours aux idées d'Auguste Comte, il conviendrait peut-être mieux de parler de scientisme. Toujours est-il que la foi dans le progrès, dans l'idée que seule la science fait reculer l'ignorance, libère l'homme de la superstition et le guide vers la clarté reste aujourd'hui encore une croyance largement partagée. Le positivisme dans les sciences humaines, et le terme même est significatif, se traduit par un ensemble de croyances dont il n'est pas inutile de rappeler les lignes de force :

1. Les activités humaines peuvent se réduire à un ensemble de faits gouvernés par des lois générales ; l'objectif des sciences humaines est de découvrir ces lois.
2. L'épistémologie des sciences naturelles et physiques doit gouverner la recherche de ces lois, il est en particulier possible de séparer le sujet et l'objet de leur étude.
3. L'étude de phénomènes complexes peut se ramener à l'étude des interactions de leurs composantes essentielles.

Dans l'entre-deux guerre le positivisme logique du Cercle de Vienne est l'aboutissement de cette démarche. Il se caractérise par un empirisme qui renonce au pourquoi et fait prévaloir le comment. Il se traduit dans l'étude du langage par l'idée que les mots sont des atomes qui désignent des objets du monde réel obéissant à des lois qui sont le reflet de la logique formelle. Cette vision scolastique du langage est toujours celle qui prévaut en IA et dans le TAL (Traitement Automatique des Langues). Elle participe de la même défiance Platonicienne à l'égard des langues trompeuses, propices aux ambiguïtés et aux paradoxes. Le langage idéal est celui des mathématiques : l'espéranto de l'esprit. Dans le langage idéal, les mots à l'image des nombres sont de simples symboles désignant de manière univoques les objets du monde et que l'on peut assembler suivant des règles de production formelles pour former des propositions qui ont un sens en ce que l'ont peut décider si elles sont vraies ou fausses. Le *Tractatus Logico-Philisophicus* de Ludwig Wittgenstein est l'aboutissement de cette pensée, son exposé définitif. Le *Tractatus* définit le monde au travers des faits et de la logique : « les faits dans l'espace logique sont le monde »[37], et affirme : « les frontières de mon langage sont les frontières de mon monde ». Le Tractatus eut un impact considérable et le positivisme logique semblait définitivement consacré envoyant aux oubliettes de l'Histoire les pauvres conjectures métaphysiques d'autrefois. Un peu à l'image de ce qu'il advint du programme d'Hilbert et la preuve que son dixième problème était insoluble, c'est de l'intérieur que vint la contradiction. Wittgenstein lui-

[37] Tractatus par. 1.13

même remit en cause le Tractatus dans ses *Recherche philosophiques* [Wittgenstein 1953] constatant la plasticité des langues et notre capacité à déterminer le sens des mots en fonction de leurs contextes : « tout signe isolé paraît mort. Qu'est-ce qui lui donne vie ? C'est dans l'usage qu'il est vivant. A-t-il en lui-même le souffle de la vie ? Ou l'usage est-il son souffle ? »[38]. Pour Wittgenstein les enfants apprennent le langage au travers de jeux : les jeux de langages au cours desquels ils sont dressés à comprendre : « Ici l'enseignement du langage n'est pas une explication mais un dressage »[39]. Les enfants au cours de ces jeux partagent les mêmes objectifs et les mêmes intérêts que leurs maîtres et c'est ainsi qu'ils apprennent leur langue non pas parce qu'ils apprennent des règles mais parce que l'élève et le maître partagent un contexte. Nous sommes si impliqués dans ces jeux de langages que nous ne pouvons nous en tenir à l'écart pour en étudier les principes, il est vain de vouloir séparer le sujet et l'objet de l'étude, le principe épistémologique des sciences objectives est inapplicable. Le modèle positiviste du langage est ruiné car tout système prétendant décrire le langage doit se tenir en dehors du langage pour le considérer objectivement ce qui est impossible car ce système doit être décrit par le langage.

Le positivisme logique qui s'est vu récusé par l'un de ses plus fameux représentants perdure pourtant au travers des systèmes formels et des ontologies qui en sont les avatars. Ce succès tient en partie à la tradition objectiviste multiséculaire de la grammaire et en partie à l'influence de l'informatique et plus particulièrement de la théorie des langages. Dans un programme informatique les mots sont des tokens pouvant être des variables ou des mots-clés du langage (*if, then, else, while, for*, etc.) leurs agencements acceptables, reconnus par l'analyseur syntaxique sont spécifiés par les règles de production d'une grammaire formelle. Le sens est alors défini par la circulation d'attributs sémantiques dans l'arbre de dérivation syntaxique. On fait ainsi correspondre à une suite d'expressions syntaxiquement correctes une liste d'instructions machines ; le parallèle avec les grammaires de Montague faisant correspondre à une phrase reconnue par une grammaire hors contexte une proposition logique, simple traduction de la structure syntaxique, est irrésistible et s'est naturellement imposé[40]. Cette approche évidemment positiviste se heurte aux classiques problèmes d'ambiguïté : ambiguïté lexicale, syntaxique et sémantique donnant lieu à toute sorte de conjectures comme la question de savoir si le mot

[38] *Recherche philosophiques* par. 432

[39] *Recherche philosophiques* par. 6

[40] On peut ainsi faire correspondre à la phrase « tous les hommes sont mortels » la proposition « pour tout x, homme(x) => mortel(x) »

« car » est un nom ou une conjonction de coordination, de reconnaître l'arbre de dérivation correct de la phrase « la petite brise la glace » ou bien encore de savoir si le « chien » est celui du fusil ou celui de son maître. En particulier, les termes atomiques dans les propositions logiques construites à partir des phrases doivent être identifiés de façon univoque afin d'en déterminer le type et donc la signification. Cette problématique fait écho à la tradition logique et grammaticale qui privilégie les signes et la syntaxe. Dans cette tradition, le sens des mots est indépendant du contexte, il se représente comme la relation entre un concept et l'objet qu'il désigne. Les ontologies actuelles répondent à l'ontologie scholastique en ce qu'elles considèrent que l'interprétation d'un mot se réduit à l'identification du concept auquel il se rapporte, elles relèvent d'une théorie de la signification. L'hégémonie des ontologies du Web sémantique en ramenant la question du sens à celle de la signification masque une autre tradition occidentale de l'étude du langage d'inspiration rhétorique et herméneutique dont les objets d'étude sont la production et l'interprétation des textes et des discours. Cette autre tradition s'est plus particulièrement intéressée à la question du sens, un texte n'ayant pas de signification mais un sens issu de son interprétation. Or en dépit du fort courant positiviste qui anime la terminologie depuis Eugen Wüster[41], dans une perspective documentaire, c'est bien le sens des textes qui importe plutôt que la signification des mots. La réhabilitation de l'approche herméneutique ouvre une voie naturelle pour contourner les apories du positivisme dans l'étude des textes en général et de leur indexation en particulier.

3.3 L'épiphanie de la sémantique interprétative des textes

A la suite de Peter Szondi, François Rastier se prononce pour une herméneutique matérielle qui recentrerait l'herméneutique sur le texte, qui contribuerait à « la réunification des sciences de la « lettre » et des sciences de l'« esprit », en précisant les contraintes linguistiques sur l'interprétation. » [Rastier 2003]. Le caractère culturel des langues suppose que la signification dépend des conditions de l'interprétation qui est elle-même soumise aux objectifs d'une pratique sociale. C'est la réintroduction de l'ordre herméneutique en linguistique qui permet de rendre compte de la dimension culturelle du texte. Par ailleurs, comme cela a déjà été dit, la compréhension est guidée par un ensemble de normes établies par la tradition. Ces normes sont établies par un corpus de textes faisant autorité dans la pratique sociale considérée, par exemple : « la référence de La Cousine

[41] Les thesaurus d'autrefois sont les ancêtres des ontologies d'aujourd'hui

Bette n'est pas directement la France louis-philipparde, mais en premier lieu, voir exclusivement, La Comédie Humaine, augmentée des romans d'Eugène Sue que Balzac voulait égaler et dépasser »[42]. Le cercle herméneutique s'applique donc à deux paliers : celui du texte et celui de l'intertexte. La prise en compte de l'extérieur du texte comme un corpus d'autres textes auquel il se réfère explicitement ou qui partagent le même genre permet de garantir la validité de son interprétation sans recours au réel ou à l'essence des choses.

Le signe Saussurien est le lieu d'une rupture épistémologique, le passage d'une sémantique référentielle à une sémantique différentielle. La valeur du signe conçue comme l'ensemble des différences avec d'autres valeurs s'oppose à la signification comme référence à des concepts ou à des objets du monde physique. Pour Saussure, en effet, la notion de valeur se distingue de celle de signification en ce que la signification considère le signe en lui-même alors que la valeur le considère comme partie de l'ensemble des signes à savoir la langue : la langue est un « système dont tous les termes sont solidaires et où la valeur de l'un ne résulte que de la présence simultané des autres » [Saussure 1916]. La conséquence de cette rupture est d'ancrer la sémantique dans la linguistique puisqu'elle opère sur les signifiés et non plus sur des concepts universels indépendants de la langue. La sémantique structurale européenne s'inscrit dans ce paradigme, le sens y est décrit par regroupements et oppositions de signifiés. Les sèmes qui sont des traits distinctifs élémentaires désignent ces regroupements (sèmes génériques), et ces oppositions (sèmes spécifiques). Les sèmes n'ont pas d'existence sémantique propre, ils sont regroupés en unités sémantiques minimales, les sémèmes. En langue, le sémème exprime le contenu sémantique du signe minimal, le morphème. L'exemple classique du 'couteau' et de la 'fourchette' illustre ce propos : ces deux sémèmes partagent le sème générique /couvert/ alors que le sème spécifique /pour couper/ oppose le 'couteau' à la 'fourchette'. Cette caractérisation des sèmes les structure en classes : le taxème est la classe de sémème minimale, le domaine regroupe des taxèmes relatifs à une dimension sociale[43], les dimension, enfin, sont des classes très générales regroupées par des oppositions comme //animé// Vs //inanimé//, //humain// Vs //animal//, etc. [44]

[42] [Rastier 2001a] Herméneutique matérielle ; interprétation et corpus.

[43] Le taxème //couvert// de l'exemple précédent est englobé dans le domaine //alimentation//

[44] Exemples venant de « Sémantique Interprétative » [Rastier 1996]

La sémantique interprétative de François Rastier reprend à son compte cet appareil conceptuel et l'étend pour décrire les interprétations d'un texte par l'ensemble des sèmes activés au cours d'une lecture. Elle l'enrichit en particulier, de la notion d'afférence qui distingue les sèmes inhérents faisant, en langue, a priori partie du sémème, des sèmes afférents qui, au contraire, lui sont affectés en fonction du contexte ou de l'entour social. Cette notion permet de rendre compte des phénomènes d'assimilation et de dissimilation. L'assimilation renforce l'homogénéité sémantique d'énoncés sémantiquement contrastés, comme par exemple, le titre des mémoires de Claude Mauriac : « Le temps immobile ». La dissimilation, au contraire, souligne l'opposition, la différence entre mots similaires, comme c'est notamment le cas des tautologies apparentes, par exemple : « les affaires sont les affaires ». Trois opérations interprétatives permettent de décrire ces phénomènes : l'inhibition d'un sème inhérent, l'activation ou la propagation d'un sème afférent.

Ainsi dans l'exemple du « temps immobile » le sème /écoulement/ inhérent à 'temps' est-il inhibé par la proximité du sème /fixité/ inhérent à 'immobile'. Au contraire dans celui de « désolé les affaires sont les affaires » un sème afférent socialement normé /impitoyable/ est activé dans la seconde occurrence de 'affaires' afin de contraster les deux occurrences. L'entour social est aussi et surtout propagateur d'afférences[45] ainsi le sème afférent /torture/ est-il propagé sans ambiguïté dans 'morts', 'cave' et 'parlé' dans le passage du célèbre discours d'André Malraux : « entre ici, Jean Moulin, avec ton terrible cortège. Avec ceux qui sont morts dans les caves sans avoir parlé, comme toi ; …» L'entour historique du texte ne laisse aucun doute sur ce à quoi on se réfère ici : La gestapo torturait dans les *caves* pour faire *parler* les suppliciés, Jean Moulin est *mort* sous la torture. Ces actualisations et virtualisations de sèmes sont le résultat de parcours interprétatifs, de suites d'opérations interprétatives qui permettent l'attribution d'un ou plusieurs sens à un texte.

L'analyse sémique d'un texte vise à rendre compte de parcours interprétatifs, à en dégager les sèmes, à définir leurs regroupements et les relations qu'ils entretiennent. Deux types de regroupements sont d'un intérêt particulier : l'isotopie et la molécule sémique. L'isotopie est la répétition d'un sème, elle « est le nom de l'accord sémantique des mots au sein d'un texte » [Pincemin 1999a]. Une molécule sémique est un groupement stable de sèmes dans une même unité linguistique, un mot ou un syntagme par exemple. La récurrence d'une

[45] Voire aussi l'analyse du désormais fameux : « Un opéra raisonnable, c'est un corbeau blanc, un bel esprit silencieux, un Normand sincère, un Gascon modeste, un procureur désintéressé, enfin un petit maître constant et un musicien sobre » (Antoine La Motte, épigraphe au livret d'*Alcyone*, de Martin Marais cité par François Rastier, Michel Ballabriga et Christophe Gérard)

molécule sémique génère un groupe d'isotopies que l'on appelle faisceau isotopique. Une proposition d'analyse du poème de Luis Cernuda, « Un español habla de su tierra »[46] illustre l'utilisation des isotopies et des molécules sémiques dans la description des parcours interprétatifs.

Las playas, parameras
Al rubio sol durmiendo,
Los oteros, las vegas
En paz, a solas, lejos;

Los castillos, ermitas,
Cortijos y conventos,
La vida con la historia,
Tan dulces al recuerdo.[47]

On peut y reconnaître une première molécule sémique (/horizontalité/, /terre/, /paix/, /solitude/)[48] qui se répète deux fois dans la première strophe et une seconde (/verticalité/, /culture/, /sérénité/, /richesse/, /solitude/)[49] se répétant quatre fois dans les deux premiers vers de la seconde strophe. Les deux derniers vers rappellent les isotopies /nature/, /culture/ et /douceur/. Ces deux premières strophes annoncent le sujet du poème, l'exil[50], au travers

[46] « Un espagnol parle de sa terre », mis en musique par Paco Ibañes.

[47] Une traduction possible serait :
Les plages, déserts
Dormant sous le soleil blond,
Les tertres, les plaines,
Paisibles, seuls*, loin ;

Les châteaux, ermitages,
fermes et couvents,
La vie avec l'histoire,
Si douces au souvenir.

* a solas veut aussi dire « en tête-à-tête »

[48] La première occurrence de cette molécule apparaît dans les deux premiers vers : /horizontalité/ porté par 'plages', 'désert' ; /terre/ porté par 'plages', 'désert' ; /paix/ porté par 'dormant' ; /solitude/ porté par 'désert' ; la deuxième dans les deux vers suivants : /horizontalité/ porté par 'plaine' ; /terre/ porté par 'tertres', 'plaines' ; /paix/ porté par 'paisibles' ; /solitude/ porté par 'seuls'

[49] Cette seconde molécule est lexicalisée dans : 'châteaux', 'ermitages', 'fermes', 'couvents', le sème /verticalité/ est inhérent aux édifices qui sont érigés par les hommes comme d'ailleurs aussi le sème /culture/, le sème /solitude/ tient à ce que ces édifices sont en général isolés, enfin le sème socialement normé /sérénité/ est afférent à 'ermitages' et les 'couvents' comme /richesse/ est afférent aux quatre occurrences.

[50] La strophe suivante précisera le sujet : « Ellos los vencedores, Caínes sempiternos, De todo me arrancaron, Me dejan el destierro », « Eux les vainqueurs, Cains éternels, M'arrachèrent de tout, Ne me laissant que l'exil »

des isotopies /solitude/ et /souvenir/[51] d'une part et /terre/ d'autre part : l'exil, en espagnol, rappelle explicitement la terre[52] par son signifiant ; il est d'ailleurs intéressant de remarquer l'importance de la langue ici, le sens est un phénomène linguistique. Les isotopies /solitude/ et /souvenir/ diffusent donc dans le poème[53] un fond correspondant au thème principal, le sujet sur lequel se détachent dans les deux premières strophes deux molécules sémiques, deux formes sémantiques correspondant aux thèmes secondaires, dans le poème, de la nature et de la culture.

Bien que n'étant pas toutes directement au cœur de la problématique de cet exposé il est difficile de ne pas mentionner les quatre composantes qui interagissent lors de la production et de l'interprétation des textes : la thématique, la dialectique, la dialogique et la tactique.[54]

1. « La thématique rend compte des contenus investis et de leurs structures paradigmatiques. »[55]
2. « La dialectique rend compte de la succession des intervalles dans le temps textuel, comme des états qui y prennent place, et des processus qui s'y déroulent »[56].
3. La dialogique rend compte des modalités. « Retenons sans exclusive les modalités ontiques, aléthiques, épistémiques, déontiques, boulestiques, évaluatives, sémiotiques. »[57]
4. « La tactique rend compte de la disposition linéaire des unités sémantiques. »[58]

Seules la thématique et la tactique seront plus amplement développées par la suite.

Une sémantique textuelle subordonne la signification au sens, la valeur en langue à la valeur en contexte. On considère généralement trois paliers d'analyse selon le contexte : les paliers macrosémantiques, mésosémantiques et microsémantiques. Le palier macrosémantique correspond au texte dans son ensemble, le palier mésosémantique à

[51] /souvenir/ est porté par 'histoire' et 'souvenir'.

[52] On retrouve la racine 'tierra', 'terre' dans 'destierro', 'exil'

[53] Le poème entier est consultable en annexe.

[54] Se reporter à l'ouvrage de François Rastier « Sens et textualité » [Rastier 1989] disponible sur le site de la revue texto

[55] Sens et textualité page 54.

[56] Sens et textualité page 66.

[57] Sens et textualité page 82.

[58] Sens et textualité page 95.

l'espace qui s'étend du syntagme à la phrase voire le paragraphe, le palier microsémantique au syntagme et à ses constituants, morphèmes et lexies. La production et l'interprétation d'un texte sont par ailleurs régies par des règles regroupées au sein de normes en fonction de leur degré de systématicité : les normes dialectales, sociolectales et idiolectales. Les normes dialectales, les plus rigoureuses, régissent les usages de la langue. Les normes sociolectales correspondent à des pratiques sociales (judiciaire, politique ou religieuse par exemple). Les normes idiolectales reflètent quant à elles aux régularités personnelles à un auteur. Les normes dialectales déterminent plus particulièrement les paliers microsémantiques et mésosémantiques, les normes sociolectales le palier macrosémantique, bien que certaines tournures ou certains emplois en relèvent évidemment. A toute pratique sociale correspond un discours, le genre assure le lien entre le texte et le discours dont il relève, le genre est une forme du discours. Par les régularités qu'elles induisent dans les textes, les normes de discours et de genre permettent, l'automatisation de certaines procédures de segmentation et de catégorisation sur des formes et des fonds sémantiques dans les textes.

Les notions de formes et fonds sémantiques se fondent sur l'hypothèse de la perception sémantique : « La compréhension d'une suite linguistique est pour l'essentiel une activité de reconnaissance des formes sémantiques, qu'elles soient déjà apprises ou construites au cours du traitement. »[59] Le découpage des textes en unités aux contours aisément identifiables a longtemps prévalu : mots, phrases, fonctions narratives, le texte est décrit comme la concaténation d'unités discrètes. « La conception rhétorique/herméneutique admet en revanche que les objectivités qu'elle construit soient continues, parfois implicites, varient dans le temps et selon leurs occurrences et leurs contextes, connaissent entre elles des inégalités qualitatives et ne relèvent pas uniformément des mêmes règles. » [Rastier 2006] Dans une problématique textuelle l'unité est le passage ; or le passage n'a pas de frontières bien définies, son étendue dépend de l'interprète et de ses objectifs. La sémantique interprétative a été ainsi amenée à considérer les textes comme des fonds, constitués de faisceaux d'isotopies sur lesquels viennent se contraster des formes que sont les molécules sémiques. L'interprétation peut ainsi être décrite comme une activité perceptive qui fait varier le rapport entre les fonds et les formes. Les processus perceptifs à l'œuvre dans ces activités d'interprétation sont la dissimilation, l'assimilation et la présomption d'isotopie précédemment décrits. La morphodynamique qui est l'étude de la

[59] Sens et textualité page 9.

dynamique des formes et des fonds sémantiques s'attache à décrire les rythmes des fonds, les contours des formes et leurs transformations et les rapports qu'entretiennent les formes et les fonds.

Dans cet exposé consacré à l'indexation, la thématique et la morphodynamique seront d'un intérêt particulier. En effet, traditionnellement, l'indexation d'un texte porte sur ses thèmes principaux, son sujet, d'autre part, un thème peut être vu comme la récurrence d'une molécule sémique dans un corpus. C'est cette connexion évidente entre l'indexation et la sémantique interprétative des textes qui sera développée dans cet exposé.

4 Isotopie et statistiques contextuelles

La question que nous devons maintenant aborder est l'articulation effective entre statistiques textuelles et le concept d'isotopie. Voici donc venu le temps de martyriser la sémantique interprétative à l'aide des multiples moyens que nous offrent les statistiques. D'autres avant nous se sont livrés à ce pénible exercice et nous nous inspirerons dans une large mesure de leur exemple. Nous pensons en particulier à Bénédicte Pincemin [Pincemin 1999b] et Mathias Rossignol [Rossignol 2005]. Notre apport sera de mettre ces travaux en perspective avec les idées d'analyse sémantique latente et des études récentes concernant l'étude des tableaux de contingences de dimension 2.

4.1 Motivations

Le problème essentiel de l'herméneutique est aujourd'hui celui de l'application, l'impartialité de l'interprétation, la rigueur de l'application. L'affaire Sokal quelque soit l'opinion que l'on a sur le procédé, montre à quel point la démarche herméneutique est fragile. Cette fragilité est bien celle de l'absence d'objectivité de l'interprétation or comme nous l'avons vu au chapitre 2 cette absence d'objectivité est inévitable dans les sciences humaines. Alan Sokal peut ironiser sur ce thème:

> « Anyone who believes that the laws of physics are mere social conventions is invited to try transgressing those conventions from the windows of my apartment. (I live on the twenty-first floor.) » [Sokal 1996]

Il n'en reste pas moins que dans l'expérience artistique les lois de la physique sont d'un intérêt tout relatif. L'impartialité et la rigueur ne sont pas l'apanage de l'objectivité même si elles se trouvent souvent confondues dans le langage courant. La sémantique interprétative offre les moyens d'analyser rigoureusement des textes, dans le chapitre 3, nous avons vu, certes rapidement, comment. Pourquoi, du point de vue de l'indexeur, ces interprétations sont-elles rigoureuses ? Parce qu'elles sont, d'une certaine manière reproductibles, parce qu'elles s'inscrivent dans un cadre formel. Ce cadre formel permet de décrire des molécules sémiques qui se détachent sur des fonds que sont les isotopies. Molécules sémiques et isotopies se décrivent à l'aide d'un appareillage formel, d'un ensemble de notations, d'une écriture spécifique en somme. Elles structurent l'interprétation dans un cadre gestaltiste [Missire 2005], [Rastier 2006b], au travers des notions de formes et de fonds, cette

correspondance naturelle entre théorie de la forme et sémantique interprétative est résumée par Christophe Gérard [Gérard 2004] à la figure :

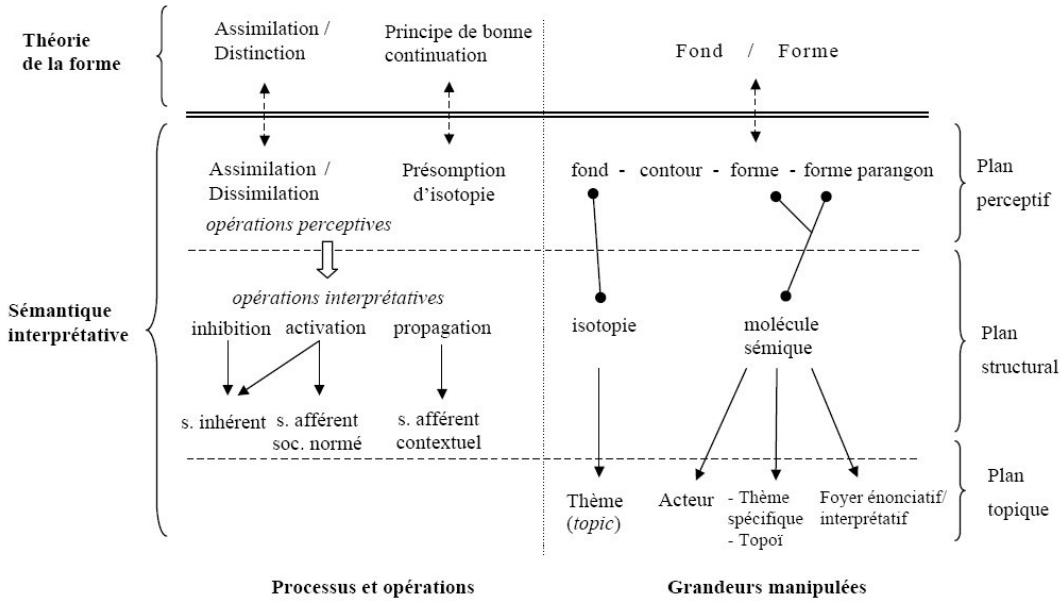

Figure 6: Les plans de la Sémantique Interprétative et leurs relations (Christophe Gérard)

Cette idée de perception sémantique [Rastier 1996] est fondamentale dans la structuration de l'interprétation car se fondant sur les notions de dissimilation, d'assimilation et d'isotopie elle trouve un écho dans les théories de la reconnaissance des formes qui utilisent les notions proches de contraste, de proximité et de similarité. Ce point d'ancrage à ces théories ouvre des perspectives nouvelles à l'analyse des textes qui rappellent très clairement d'autres travaux sur la reconnaissance des formes visuelles. En particulier l'idées de formes se détachant sur des fonds rappelle irrésistiblement le théâtre d'ombres chinoises ou les lanternes magiques. Mais l'idée de perception sémantique est d'une certaine façon obligée d'aller tout de suite au delà des purs concepts d'analyse d'image, en effet, l'analyse d'image peut se cantonner à l'analyse d'images statiques, le texte est quand à lui d'essence purement dynamique, l'analyse de texte ressemble en ce sens plus à l'analyse de vidéos qu'à l'analyse d'images fixes. François Rastier dans [Rastier 2006b] illustre sa vision de cette dynamique au travers d'un certain nombre de figures.

La figure 7 illustre la nécessité de redéfinir le signe comme une passage. Toujours dans [Rastier 2006b][60], François Rastier rappelle les deux problématiques de la linguistique occidentale :

> *Les deux problématiques.* — Deux problématiques, logico-grammaticale et rhétorique/herméneutique, se partagent l'histoire des idées linguistiques occidentales [1]. En bref, nous appellerons la première problématique du *signe* et la seconde problématique du *texte*. Convenons que la *signification* est attribuée aux signes et le *sens* aux textes. Si l'on approfondit cette distinction, un signe, du moins quand il est isolé, n'a pas de sens ; et corrélativement un texte n'a pas de signification.
>
> La signification résulte en effet d'un processus de décontextualisation, comme on le voit en sémantique lexicale et en terminologie ; d'où son enjeu ontologique, puisque traditionnellement on caractérise l'Etre par son identité à soi. Le sens suppose en revanche une contextualisation maximale, aussi bien par l'étendue linguistique — le contexte, c'est tout le texte — que par la situation, <99> définie par une histoire et une culture, bien au-delà du *hic et nunc* seul considéré par la pragmatique. Aussi, alors que la signification est traditionnellement conçue comme une *relation*, le sens peut être représenté comme un *parcours* au sein du texte et de l'intertexte.
>
> *Unités ou formes ?* — Les divergences entre problématiques apparaissent clairement à propos des unités textuelles. La conception logico-grammaticale tend à faire de l'unité un élément de « vocabulaire » textuel : à l'image d'une phrase considérée comme un enchaînement de mots, un texte résulterait d'un enchaînement d'unités : propositions, séquences, fonctions narratives, etc. : la linguistique textuelle a ainsi conçu le texte comme une suite structurée de propositions, la narratologie greimassienne a représenté le discours par une concaténation de fonctions narratives. Ces unités sont considérées comme discrètes et localisables, ce qu'atteste, par exemple, le nom de *séquence*

pour plus loin arriver à la conclusion fondamentale dans le reste de ce chapitre:

> Redéfinir le signe comme passage. — Il convient de proposer une redéfinition du signe qui s'accorde avec la problématique textuelle. L'unité, quelle que soit sa taille et son palier de description, peut être redéfinie comme un passage : or un passage n'a pas de bornes fixes et dépend évidemment du point de vue qui a déterminé sa sélection.
>
> Définir le signe comme passage, c'est élaborer une définition purement relationnelle et donc contextuelle. Puisque la parole commande la langue, le signe est d'abord un « segment de parole » [3] : au plan du signifiant, c'est un extrait — entre deux blancs, s'il s'agit d'une chaîne de caractères ; entre deux pauses ou ponctuations, s'il s'agit par exemple d'une période. L'extrait peut renvoyer aux étendues connexes, par exemple par des règles d'isophonie, ou de concordance de morphèmes : ce sont des cooccurrents expressifs.
>
> Au plan du signifié, le passage est un fragment qui pointe vers ses contextes gauche et droit, proche et lointain. Cela vaut pour le contenu de la lexie comme pour celui du syntagme ou de la période, de la section, etc. Les méthodes statistiques de la linguistique de corpus permettent aujourd'hui de qualifier les unités contextuelles, qui sont elles aussi des passages de taille variable définis comme corrélats sémantiques :

[60] Dans cette première section consacrée aux statistiques, nous citons abondamment cet article de François Rastier qui nous paraît essentiel à la justification de notre approche statistique et nous en reproduisons des passages entiers plutôt que de les paraphraser, de la même façon plutôt que de recopier ses figures à peu près à l'identique nous avons fait le choix de les reproduire, nous espérons qu'il nous en tiendra pas rigueur et qu'il n'y verra ni servilité ni désinvolture.

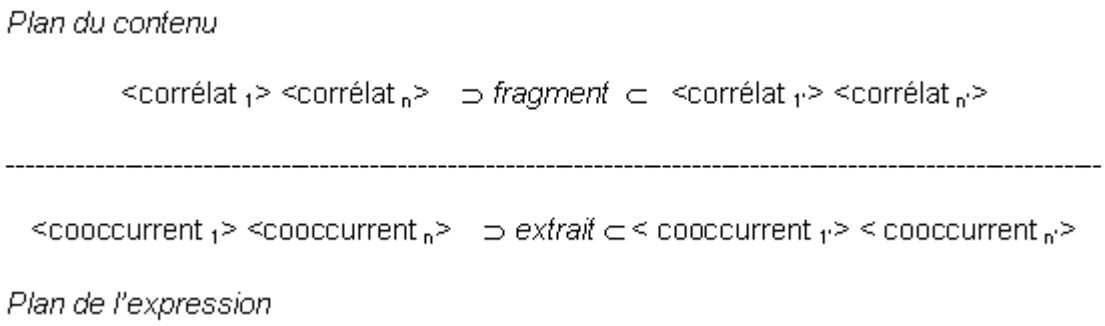

Figure 7: Le passage et ses contextes (François Rastier)

C'est cette constatation qui guidera notre choix de l'étude statistique des mots aux sein de contextes. Ces contextes resteront encore bien grossièrement définis comme des paragraphes et nous ne retiendrons rien de leur structure syntaxique. Nous ramènerons en fait à la simple analyse statistique des cooccurences de leurs mots. On peut imaginer bien d'autres approches tant au niveau de la caractérisation des passages qu'à celui de la prise en compte de leur structure et de leurs unités minimales. Il faut bien le dire, notre approche est minimale de ce point de vue, elle nous permettra pourtant de mettre en évidence des formes récurrentes se détachant sur un fond isotopique. L'idée de François Rastier est parfaitement résumée à la figure 8, voici comment nous tenterons de la mettre en oeuvre dans un programme d'analyse interprétative des données tel que nous l'avons définie dans le chapitre deux. Tout d'abord précisons nos présupposés :

1. Les contours que nous prendrons en compte sont des mots autour desquels se regroupent des cooccurrences remarquables au sein de passages[61]. Dans l'esprit du chapitre deux, nous utiliserons un filtre statistique pour rechercher de tels mots et les proposer à un interprète[62].

2. Toujours dans ce même esprit nous proposerons un interprète des formes qui seront des classes de passages regroupés selon des contours communs.

3. Le fond étant vu comme l'ensemble des contours, leurs répétitions seront selon nous porteurs d'isotopies. La classification des textes en fonction des contours apparaissant dans les formes retenues permettra à l'interprète de valider[63] ces présomptions d'isotopie et par la même une indexation de ces mêmes textes.

[61]Ici de paragraphes.

[62]En IA cet interprète serait appelé un méta programme ou quelque chose du genre, or il se trouve que notre préjugé est que le seul méta programme concevable est l'esprit humain..

[63]Ou d'invalider

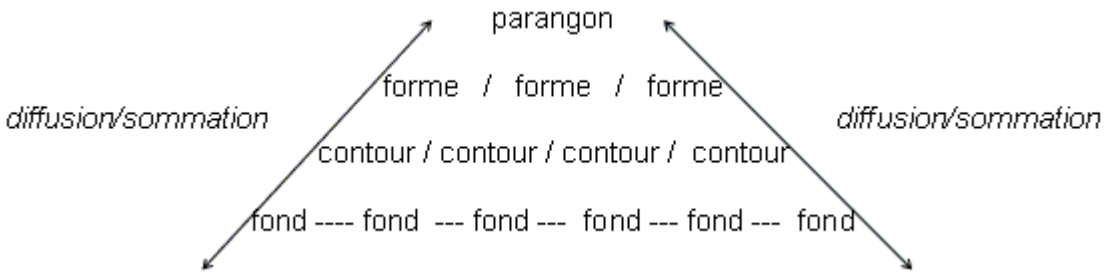

Figure 8: Médiations entre formes et fonds sémantiques (François Rastier)

Ce chapitre est donc plus particulièrement dédié aux aspects calculatoires de ce processus. Nous nous attarderons donc sur le filtrage statistique des contours, sur la classification des passages où ils apparaissent et la classification finale des textes. Ceci étant, la lecture de ce chapitre ne doit pas faire oublier que nous nous inscrivons dans un processus interprétatif tourné vers la compréhension humaine qui est seule garante de la qualité des résultats.

La suite de ce chapitre se présente comme suit. Nous commencerons par un bref rappel du modèle vectoriel de Gerard Salton et de l'analyse sémantique latente. Nous poursuivrons par une présentation d'une méthode d'analyse de matrices de cooccurrence basée sur une méthode exacte du calcul du test de signifiance de Fisher plutôt que sur les propriétés asymptotiques du ratio de vraisemblance proposé par Ted Dunning [Dunning 1994], nous présenterons en particulier une méthode de factorisation des cooccurrences s'inspirant de l'analyse en composantes principales. Cette factorisation nous permettra de représenter les mots et des passages comme des vecteurs sur d'autres mots, les facteurs. Cette représentation nous permettra de définir des fonctions noyau dans le but de mettre en évidence des classes de passages correspondant à des molécules sémiques, ces molécules sémiques permettant enfin de mettre en évidence les isotopies sur lesquelles elles se détachent. Tout ceci peut sembler tenir du miroir aux alouettes et la pratique d'une telle approche, pour fournir des résultats satisfaisants nécessitera le plus souvent un minutieux travail de réglage. Mais nous avons déjà suffisamment insisté sur ce point dans au chapitre 2 pour ne plus nous y appesantir ; disons simplement que si il est possible d'envisager un chaîne de traitement purement automatique l'intérêt de notre approche est de faire valoir le point de vue de l'indexeur, son interprétation à l'aide d'outils statistiques.

4.2 Modèle vectoriel et sémantique latente

Gerard Salton a été le pionnier incontesté des systèmes d'interrogation de bases de données textuelles [Salton, Yang 1975]. Dans sa version la plus grossière, le modèle

vectoriel correspond à ce que l'on appelle un système de fichier inversé, c'est-à-dire un modèle de données où à chaque mot d'un corpus correspond la liste des fichiers dans lesquels il apparaît ainsi que sa fréquence d'apparition dans chacun d'eux. L'expression de modèle vectoriel vient de ce que chaque texte, ou d'une manière plus générale, chaque passage est un vecteur dans l'espace des mots : ainsi la locution « modèle vectoriel » aura les coordonnées 1 et 1 sur les dimensions 'modèle' et 'vectoriel' et 0 sur les autres dimensions. L'attrait d'une telle représentation tient à sa simplicité et à l'aisance d'interrogation d'un fonds documentaire qu'elle offre : à la requête 'modèle' **et** 'vectoriel' correspondra l'intersection de la liste des documents contenant le mot 'modèle' avec celle des documents contenant le mot 'vectoriel'. Ce modèle est aussi appelé sac de mots car la disposition des mots les uns par rapport aux autres au sein d'un même passage est perdue. D'autres défauts sont bien connus : les mots très fréquents induisent des réponses bruyantes, les mots qui le sont peu des réponses incomplètes. En particulier deux documents proches peuvent ne pas partager une grande quantité de mots. Comment, par exemple, rapprocher les documents contenant l'expression « sac de mots » et l'expression « modèles vectoriels » ? Ainsi de multiples améliorations ont été proposées dont les plus importantes furent : la pondération par fréquence de documents inverse, la densité spatiale et l'analyse de sémantique latente. La pondération par fréquence de document inverse se base sur l'idée qu'un mot apparaissant dans tous les documents est peu informatif. Une formule simple de pondération basée sur cette idée est donnée par la formule :

$$idf(w) = \log\left(\frac{l}{df(w)}\right) \quad (1)$$

$df(w)$ est le nombre de documents où apparaît le mot w et l le nombre de document du fonds documentaire ainsi la fréquence de document inverse est nulle pour un mot apparaissant dans tous les documents. Kenneth Church et William Gale montrèrent dans un article fameux [Church, Gale 1995] que cet indicateur est une excellente mesure de la déviation de la loi de Poisson, c'est-à-dire, de l'apparition d'un mot dans un document par hasard. Il est intéressant de noter que dans la plupart des systèmes des listes de mots indésirables « stop words » permettent d'éviter l'indexation par des termes bruyants, en particulier il est d'usage courant d'exclure les mots grammaticaux. Il faut noter que la normalisation des mots par 'stemming' ou lemmatisation permet de diminuer le silence du système : 'vectorielle', 'vectoriels', 'vectoriel', 'Vectorielle', 'Vectoriels' et 'Vectoriel' peuvent être ramenée au lemme 'vectoriel' ou au stem 'vector-'. La nominalisation permet

de détecter des expressions canoniques ‘calibrer une pièce’ peut ainsi être normalisée en ‘calibrage de pièce’. Notons les abus possibles de ce genre de procédé « ces émeutes ont coûté la vie à des dizaines de personnes » pouvant être indûment indexé par « coût de la vie ». Un autre type d'indicateur, dont il est peu fait mention dans la littérature, est la valeur discriminante : pour une fonction noyau k et un ensemble de documents Δ cet indicateur est basé sur la quantité $Q=\sum_{(d_i,d_j)\in\Delta^2} k(d_i,d_j)$. Moins cette quantité est grande moins les documents de Δ ont tendance à se ressembler. Si l'on supprime des documents Δ le terme t on fait correspondre la quantité Q_t qui se calcule comme Q à ceci près que le terme t ne rentre pas en compte dans la fonction noyau k. La valeur discriminante du terme t est donnée par la formule $DV_t=Q_t-Q$ [Salton, Yang 1975]. Le terme t est dit discriminant pour l'ensemble Δ si $DV_t>0$ car cela veut dire que si on le retire de cet ensemble les documents ont tendance à plus se ressembler, le terme t tend donc à séparer les documents.

D’autres techniques ont été utilisées pour améliorer le système initial comme par exemple l’utilisation des liens hypertextuels sur le Web, une forme de filtrage collaboratif passif, que nous ne développerons pas plus avant car cela nous éloignerait de notre propos.

L’analyse sémantique latente qui fut décrite pour la première fois en 1988 [Deerwester et al 1988] nous semble mieux correspondre à nos préoccupations. L’idée initiale a été de surmonter l’imprécision et l’incomplétude du système originel en utilisant une analyse factorielle de la matrice croisant les passages avec les termes qu’ils contiennent. Quelques notations s’imposent ici que nous aurons l’occasion d’utiliser par la suite. D’une façon très générale nous étudions une liste de *n* passages contenant *p* termes distincts ; un passage pouvant être un paragraphe, un syntagme, une phrase voire un document entier ; un terme pouvant être un mot, un lemme, un stem ou une locution rencontrés dans un passage ainsi l’objet de notre étude se présente comme une matrice $C=(c_{i,k})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant k\leqslant p}}$ où le terme $c_{i.k}$ est une fonction de la fréquence du terme *k* dans le passage *i*. Il est d’usage courant de considérer $c_{i,k}=t_{i,k}idf_k$ où $t_{i,k}$ représente la fréquence du terme *k* dans le passage *i* et idf_k la fréquence de document inverse du terme *k*. L’idée initiale de l’indexation

sémantique latente (LSI[64]) est de décomposer la matrice C croisant les passages et les termes en un produit de trois matrices $C = T_{pr} S_{rr} X_{rn}$ où les colonnes des matrices rectangulaires T_{pr} et X_{rn} sont orthonormales[65], et la matrice S_{rr} une matrice diagonale dont les termes diagonaux sont les racines carrées des r premières valeurs propres de la matrice $C^t C$. Au-delà du formalisme il faut retenir que chaque passage est projeté sur les facteurs principaux de la matrice des cooccurrences des termes dans les passages $C^t C$.

On voit qu'il n'y a rien de neuf sous le soleil dans la mesure où c'est cette même idée qui est à l'origine de l'analyse factorielle des correspondances comme le fait justement remarquer Laurence Favier [Favier 1998] dans sa thèse. Cependant, à l'origine tout du moins, les objectifs des deux approches n'étaient pas les mêmes. L'une (LSI) s'intéressait à l'amélioration du système Saltonien l'autre s'intéressait plus à la problématique de la visualisation des données. Techniquement, même si les deux méthodes dérivent de l'analyse en composantes principales l'analyse des correspondances s'intéresse aux variables qualitatives en utilisant la métrique du χ^2 ce qui n'est pas le cas de la LSI. Les avantages escomptés de la LSI, et en partie obtenus, sont la prise en compte implicite des relations entre mots n'apparaissant pas dans les mêmes passages, améliorant ainsi les performance du système. En particulier une requête peut être représentée par un vecteur R_{p1} dans l'espace des mots à l'image de C puis transformée en un vecteur Y_{r1} vérifiant la relation $R_{p1} = T_{pr} S_{rr} Y_{r1}$ les réponses à la requête étant les vecteurs colonnes X_i de la matrice X_{rn} ayant le cosinus le plus élevé avec Y_{r1} [66]. L'indexation sémantique latente n'est pas massivement utilisée dans les moteurs de recherche : quelqu'un a prétendu cependant que Google utiliserait un système apparenté pour le tri de ces réponses mais cela reste du domaine de la rumeur dont le Web n'est pas avare. Disons plutôt que le niveau de service offert par les moteurs actuels ne nécessite pas la mise en place d'une telle

[64] LSI pour « Latent Semantic Indexing ».

[65] à savoir que $X_{rn}\,{}^t X_{rn} = I_r$ et ${}^t T_{pr} T_{pr} = I_r$ où ${}^t A$ est la matrice transposée de A et I_r la matrice identité à r colonnes, les indices rn et pr indiquent la dimensionnalité des matrices rectangulaires (r lignes et n colonnes ou p lignes et r colonnes), un indice unique r indique la dimension d'un matrice carrée (r lignes et r colonnes).

[66] $\cos(X_i, Y_{r1}) = \dfrac{{}^t X_i Y_{r1}}{\|X_i\| \|Y_{r1}\|}$

machinerie et que l'étude des liens hypertextuels offre d'incomparables avantages en terme d'efficacité, de facilité de programmation et de stockage. Il n'en reste pas moins que ceux qui s'intéressent au sens et en particulier au sens en passage cette voie est loin d'avoir épuisé ses ressources.

Revenons donc à cette décomposition de la matrice *C* et en particulier à l'étude de la matrice de cooccurrences C^tC, notons tout d'abord que la matrice $P_{pr}=T_{pr}S_{rr}$ qui en est déduite est en quelque sorte une matrice de proximité entre termes et facteurs. En fait il serait tout à fait possible de définir une fonction noyau entre passages en utilisant la projection $P_{pr}C_i$ du passage C_i sur les facteurs principaux comme suit : $k(C_i, C_j)={}^tP_{pr}{}^tC_iP_{pr}C_j$, la matrice de projection pouvant être fort éloignée de la projection traditionnelle $P_{pr}=T_{pr}S_{rr}$. C'est d'ailleurs ce qui a été fait par Siolas et d'Alché-Buc [Alché-Buc, Siolas 2000] qui utilisent une matrice de proximité terme par terme dérivée d'un réseau sémantique où la proximité $p_{k1,k2}$ en deux termes *k1* et *k2* dans la matrice de proximité est une fonction inverse de la longueur du plus court chemin reliant les deux termes dans le réseau sémantique. On le voit, les variations sur ce thème sont potentiellement importantes. Nous voudrions apporter notre pierre à cet édifice en tenant compte d'une remarque de Ted Dunning dans son fameux papier [Dunning 1994].

> « Even recent and very innovative work such as that using Latent Semantic Indexing and Pathfinder Networks has not addressed the statistical reliability of the internal processing. »

En effet, peu de cooccurrences dans la matrice C^tC sont statistiquement signifiantes. Et l'on peut se poser la question de savoir l'impact du filtrage de cette matrice de cooccurrences sur le système dans son ensemble.

4.3 De la signification des cooccurrences de mots dans la description des mots eux-mêmes

La littérature sur les cooccurrences de mots est pléthorique et il peut paraître présomptueux de vouloir ajouter un commentaire de plus sur le sujet. Cependant nous n'avons pas trouvé de réponses au commentaire de Dunning concernant la LSI et de récents travaux de Robert C. Moore [Moore 2004] sur les collocations nous ont convaincu qu'il y avait encore là matière à réflexion. Nous commencerons par rappeler les notions de base liées aux collocations puis nous présenterons une méthode de calcul exact du test Fisher permettant

un filtrage efficace des collocations les moins significatives. Une méthode dérivée des matrices de diffusion nous permettra de définir une fonction noyau à partir de la matrice des cooccurrences filtrées.

Commençons par rafraîchir notre mémoire sur les tests statistiques, en nous rappelant l'exemple du détecteur de fumée rendu fameux par le livre de Gonick et Smith « The cartoon guide to statistics » [Gonick et al 1994] : rien de plus agaçant, en effet, qu'un détecteur qui se déclenche chaque fois que l'on fait brûler un toast. Dans le langage statistique on parle d'erreur de première espèce : une alarme sans feu. Un moyen efficace de supprimer les erreurs de première espèce est de retirer les batteries du détecteur de fumée, on accroît malheureusement alors les risques d'erreurs de seconde espèce : un feu sans alarme !

On peut résumer ce dilemme par le tableau suivant :

	Pas d'incendie	Incendie
Pas d'alarme	Pas d'erreur	Erreur de seconde espèce
Alarme	Erreur de première espèce	Pas d'erreur

En termes statistiques, on parle de comparaison d'hypothèses, l'hypothèse nulle (pas d'incendie) et son alternative (l'incendie) et la question que l'on se pose est de savoir si l'on rejette (alarme) ou pas (pas d'alarme) l'hypothèse nulle (pas d'incendie). Ce qui se résume par le nouveau tableau :

	Réalité	
	Hypothèse nulle (H_0)	Alternative (H_a)
Acceptation de l'hypothèse nulle	Pas d'erreur	Erreur de seconde espèce
Rejet de l'hypothèse nulle	Erreur de première espèce	Pas d'erreur

Comme nous l'avons vu, la théorie de la décision (c'est ainsi que l'on appelle cette branche des statistiques) est un nécessaire compromis entre les erreurs de première et de seconde espèce. L'on a coutume de noter α la probabilité de commettre une erreur de première espèce et β une erreur de seconde espèce. On appelle $1 - \beta$ la puissance du test, c'est-à-dire sa capacité à correctement rejeter l'hypothèse nulle. Évidemment, β dépend de α (retirer les piles du détecteur de fumée n'est pas nécessairement une bonne idée) on évalue en général la puissance du test en étudiant la courbe liant l'état réel du système surveillé et la probabilité de rejeter l'hypothèse nulle.

Ceci étant quel est le lien entre les détecteurs de fumée et les cooccurrences de mots ? Élémentaire, l'hypothèse nulle est : « ces deux mots apparaissent dans ce passage par pur

hasard » et l'on cherche un test nous permettant de décider de rejeter cette hypothèse avec un seuil α bas de se tromper ou, ce qui revient au même, un niveau de confiance 1 - α élevé de le faire.

C'est donc la quête de ce test qui a longtemps occupé la communauté des statistiques textuelles ces dernières années en particulier dans le domaine de la lexicographie et de l'alignement de corpus bilingues. Pour les collocations, le test étalon, si l'on peut dire, est le test d'indépendance de Fisher pour les tables de contingence de dimension 2. Ce test est traditionnellement considéré comme impossible à calculer pour des échantillons de grande taille et on utilise généralement des tests approximant le χ^2 comme le logarithme du rapport de vraisemblance pour tester des hypothèses d'indépendance (ces deux mots apparaissent-ils ensemble par hasard ?). Cependant, comme c'est souvent le cas pour les cooccurrences de mots ces suppositions ne sont pas réalistes car les fréquences de cooccurrences sont faibles tandis que les valeurs marginales sont élevées en raison de la distribution de Zipf des mots dans un corpus.

Nous nous laissons quelque peu emporter ici : qu'entend-on par distribution de Zipf ? Que sont les tables de contingence ? Et que sont les valeurs marginales ?

La loi de Zipf qui remonte à 1949 et qui traite du comportement humain en général et du principe du moindre effort, nous enseigne que dans tout texte il y a peu de mots fréquents et beaucoup de mots n'apparaissant qu'une seule fois, en fait Zipf avait formulé quelque chose de plus complexe mais le fait frappant dans toute statistique textuelle est bien que dans un texte plus de la moitié des mots n'apparaissent qu'une fois (on leur a même donné un nom, on les appelle *hapax*), les mots n'apparaissant que deux fois sont aussi très nombreux, rendant le nombre de mots décents aux yeux du statisticien ridiculement faible. Il faut amasser un corpus de blogs de 60 Mo pour collecter 10000 mots apparaissant plus de 5 fois. Il faut modérer cette terrible constatation en remarquant que si l'on remplace dans un texte les mots apparaissant moins de 5 fois par un même signe sans signification (un petit cercle par exemple) le texte reste compréhensible, ce qui semble bien confirmer le rôle de l'isotopie dans la compréhension des textes, c'est ce principe qui nous permet d'imaginer ce qui pourrait se trouver à la place de ces signes vides de sens. Toujours est-il que la loi de Zipf pose bien des problèmes à l'honnête statisticien qui aime les distributions en cloche, qu'il appelle d'ailleurs normales, et qui se trouve dépourvu d'une grande partie de son arsenal mathématique devant de telles bizarreries statistiques. En tous cas, cette

distribution particulière explique que si les mots fréquents sont rares, plus rares encore (au regard des potentialités) sont leurs cooccurrences. Venons-en aux tableaux de contingences et leurs valeurs marginales élevées. Le tableau ci-dessous montre une table de contingence de dimension 2, représentant la distribution de deux caractéristiques A et B de taille a et b respectivement dans une population de taille n.

A x B	A	Non A & B	Marge
B	x	b – x	b
A & Non B	a - x	n – a – b + x	n – b
Marge	a	n – a	n

Dans cette table x représente le nombre d'observations où les caractéristiques A et B apparaissent simultanément, b - x le nombre d'observations où la caractéristique B apparaît mais pas la caractéristiques A, a - x le nombre d'observations où la caractéristique A apparaît mais pas la caractéristiques B enfin n – a – b + x le nombre d'observations où ni la caractéristique A ni la caractéristique B n'apparaissent. Par la suite nous noterons T(x, a, b, n) une telle table. Les valeurs marginales sont n, a, n – a, b et n – b et l'on conçoit aisément qu'elles peuvent être élevées si A et B représentent des mots et que la taille et le nombre de passages dans lesquels ils apparaissent sont grands.

C'est un fait bien connu que la probabilité d'observer une telle table de contingence suit une distribution hypergéométrique :

$$f(x)=\frac{\binom{a}{x}\binom{n-a}{b-x}}{\binom{n}{b}}=\frac{a!\,b!\,(n-b)!\,(n-a)!}{n!\,x!\,(a-x)!\,(b-x)!\,(n-a-b+x)!} \quad (2)$$

Par exemple une biologiste étudie une population de **n** animaux dans une zone géographique précise. Elle prélève **b** animaux puis les relâche. Plus tard elle prélève **a** animaux dans la même population : *f(x)* représente la probabilité qu'elle en ait marqué **x** parmi ces **a** la fois d'avant. Un autre exemple, plus proche de nos centres d'intérêt est celui de passages extraits d'un corpus textuel où l'on compte les collocations de mots dans ces passages : si **n** est le nombre de toutes les collocations observées et si pour deux mots $\mathbf{m_1}$ et $\mathbf{m_2}$, **b** est le nombre de collocations contenant le premier mot et a le nombre de collocations

contenant le second : *f(x)* représente la probabilité sous l'hypothèse d'indépendance que deux mots apparaissent dans le même passage.

Nous voici donc au cœur du problème. Dans un article qui a fait date Ted Dunning [Dunning 1994] a mis en évidence les insuffisances relatives aux présupposés de distributions normales liés aux événements rares dans des tests comme l'information mutuelle. Il a proposé au contraire d'utiliser des statistiques ne relevant pas de la loi normale, en particulier l'utilisation de la loi binomiale comme modèle statistique d'apparition d'un mot dans un passage, c'est-à-dire que l'apparition d'un mot après un autre peut se modéliser comme un jeu de pile ou face avec une pièce truquée (présentant beaucoup plus de faces que de piles par exemple), justifiant l'hypothèse d'indépendance sous-jacente par la constatation que l'influence d'apparition d'un mot sur un autre s'estompe rapidement avec la distance qui les sépare. Enfin, s'appuyant sur cette modélisation il utilise un argument de maximum de vraisemblance pour définir un test de rejet de l'hypothèse nulle ne faisant pas appel à un argument de normalité. Ce test connu en anglais sous le nom de *log-likelihood-ratio* (LLR) que l'on peut traduire par *logarithme du rapport de vraisemblance* a été très largement utilisé pour ses excellents résultats. Cependant le seuil de rejet de l'hypothèse nulle par ce test est resté une décision tout à fait empirique laissée à la discrétion de chacun. Récemment, Robert C. Moore [Moore 2004] est revenu sur l'article de Ted Dunning le questionnant sur trois points fondamentaux :

1. Qu'est ce que la signifiance d'évènements rares ?
2. Quelle est la nature exacte du LLR et quel est son rapport avec l'information mutuelle ?
3. Quel est le rapport avec le test exact de Fisher ?

Nous ne rentrerons pas dans les détails des deux premières questions. Disons cependant que si l'on sait calculer rapidement et précisément le test exact de Fisher on peut calculer le bruit et la précision dus au corpus de n'importe quel test LLR y compris. Que le LLR est étroitement lié à l'information mutuelle mais surtout qu'une régression des moindres carrés montre que le LLR est presque parfaitement linéairement dépendant du test exact de Fisher. Ce résultat est extrêmement important car de nouvelles approches basées sur des approximations extrêmement précises et rapides de la fonction Γ permettent des calculs précis de la probabilité (1) pour des tables de contingences de dimension 2, nous

présentons ici une telle méthode. La partie droite de l'équation (2) suggère qu'il existe une relation *f(x+1)* et *f(x)* :

$$f(x+1)=f(x)\frac{(a-x)(b-x)}{(x+1)(n-a-b+x+1)} \text{ et } f(0)=\frac{(n-b)!(n-a)!}{n!(n-a-b)!} \quad (3)$$

Quand n et les marges sont élevées *f(0)* ne peut être calculé directement, heureusement des méthodes de calcul efficaces du logarithme de la fonction Γ existent (par exemple l'approximation de Lanczos [Pugh 2004]). Et comme *n! = n Γ(n)* on peut efficacement calculer *f(0)* :

$$f(0)=e^{\log(\Gamma(n-b+1))+\log(\Gamma(n-a+1))-\log(\Gamma(n+1))-\log(\Gamma(n-a-b+1))} \quad (4)$$

Nous serons particulièrement intéressés dans la suite de l'exposé par probabilité d'observer par chance ν fois ou plus les caractères A et B à la fois[67], cette probabilité calculée en utilisant la relation de récurrence (3) donne lieu à la formule :

$$p(T(x,a,b,n);x\geqslant\nu)=\sum_{x=\nu}^{x=min(a,b)-\nu} f(x)=\sum_{x=\nu}^{x=min(a,b)-\nu}\prod_{i=0}^{x}\frac{(a-i-\nu)(b-i-\nu)}{(i+\nu+1)(n-a-b+i+\nu+1)} \quad (5)$$

Il est par conséquent possible de calculer cette probabilité pour des tables de contingences de dimension 2 même lorsque n est grand puisque le produit intérieur de la formule peut être accumulé au cours du calcul, de plus le calcul peur être stoppé lorsque le terme le plus interne atteint zéro du fait des limitations machine.

Nous avons testé cette formule pour de petites valeurs ne nécessitant pas l'utilisation de la fonction Γ et la relation de récurrence et trouvé des valeurs semblables, par exemple :

- P (T (10, 25, 50, 165)) = 0.1806094626960073 avec cette formule contre 0.18060946 pour un calculateur Web[68].

Pour de plus grandes valeurs nous avons vérifiés que P (T (0, a, b, n)) = 1 nous avons par exemple obtenu :

- P (T (0, 10000, 100000, 4000000)) = 0.999999997040778.

Nous avons donc la possibilité de calculer la probabilité d'observer par chance une collocation dans un ensemble de passages grâce au test exact de Fisher et comme le souligne Moore dans son article sur la signifiance des événements rares [Moore 2004],

[67] souvent appelée *p-value* dans la littérature statistique de langue anglaise.

[68] http://www.psych.ku.edu/preacher/fisher/fisher.htm

nous n'avons pas de raisons, vu la faible pénalité de temps de calcul, d'utiliser le rapport de vraisemblance, l'information mutuelle ou d'autres tests pour décider du seuil de signifiance des cooccurrences observées.

Dans la suite de cet exposé c'est le test de Fisher qui sera utilisé pour filtrer les cooccurrences ayant une probabilité d'erreur de première espèce supérieure à un seuil α d'apparaître.

Le filtrage de la matrice de cooccurrence C^tC du paragraphe précédant pour un niveau de vraisemblance 1 - α donné a, comme on peut se l'imaginer, un impact important sur la décomposition LSI traditionnelle. Avant d'en aborder toutes les implications nous voudrions montrer qu'il est possible de factoriser la matrice résultante de façon très simple. Mais tout d'abord précisons notre modèle : nous considérons un ensemble de N passages dans un ensemble de M documents que nous appellerons généralement le corpus. Nous voulons mettre en évidence, sur la base de collocations signifiantes, des classes de passages correspondants peu ou prou à des molécules sémiques qui, une fois réintroduites dans leurs documents d'origine, mettraient en évidence les isotopies sur lesquelles elles se détachent. Pour nous aider à identifier les classes de passages nous nous appuyons sur une factorisation de la matrice de cooccurrence en facteurs/pivots principaux respectant en cela la démarche initiale de la LSI. La méthode de classification des passages se base sur une fonction noyau comme cela a déjà été évoqué dans la section 2. Cette fonction noyau est construite comme une matrice de diffusion de la matrice des collocations filtrées.

Le filtrage de la matrice de collocations se base donc sur le test exact de Fisher pour des tableaux de contingences T (c (x, y), c (x, .), c (y, .), c (., .)) où :

- c (x, y) est le nombre de collocations du terme x et du terme y,
- c (x, .) est le nombre de collocations du terme x avec un quelconque autre terme,
- c (y, .) est le nombre de collocations du terme y avec un quelconque autre terme,
- c (., .) est le nombre total de collocations

Ces collocations sont comptées passage après passage. Ainsi, si $c_k(x, y)$ est le nombre de collocations dans le passage *k :*

$$c(x, y) = \sum_k c_k(x, y) \quad (6)$$

et de la même façon :

$$c(x,.)=\sum_k c_k(x,.) \quad (7)$$

et

$$c(.,.)=\sum_k c_k(.,.) \quad (8)$$

On a aussi :

$$c_k(x,y)=\begin{cases}|x|_k|y|_k & \text{si } x\neq y\\ \dfrac{|x|_k(|x|_k-1)}{2} & \text{sinon}\end{cases} \quad (9),$$

$|x|_k$ étant le nombre de mots x dans le passage k. De même :

$$x_k(x,.)=\sum_{y\in Ctx_k} c_k(x,y)$$

Ctx_k est l'ensemble des termes du passage k.

De la matrice C^tC nous retiendrons la matrice $B=(b_{i,j})_{\substack{1\leqslant i\leqslant p\\ 1\leqslant j\leqslant p}}$ où $b_{i,j}=c(x_i,y_j)\alpha_{i,j}$ ou $\alpha_{i,j}$ est une fonction du test exact de Fisher et d'un niveau de signifiance α préétabli par exemple si $P_{i,j}$ est la probabilité $p(T(x,c(x_i,.),c(x_j,.),c(.,.);x\geqslant c(x_i,x_j)))\geqslant\alpha$ on pourra utiliser :

$$\alpha_{i,j}=\begin{cases}\dfrac{\alpha-p_{i,j}}{\alpha} & \text{si } p_{i,j}\leqslant\alpha\\ 0 & \text{sinon}\end{cases} \quad (10)$$

La matrice *B* peut être vue comme un graphe pondéré non orienté dont les sommets sont des mots et les arcs des relations de cooccurrences signifiantes entre ces mots le poids de la relation entre un mot *i* et un mot *j* étant la valeur $b_{i,j}$. Un vecteur colonne b_j de la matrice *B* est une représentation dans un espace à *p* dimensions des collocations signifiantes entre le mot *j* et les autres mots. Une technique mathématique universellement employée pour étudier un ensemble de points (on dit un *nuage de points*, ce qui est plus joli) dans un espace est de rechercher ses axes principaux d'élongation. Ainsi, un ensemble de points régulièrement contenu dans une ellipse et ayant tous la même pondération auront pour axe principaux d'élongation les deux axes principaux de l'ellipse, pour un cercle toute paire d'axes perpendiculaires se croisant au centre du cercle fera l'affaire, bien sûr si le nuage de

points s'étire le long d'une parabole les axes principaux paraîtrons moins significatifs quoiqu'en utilisant une fonction noyau différente du produit scalaire usuel $\langle x, y \rangle$ mais $(\langle x, y \rangle + 1)^3$ par exemple les axes principaux s'étireront le long des deux bras de la parabole. Dans un article récent, Shawn Martin [Shawn 2006] reprend des idées développées par N. Cristianini, J. Shawe-Taylor, H. Lodhi et A. Smola dans [Cristianini et al 2001] et le livre « Kernel Methods for Pattern Analysis » [Cristianini et al 2004] ces idées tournent autour de l'articulation de la procédure d'orthonormalisation de Gram/Schmidt et de la recherche du vecteur de norme maximal dans un nuage de points pour un produit scalaire donné (une fonction noyau donnée). La procédure fonctionne comme suit : à une étape donnée on cherche parmi les vecteurs non traités celui de norme maximale, on transforme alors les vecteurs restants grâce à la procédure d'orthonormalisation de Gram/Schmidt de façon à ce qu'ils soient orthogonaux à ce vecteur de norme maximale ainsi qu'aux autres vecteurs déjà sélectionnés, puis on calcule leurs normes dans l'hyperplan orthogonal résultant et ainsi de suite. Cette procédure est à peu près équivalente à une analyse en composantes principales à ceci près que les facteurs principaux sont représentés par des points du nuage étudié et que la procédure est moins lourde qu'une procédure classique de décomposition spectrale (recherche des valeurs propres et vecteurs propres de la matrice d'inertie thème très connus en mathématiques dont on trouvera foison de cours élémentaires dans le commerce ou sur la Toile). On peut arrêter cette procédure lorsque l'on atteint un nombre de facteurs prédéfinis par avance où lorsque les facteurs sélectionnés forment une base du nuage de points. On désignera par AACP (pour Approximation d'une Analyse en Composantes Principales) ce type d'approche en général.

Dans le graphe représenté par la matrice *B* des cooccurrences significatives cette procédure peut être appliquée pour trouver les nœuds principaux du graphe ou, selon un autre point de vue, les mots autour desquels se forment les collocations le plus grands nombre de collocations significatives. Plusieurs fonctions noyau peuvent être utilisées pour ce faire, par exemple :

- $k(b_i, b_j) = b_{i,j} + \lambda \sum_{k=1}^{p} b_{i,k} b_{j,k}$ où b_i et b_j représentent des vecteurs colonnes de la matrice *B* où ce qui revient au même des nœuds du graphe des cooccurrences

signifiantes et λ un facteur d'affaiblissement, il faut noter cependant que pour que k soit un noyau il faut que ce facteur soit inférieur à l'inverse du rayon spectral de la matrice B on se référera à l'ouvrage déjà cité « Kernel Methods for Pattern Analysis » pour une démonstration.

Ce type de fonction noyau est très connu, on parle de *fonction noyau de diffusion*, ce terme se comprenant mieux si on considère que deux noeuds d'un graphe ont tendance à se ressembler si ils sont liés à des nœuds communs la ressemblance se diffuse au travers du graphe selon le principe que les amis de mes amis sont mes amis. La matrice de Gram K de cette fonction noyau s'exprime simplement par :

$$K = B + \lambda B^2 \quad (11)$$

Nous pouvons cependant utiliser d'autres fonctions noyau pour effectuer une AACP. Nous retiendrons en particulier la fonction suivante :

$$k(b_i, b_j) = \begin{cases} b_{i,j}^2 & \text{si } i \neq j \\ \sum_{l=1}^{p} b_{i,j}^2 & \text{sinon} \end{cases} \quad (12)$$

Si l'on se souvient qu'une fonction noyau est un produit scalaire dans un espace de plus grande dimension[69] que l'espace de départ, la fonction qui nous intéresse opère dans l'espace des matrices symétriques de dimension p sur un ensemble défini par le projection des colonnes de la matrice B par la fonction ϕ :

$$\Phi(b_i) = \begin{pmatrix} 0 & . & 0 & b_{i,1} & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & . & 0 & . & 0 & . & . & . & . \\ b_{i,1} & . & . & b_{i,i} & . & . & . & . & b_{n,1} \\ . & . & 0 & . & 0 & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & . & 0 & b_{n,1} & 0 & . & . & . & 0 \end{pmatrix} \quad (13)$$

ou bien encore:

[69] Et même très souvent infini comme c'est le cas pour la fonction radiale que nous avons utilisé dans l'exemple des anneaux de points concentriques de la section 2.

$$\Phi(b_i)=(x_{r,s})_{\substack{1\leqslant r\leqslant p\\1\leqslant s\leqslant p}} \text{ où } x_{r,s}=\begin{cases} b_{i,s}^2 \text{ si } i=r \\ b_{i,s}^2 \text{ si } i=r \\ 0 \text{ sinon} \end{cases} \quad (14)$$

ainsi $k(b_i,b_j)-\frac{1}{2}\langle\Phi(b_i),\Phi(b_j)\rangle$ où $\langle .,. \rangle$ est le produit scalaire de Frobenius.

Pour cette fonction noyau, deux sommets du graphe des cooccurrences, représenté par la matrice *B,* sont orthogonaux s'ils ne sont pas connectés. Nous en arrivons donc au but de cette présentation technique : l'utilisation des cooccurrences significatives pour représenter les mots du corpus par certains d'eux (dont on voudrait qu'ils soient les plus significatifs). Comme nous l'avons vu les facteurs d'une AACP sont des mots autour desquels se forment les cooccurrences signifiantes de la matrice *B* et donc tout mot du réseau de cooccurrences va pouvoir se projeter sur ces mots facteurs comme on le fait usuellement dans toute analyse en composantes principales avec les facteurs principaux, on reporte en annexe les détails de l'algorithme d'une extrême simplicité dans ce cas particulier préférant une représentation visuelle plus intuitive.

Soit le graphe représenté par la matrice à la figure 9: pour plus de simplicité tous les poids sont égaux à 1. Mais d'autres poids pourraient être utilisés, la figure 10 représente le graphe lui-même.

Selon la fonction noyau précédemment définie, le sommet de plus grande norme de ce graphe est le numéro 14. Le sous-graphe orthogonal à ce sommet au sens de cette fonction est obtenu en le retirant du graphe initial comme cela est montré à la figure 6. Le sommet de plus grande norme dans ce nouveau graphe est le sommet 1 comme montré à la figure , puis à la figure 11 le sommet 1 est retiré et ainsi de suite jusqu'à l'obtention du nombre de facteurs désiré ou l'isolement de tous les sommets.

Si l'on ne conserve que les facteurs 14, 1 et 13 : le point 2 sera représenté par le triplet (1, 1, 0) et le point 8 par le triplet (1, 0, 1). D'une façon générale chaque point du graphe peut ainsi être projeté sur ces facteurs. Ainsi, dans le cas qui nous intéresse chaque mot pourra être projeté sur un ensemble de mots facteurs nous permettant ainsi de représenter chaque mot par projection sur un ensemble de mots facteurs dont l'importance se décide en fonction du nombre de cooccurrences signifiantes qu'il peut agréger.

$$\begin{pmatrix}
0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0
\end{pmatrix}$$

Figure 9: Matrice B

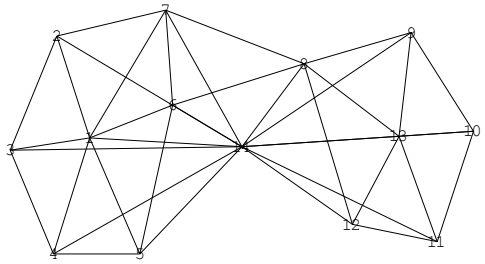

Figure 10: *Graphe Correspondant à la matrice B*

14

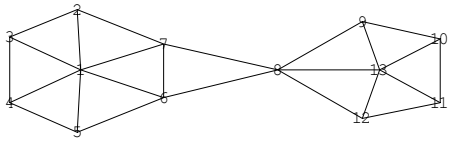

Figure 11: *Graphe correspondant au sous-graphe orthogonal au sommet 14*

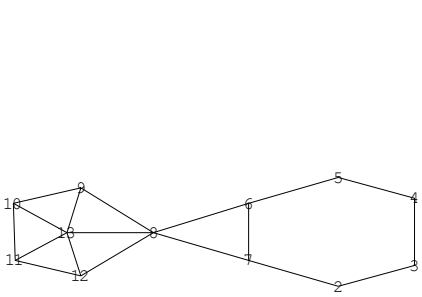

Figure 12 : Graphe correspondant au sous-graphe orthogonal au sommet 14 et 1

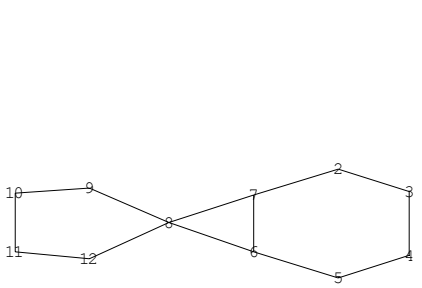

Figure 13: Graphe correspondant au sous-graphe orthogonal au sommet 14, 1 et 13

À proprement parler les facteurs que nous mettons en évidence ne sont pas orthogonaux puisqu'il existe une relation de cooccurrence entre le nœud 14 et les noeuds 1 et 13. Il en serait de même avec une approche plus conventionnelle comme celle décrite dans [Cristianini et al. 2004]. En effet quand on parle des facteurs 1 et 13 on ne parle pas des noeuds 1 et 13 du graphe original mais de la projection du noeud 1 sur l'hyperplan orthogonal au noeud 14 au sens de la fonction noyau choisie pour le facteur 1 puis de la projection du noeud 13 sur l'hyperplan orthogonal aux facteurs 14 et 1 pour le facteur 13. En d'autres termes comme dans la procédure d'orthonormalisation de Gram/Schmidt. Cet état de choses est illustré à la figure 12 à partir des deux points A et B de coordonnées respectives (1,1) et (0,1) on construit en partant du point A un second point déduit de B de coordonnées $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ qui est la projection de B sur la droite passant par l'origine perpendiculaire à A. De la même manière si la cooccurrence ('voyage', 'ville') est

significative il se peut très bien (et c'est d'ailleurs le cas dans notre corpus d'étude) que 'voyage' et 'ville' soient deux facteurs lesquels s'expriment un nombre très important de collocations, si le mot 'voyage' est de norme supérieure à 'ville' au sens de la fonction noyau retenue alors le mot 'ville' aura une composante sur le mot voyage, alors que le facteur 'ville' sera orthogonal au facteur 'voyage'.

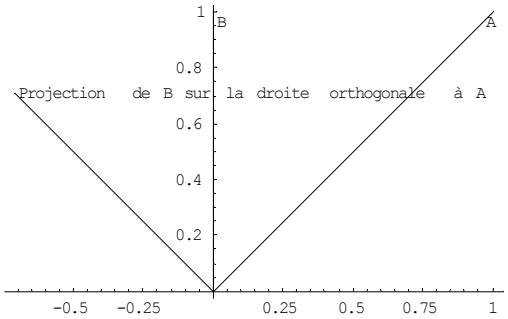

Figure 14: Exemple de procédure d'orthonormalisation de Gram/Schmidt

Si l'on revient maintenant au problème de Ted Dunning concernant l'utilisation des tests statistiques en LSI, nous pensons y avoir répondu dans la mesure où nous avons bien décrit une méthode permettant la décomposition : $C = T_{pr} S_{rr} X_{rn}$ de la LSI puisque si chaque mot peut être décomposé sur les facteurs de la matrice *B* il en est de même des passages considérés comme des ensembles de mots donnant lieu au produit $S_{rr} X_{rn}$ pour les *n* passages et les *r* facteurs, la matrice $T_{pr} S_{rr}$ étant la matrice de projection termes sur les facteurs. Comme nous le verrons dans la partie consacrée à l'application de la méthode ce résultat est intéressant en soit, mais il est temps de revenir à la sémantique interprétative et de comprendre comment la LSI, vue sous cet angle, va nous permettre de regrouper des classes de passages dans le but de tenter de mettre en valeur de possibles molécules sémiques. Avant d'aborder ce sujet, résumons ce que nous avons mis en évidence dans cette section.

La sémantique latente permet une représentation plus riche des passages en utilisant la notion de cooccurrence de termes et la projection sur des axes principaux afin de mettre en évidence des relations entre passages qui resteraient cachées dans le modèle traditionnel de Salton. Ces cooccurrences sont cependant utilisées sans tenir compte de leurs signifiances statistiques. L'utilisation de méthodes performantes d'évaluation de la fonction Γ permet des calculs précis et rapides du test exact de Fisher pour des tables de contingence de

dimension 2. Cette possibilité offre une utilisation traditionnelle des seuils de signifiance des erreurs de première espèce plutôt que des seuils empiriques de la fonction de maximum de vraisemblance pour filtrer les cooccurrences signifiantes. Les collocations signifiantes retenues peuvent être factorisées en utilisant une fonction noyau permettant de décomposer un ensemble de passages selon le schéma classique de la sémantique latente tout en tenant compte de la signifiance statistique des cooccurrences sur lesquelles elle se base. Par ailleurs la méthode utilisée représente naturellement les facteurs principaux comme des mots significatifs du corpus.

4.4 Classes de passages, molécules sémiques et isotopies

Nous arrivons donc à la liaison entre sémantique latente et sémantique interprétative. Plusieurs raisons nous ont guidé dans cette démarche. Outre le fait que la notion d'isotopie ne dépend pas directement du niveau syntaxique et qu'un modèle dérivé de celui de Salton et de l'étude des cooccurrences permet donc d'en modéliser certains aspects, il nous a semblé intéressant de se faire une idée des améliorations que la sémantique interprétative pourrait apporter aux systèmes d'interrogation et plus largement d'exploration et de cartographie des bases de données textuelles. Nous avons délibérément ignoré l'information extra textuelle comme les liens hypertextuels non pas que nous en sous-estimions l'importance mais parce que l'étude des graphes de liens hypertextuels est en soi un vaste champ d'étude : la tâche aurait été insurmontable. Par ailleurs, rien dans nos présupposés n'empêche l'intégration future de cette information. Dans cette partie, nous nous intéresserons donc principalement à la classification de X_{rn} autrement dit à la classification des passages en fonction des facteurs. Chaque passage est représenté par un vecteur colonne X_i de dimension *r* de X_{rn} ; une fonction noyau usuelle permettant de comparer ces vecteurs est le cosinus :

$$\cos(X_i, Y_j) = \frac{\langle X_i, X_j \rangle}{\|X_i\| \|Y_{rl}\|}$$

où $\langle ., . \rangle$ est le produit scalaire usuel de $\mathbb{R}^r$ et sa $\|.\|$ norme.

Comme nous l'avons vu à la section 2, nous utilisons une méthode basée sur la densité d'une fonction noyau, formellement, si *k* est une telle fonction noyau et θ seuil de densité

prédéterminé, on cherche une partition Π de n objets telle que la densité intraclasse soit supérieure ou égale à ce seuil et la densité interclasse inférieure, on cherche donc :

$$\max_{\Pi \in \text{ partitions}} (\Omega(\Pi)) \text{ où } \Omega(\Pi) = \sum_{\substack{C \in \Pi \\ |C|>1}} \sum_{x \in C} \left(\frac{\sum_{y \in C, y \neq x} k(x,y)}{|C|-1} - \theta \right) \text{[70]} \quad (15)$$

sous les contraintes :

$$\frac{\sum_{y \in C, y \neq x} k(x,y)}{|C|-1} \geqslant \theta, \ \forall C \in \Pi \text{ et } |C|>1 \quad (16)$$

$$\frac{\sum_{x \in C_i, y \in C_j} k(x,y)}{|C_i||C_j|} < \theta, \ \forall (C_i, C_j) \in \Pi \times \Pi \quad (17)$$

où Π est une partition de l'ensemble d'objets à classifier, soit les vecteurs colonnes X_i de X_{rn} dans le cas qui nous intéresse, la fonction noyau k étant la fonction cosinus. La contrainte (16) assure que la densité de chaque classe de cardinal supérieur à 1 est supérieure à θ, ainsi pour toute partition dont au moins une classe contient plus d'un élément dans (15) $\Omega(\Pi)$ est strictement supérieure à θ. La contrainte (17) permet de d'assurer que la densité entre deux classes est toujours inférieure à θ.

La partition est calculée par les algorithmes 1 et 2. L'algorithme 2 construit une partition Π satisfaisant la condition (16) en affectant progressivement les éléments à classifier aux classes existantes selon que leurs valeurs moyennes de la fonction noyau est supérieure au seuil θ ou créant une nouvelle classe si aucune classe existante ne satisfait à cette condition.

L'algorithme 2 permet d'améliorer cette partition initiale en remarquant que si un élément x appartenant à une classe C_x, le transfert de x à une nouvelle classe C change la valeur de la fonction objective Ω dans (15) par :

[70] $|C|$ est le cardinal de la classe C

$$\begin{cases} \theta - \dfrac{\sum_{y \in C_x, y \neq x} k(x,y)}{|C_x| - 1} & \text{si } x \text{ est transféré dans un nouveau cluster} \\ \dfrac{\sum_{y \in C, y \neq x} k(x,y)}{|C|} - \theta & \text{si } C_x = \{x\} \\ \dfrac{\sum_{y \in C, y \neq x} k(x,y)}{|C|} - \dfrac{\sum_{y \in C_x, y \neq x} k(x,y)}{|C_x| - 1} & \text{sinon} \end{cases} \quad (18)$$

L'algorithme 2 calcule itérativement pour chaque élément x l'amélioration qu'un possible transfert de x apporterait à la fonction Ω :

1. Si $\dfrac{\sum_{y \in C_x, y \neq x} k(x,y)}{|C_x| - 1} < \theta$ et $\dfrac{\sum_{y \in C} k(x,y)}{|C|} < \theta$ alors x est transféré dans une nouvelle classe.
2. Si le transfert de x apporte une amélioration à la fonction Ω alors x est transféré dans la classe induisant la meilleure amélioration.

L'algorithme se termine lorsque plus aucun transfert n'est possible ou quand un nombre limite de transferts est atteint. Enfin, afin de satisfaire la contrainte (17) les classes dont la densité interclasse est plus grande que le seuil θ sont progressivement jointes.

Π= Φ
Pour chaque x in X {
 meilleurGain = 0
 Pour chaque C dans Π {
 gain = $\frac{1}{|C|} \sum_{y \in C,\, y \neq x} k(x, y)$
 Si(meilleurGain < gain) {
 meilleurGain = gain
 meilleureClasse = C
 }
 }
 Si(meilleurGain > θ) {
 C = meilleureClasse
 $\Pi = (\Pi - \{C\}) \cup (C \cup \{x\})$ /* Affectation de x à C */
 } Sinon {
 C = □/* Création d'une classe vide */
 $\Pi = \Pi \cup (C \cup \{x\})$ /* Affectation de x à C */
 }
}

Algorithme 1 Recherche d'une partition initiale.

Π= $\Pi_{initial}$
Faire {
 Pour chaque x dans X {
 Soit C_x la classe de x
 $\text{contributionCourante} = \frac{1}{|C_x| - 1} \sum_{y \in C_x,\, y \neq x} k(x, y)$
 meilleurGain = 0
 Pour chaque C dans Π {
 gain = $\frac{1}{|C|} \sum_{y \in C,\, y \neq x} k(x, y) - \text{contributionCourante}$
 Si(meilleurGain < gain) {
 meilleurGain = gain
 meilleureClasse = C
 }
 }
 Si(bestGain > θ) {
 C = meilleureClasse
 $\Pi = (\Pi - \{C\}) \cup (C \cup \{x\})$ /* Affectation de x à C */
 } Sinon {
 C = □/* Création d'une classe vide */
 $\Pi = \Pi \cup (C \cup \{x\})$ /* Affectation de x à C */
 }
 }
}Tant que (Π a changé et que le nombre d'itérations n'est pas trop élevé)

Algorithme 2 Amélioration d'une partition initiale.

Comme nous l'avons vu dans la section 2 avec un exemple simple cette méthode de partitionnement regroupe les objets en classes homogènes selon la fonction noyau mais l'interprétation des partitions obtenues dépend de la nature des objets classifiés, de la fonction noyau utilisée et bien sûr et surtout de ce que l'on cherche à mettre en évidence. Ce que l'on cherche à mettre en évidence ce sont des classes de passages que l'on pourrait caractériser par des sèmes, cette caractérisation correspondant à son tour à une classe de molécule sémique. Il y a loin de la coupe aux lèvres mais notre hypothèse se fonde, comme cela a été constaté ailleurs [Yarowski 1995], sur le fait que les collocations signifiantes font souvent sens, ainsi peut-on espérer que les facteurs principaux du réseau de collocations sont de bons indicateurs du sens des passages qu'ils décrivent et que les regroupements qu'ils provoquent renforcent ces indications. Notre objectif est de décrire la nature exacte du sens révélé par ces regroupements à l'aide de molécules sémiques : ou de façon plus simplificatrice d'un ensemble d'étiquettes pondérées. A première vue le saut est brutal entre molécules sémiques et jeux d'étiquettes pondérées, et il convient de nous justifier sur ce point. Que nous dit le petit glossaire du sémanticien[71] à ce propos ?

Molécule sémique : groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier. Un « thème », quand il peut être défini sémantiquement, n'est autre qu'une molécule sémique.

Sème : élément d'un sémème, défini comme l'extrémité d'une relation fonctionnelle binaire entre sémèmes. Le sème est la plus petite unité de signification définie par l'analyse.

Partant de là, il nous semble justifiable de décrire une molécule sémique par autre chose que des sèmes, surtout si, comme c'est ici le cas, l'on se trouve enfermé dans le cadre étroit d'une machine. En effet, ce qui nous semble important dans le cadre de l'indexation, c'est la notion de stabilité. D'autre part, dans la définition du sème elle-même rien n'est directement dit de sa représentation formelle. En particulier la notation entre traits obliques n'oblige en rien la représentation formelle qui se situe à un autre niveau. La représentation par des graphes conceptuels pour pratique qu'elle soit n'est peut-être pas non plus un passage obligatoire. Toute représentation formelle du sens en est un appauvrissement et comme le faisait remarquer Greimas [Greimas 1983] il n'est pas facile de dire des « choses sensées sur le sens ».

[71] http://www.revue-texto.net/Reperes/Glossaires/Glossaire_fr.html

Qu'on nous pardonne donc cette représentation d'une molécule sémique qui comme l'on s'en doute n'est pas dénuée d'arrière-pensées. Notre recherche de stabilité dont il vient d'être fait mention se fait au travers de regroupements de passages en classes grâce à leur description par les facteurs principaux du graphe des cooccurrences signifiantes. Or cette description est essentiellement une représentation vectorielle indexée par les facteurs, c'est-à-dire une suite ordonnée d'étiquettes (les facteurs) pondérées (la projection des mots des passages sur ces facteurs). Ce sont ces vecteurs qui sont en fait classifiés et ce sont des fonctions de ces vecteurs qui servent à représenter les classes.

Mais il est une autre raison que nous pensons plus profonde et qui explique le lien entre sémantique latente et sémantique interprétative : une molécule sémique est un groupement stable de sèmes. Sans vouloir forcer le trait[72] on peut, d'une certaine façon, faire l'analogie entre cooccurrences de sèmes et cooccurrences de mots. Ainsi à une ressemblance induite par la projection sur des facteurs communs de cooccurrences de mots signifiantes il est possible de faire correspondre une ressemblance induite par une cooccurrence de sèmes. C'est d'ailleurs un argument des tenants de la sémantique latente lorsqu'ils mettent en avant que la projection des mots sur des facteurs principaux permet de reconnaître des ressemblances entre des documents qui ne partagent pas de mots significatifs[73].

Bien qu'il soit fait mention dans la littérature sur la sémantique latente de la possibilité d'utilisation d'unités plus courtes que le document entier, nous n'avons pas connaissance de développement dans ce sens. Notre approche au contraire, en cherchant à mettre en évidence des molécules sémiques privilégie, comme le fait Mathias Rossignol [Rossignol 2005], des passages plus restreints. Bien qu'utilisant des techniques différentes, il existe des similarités importantes entre son approche et la nôtre. Sans vouloir être outrageusement réducteur, elle consiste à classifier le vocabulaire d'un corpus thématique selon une technique de vraisemblance du lien, c'est-à-dire selon une analyse statistique de la pertinence des regroupements. Il classifie alors les paragraphes en fonction de ces classes de mots afin d'améliorer la classification initiale du vocabulaire. Un graphe synthétise les résultats obtenus, les parties denses du graphe obtenu mettant en évidence des domaines sémantiques. D'une façon générale, c'est la dualité entre classes de mots et classes de passages (syntagmes, paragraphes ou document) qui met en évidence des relations sémantiques au sein du corpus. C'est cette même dualité, quoique envisagée de façon

[72] Sans jeu de mots.

[73] C'est à dire n'appartenant pas à une liste d'exclusion comme les mots grammaticaux, etc.

différente, que nous voulons exploiter. C'est en fait une façon naturelle de procéder du point de vue de la sémantique interprétative qui ne considère de sens qu'en contexte. Nous utilisons le terme générique de passage dans la mesure où nous pensons que cette démarche peut être reproduite à différents paliers ; nous nous sommes cependant restreint, dans l'objectif d'indexation qui est le notre, au niveau méso-sémantique, c'est-à-dire d'un point de vue plus terre à terre toujours dicté par la contrainte machine au titres, aux paragraphes et aux énumérations quand il nous est possible de les détecter par des expressions rationnelles.

Mais pourquoi, dans un objectif d'indexation, s'évertuer ainsi à vouloir traiter ce niveau de passage ?

Parce que notre but final est la caractérisation des documents par les isotopies génériques sur lesquelles se détachent les molécules sémiques. Les molécules sémiques peuvent être vues, en effet comme des formes qui se détachent sur un fond que sont les isotopies. Or, en indexation, c'est ce fond qui nous intéresse, plus particulièrement si il est récurrent. Un peu à l'image de l'analyse des photos satellites, nous nous intéressons à la texture du sol : cette zone représente-t-elle un champs de colza, un marécage, une forêt, un désert de sable, ou autre chose encore ? Traduit dans le langage de la sémantique latente et de la classification nous cherchons à caractériser les documents par la récurrence de *facteurs caractéristiques* de *classes de passage*. Dans un processus automatisé à l'extrême, ce qui une fois encore n'est pas ce que l'on recherche, les facteurs tiendraient le rôle de sèmes, les classes de passages celui de molécules sémiques et les facteurs caractéristiques de ces classes celui d'isotopie. Pour caricaturale qu'elle soit, cette description a le mérite de synthétiser notre démarche. Ce qui compte cependant avant tout à nos yeux est l'interprétation humaine, qui permet d'apporter un sens réel à cette indexation mécanique comme nous avons voulu le souligner dans la section 2. La méthode d'indexation que nous proposons est donc un compromis entre qualité et efficacité : qualité apportée par l'interprétation humaine, efficacité apportée par la machine ; une interprétation sous contrainte. Nous envisageons l'intervention humaine lors de différentes étapes du traitement :

- la définition des passages et la sélection du vocabulaire,
- la validation des facteurs,
- la caractérisation des classes de passages et leur validation,

- la validation de la classification des documents en fonction des isotopies retenues grâce à la caractérisation des classes de passages.

Ces interventions humaines impliquent des allers-retours entre elles : le rejet d'un facteur pouvant impliquer le rejet d'un ou plusieurs mots du vocabulaire. De la même façon le rejet d'une classe de passages considérée comme non pertinente peut remettre en évidence la pertinence d'un facteur ou d'une partie de ses collocations associées. Enfin des isotopies provoquant des regroupements incompréhensibles de documents peuvent amener à revoir des classes de passages. Ce va-et-vient entre ces différentes étapes peut être long et fastidieux selon l'exigence de qualité attendue, la nature du corpus et l'objectif de l'indexation selon la typologie de Hanne Albrechtsen présentée dans la section 3.

Quel apport attend-on de ce processus ?

Tout d'abord une présentation ergonomique des résultats d'une requête sur un fonds documentaire textuel, une interface intelligente entre un moteur de recherche et l'utilisateur mais en même temps la réutilisation des résultats de recherches précédentes par d'autres utilisateurs, une forme de filtrage collaboratif. En effet, une fois jugée acceptable une analyse peut-être figée sous la forme d'un classificateur automatique à vaste marge apposant des étiquettes d'isotopie à un texte. De tels classificateurs peuvent être combinés entre eux pour en former de nouveaux. On peut envisager ainsi des pendants aux traditionnelles ontologies. Mais au-delà de ces considérations nous pensons qu'une intégration de ce type d'approches au cœur même des moteurs de recherche permettrait d'en améliorer qualitativement les performances.

Nous avons jusque là présenté la méthode générale de partitionnement des passages et la justification de notre démarche. Il est maintenant temps d'entrer dans le détail de notre méthode d'analyse de ses partitions, de la façon de les caractériser, d'en caractériser les classes et des liens qu'elles entretiennent. Nous allons aussi montrer que le modèle permet de traiter de très grands volumes de données. Ce sera l'objet des prochaines sections.

4.5 Analyse d'une partition de passages

Nous avons abordé dans la section 2 les principes généraux d'analyse d'une partition nous proposons dans cette partie d'en approfondir l'application dans le cas de l'analyse d'une partition de passages en fonction des facteurs. Notre approche sera plutôt formelle et

technique, une application effective de ce type d'analyse étant reporté à la section 6. Un premier point souvent accepté comme allant de soi est l'utilisation de la fonction cosinus comme fonction noyau quand d'autres fonctions noyau pourraient tout à fait être utilisées. Gerard Salton en donnait la justification suivante en 1975:

> « Instead of Identifying each document by a complete vector originating at the 0 point in the coordinate system, the relative distance between the vectors is preserved by normalizing all vector lengths to one and considering the projection of the vectors onto the envelope of the space represented by the unit sphere. In that case, each document may be depicted by a single point whose position is specified by the area where the corresponding document vector touches the envelope of the space. Two documents with similar index terms are then represented by points that are very close together in the space. » [Salton, Yang 1975]

On voit deux idées principales se dégager ici : tout d'abord une volonté de normalisation de la représentation des documents de façon à les rendre indépendants de leurs tailles ensuite un souci de clarté de cette même représentation. En effet, un avantage évident de représenter les documents comme des points sur l'enveloppe de l'hypersphère de rayon 1 permet une représentation visuelle aisée de la proximité de deux documents (comme la représentation de figure 1 par exemple). D'autre part cette normalisation à l'immense avantage de simplifier considérablement les calculs de similarité entre documents d_i et d_j qui se trouvent ramenés à de simples produits scalaires la distance s'en déduisant naturellement comme la norme de leur différence :

$$\|d_i - d_j\|^2 = \|d_i\|^2 + \|d_j\|^2 - 2\|d_i\|\|d_j\|\cos(d_i, d_j)$$

soit $2 - 2\langle d_i, d_j \rangle$ puisque ce sont des vecteurs normalisés. Entre autres choses, certaines formules des algorithmes 1 et 2 s'en trouvent considérablement simplifiées, ainsi :

$$\frac{1}{|C|} \sum_{y \in C,\, y \neq x} k(x, y) = \langle x, \frac{1}{|C|} \sum_{y \in C,\, y \neq x} y \rangle \quad (19)$$

l'expression $\frac{1}{|C|} \sum_{y \in C} y$ est en fait le centre de gravité des documents de la classe C, centre de gravité que l'on appelle *centroid* en anglais et que l'on appelle donc souvent *centroïde* en français parce que c'est plus court et plus mystérieux. Toujours est-il que le centre de gravité dans un modèle vectoriel est d'une grande utilité et permet de simplifier nombre de calculs, nous noterons G_C le centre de gravité de la classe C.

La densité d'une classe C telle que nous l'utilisons dans l'algorithme de partitionnement déjà présenté s'exprime par la formule :

$$\langle G_C, G_C \rangle = \|G_C\|^2 \quad (20)$$

La densité interclasse entre deux classes C_i et C_j s'exprime comme :

$$\langle G_{C_i}, G_{C_j} \rangle \quad (21)$$

La similarité moyenne d'un passage c avec les passages d'une classe C s'exprime comme :

$$\langle c, G_{C_j} \rangle \quad (22)$$

Ces premiers indicateurs permettent de regrouper les classes dont la densité interclasse est supérieure au seuil de densité θ de l'algorithme 1. Ils permettent aussi de construire le graphe de similarité entre classes permettant des analyses du type de celle décrites dans la section 2 :

- détection de composantes connexes,
- analyse de centralité et de densité des classes.

La similarité moyenne d'un objet, ici d'un passage, à une classe permet de lever une ambiguïté longtemps entretenue entre classification floue et classification non-floue : en effet, grosso modo, la décision d'affecter un objet à une classe dans les algorithmes 1 et 2 se fait en fonction de la valeur $\langle c, G_{C_j} \rangle$ aussi lorsque l'algorithme se termine c'est une pure décision quant à la présentation des résultats que de dire que tel passage c appartient à une classe ou de dire, pour chaque classe C d'une partition Π, quel est son degré d'appartenance à cette classe, soit le ratio :

$$\frac{\langle c, G_{C_j} \rangle}{\sum_{C' \in \Pi} \langle c, G_{C'} \rangle} \quad (23)$$

On le voit la différence entre partition floue et partition non floue est tout à fait artificielle. Dans une partition non floue l'appartenance d'un passage à une classe correspond à la sélection de la classe pour laquelle le degré d'appartenance donné par la précédente formule est le plus élevé. C'est d'ailleurs cette propriété qui permet de calculer la densité interclasse $\langle G_{C_i}, G_{C_j} \rangle$. La méthode de segmentation utilisée n'est à proprement parler qu'une

permutation simultanée des lignes et des colonnes de la matrice de Gram de la fonction noyau agglutinant les cellules de valeur plus élevée de cette matrice le long de la diagonale. Les classes finales étant des agrégats le long la diagonale dont la valeur moyenne dépasse un certain seuil comme on peut le constater sur la figure 6 et surtout 4 où les 3 classes de la figure 1 correspondent à des composantes connexes de la figure 4.

Dans la section 2 nous faisions allusion aux thèmes caractérisant une classe, thèmes que l'on ne doit pas confondre avec ceux de la sémantique. Une classe de passage sera caractérisée par des facteurs du graphe de cooccurrences (on identifie les thème caractérisant une classe aux facteurs). Que signifie les notions d'intensité et de spécificité d'un facteur?

- Un facteur spécifique à une classe caractérise cette classe par rapport au tout, elle est un élément de contraste. Il affecte peu d'éléments en dehors de la classe.
- L'intensité d'un facteur marque son importance dans la classe, il affecte la plupart des éléments de la classe.

Formellement si, G_C est le centre de gravité de la classe C, alors il s'exprime comme un vecteur $(f_{C,i})_{1\leqslant i\leqslant m}$ où $f_{C,i}$ est la valeur du $i^{\text{ème}}$ facteur pour le centroïde[74] G_C soit $f_{C,i}=\frac{1}{|C|}\sum_{c\in C} f_{c,i}$ où $f_{c,i}$ est la valeur du $i^{\text{ème}}$ facteur pour le passage c. La spécificité du $i^{\text{ème}}$ facteur pour la classe C s'exprime par la formule :

$$\frac{|C| f_{C,i}}{\sum_{C\in \Pi} |C'| f_{C',i}} \quad (24)$$

En effet si seule la classe C contient le $i^{\text{ème}}$ facteur cet indicateur vaut 1 qui est la valeur maximale et si à l'autre extrême elle ne le contient aucunement il vaut 0.

L'intensité quant à elle s'exprime par:

$$\frac{\sum_{c\in C} f_{c,i}^{\,2}}{|C|} \quad (25)$$

En si le $i^{\text{ème}}$ facteur indexe tous les passages de la classe C à l'exclusion de tous les autres, son intensité est maximal dans cette classe et l'indicateur vaut 1 valeur maximale ; à l'autre

[74]Le mot nous a échappé

extrême si il n'indexe aucun passage de la classe il vaut 0[75]. Pour terminer sur les indicateurs de facteurs, notons que l'on peut emprunter à Gerard Salton son argument sur les termes discriminant si l'on tient compte de ce qu'ici les termes sont remplacés par des facteurs. Bien sûr l'objectif n'est pas exactement le même, le sien est qu'en dehors de toute autre information que les comptages des termes dans les documents il cherche les termes séparant le plus les documents. Ici nous cherchons les facteurs séparant au mieux les centres de gravité des classes et donc la quantité $Q = \sum_{(d_i, d_j) \in \Delta^2} k(d_i, d_j)$ de 4.1 se transforme en $Q = \sum_{(C_i, C_j) \in \Pi} \langle G_{C_i}, G_{C_j} \rangle$; en d'autres termes[76] un facteur discriminant est un facteur qui minimise la densité interclasse, nous souscrivons bien évidemment à cette définition. Les termes eux-mêmes peuvent être caractérisés de manière équivalente. La spécificité d'un terme se traduit par la formule :

$$\frac{t_C}{\sum_{C \in \Pi} t_{C'}} \quad (26)$$

où t_C est la fréquence du terme t dans la classe C : ainsi, si le terme t n'apparaît que dans la classe C, l'indicateur atteint sa valeur maximale 1 et si il en est absent sa valeur minimale 0. L'intensité d'un terme se traduit par la formule :

$$\sum_{c \in C} \frac{t_c}{|c|} \quad (27)$$

où t_c est la fréquence du terme t dans le passage c et $|c|$ le nombre de termes du passage c. Donc dans le cas, peu réaliste il est vrai, où tous les passages de la classe ne contiennent que le terme t, cet indicateur atteint la valeur maximale 1, et 0 si il n'apparaît nulle part dans la classe. Enfin suivant le même argument que précédemment, un terme est discriminant au sens d'une partition si il minimise la densité interclasse.

L'ensemble de ces indicateurs permettent de mettre en oeuvre le type d'analyse décrit dans la section 2. Ils seront utilisés dans la section 5.

[75]Le facteur est au carré car les vecteurs étant normalisés la somme de leurs carrés vaut 1

[76]Le terme 'terme' est décidément un terme rare du point de vue de Zipf

4.6 Classification des documents en fonction des isotopies

Nous arrivons à ce que nous considérons être l'aspect essentiel de notre thèse : l'indexation par isotopie vue comme une extension de la sémantique latente fondée sur la théorie linguistique de la sémantique interprétative. La sémantique latente est une amélioration empirique du modèle de Salton basée sur le constat que les cooccurrences de mots permettent rendre compte de similarités entre documents, même lorsque ces derniers ne partagent pas de mots significatifs, l'utilisation des facteurs principaux les plus importants du nuage de documents à indexer permettant d'atténuer le bruit. Un concept fondamental de la sémantique interprétative est la notion d'isotopie, l'isotopie est le résultat de l'interprétation d'un texte. Elle se traduit par la récurrence d'unités minimales de sens appelés sèmes. Elle correspond à une attente du lecteur, qu'on appelle généralement présomption d'isotopie, elle est une manifestation du cercle de la compréhension, le concept central pour rationaliser le processus d'interprétation d'un texte. Les isotopies ne peuvent donc pas, a priori, être décrites dans des dictionnaires ou autres ontologies, seules les rencontres entre des lecteurs et des textes décrivent des isotopies. L'isotopie peut donc sembler être d'un intérêt tout relatif pour le traitement automatique des textes et c'est en partie vrai. A y regarder de plus près cependant, on peut penser que ce qui est attendu par le lecteur, ce qui structure le texte donc, se répète souvent et que d'une certaine façon ceci doit se refléter dans la fréquence des mots. Pour s'en convaincre nous nous proposons d'effectuer l'expérience de la section 4.2 à propos de la loi de Zipf sur le roman de Pierre Loti « Aziyadé » [Loti 1879]. Ce roman disponible gratuitement sur l'« Association de Bibliophiles Universels »[77](ABU) compte un vocabulaire de 13427 mots licence ABU comprise, seulement 3789 d'entre eux apparaissent plus de 2 fois. Mais en regardant par l'autre bout de la lorgnette on s'aperçoit que ces 3789 mots couvrent 80%[78] du texte, en d'autres mots 80% de « Aziyadé » a été écrit avec moins de 29% de son vocabulaire. Voici deux extraits de cette version ABU de « Aziyadé » ne comportant que les mots n'apparaissant que deux fois, le caractère '@' représentant des hapax :

[77] http://abu.cnam.fr/

[78]Dans cette expérience, un mot est une chaîne de caractères ne contenant pas de caractères blancs ou de caractères de fin de ligne et entourés de caractères blancs ou de de caractères de fin de ligne : ainsi « l'horrible » est considéré comme un mot, c'est un ici un hapax alors qu'avec un traitement un peu plus propre *horrible* n'aurait pas été un hapax, il en est de même pour « voisin, ». Cette excès de grossièreté du traitement informatique ne fait que souligner le propos.

> « ... Une belle journée de @ un beau soleil, un ciel pur... Quand les canots étrangers arrivèrent, les @ sur les quais, mettaient la dernière main à leur oeuvre : six pendus exécutaient en présence de la foule @ @ @ Les fenêtres, les toits étaient @ de spectateurs ; sur un balcon @ les autorités turques @ à ce spectacle @ »
>
> « Je traversais @ au soir Stamboul à cheval, pour aller chez @ la grande fête du @ grande féerie orientale, dernier @ @ : toutes les mosquées @ ; les minarets @ leur extrême pointe ; des versets du Koran en lettres @ dans l'air ; des milliers d'hommes criant à la fois, au bruit @ le nom @ d'Allah ; une foule en habits de fête, @ dans les rues des @ de feux et de lanternes ; des femmes @ par @ vêtues de soie, d'argent et d'or. »

Voici les mêmes extraits dans leur totalité cette fois :

> « ... Une belle journée de mai, un beau soleil, un ciel pur... Quand les canots étrangers arrivèrent, les bourreaux, sur les quais, mettaient la dernière main à leur oeuvre : six pendus exécutaient en présence de la foule l'horrible contorsion finale... Les fenêtres, les toits étaient encombrés de spectateurs ; sur un balcon voisin, les autorités turques souriaient à ce spectacle familier. »
>
> « Je traversais hier au soir Stamboul à cheval, pour aller chez Izeddin-Ali. C'était la grande fête du Baïram, grande féerie orientale, dernier tableau du Ramazan : toutes les mosquées illuminées ; les minarets étincelants jusqu'à leur extrême pointe ; des versets du Koran en lettres lumineuses suspendus dans l'air ; des milliers d'hommes criant à la fois, au bruit du canon, le nom vénéré d'Allah ; une foule en habits de fête, promenant dans les rues des profusions de feux et de lanternes ; des femmes voilées circulant par troupes, vêtues de soie, d'argent et d'or. »

Il est frappant de remarquer que le sens général du texte est tout à fait compréhensible, sans les hapax. On rate certainement le sens de « les autorités turques souriaient à ce spectacle familier », mais le propos général reste compréhensible plus particulièrement dans le second extrait plus près d'un thème récurrent chez Pierre Loti « la féerie orientale » justement. Ces paragraphes de l'oeuvre de Pierre Loti correspondent tout à fait à ce que l'on entend par passage. Nous sommes conscient des limites de cet argument, notre compréhension est aussi guidée par l'ordre des mots et certains mots grammaticaux sont fréquents ; toutes choses dont on ne tient pas compte dans notre modèle. Par ailleurs le filtrage des cooccurrences va encore réduire la taille du vocabulaire effectif mais il reste clair que seule une très petite partie du vocabulaire est nécessaire pour comprendre un texte dans son ensemble et quiconque a fait l'apprentissage d'une langue étrangère en sera conscient : c'est d'ailleurs le phénomène d'isotopie qui nous permet de combler les trous en nous appuyant sur les mots d'un usage plus courant dont très souvent à notre insu la simple cooccurrence déclenche notre intérêt lorsque nous sommes en écoute flottante ou lorsque nous décrochons au cours d'une lecture barbante[79].

[79]Merci de votre attention.

Ce n'est donc pas sans fondement empirique que nous avançons un argument liant l'isotopie aux mots fréquents, malgré son aspect éminemment interprétatif. Qu'en est-il maintenant de la liaison fondamentale pour notre propos entre cooccurrences et molécules sémiques ? Si nous reprenons les deux paragraphes de *Aziyadé* déjà cité, les cooccurrences : « belle journée », « beau soleil », « ciel pur » et « oeuvre », « spectateur », « spectacle » dans le premier paragraphe font sens sur les thèmes de l'euphorie et du spectacle. Comme « fête », « féerie », « étincelants », « feux », « lanternes » d'une part et « Koran », « Allah », « mosquées », « minarets » dans le deuxième extrait rappellent irrésistiblement les thèmes de la fête et de l'islam. Notre pari est que, à des classes de passages regroupés sur les facteurs de telles cooccurrences correspondent des molécules sémiques thématiques. Le problème réside dans la description de ces classes : est-ce qu'à la récurrence de pondérations des facteurs du graphe de cooccurrences décrivant les classes de passage correspondent des isotopies ou plus précisément des pré-isotopies ? Un élément de réponse nous semble résider dans leur capacité à regrouper les documents en classes cohérentes en fonction de classes thématiques[80]. Pour ce faire nous allons nous inspirer des caractérisations d'isotopie proposées par Ludovic Tanguy dans [Tanguy 1997]. Il y propose de caractériser une isotopie en fonction de son poids et de son volume :

- « Poids : le nombre d'épisémèmes supportant l'isotopie »
- « Volume : nombre total d'épisémèmes compris entre le premier et le dernier épisémème de l'isotopie »

Ludovic Tanguy introduit, toujours dans [Tanguy 1997], la notion d'épisémème pour tenir compte de la position des sémèmes dans le texte, bien que nous ne soyons pas directement concernés par cette considération nous rapportons sa définition pour éclairer les définitions de poids et de volume d'une isotopie selon lui :

> « L'ensemble des positions des termes interprétés d'un texte forme l'ensemble *E* des *épisémèmes*. L'identité inhérente à *E* se base *strictement* sur la position. L'identité des chaînes de caractères n'est donc plus qu'un phénomène périphérique et non suffisant. L'ensemble des sémèmes *S* devient donc un ensemble extensionnel basé sur *E* : un sémème sera un ensemble d'épisémèmes. »

Notre unité minimale d'analyse n'étant pas le mot, l'épisémème sur le plan sémantique, mais le passage, i.e la molécule sémique sur le plan sémantique, nous proposons de caractériser l'isotopie de façon différente tout en respectant les idées initiales de couverture et d'intensité sous-jacentes.

[80]De classes de passages.

4 Isotopie et statistiques contextuelles

Dans un sémème ou une molécule sémique, l'unité de sens est le sème. Dans un passage projeté sur les facteurs du graphe de cooccurrences ce sont les projections elles-mêmes. Comme nous l'avons vu la difficulté réside dans l'identification du lien entre sèmes et projections sur des facteurs. Autrement dit dans l'interprétation des projections sur les facteurs. Par exemple, dans *Aziyadé* le premier facteur est Loti[81] et les mots qui cooccurrent le plus significativement avec lui sont, sans filtrage préalable :

- Allah ,
- Aziyadé,
- Achmet,
- fini,
- disait-elle.

Voici quatre passages qui expliquent la nature apparemment obscure de ces collocations

> « Il était en effet très petit, le plus petit doigt d'**Aziyadé**. Son ongle, très rose à la base, dans la partie qui venait de pousser, était à sa partie supérieure teint tout comme les autres d'une couche de henné, d'un beau rouge orange. Eh bien, dit-elle, de même, et à plus forte raison, **Loti**, les créatures d'**Allah**, qui sont beaucoup plus nombreuses, ne sont pas toutes semblables ; toutes les femmes ne sont pas les mêmes, ni tous les hommes non plus ... »

> « Alors j'entendis sa voix. Pour la première fois, elle parlait et je comprenais, ravissement encore inconnu! Et je ne trouvais plus un seul mot de cette langue turque que j'avais apprise pour elle ; je lui répondais dans la vieille langue anglaise des choses incohérentes que je n'entendais même plus! Severim seni, Lotim ! (Je t'aime, **Loti**, **disait-elle**, je t'aime!) »

> « La place du Sultan Sélim est entourée d'une antique muraille, dans laquelle s'ouvrent de loin en loin des portes ogivales. Les promeneurs y sont rares, et quelques tombes s'y abritent sous des cyprès; on est là en bon quartier turc, et on peut aisément s'y tromper de deux siècles. Moi, disait **Achmet** d'un air frondeur, je sais bien ce que je ferai, **Loti**, quand tu seras parti : je mènerai joyeuse vie et je me griserai tous les jours; un joueur d'orgue me suivra, et me fera de la musique du matin jusqu'au soir. Je mangerai mon argent, mais cela m'est égal (zarar yok). Je suis comme **Aziyadé**, quand tu seras parti, ce sera **fini** aussi de ton **Achmet**. »

> « **Aziyadé** arriva le soir, me racontant combien elle avait été inquiète, et combien de fois elle avait dit pour moi : **Allah**! Sélamet versen **Loti** ! (**Allah**! protège **Loti**!) »

En fait le mot 'Loti' est un facteur, non pas à cause d'une forme d'hypertrophie de l'ego de l'auteur mais parce qu'au contraire il met en relief les autres personnages principaux du roman Aziyadé et Achmet au travers de ce qu'ils disent du héros Loti, officier de sa

[81]Le narrateur protagoniste.

Majesté. C'est à nos yeux une illustration exemplaire de l'opposition entre le sens et la signification souvent souligné par François Rastier.

> Quoi qu'il en soit, la tradition occidentale retient deux façons principales de définir le contenu linguistique :
>
> 1. La *signification* est conçue comme relation entre les plans du signe (signifiant, signifié) ou les corrélats du signe (concept, référent). Même orientée, cette relation reste statique, typée, susceptible d'une expression logique. Dans la sémiotique de tradition logico-grammaticale sur laquelle on s'appuie alors, l'interprétation se définit comme l'identification d'une relation de représentation, simple ou complexe.
>
> 2. Le *sens* est défini comme parcours entre les deux plans du texte (contenu et expression), et au sein de chaque plan. Un parcours est un processus dynamique, obéissant à des paramètres variables selon les situations particulières et les pratiques codifiées. Si bien que le sens n'est pas donné, mais résulte du parcours interprétatif normé par une pratique. [Rastier 1999]

Or ce qui nous intéresse est bien le sens, et, en l'occurrence, le facteur 'Loti' rend compte du sens que nous venons de mettre en évidence : « les amis de Loti parlent de Loti » que l'on fige ainsi, bien sûr, mais qui ne peut être compris au départ que dans un parcours interprétatif. Donc les facteurs ne sont pas des sèmes très orthodoxes, mais il n'est pas honteux de leur reconnaître une certaine dignité sémantique[82].

Ceci étant, d'ailleurs, ne serait-il pas envisageable, dans un corpus, de définir le sème à partir des passages dans lesquels ils apparaissent? C'est d'une certaine façon la position implicite du modèle DSIR lui aussi basé sur la sémantique latente quand il est dit dans [Besançon et al. 2000]:

> « Deux unités linguistiques sont sémantiquement similaires si leurs contextes textuels sont similaires. »

Ce qui est en jeu n'est pas tant le statut du sème que sa représentation machine. Une analogie est peut être intéressante arrivés à ce point de la discussion : la théorie du traitement du signal et son implémentation informatique. En étudiant le phénomène de propagation de la chaleur Jean-Baptiste Fourier fit l'hypothèse[83] que toute fonction périodique se décompose en une somme[84] de fonction sinusoïdales (de sinus et de cosinus)[85]. La théorie de Fourier est particulièrement intéressante car elle permet de

[82]Ce que font d'ailleurs allègrement les tenants de la sémantique latente sans même s'embarrasser de considérations de signifiance sur de simples considérations empiriques.

[83]Hypothèse vérifiée depuis

[84]Une série pour être plus précis car la dite somme peut être infinie.

[85]Ce qui lui valut les foudres du grand Lagrange qui enterra son mémoire « *sur la propagation de la Chaleur dans les corps solides* ». Il reçut finalement le prix de l'institut en 1812

résoudre dans le domaine des fréquences des équations qui seraient impossibles de résoudre dans le domaine du temps, elle est à la base du traitement du signal. Elle permet, par exemple, d'expliquer comment un son se décompose en fréquences harmoniques. Il existe des programmes informatiques de traitement du signal, le principal problème de ces programmes est la transformation d'un signal analogique en signal numérique. L'échantillonnage permet de transformer un signal continu en signal discret, en une suite de valeurs. Les programmes de traitement numérique du signal utilisent des signaux discrets et dépendent donc de la qualité de l'échantillonnage. De façon analogue la qualité des programmes traitant le sens est tributaire de la qualité de la représentation machine du sens, de son échantillonnage en quelque sorte. Mais l'analogie ne s'arrête pas là, de la même façon qu'un son se décompose en harmoniques, le sens d'un texte se décompose en isotopies[86]. L'échantillonnage que nous utilisons, le paragraphe, est plutôt grossier ; il nous permet cependant de mettre en évidence certaines harmoniques : les facteurs du graphes de cooccurrences. C'est la récurrence de ces facteurs qui traduit pour nous le phénomène d'isotopie et comme les intensités de ses harmoniques caractérisent un son, les intensités de ses isotopies caractérisent le sens d'un texte, en assurent la cohésion. Si le $i^{\text{ème}}$ facteur du passage c est noté $f_{c,i}$ l'intensité d'un texte T s'exprime par :

$$\frac{\sum_{c \in T} {f_{c,i}}^2}{|T|} \quad (28)$$

où $|T|$ est le nombre de passages du texte T. On voit que cette formule est très proche de la définition de la densité de Ludovic Tanguy tout en correspondant à l'intensité d'un facteur donnée à l'équation (25).

Une texte peut dès lors être représenté par l'intensité de ses isotopies et donc comme une vecteur dans l'espace des facteurs ainsi, de la même façon que l'on peut classifier les passages représentés comme des points sur l'hypersphère unité de l'espace des facteurs en utilisant la fonction cosinus on peut classifier les documents représentés comme des points sur l'hypersphère unité de l'espace des facteurs et les mêmes critères d'analyse comme la spécificité, l'intensité et la valeur discriminante valent pour les documents indexés par des isotopies.

Avant de dire quelques mots sur l'interrogation et l'apprentissage, faisons une dernière fois le point sur les motivations de notre approche. Notre propos est multiple ;

[86]Ce qui ne signifie pas qu'il s'y réduise.

1. faire le lien entre une approche empirique la sémantique latente utilisée avec succès en IR[87] et une théorie linguistique la sémantique interprétative,
2. proposer un filtrage de la matrice de cooccurrences de la sémantique latente basée sur un calcul précis et rapide du test exact de Fisher,
3. proposer une méthode simple et efficace de factorisation du graphe de cooccurrence basée sur une approximation d'une analyse en composantes principales,
4. proposer une méthode générale de classification des passages en fonctions des facteurs ou des documents en fonction des isotopies pour interpréter les résultats obtenus et suggérer des modifications de la sélection du vocabulaire initial, des facteurs ou des passages.

4.7 Interrogation et classificateur

Nous ne pouvons quitter cette section sans ajouter quelques mots sur l'interrogation d'une base de donnée documentaire bâtie sur ce modèle. L'indexation peut être complètement automatisée, mais encore une fois ce n'est absolument pas notre but, notre but est de pouvoir enregistrer une interprétation de façon à la partager. Cette interprétation peut se réduire au choix d'un corpus d'apprentissage, un ensemble de textes dont on sait qu'ils ont une cohérence sémantique. On peut améliorer les résultats obtenus en filtrant le vocabulaire sur la base de connaissances a priori sur le sujet du corpus indexé. Un réglage fin du nombre de facteurs ou du seuil du test de Fisher permet de moduler le nécessaire compromis entre la finesse et la robustesse de l'indexation. Des classifications des passages et des documents permet d'offrir des thématiques prédéfinies destinées à aider des utilisateurs dans leur prise de connaissance du contenu ou à échanger des points de vue dans un système collaboratif. On peut au travers de ces opérations littéralement sculpter une indexation mais au bout du compte comment interroge-t-on le fonds documentaire ainsi indexé et comment ajoute-t-on de nouveaux documents ? Il y deux types d'interrogation :

1. la navigation dans des classifications existantes ou dans des classifications faites à la demande,[88]

[87]Pour *Information Retrieval* qui est si abondamment utilisé que nous n'avons pas cherché de traduction.

[88]La rapidité de la procédure de classification le permet aisément.

2. l'interrogation par mots clefs.

Nous avons suffisamment parlé de la classification pour n'y pas revenir, l'interrogation par mots clefs quant à elle, se décompose en trois cas pour chaque mot clef ;

1. le mot se projette sur les facteurs existants,
2. le mot existe dans le fond mais ne se projette pas sur les facteurs,
3. le mot n'existe pas dans le fonds.

Bien sûr, si des mots se projettent sur les facteurs on peut les considérer comme formant un passage et utiliser le cosinus pour chercher des passages proches et, par extension, les documents ayant des isotopies proches ; qui plus est, s'il existe une classification préexistante on peut utiliser la formule (22) pour trouver les classes les plus proches.

Si aucun des mots ne se projette sur les facteurs mais si certains apparaissent dans le fonds, on peut utiliser les documents les contenant comme réponses ou comme requêtes pour élargir la recherche, comme précédemment. Si une classification existe on a tout avantage à utiliser les centroïdes pour suggérer de nouveaux axes de recherche.

Si aucun des mots n'existe dans le fonds, il faut revenir a une classification existante ou en créer une nouvelle.

L'ajout d'un document au fonds suit une procédure analogue à l'interrogation par mots clefs pour savoir à quel point ses passages se projettent sur les facteurs. Si peu de passages se projettent, on obtiendra des isotopies de faible intensité indiquant que le document est étranger au fonds ou qu'une nouvelle indexation est nécessaire. D'une façon générale, on peut confronter un même document à plusieurs indexations qui se comporteront comme des classificateurs automatiques indiquant la plus ou moins grande adéquation du nouveau document aux thématiques correspondantes.

C'est certainement un aspect novateur de notre approche qui se rapproche plus de Yahoo que de Google dans la mesure où nous privilégions l'interprétation sur le traitement automatique. Nous sommes persuadés qu'à terme de telles indexations interprétatives basées sur des réseaux collaboratifs comme *del.icio.us*[89] permettrons de meilleures performances (qualitatives). Le problème étant d'offrir des interfaces de navigation attrayantes permettant de capturer des navigations qui pourraient être utilisées comme des

[89]http://del.icio.us/

systèmes de filtrage collaboratif passifs[90]. En effet il est assez simple de proposer à des utilisateurs les facteurs principaux des documents rendus par une requête standard afin de suggérer en même temps la présentation des résultats de nouveaux axes de recherche et enregistrer les appariements facteurs requêtes pour créer de nouvelles thématiques[91]. D'une manière générale, nous pensons que notre approche serait d'un très grand intérêt pour le WebMining par l'association de données sémantiques aux analyses de logs de serveurs.

[90]Dont le fameux « les personnes qui ont acheté ce livre ont aussi acheté celui-ci et celui-là » d'Amazon est l'exemple typique.

[91]Ce qui est une forme de *relevance feedback* pas trop contraignante

5 Application

Nous avons choisi comme corpus : le roman de Pierre Loti *Aziyadé*. C'est à dessein que nous avons choisi un roman assez court comme corpus de démonstration : il est en effet facile de se faire une idée de son contenu, il suffit de le lire, et sa taille reste significative, tout en permettant des traitements rapides permettant de tester différentes hypothèse en un laps de temps assez court ; un cycle complet de traitement n'excédant pas une heure.

5.1 Aziyadé

Aziyadé est un roman de jeunesse, c'est en fait le premier roman de Louis Marie Julien Viaud qui, étant tenu à un devoir de réserve du fait de sa condition d'officier choisit comme pseudonyme le surnom que lui donnaient les suivantes de la reine Pomaré lors de son séjour à Tahiti[92]. Comme on sait, Loti restera sa vie durant un grand amoureux de la Turquie[93]: il trouvera plus tard dans *Fantôme d'Orient* ce premier roman malhabile on y trouve en tous cas campé un type de héros qui n'est pas sans rappeler Lord Jim ou, plus récemment et dans un tout autre genre, Corto Maltese, bref un type de héros qui aura fait rêver beaucoup de jeunes gens. Le roman est une transcription très proche de son journal, des passages entiers ont été retranscrits tels quels. On découvre, au-delà de l'aventure sans issue avec une esclave de harem[94], le véritable grand amour de Loti, la Turquie. Bien sûr, le roman a vieilli et l'on pourra plus sûrement être aujourd'hui choqué par certaines considérations sur les grecs immondes et sur une négresse qui ressemble à un macaque, que par les pratiques homosexuelles des turques, à peine évoquées, mais qui firent grand bruit en son temps. La version électronique que nous avons utilisée est celle que l'on peut trouver sur le site du projet Gutenberg[95].

[92] Roti veut dire Rose, mais le *r* roulé est devenu un *l* dans l'oreille du bel aspirant.

[93] Et un grand amoureux tout court

[94]Il semble qu'Aziyadé ait été plus vraisemblablement une esclave qu'une épouse de harem du vieil Abeddin Cf. la préface de Claude Martin dans [Loti 1991] où malgré les suppositions d'Edmond de Goncourt et d'André Gide il semble clair que la véritable Aziyadé ait bien été une jeune Circassienne et non un jeune Circassien.

[95]la version de l'ABU dont nous avons déjà parlé étant décidément de trop mauvaise qualité, de nombreux mots sont accolés certainement du fait d'une mauvaise gestion des caractères de fin de ligne

5.2 Prétraitements

```
<html>
<head><
title>
        1 SALONIQUE JOURNAL DE LOTI </title></head>
<body>
        <h2>
        1 SALONIQUE JOURNAL DE LOTI Chapitre I
        </h2>
        <p>  16 mai 1876. </p>
        <p>  ... Une belle journée de mai, un beau soleil, un ciel pur ...
Quand les  canots étrangers arrivèrent, les bourreaux, sur les quais,
mettaient la dernière main à leur oeuvre : six pendus exécutaient en
présence de la foule l'horrible contorsion finale ... Les fenêtres, les toits
étaient encombrés de spectateurs ; sur un balcon voisin, les autorités
turques souriaient à ce spectacle familier. </p>
        <p>  Le gouvernement du sultan avait fait peu de frais pour
l'appareil du supplice ; les potences étaient si basses que les pieds nus
des condamnés touchaient la terre. Leurs ongles crispés grinçaient sur le
sable. </p>
</body>
</html>
```

Figure 15 Balisage HTML du chapitre 1 d'Aziyadé

Le texte du projet Gutenberg se présente sans balises d'aucune sorte, nous l'avons donc, à l'aide d'un script Perl, transformé en 154 documents HTML : soit autant de documents qu'il y a de chapitres. Nous avons balisé le texte de chaque document de façon à repérer les titres et paragraphes du texte, ainsi qu'on peut le voir à la figure 15 et dont le rendu dans un navigateur est visible à la figure 16. Aucun autre prétraitement n'a été fait.

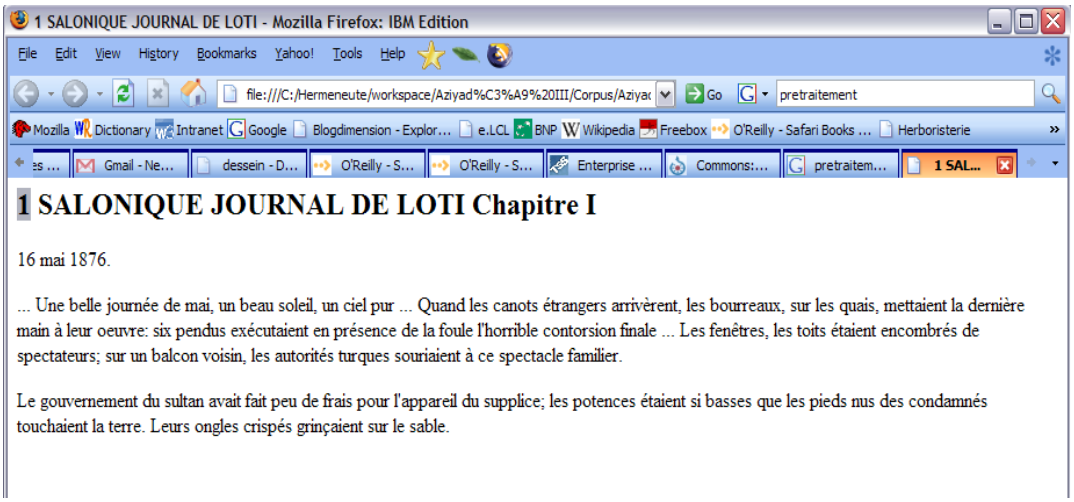

Figure 16 : Chapitre 1 d'Aziyadé vu dans un navigateur

5.3 Analyse

L'analyse se décompose en quatre phases :

1. La collecte du vocabulaire du corpus et la sélection des mots signifiants,
2. le calcul des facteurs,
3. la classification des passages et leur analyse,
4. la classification des documents en fonction des classes de passages.

Ces quatre phases ne se déroulent pas selon un ordre figé, la sélection des mots signifiants, par exemple, peut être revue selon les résultats d'une classification des passages. De même, certains facteurs pourront être inhibés lors des phases de classification. N'oublions pas que nous nous inscrivons dans un processus interprétatif. Si la méthode décrite au chapitre 2 et les programmes utilisés au chapitre 4, encadre l'interprétation l'interprète n'en conserve pas moins une grande marge de liberté quand aux choix du vocabulaire, des facteurs et de la validation des classes. Le *jeu* auquel se prête l'analyste, l'interprète consiste donc à jouer de cet espace de liberté pour aboutir à un résultat satisfaisant, une interprétation acceptable.

Nous présentons ici le résultat d'une analyse, pour gagner du temps nous avons procédé à certaines simplifications que l'on pourra considérer comme abusives, elle n'entament cependant pas à nos yeux la démarche générale et l'analyse reste représentative de ce que

nous voulons démontrer. Nous avons éliminé, manuellement, certains mots grammaticaux fréquents, certains adverbes, auxiliaires et verbes support. Nous n'avons pas non plus tenu compte des hapax et 4053 mots ont été finalement retenus sur 8337. La définition initiale d'un passage fut de retenir les mots du vocabulaire sélectionné compris entre deux balises d'un même type (ici <h2> et </h2> ou et). Les cooccurrences retenues sont celles passant le test de Fisher exact pour un niveau de confiance de 99% (valeur 0.01 du paramètre α). Les mots retenus sont projetés sur les 500 premiers facteurs qui épuisent 66% des cooccurrences précédemment filtrées. Au cours de notre étude certains mots ont été éliminés. On trouvera en annexe 1 la liste des 50 premiers facteurs accompagnés de leurs cooccurrents ayant une contribution supérieure à 0.1, les valeurs de contribution et de discrimination étant celles décrites au chapitre 4. Comme nous l'avions déjà indiqué 'Loti' est le premier facteur car il regroupe les personnages principaux du roman. 'Samuel' n'apparaît pas directement dans la liste mais plus loin comme un des 50 premiers facteurs, 'Aziyadé' et 'Achmet' partagent bien évidemment cet honneur. Ces cinquante premiers facteurs laissent déjà transparaître la thématique du roman :

- Les protagonistes donc, Aziyadé, Achmet, Samuel et Loti.
- Des personnages secondaires regroupés autour du facteur 'madame' : la chatte qui terrorise Samuel 'bir madame kédi' (une madame chat) et par extension les chats 'kédis' ont une certaine importance, c'est en tout cas un thème de fond récurrent (la visite à Ankara, Angora capitale des chats par exemple). Un autre personnage secondaire participe à la création de ce facteur : la tenancière du café où se rendent Samuel et Achmet « leur madame était une vieille coquine », etc.
- Le thème de l'islam : 'Allah', 'minaret', 'Eyoub'. Mais aussi celui de la Turquie et de l'Orient : 'Sultan', 'marbre', 'café', 'Stamboul'
- L'amour : 'aime', 'bague' (le cadeau d'Aziyadé), 'jaloux', 'Eyoub'...
- La perdition : 'abîme', 'fond'
- Le facteur 'cyprès' est intéressant car il évoque aussi bien les cimetières et les mosquées que les paysages proprement dits.
- La crise politique de l'empire ottoman est la toile de fond du roman : les facteurs 'patrie', 'sultan', 'histoire' rappellent la constitution turque de 1876, la guerre

russo-turque de 1877, l'arrivée au pouvoir d'Abd-Ul-Hamid II, les manifestations des softas[96].

Mais plus que des thèmes épars ce sont les regroupements qu'ils opèrent qui nous intéressent. Nous avons donc classifié les passages en fonction des facteurs. En général, pour ce type de classification, nous utilisons seuil de densité de 0,01 (le seuil θ de l'algorithme de classification).

La première matrice de densité que nous avons obtenue est présentée à la figure 17 : on remarque immédiatement que la classe 0 (la première), est dense et isolée c'est-à-dire sans relation avec les autres classes, tandis que les autres classes sont peu denses et centrales. À y regarder de plus près cette classe regroupait sous le titre 'Chapitre'[97], les passages regroupés dans cette classe étaient en fait les titres de chaque chapitre qui, dans ce corpus sont de la forme 'SALONIQUE JOURNAL DE LOTI Chapitre I' ou bien encore 'EYOUB À DEUX Chapitre LV', considérant que ces passages étaient de peu d'intérêt, nous avons décidé de ne plus considérer les passages entre les balises <h2> et </h2>. Par ailleurs dans cette première expérimentation nous avions calculé 1000 facteurs ce qui explique le grand nombre de classes.

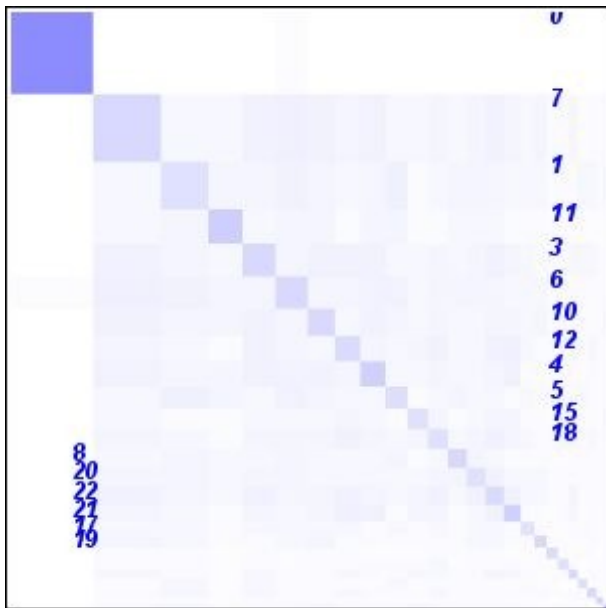

Figure 17 : Première matrice de densité

[96]Un softa est un étudiant d'une école musulmane attachée à une mosquée, une madrassah.

[97]Le titre d'une classe, affecté automatiquement est en fait le facteur de plus forte contibution de la classe.

Après quelques réglages, comme par exemple le refus d'accepter le terme 'eût' comme facteur[98], nous sommes finalement arrivé à la classification que nous allons maintenant présenter. La matrice de densité de la figure 18 exhibe 10 classes que nous allons analyser une par une. Les cinq premières classes sont très centrales et liées entre elles suggérant des regroupements liés et diffus dans le texte. Les cinq dernières sont au contraire beaucoup plus isolées en particulier la plus petite qui, comme on le verra, regroupe des passages figés. Afin de faciliter la présentation de cette analyse nous avons utilisé l'application Eclipse BIRT[99] pour générer automatiquement un rapport de classification. C'est la structure de ce rapport qui est utilisé dans les pages suivantes : pour chaque classe nous avons donc les dix passages et les dix facteurs de meilleure contribution.

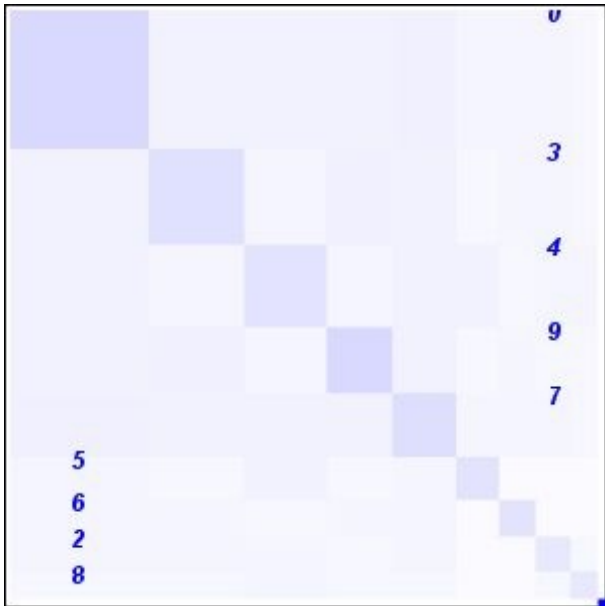

Figure 18 : Matrice de densité finale

[98]Ce sont toujours des considérations de temps qui ont déterminé notre décision ici, a priori, ce facteur semblait ne se constituer qu'autour de phénomènes syntaxiques et il nous a semblé plus simple de l'écarter que de tenter d'affiner nos traitements.

[99]Business Intelligence Reporting Tool

5.3.1 **Classe 'Loti', 274 passages, numéro 0.**

Cette classe se constitue autour des personnages principaux du roman, principalement autour des passages où Aziyadé s'adresse à Loti, comme le montrent, les extraits ci dessous. Le thème, principal du roman apparaît clairement ici. L'aventure sans issue entre Loti et Aziyadé, la Turquie (au travers des nombreuses expressions turques), Achmet l'ami turc[100].

> ...eut tout épuisé : Benim djan senin, Loti. (Mon âme est à toi, Loti.) Tu es mon Dieu, mon frère, mon ami, mon amant ; quand tu seras parti, ce sera fini d'Aziyadé ; ses yeux seront fermés, Aziyadé sera morte. Maintenant, fais ce que tu voudras...
>
> ...Aziyadé fut assise dans notre barque...
>
> ...Aziyadé arriva le soir, me racontant combien elle avait été inquiète, et combien de fois elle avait dit pour moi : Allah ! Sélamet versen Loti ! (Allah ! protège Loti...
>
> ...Aziyadé, au contraire, passa...
>
> ...Aziyadé, brisée de fatigue, s'endormit au son de sa voix lugubre, en pleurant à chaudes larmes...
>
> ...marié, Loti, disait-elle, cela ne fait rien. Je ne serai plus ta maîtresse, je serai ta soeur. Marie-toi, Loti ; c'est secondaire, cela ! J'aime mieux ton âme. Te revoir seulement, c'est tout ce que je demande à Allah. Après cela, je serai presque heureuse encore, je vivrai pour t'attendre, tout ne sera pas fini pour Aziyadé...
>
> ...Aziyadé avait pleuré plusieurs heures ; mais ses larmes étaient moins amères. L'idée de me revoir commençait à prendre consistance dans son esprit et la rendait plus calme. Elle commençait à dire : « Quand tu seras de retour ... » Je ne sais pas, Loti, disait-elle, si tu reviendras, Allah seul le sait ! Tous les jours je répéterai : Allah ! sélamet versen Loti ! (Allah ! protège Loti !) et Allah ensuite fera selon sa volonté. Pourtant, reprenait-elle avec sérieux, comment pourrais-je t'attendre un an, Loti ? Comment cela se pourrait-il, quand je ne sais plus rester un jour, non pas même une heure, sans te voir. Tu ne sais pas, toi, que les jours où tu es de garde, je vais me promener en haut du Taxim, ou m'installer en visite chez ma mère Béhidjé, parce que de là on aperçoit de loin le Deerhound. Tu vois bien, Loti, que c'est impossible, et que, si tu reviens, Aziyadé sera morte...
>
> ...air mystique, Achmet et Aziyadé m'apprennent que ces chiens hurlent ainsi pour demander à Allah un certain pain mystérieux qui leur est dispensé dans certaines circonstances solennelles, et que les hommes ne peuvent voir...
>
> ...Achmet était très important et très solennel : nous accomplissions tous deux une expédition pleine de mystère, et lui était nanti des instructions d'Aziyadé, tandis que moi, j'avais juré de me laisser mener et d'obéir...
>
> ...Aziyadé m'exprime quelque idée neuve, quelque notion nette et profonde sur des choses qu'elle semblerait devoir ignorer absolument, et que je lui demande : « Qui t'a appris cela, ma chérie ? » Aziyadé répond : « C'est ma mère Béhidjé...

[100] Achmet était arménien en fait.

Les facteurs confirment cette première impression : Loti, Achmet, Aziyadé et Samuel arrivent dans les cinq premiers facteurs en terme de contribution. On remarquera en particulier la spécificité extrêmement élevée de Loti et Achmet indiquant que ces facteurs contribuent à peu près uniquement à cette classe. Le facteur Allah y tient une grande importance car les dialogues d'Achmet et d'Aziyadé sont émaillés de références à Dieu.

Facteur	Intensité	Spécificité	Contribution
Loti	0.215	0.811	58.95
Achmet	0.19	0.815	52.09
Allah	0.137	0.727	37.469
Aziyadé	0.073	0.619	20.006
Samuel	0.049	0.398	13.469
Deerhound	0.047	0.591	12.808
commençait	0.046	0.728	12.497
heure	0.04	0.478	11.003
guerre	0.037	0.386	10.09
nuit	0.032	0.244	8.724

5.3.2 **Classe 'minarets', 189 passages, numéro 3.**

Cette seconde classe par la taille, est elle aussi porteuse de thèmes importants du roman, centrés autour de descriptions de paysages. Bien évidemment, les minarets et les mosquées y tiennent une grande importance mais aussi la mer et les cyprès. Le thème de la mer s'explique aisément, Loti est un officier la flotte de sa gracieuse majesté[101] et ses premières rencontres avec Aziyadé se font à bord de caïques dans la rade de Salonique, le thème du caïque est d'ailleurs omniprésent dans le roman, même si le facteur correspondant n'arrive qu'en cent soixante troisième position et ne se trouve de ce fait pas reporté en annexe. Le thème du cyprès est intéressant, il est souvent associé aux mosquées et aux cimetières, associations que l'on retrouve d'ailleurs dans *Fantôme d'Orient* [Loti 1991].

> ...oublierai jamais l'aspect qu'avait pris, cette nuit-là, la grande place du Séraskiérat, esplanade immense sur la hauteur centrale de Stamboul, d'où, par-dessus les jardins du sérail, le regard s'étend dans le lointain jusqu'aux montagnes d'Asie. Les portiques arabes, la haute tour aux formes bizarres étaient illuminés comme aux soirs de grandes fêtes. Le déluge de la journée avait fait de ce lieu un vrai lac où se reflétaient toutes ces lignes de feux ; autour du vaste horizon surgissaient dans le ciel les dômes des mosquées et les minarets aigus, longues tiges surmontées d'aériennes couronnes de lumières...
>
> ...ciel pur et la mer bleue du Levant. Là-bas, quelque chose se dessine ; l'horizon se frange de mosquées et de minarets ; mon coeur bat, c'est Stamboul...
>
> ...maison était située en un point retiré de Péra, dominant de haut la Corne d'or et le panorama lointain de la ville turque ; la splendeur de l'été donnait du charme à cette habitation. En travaillant la langue de l'islam devant ma grande fenêtre ouverte, je planais sur le vieux Stamboul baigné de soleil. Tout au fond, dans un bois de cyprès, apparaissait Eyoub, où il eût été doux d'aller avec elle cacher son existence, point mystérieux et ignoré où notre vie eût trouvé un cadre étrange et charmant...
>
> ...À mes pieds, les vieilles cases arméniennes sont obscures et endormies ; j'ai vue sur un très profond ravin, au bas duquel un bois de cyprès séculaires forme une masse absolument noire ; ces arbres tristes ombragent d'antiques sépultures de musulmans ; ils exhalent dans la nuit des parfums balsamiques. L'immense horizon est tranquille et pur ; je domine de haut tout ce pays. Au-dessus des cyprès, une nappe brillante, c'est la Corne d'or ; au-dessus encore, tout en haut, la silhouette d'une ville orientale, c'est Stamboul. Les minarets, les hautes coupoles des mosquées se découpent sur un ciel très étoilé où un mince croissant de lune est suspendu ; l'horizon est tout frangé de tours et minarets, légèrement dessinés en silhouettes bleuâtres sur la teinte pâle de la nuit. Les grands dômes superposés des mosquées montent en teintes vagues jusqu'à la lune, et produisent sur l'imagination l'impression du gigantesque...
>
> ...choeur passa, et se perdit dans l'éloignement. Par ma fenêtre grande ouverte, on ne voyait que la vapeur du matin, le vide immense du ciel ; et puis, tout en haut, quelque chose se dessina en rose, un dôme et des minarets ; la silhouette de la ville turque s'esquissa peu à peu, comme suspendue dans l'air ... Alors, je me rappelai que j'étais à Stamboul, et qu'elle avait juré d'y venir...

[101] Louis Marie Julien Viaud, le véritable Loti, était quant à lui officier la « Royale ».

> ...haut de la djiami d'Orkhan, la vue plane sur le golfe d'Ismidt aux eaux bleues, sur les fertiles plaines d'Asie, et sur l'Olympe de Brousse qui dresse là-haut tout au loin sa grande cime neigeuse...
>
> ...étoiles s'allument dans le ciel pur ; la lune éclaire la rue large et déserte, les arcades arabes et les vieilles tombes. De loin en loin un café turc encore ouvert jette une lueur rouge sur les pavés gris ; les passants sont rares et circulent le fanal à la main ; par-ci par-là, de petites lampes tristes brûlent dans les kiosques funéraires. Je vois pour la dernière fois ces tableaux familiers ; demain, à pareille heure, je serai loin de ce pays...
>
> ...journées à errer sur ce chemin de Monastir. C'était une campagne nue et triste, où l'oeil s'étendait à perte de vue sur des cimetières antiques ; des tombes de marbre en ruine, dont le lichen rongeait les inscriptions mystérieuses ; des champs plantés de menhirs de granit ; des sépultures grecques, byzantines, musulmanes, couvraient ce vieux sol de Macédoine où les grands peuples du passé ont laissé leur poussière. De loin en loin, la silhouette aiguë d'un cyprès, ou un platane immense, abritant des bergers albanais et des chèvres ; sur la terre aride, de larges fleurs lilas pâle, répandant une douce odeur de chèvrefeuille, sous un soleil déjà brûlant. Les moindres détails de ce pays sont restés dans ma mémoire...
>
> ...Stamboul avait un aspect inaccoutumé ; les hodjas dans tous les minarets chantaient des prières inconnues sur des airs étranges ; ces voix aiguës, parties de si haut, à une heure insolite de la nuit inquiétaient l'imagination ; et les musulmans, groupés sur leurs portes, semblaient regarder tous quelque point effrayant du ciel...
>
> ...foule se presse sous un soleil brûlant ; c'est bien le printemps, pour tout de bon, qui arrive comme moi je m'en vais. La grande lumière de midi ruisselle sur tout cet ensemble de murailles, de dômes et de minarets, qui couronnent là-haut Stamboul ; elle s'éparpille sur une foule bariolée, vêtue des couleurs...

Le facteur 'sultan' est peu intense mais très spécifique, il n'apparaît pas en tant que mot dans les dix premiers passages de la classe on le retrouve pourtant dans le passage suivant où l'on voit l'association avec le thème de la mer dans cette superbe description du sacre d'Abd-Ul-Hamid II.

> ...caïques du sultan sont conduits chacun par vingt-six rameurs. Leurs formes ont l'élégance originale de l'Orient ; ils sont d'une grande magnificence, entièrement ciselés et dorés, et portent à l'avant un éperon d'or. La livrée des laquais de la cour est verte et orange, couverte de dorures. Le trône du sultan, orné de plusieurs soleils, est placé sous un dais rouge...[102]

L'association entre la lune de l'islam et le minaret, est frappante.

Facteur	Intensité	Spécificité	Contribution
minarets	0.155	0.639	29.387
pays	0.108	0.5	20.437
mer	0.092	0.524	17.332
Stamboul	0.09	0.513	17.091
sultan	0.07	0.616	13.162
cyprès	0.062	0.719	11.628
loin	0.061	0.35	11.52

[102] On aura déjà noté que les passages débutent et se terminent par un mot du vocabulaire retenu, d'où par exemple ici l'absence d'article devant caïque.

ciel	0.058	0.599	10.928
horizon	0.057	0.682	10.782
lune	0.051	0.694	9.548

Classe 'sens', 163 passages, numéro 4.

Le thème des sens et de la débauche apparaît ici. Le dandy blasé profitant de l'innocence de la jeune fille est un motif en peu rebattu qui est la base de l'intrigue amoureuse du roman. C'est principalement dans ses échanges épistolaires avec son ami Plumckett et sa soeur que s'exprime cette posture.

> ...pauvre ami, le temps et la débauche sont deux grands remèdes ; le coeur s'engourdit à la longue, et c'est alors qu'on ne souffre plus. Cette vérité n'est pas neuve, et je reconnais qu'Alfred de Musset vous l'eût beaucoup mieux accommodée ; mais, de tous les vieux adages, que, de génération en génération, les hommes se repassent, celui-là est un des plus immortellement vrais. Cet amour pur que vous rêvez est une fiction comme l'amitié ; oubliez celle que vous aimez pour une coureuse. Cette femme idéale vous échappe ; éprenez-vous d'une fille de cirque qui aura de belles formes...
>
> ...singulier de l'histoire est encore ceci, c'est que je l'aime. La « petite fleur bleue de l'amour naïf » s'est de nouveau épanouie dans mon coeur, au contact de cette passion jeune et ardente. Du plus profond de mon âme, je l'aime et je l'adore...
>
> ...exalte au point de vue religieux d'abord, tant et si bien, que la pauvre petite abandonnée verse souvent des larmes très amères sur son amour pour un infidèle...
>
> ...songé qu'il fallait qu'il l'aimât bien, elle, l'esclave achetée, l'obscure Aziyadé, puisque, pour la contempler, il risquait si témérairement sa tête. Elle ne se doutait pas, la pauvre petite, que ce garçon si jeune de visage avait déjà abusé de toutes les choses de la vie, et ne lui apportait qu'un coeur blasé, en quête de quelque nouveauté originale ; elle s'était dit qu'il devait faire bon être aimée ainsi, et tout doucement elle avait glissé sur la pente qui devait l'amener dans les bras du giaour...
>
> ...milieu de froissements continuels, conserver les illusions, l'enthousiasme et la fraîcheur morale de la jeunesse ? Non, vous le savez bien ; j'ai renoncé aux plaisirs de mon âge, qui ne sont déjà plus de mon goût, j'ai perdu l'aspect et les allures d'un jeune homme, et je vis désormais sans but comme sans espoir ... Est-ce à dire pourtant que j'en sois réduit au même point que vous, dégoûté de tout, niant tout ce qui est bon, niant la vertu, niant l'amitié, niant tout ce qui peut nous rendre supérieurs à la brute ? Entendons-nous, mon ami ; sur ces points, je pense tout autrement que vous. J'avoue que, malgré mon expérience des choses de ce monde (puissiez-vous n'en jamais acquérir une pareille, il en coûte trop cher !), je crois encore à tout cela, et à bien d'autres choses...
>
> ...âme de Marguerite, son âme était pure et vierge, bien que son corps d'enfant, acheté par un vieillard...
>
> ...femme ? Bien souvent cela tient uniquement à ce que la courbe de son nez, l'arc de ses sourcils, l'ovale de son visage, que sais-je ? ont ce je ne sais quoi auquel correspond en vous un autre je ne sais quoi qui fait le diable à quatre dans votre imagination. Ne vous récriez pas ! la moitié du temps, votre amour...
>
> ...calme de la mer, ce ciel pâle de mars me serrent le coeur. Je souffre bien, mon Dieu ; c'est une angoisse comme si je l'avais vue mourir. J'embrasse ce qui me vient d'elle ; je voudrais pleurer...

> ...regrette Samuel aussi, le pauvre Samuel, qui jouait si gratuitement sa vie pour moi, et qui va pleurer mon départ comme un enfant. C'est ainsi que je me laisse aller encore et prendre à toutes les affections ardentes, à tout ce qui y ressemble, quel qu'en soit le mobile intéressé ou ténébreux ; j'accepte, en fermant les yeux, tout ce qui peut pour une heure combler le vide effrayant de la vie, tout ce qui est une apparence d'amitié ou d'amour...
>
> ...femme un charme moral, une délicatesse de sentiment, une élévation de caractère qui sont la vraie cause de votre amour ... Hélas ! gardez-vous bien de confondre ce qui est en elle et ce qui est en vous. Toutes nos illusions viennent de là : attribuer ce qui est en nous et nulle part ailleurs à ce qui nous plaît. Faire une châsse à la femme que l'on aime et prendre son ami pour un homme de génie...

Donc bien entendu les facteurs 'sens', 'âme', 'femme', 'aime', 'infini' et 'abîme'.

Facteur	Intensité	Spécificité	Contribution
sens	0.089	0.573	14.558
âme	0.087	0.587	14.139
temps	0.081	0.571	13.282
femme	0.08	0.508	12.993
aime	0.07	0.467	11.478
abîme	0.067	0.352	10.912
heureux	0.061	0.478	9.878
infini	0.058	0.307	9.502
garçon	0.051	0.312	8.264
personne	0.05	0.319	8.14

5.3.3 **Classe 'Eyoub', 130 passages, numéro 9.**

Le quartier d'Eyoub, lieu des amours de Loti et Aziyadé tient bien sûr une place importante dans le roman, c'est d'ailleurs le titre le titre de la partie centrale du roman 'Eyoub à deux'. Malheureusement les huit extraits correspondant à des en-têtes de courriers sélectionnés automatiquement par le système ne rendent pas justice à cette classe. Nous aurions par exemple préféré :

> ...soir, on nous trouve, comme deux bons Orientaux, fumant notre narguilhé sous les platanes d'un café turc, ou bien nous allons au théâtre des ombres chinoises, voir Karagueuz, le Guignol turc qui nous captive. Nous vivons en dehors de toutes les agitations, et la politique n'existe...

En effet, ce passage se trouve dans cette classe par le thème du spectacle et de la description des moeurs turques, on y trouve aussi :

> ...Ismidt est une grande ville turque, assez civilisée, située au bord d'un golfe admirable ; les bazars y sont animés et pittoresques. Il est interdit aux habitants de se promener après huit heures du soir, même en compagnie d'une lanterne...

ou encore :

> ...café turc, chez le cafedji Suleïman, on élargit le cercle autour du feu, quand j'arrive le soir, avec Samuel et Achmet. Je donne la main à tous les assistants, et je m'assieds pour

> écouter le conteur des veillées d'hiver (les longues histoires qui durent huit jours, et où figurent les djinns et les génies). Les heures passent là sans fatigue et sans remords ; je me trouve à l'aise au milieu d'eux, et nullement...

C'est une dimension importante du roman, que cette double identité Loti Arif, ces déguisements qui rappellent René Caillié visitant Tombouctou presque cinquante années auparavant[103] en se faisant passer pour un pèlerin musulman. Loti est fier d'habiter Eyoub comme un musulman :

> La mosquée d'Eyoub, située au fond de la Corne d'or, fut construite sous Mahomet II, sur l'emplacement du tombeau d'Eyoub, compagnon du prophète. L'accès en est de tout temps interdit aux chrétiens, et les abords mêmes n'en sont pas sûrs pour eux.

La sélection automatique ne rend donc pas compte de la richesse de cette classe mais cela fait partie des limitations des systèmes automatiques que nous avons trop souvent soulignés pour qu'il soit nécessaire d'y revenir. Nous avions le choix de filtrer ces en-têtes qui sont des attracteurs discutables mais d'un autre coté, comme nous l'avons déjà dit, la dimension épistolaire du roman est indiscutable et l'opposition entre Brightbury et Eyoub renvoie aussi clairement à l'opposition entre chrétienté et islam qui est bien une composante importante du roman. Est-ce le fait du hasard ? C'est peu probable, la cohérence de cette classe nous a paru justifiable et nous n'avons pas filtré ces passages.

> ...heure de la prière du soir, un soir d'hiver. Le muezzin chantait son éternelle chanson, et nous étions enfermés tous deux dans notre mystérieux logis d'Eyoub...

[103] 1828

...Eyoub, 27 septembre 1876...

...Eyoub, décembre 1876...

...Eyoub, le 4 décembre 1876...

...Eyoub, 15 novembre 1876...

...Eyoub ..., 1876...

...Septembre 1876...

...Eyoub, décembre 1876...

...soir, nous remontions en caïque la Corne d'or ; jamais nous n'avions tant couru Stamboul ensemble en plein jour. Elle paraissait ne plus se soucier d'aucune précaution, comme si tout était fini pour elle, et que le monde lui fût indifférent...

...Londres, juin 1876...

L'opposition entre les facteurs 'Eyoub' et 'Brightbury', 'Eyoub' et le 'Phanar' [104], renforce notre interprétation. Enfin le facteur 'Kars' confirme cette impression, en effet, le roman se termine par ces mots :

> Parmi les morts de la dernière bataille de Kars, on a retrouvé le corps d'un jeune officier de la marine anglaise, récemment engagé au service de la Turquie sous le nom de Arif-Ussam effendi. « Il a été inhumé parmi les braves défenseurs de l'islam (que Mahomet protège !), aux pieds du Kizil-Tépé, dans les plaines de Karadjémir. » FIN

Notons enfin la présence du facteur 'caïque' qui s'inscrit dans cette opposition symbolique entre l'échelle du Phanar et celle d'Eyoub lors des fréquentes traversées en caïque de la Corne d'Or par les deux amants.

Facteur	Intensité	Spécificité	Contribution
Eyoub	0.219	0.627	28.418
Brightbury	0.136	0.889	17.62
Salonique	0.133	0.642	17.232
caïque	0.122	0.685	15.924
nuit	0.063	0.228	8.149
chanson	0.052	0.456	6.745
Extrait	0.048	0.663	6.303

[104] Le Phanar est le quartier grec, citons ici un passage du roman qui éclaire la vison que Loti avait de ce quartier :

« C'était Noël à la grecque ; le vieux Phanar était en fête.

Des bandes d'enfants promenaient des lanternes, des girandoles de papier, de toutes les formes et de toutes les couleurs ; ils frappaient à toutes les portes, à tour de bras, et donnaient des sérénades terribles, avec accompagnement de tambour.

Achmet, qui passait avec moi, témoignait un grand mépris pour ces réjouissances d'infidèles.

Le vieux Phanar, même au milieu de ce bruit, ne pouvait s'empêcher d'avoir l'air sinistre. »

Et plus loin :

« On voyait cependant s'ouvrir toutes les petites portes byzantines, rongées de vétusté, et dans leurs embrasures massives apparaissaient des jeunes filles, vêtues comme des Parisiennes, qui jetaient aux musiciens des piastres de cuivre.

Ce fut bien pis quand nous arrivâmes à Galata ; jamais, dans aucun pays du monde, il ne fut donné d'ouïr un vacarme plus discordant, ni de contempler un spectacle plus misérable.

C'était un grouillement cosmopolite inimaginable, dans lequel dominait en grande majorité l'élément grec. L'immonde population grecque affluait en masses compactes ; il en sortait de toutes les ruelles de prostitution, de tous les estaminets, de toutes les tavernes. Impossible de se figurer tout ce qu'il y avait là d'hommes et de femmes ivres, tout ce qu'on y entendait de braillements avinés, de cris écoeurants.

Et quelques bons musulmans s'y trouvaient aussi, venus pour rire tranquillement aux dépens des infidèles, pour voir comment ces chrétiens du Levant sur le sort desquels on a attendri l'Europe, par de si pathétiques discours, célébraient la naissance de leur prophète.

Tous ces hommes qui avaient si grande peur d'être obligés d'aller se battre comme des Turcs, depuis que la Constitution leur conférait le titre immérité de citoyens, s'en donnaient à coeur joie de chanter et de boire. »

Kars	0.047	0.515	6.136
Phanar	0.042	0.365	5.516
caïques	0.042	0.589	5.45

5.3.4 **Classe 'vieille', 126 passages, numéro 7.**

Le premier passage exprime bien le problème posé par cette classe le sème /vieux/ s'applique indifféremment aux personnes et aux monuments. Est-ce là encore le fait du hasard ? Nous ne le croyons pas. Il y a bien un thème, l'opposition entre la vieille Kadidja[105] et la jeune d'Aziyadé mais plus profondément le thème de la coutume, des racines ancestrales, de la sagesse que traduit bien le terme turque 'eski' que Loti rapporte dans ce passage :

> J'examinai les vieillards qui m'entouraient : leurs costumes indiquaient la recherche minutieuse des modes du bon vieux temps ; tout ce qu'ils portaient était eski, jusqu'à leurs grandes lunettes d'argent, jusqu'aux lignes de leurs vieux profils. Eski, mot prononcé avec vénération, qui veut dire antique, et qui s'applique en Turquie aussi bien à de vieilles coutumes qu'à de vieilles formes de vêtement ou à de vieilles étoffes. Les Turcs ont l'amour du passé, l'amour de l'immobilité et de la stagnation.

Donc si la classe est polysémique, certains passages ne se retrouvent là qu'à cause d'associations purement formelles elle traduit bien un thème malgré tout.

> ...pied des vieilles murailles du quartier byzantin de Constantinople, lieu qui n'est fréquenté à pareille heure par aucun être humain. Deux femmes pourtant s'y tenaient blotties, deux ombres à tête blanche, cachées dans certain recoin obscur qui nous était familier, sous le balcon d'une maison en ruine ... C'étaient Aziyadé, et la vieille, la fidèle Kadidja...
>
> ... À Achmet, fils d'Ibrahim, qui demeure à Yedi-Koulé, dans une traverse donnant sur Arabahdjilar-Malessi, près de la mosquée. C'est la troisième maison après un tutundji, et à côté il y a une vieille Arménienne qui vend des remèdes, et, en face, un derviche...
>
> ...expédiai en même temps dans la campagne trois enfants chargés de me rapporter des branches de verdure, et des gerbes, de pleins paniers de narcisses et de jonquilles. Je voulais que la vieille maison prît ce jour-là pour son retour un aspect inaccoutumé de joie et de fête...
>
> ...intimité de la jeune femme obscure et de la vieille cadine, rigide et fière, de noble souche et de grande maison...
>
> ...maison brûlerait. Cours vite, Arif ! me répondit un vieux Turc, cours vite, Arif ! c'est ta maison...
>
> ...hiver ; une pluie froide et un grand vent battent les vitres de ma triste case ; on n'entend plus d'autre bruit que celui qu'ils font, et la vieille lampe turque pendue au-dessus de ma tête est la seule qui brûle à cette heure dans Eyoub. C'est un sombre pays qu'Eyoub, le coeur de l'islam ; c'est ici qu'est la mosquée sainte où sont sacrés les sultans ; de vieux derviches farouches et les gardiens des saints tombeaux sont les seuls habitants de ce quartier, le plus musulman...
>
> ...femme à précautions : un aimable eunuque, habitué sans doute aux escapades de sa maîtresse, se tenait, à toute éventualité, près de la porte de ma maison...

[105] Le chaperon complice d'Aziyadé

> ...vieille masure en était pleine, et qu'elles s'y livraient, la nuit, des batailles rangées fort meurtrières. Tchok setchan var senin evdé, Lotim ! disait-elle souvent. (Il y a beaucoup de souris dans ta maison, Loti...
>
> ...Turquie, des escaliers pour monter sur les toits, et, moi qui te parle, ayant un jour eu l'idée de me promener sur ma maison, je me suis vu passer dans mon quartier pour un garçon excentrique...
>
> ...case de Kadidja ; mais la vieille avait déménagé, et personne ne put m'indiquer sa demeure...

On voit apparaître au travers des facteurs les plus importants la difficulté à interpréter cette classe. Si on peut penser qu'Arif est plus *eski* que Loti, le statut de 'porte', 'regarde' et 'tête' est plus problématique.

A part deux autres passages le mot *regarde* n'apparaît que dans les passages suivants :

> Samuel est enveloppé comme un pacha dans mon manteau de fourrure, que je lui abandonne ; sa belle tête est pâle et triste ; il **regarde** en silence défiler les quartiers de Stamboul, les places immenses et désertes où poussent l'herbe et la mousse, les minarets gigantesques, les **vieilles mosquées** décrépites, blanches sur le ciel gris, les **vieux monuments** avec leur **cachet d'antiquité** et de délabrement, qui s'en vont en ruine comme l'islamisme
>
> Je **regarde** ce **vieux portique** noir, là-bas, et cette rue déserte qui s'enfonce dans un bas-fond sombre. C'est là qu'elle habite, et, en m'avançant de quelques pas, je verrais encore sa demeure.
>
> on ne sait jamais si, des fenêtres d'une **maison turque**, quelqu'un vous **regarde** ou ne vous **regarde** pas.
>
> Je me cache à moitié derrière un pan de **muraille**, je **regarde** cette **maison**, et mon coeur bat terriblement.

Donc presque toujours associé à de vieux édifices ou au mot *maison* qui lui-même désigne souvent comme une vieille maison.

Le facteur 'porte' est le pire car très polysémique mais il s'agit bien en général de la porte d'un édifice par exemple :

> il arriva que je m'arrêtai devant la **porte** fermée d'une vieille mosquée, pour regarder se battre deux cigognes.
>
> Si vous aviez pu suivre aujourd'hui votre ami Loti dans les rues d'un **vieux** quartier solitaire, vous l'auriez vu monter dans une **maison** d'aspect fantastique. La **porte** se referme sur lui avec mystère.
>
> Loti trouve qu'il n'est pas mal en effet, et sourit tristement à cette toilette qui pourrait lui être fatale ; et puis il disparaît par une **porte** de derrière et traverse toute une ville saugrenue, des bazars d'Orient et des **mosquées** ;

Comme on peut le constater en annexe le mot vieille est le premier cooccurrent du facteur 'tête'.

Facteur	Intensité	Spécificité	Contribution
vieille	0.123	0.606	15.554
Arif	0.101	0.686	12.708
tête	0.097	0.509	12.262
mère	0.09	0.509	11.359
porte	0.085	0.61	10.669
regarde	0.081	0.501	10.17
madame	0.075	0.377	9.499
vieux	0.057	0.501	7.139
ans	0.055	0.284	6.975
quartier	0.055	0.381	6.908

Devant tant de difficultés à se faire une idée claire de la cause de ses regroupements nous avons reclassifié cette classe avec seuil de densité plus élevé, 0,06 en l'occurrence. La matrice de densité de cette classification est présentée à la figure 19.

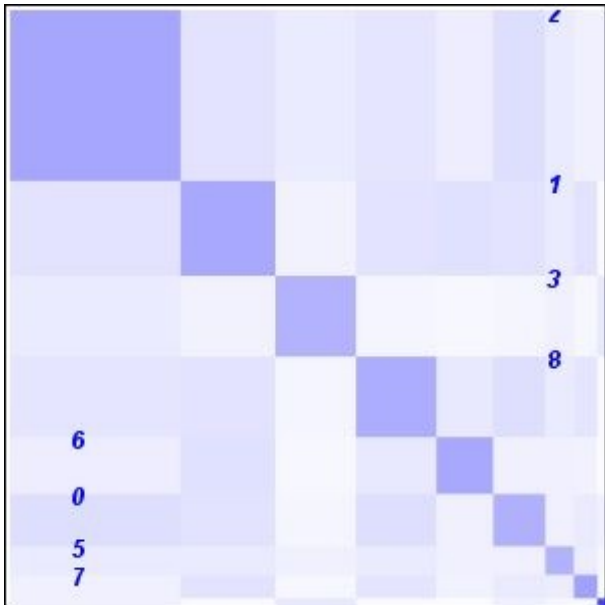

Figure 19 : Classification de la classe 'vieille'

Classe Arif, 36 passages, numéro 2

Les facteurs 'Arif' et 'maison' sont liés du fait que dans le roman l'identité d'Arif se fonde sur sa maison d'Eyoub :

> Arif et Loti étant deux personnages très différents, il suffirait, le jour du départ du Deerhound, qu'**Arif** restât dans sa **maison** ; personne sans doute ne viendrait l'y chercher ; seulement, Loti aurait disparu, et disparu pour toujours.

D'autre part si les facteurs 'maison' et 'vieille' ne sont pas liés directement : 'maison' est lié au facteur 'Kadidja' car Kadidja est une vieille femme.

> Ce soir, avait dit **Kadidja** (la **vieille** négresse qui, à Salonique, accompagnait la nuit Aziyadé dans sa barque et risquait sa vie pour sa maîtresse), ce soir, un caïque l'amènera à l'échelle d'Eyoub, devant ta **maison**.
>
> Je me retrouvai appuyé contre une fontaine de marbre, près de la **maison** peinte de tulipes et de papillons jaunes qu'Aziyadé avait habitée ; j'étais assis et la tête me tournait ; les **maisons** sombres et désertes dansaient devant mes yeux une danse macabre ; mon front frappait sur le marbre et s'ensanglantait ; une **vieille** main noire, trempée dans l'eau froide de la fontaine, faisait matelas à ma tête ... Alors, je vis la **vieille Kadidja** près de moi qui pleurait ; je serrai ses mains ridées de singe ; elle continuait de verser de l'eau sur mon front ...

Dans ce dernier extrait la première maison à laquelle il est fait référence est celle d'Aziyadé, la vielle Kadidja est le lien entre la maison d'Arif et celle d'Aziyadé. On comprend ainsi mieux la sélection automatique des passages et des facteurs qui suivent :

> ...maison brûlerait. Cours vite, Arif ! me répondit un vieux Turc, cours vite, Arif ! c'est ta maison...
>
> ...femme à précautions : un aimable eunuque, habitué sans doute aux escapades de sa maîtresse, se tenait, à toute éventualité, près de la porte de ma maison...
>
> ...cache à moitié derrière un pan de muraille, je regarde cette maison, et mon coeur bat terriblement...
>
> ...disait-il, si même elle sortait, tu n'as plus de maison pour la recevoir. Où trouverais-tu, Loti, dans Stamboul, l'hospitalité pour toi et la femme d'un autre ? Si elle te voit ou si les femmes lui disent que tu es là, elle se perdra comme une folle, et, demain, tu la laisseras dans la rue. Cela t'est égal, à toi qui vas partir ; mais, Loti, si tu fais cela, je te déteste et tu n'as pas de coeur...
>
> ...conversation ne vous touchera guère plus que celle d'une araignée qui vous raconterait qu'un plumeau dévastateur lui a détruit une partie de sa toile ; ou que celle d'un crapaud qui vous annoncerait qu'il vient d'hériter d'un gros tas de plâtras dans lequel il pourra gîter tout à l'aise. (Un monsieur me disait aujourd'hui qu'il avait fait de mauvaises récoltes, et qu'il avait hérité d'une maison de campagne...

Facteur	Intensité	Spécificité	Contribution
Arif	0.311	0.882	11.213
vieille	0.269	0.623	9.692

Classe 'mère', 20 passages, numéro 1

La mère est une vieille femme, c'est en tous cas ainsi qu'elle apparaît au narrateur. À part « la vieille femme à tête de sorcière », l'importance du facteur 'tête' est moins claire ici même si des expressions figurées comme « baisser la tête » se retrouve dans les lettres que Loti reçoit de sa soeur qui lui parle de sa mère :

> ...Tant que je conserverai ma chère vieille mère, je resterai en apparence ce que je suis aujourd'hui. Quand elle n'y sera plus, j'irai te dire adieu, et puis je disparaîtrai sans laisser trace de moi-même...
>
> ...« Ma mère Béhidjé » est une très extraordinaire vieille femme, octogénaire et infirme, fille et veuve de pacha, plus musulmane que le Koran, et plus raide que la loi...
>
> ...Achmet avait pris sa course pour aller réveiller une vieille femme à tête de sorcière qui l'arrêta enfin avec des plantes et de la cendre...
>
> ...dernière phrase était à peine articulée, et, suivant son habitude, plutôt mimée que parlée. Pour dire : « Je serai vieille », elle cassait sa voix jeune, et, pendant quelques secondes, elle se ramassait sur elle-même comme une petite vieille, courbant son corps si plein de jeunesse ardente et fraîche. Zarar yok (cela ne fait rien), était la conclusion. Cela ne fait rien, Loti...
>
> ...pavés, sur l'herbe verte, apparut une tournure de vieille, rasant les murailles ; sous les plis de son manteau passaient ses jambes maigres et nues, d'un noir d'ébène ; elle trottinait tête basse, et se parlait à elle-même ... C'était Kadidja...

Facteur	Intensité	Spécificité	Contribution
mère	0.377	0.665	7.549
tête	0.283	0.462	5.669

Classe 'ami', 17 passages, numéro 8

Comme 'maison' est associé à 'Arif', 'case' et 'toit' sont associés à 'ami' et la case est souvent appelée la *vieille case*.

> ...heureux de me retrouver dans cette petite case perdue, qu'un instant j'avais eu peur de quitter...
>
> ...beau dimanche de janvier, rentrant à la case par un gai soleil d'hiver, je vis dans mon quartier cinq cents personnes et des pompes. Qu'est-ce qui brûle ? demandai-je avec impatience...
>
> ...veilleur de nuit m'engagea à rentrer dans ma case, après s'être informé du motif de ma promenade...
>
> ...approchais lentement de notre case. Les portes étaient enfoncées, les vitres brisées, la fumée sortait par le toit ; tout était au pillage, envahi par une de ces foules sinistres qui surgissent à Constantinople dans les heures de bagarre. J'entrai chez moi, il pleuvait de l'eau noire mêlée de suie, du plâtre calciné et des planches enflammées...

...habitués de la case, Suleïman, le vieux Riza, les derviches Hassan et Mahmoud, contemplaient ce spectacle avec stupéfaction...

Facteur	Intensité	Spécificité	Contribution
ami	0.268	0.902	4.557
toit	0.201	0.698	3.424

Classe 'vieux', 17 passages, numéro 3

Le sordide passage du vieux Kaïroullah à qui Loti Marketo demande des services d'entremetteur puis pire encore...

...vieux Kaïroullah réfléchit un instant et répondit...

...vieux Kaïroullah réfléchit longuement à ma demande et répondit...

...Vieux Kaïroullah, dis-je, amène-moi des femmes...

...Vieux Kaïroullah, dis-je, ton fils est plus beau...

...monsieur Marketo, me dit le vieux Kaïroullah en se penchant à mon oreille...

Facteur	Intensité	Spécificité	Contribution
vieux	0.3	0.714	5.101
ans	0.242	0.591	4.12

Classe 'fontaine', 12 passages, numéro 6

Le thème de la fontaine et de l'eau claire est lié à la vieille Kadidja, aux souvenirs...

...retrouvai appuyé contre une fontaine de marbre, près de la maison peinte de tulipes et de papillons jaunes qu'Aziyadé avait habitée ; j'étais assis et la tête me tournait ; les maisons sombres et désertes dansaient devant mes yeux une danse macabre ; mon front frappait sur le marbre et s'ensanglantait ; une vieille main noire, trempée dans l'eau froide de la fontaine, faisait matelas à ma tête ... Alors, je vis la vieille Kadidja près de moi qui pleurait ; je serrai ses mains ridées de singe ; elle continuait de verser de l'eau sur mon front...

...souvenirs lointains ; puis peu à peu les images vinrent, plus nettes et plus précises, je m'y retrouvai...

...jour, pour aller, de la tête aux pieds, nous laver en plein vent, dans l'eau claire d'une fontaine...

...eau, la pâpoutch, ou sur la vase, ou bien encore sur la tête d'un chat...

...belle nuit de Noël, bien claire, bien étoilée, bien froide...

Facteur	Intensité	Spécificité	Contribution
fontaine	0.278	0.698	3.334
eau	0.238	0.845	2.852

Classe 'rue', 11 passages, numéro 0

Les vieilles rues dans les vieux quartiers de l'antique Stamboul, vraiment *eski*.

...passâmes Dolma-Bagtché, Sali-Bazar, Top-Hané, le bruyant quartier de Galata, et puis le pont de Stamboul, le triste Phanar et le noir Balate. A Eyoub enfin, dans une vieille rue turque, devant un Conak antique, à la mine opulente et sombre, les trois femmes s'arrêtèrent...

...rue verticale du quartier turc de Djianghir, sur les hauteurs du Taxim, habite la vieille Béhidjé-hanum. Son appartement, qui déjà surplombe des précipices, porte deux shaknisirs en saillie, soigneusement grillés de lattes de frêne...

...cinquième heure aux horloges turques ; les veilleurs de nuit frappent le sol de leurs lourds bâtons ferrés. Les chiens sont en révolution dans le quartier de Galata et poussent là-bas des hurlements lamentables. Ceux de mon quartier gardent la neutralité et je leur en sais gré ; ils dorment en monceaux devant ma porte. Tout est au grand calme dans mon voisinage ; les lumières s'y sont éteintes une à une, pendant ces trois longues heures que j'ai passées là, étendu devant ma fenêtre ouverte...

...regarde ce vieux portique noir, là-bas, et cette rue déserte qui s'enfonce dans un bas-fond sombre. C'est là qu'elle habite, et, en m'avançant de quelques pas, je verrais encore sa demeure...

...approchais toujours. J'étais dans la rue sombre qui monte à Mehmed-Fatih, la rue...

Facteur	Intensité	Spécificité	Contribution
rue	0.285	0.557	3.132
quartier	0.209	0.333	2.299

Classe 'saurai', 6 passages, numéro 5

Où il faut savoir souffrir... Le thème de la vieillesse diffuse au travers de vieillard, mère, vieux, mortel...

...lune éclairait des murailles nues, un plancher nu, une chambre vide ; les meubles absents, les tables de planches grossières dépouillées de leurs couvertures de soie, éveillaient des idées de misère, de froid et de solitude ; les chiens hurlaient au-dehors de cette manière lugubre qui, en Turquie comme en France est réputée présage de mort ; le

vent sifflait à notre porte, ou gémissait tout doucement comme un vieillard qui va mourir...

...sentais derrière moi la haine exaspérée de cette créature, qui adorait sa maîtresse que j'avais fait mourir. J'avais peur de me retourner pour la voir, peur de l'interroger, peur d'une preuve et d'une certitude, et je marchais toujours, comme un homme...

...force me fut de revenir dormir seul, dans ma chambre sans vitres ni portes, roulé, par un froid mortel, dans des couvertures mouillées qui sentaient le roussi. Je dormis peu, et mes réflexions furent sombres ; cette nuit fut une des nuits désagréables de ma vie...

...traîne le désespoir dans lequel m'a mise ta lettre ... Tu veux disparaître !... Un jour, peut-être prochain, où notre bien-aimée mère nous quittera, tu veux disparaître, m'abandonner pour toujours. Table rase de tous nos souvenirs, engloutissement de notre passé, la vieille case de Brightbury vendue, les objets chéris dispersés, et toi qui ne seras pas mort ... ! qui seras là quelque part à végéter sous la griffe de Satan, quelque part où je ne saurai pas, mais où je sentirai que tu vieillis et que tu souffres !... Que Dieu plutôt te fasse mourir ! Alors, je te pleurerai ; alors, je saurai qu'il faut ainsi que le vide se fasse, j'accepterai, je souffrirai, je courberai la tête...

...monte de vieux escaliers sombres, couverts de somptueux tapis de Perse ; le haremlike s'entr'ouvre doucement et des yeux de femmes vous observent, par l'entrebâillement d'une porte...

Facteur	Intensité	Spécificité	Contribution
saurai	0.149	0.351	0.892
souffrir	0.145	0.503	0.873

Classe 'saison', 5 passages, numéro 7

Saison, printemps, joie, poésie... A part la 'vieille poésie' le thème de la vieillesse n'est pas très explicite.

...Extrait d'une vieille poésie orientale...

...habits turcs. Je cours à Azarkapou. Je monte dans le premier caïque qui passe. Le caïqdji me reconnaît...

...Cher frère aimé, je veux, moi aussi, te souhaiter la bienvenue dans notre pays. Fasse Celui auquel je me confie que tu t'y trouves bien et que notre tendresse adoucisse tes peines ! Il me semble que nous ne négligerons rien pour cela, nous sommes pleins de la joie de ton retour...

...Achmet n'est plus là, à son poste, caracolant à Top-Hané sur son cheval blanc. Galata même est mort ; on voit que quelque chose de terrible comme une guerre d'extermination se passe au-dehors...

...printemps, les amandiers fleurissent, et moi, je vois avec terreur, chaque saison qui m'entraîne plus avant dans la nuit, chaque année qui m'approche du gouffre ... Où vais-je, mon Dieu ?... Qu'y a-t-il après ? et qui sera près de moi quand il faudra boire la sombre coupe...

Facteur	Intensité	Spécificité	Contribution
saison	0.245	0.282	1.226
printemps	0.192	0.42	0.958

Classe 'Dimitraki', 2 passages, numéro 4

Où l'on retrouve une obsession de Loti qui décidément ne portait pas les grecs dans son coeur.

> ...vieux Dimitraki exerçait l'invraisemblable métier de tatoueur pour marins grecs. Il avait une légèreté de touche, et une sûreté...
>
> ...À Azar-kapou, je dus le suivre dans d'immondes ruelles de truands, boueuses, noires, sinistres, occupées par des marchands de goudron, de vieilles poulies et de peaux de lapin ; de porte en porte, nous demandions un certain vieux Dimitraki, que nous finîmes par trouver, au fond d'un bouge inénarrable...

Facteur	Intensité	Spécificité	Contribution
Dimitraki	0.608	0.897	1.216
ruelles	0.305	0.721	0.609

Conclusion de l'analyse de la classe 'vieille'

Le thème de la vieillesse, des vieilles pierres, des vieillards dignes ou indignes, de la mère et de la vieille Kadidja et tout ce qui est *eski* se trouvent regroupés dans cette classe et même si parfois les facteurs de regroupements ne sont pas de la plus grande limpidité nous n'avons pas tenté d'améliorer cette classe 'importante.

5.3.5 **Classe 'dévouement', 85 passages, numéro 5.**

Les thèmes du dévouement et de l'amitié tiennent du poncif dans les échanges de Loti avec Plumckett ou sa soeur qui regorgent de considérations sur la morale et l'âme.

> ...À la quatrième page de votre papier, votre main courait un peu vite sans doute, quand vous avez écrit : « une affection et un dévouement illimités. » Si vous avez pensé cela, vous voyez bien, mon cher ami, qu'il y a encore chez vous de la jeunesse et de la fraîcheur, et que tout n'est pas perdu. Ces belles amitiés-là, à la vie, à la mort, personne plus que moi n'en a éprouvé tout le charme ; mais, voyez-vous, on les a à dix-huit ans ; à vingt-cinq, elles sont finies, et on n'a plus de dévouement que pour soi-même. C'est désolant, ce que je vous dis là, mais c'est terriblement vrai...
>
> ...cher William : je vous ai écrit longuement. Je ne crois nullement à votre affection, pas plus qu'à celle de personne ; mais vous êtes, parmi les gens que j'ai rencontrés deçà et delà dans le monde, un de ceux avec lesquels je puis trouver du plaisir à vivre et à échanger mes impressions. S'il y a dans ma lettre quelque peu d'épanchement, il ne faut pas m'en vouloir : j'avais bu du vin de Chypre...
>
> ...cher ami...
>
> ...cher ami...
>
> ...roulette ne donne plus, et nous voilà fort pauvres tous deux, mais si insouciants que cela compense ; assez jeunes d'ailleurs pour avoir pour rien des satisfactions que d'autres payent fort cher...
>
> ...connu tout ce dont je viens de parler, et à qui tout cela manque, est fort à plaindre...
>
> ...cher Loti...
>
> ...cher Loti...
>
> ...cher Loti...
>
> ...assez heureux pour vous inspirer quelque affection ; je vous en remercie. Nous aurons, si vous voulez bien, ce que vous appelez une amitié intellectuelle , et nos relations nous aideront à passer le temps maussade de la vie...

Les facteurs 'amis' et 'dévouement' s'interprètent assez facilement. Les facteurs 'pauvres', 'gens' et 'capables' se constitue autour de considérations sur les 'pauvres gens' capables de dévouement ou des médecins qui se dévouent pour les 'pauvres malades' comme par exemple :

> Tout cela n'empêche pas, mon ami, qu'il n'y ait sur cette terre de fort braves gens, des gens foncièrement honnêtes, organiquement bons, faisant le bien pour la satisfaction intime qu'ils en retirent : ne volant pas et n'assassinant pas, lors même qu'ils seraient sûrs de l'impunité, parce qu'ils ont une conscience qui est un contrôle perpétuel des actes auxquels leurs passions pourraient les pousser ; des gens capables d'aimer, de se dévouer corps et âme, des prêtres croyant en Dieu et pratiquant la charité chrétienne, des médecins bravant les épidémies pour sauver quelques **pauvres** malades, des soeurs de charité allant au milieu des armées soigner de **pauvres** blessés, des banquiers à qui vous

> pourrez confier votre fortune, des amis qui vous donneront la moitié de la leur ; des gens, moi par exemple sans aller chercher plus loin, qui seraient peut-être **capables**, en dépit de tous vos blasphèmes, de vous offrir une affection et un **dévouement** illimités.

Le facteur 'Marketo' s'explique mal ici, Marketo est le nom sous lequel Loti est connu dans la communauté juive de Stamboul le mot est principalement utilisé dans le navrant épisode du vieux Kaïroullah et nous n'avons pas trouvé d'explications convaincantes de sa présence ici.

Les facteurs 'mal' et 'monde' sont des facteurs diffusant énormément du fait de la haute fréquence des mots correspondants, ils peuvent être une marque du genre épistolaire qui prédomine dans cette classe.

Le facteur 'rentrais' ne se constitue qu'autour de cet unique passage, l'usage de la première personne marque le genre épistolaire :

> Vous avais-je dit, mon cher ami, que j'étais malheureux ? Je ne le crois pas, et assurément, si je vous ai dit cela, j'ai dû me tromper. Je rentrais ce soir chez moi en me disant, au contraire, que j'étais un des heureux de ce monde, et que ce monde aussi était bien beau. Je rentrais à cheval par une belle après-midi de janvier ; le soleil couchant dorait les cyprès noirs, les vieilles murailles crénelées de Stamboul, et le toit de ma case ignorée, où Aziyadé m'attendait.

Le facteur 'harem' est certainement présent du fait de passages sur le sort des femmes dans les harems que dénonçait Loti, comme par exemple :

> ...femmes turques, les grandes dames surtout, font très bon marché de la fidélité qu'elles doivent à leurs époux. Les farouches surveillances de certains hommes, et la terreur du châtiment sont indispensables pour les retenir. Toujours oisives, dévorées d'ennui, physiquement obsédées de la solitude des harems, elles sont capables de se livrer au premier venu, au domestique qui leur tombe sous la patte, ou au batelier qui les promène, s'il est beau et s'il leur plaît. Toutes sont fort curieuses des jeunes gens européens, et ceux-ci en profiteraient quelquefois s'ils les avaient, s'ils l'osaient, ou si plutôt ils étaient placés dans des conditions favorables pour le tenter. Ma position à Stamboul, ma connaissance de la langue et des usages turcs, ma porte isolée tournant sans bruit sur ses vieilles ferrures, étaient choses fort propices à ces sortes d'entreprises ; et ma maison eût pu devenir sans doute, si je l'avais désiré, le rendez-vous des belles désoeuvrées des harems...

Facteur	Intensité	Spécificité	Contribution
dévouement	0.095	0.508	8.115
ami	0.091	0.396	7.721
pauvres	0.083	0.593	7.034
capables	0.079	0.509	6.749
monde	0.07	0.451	5.947
Marketo	0.064	0.556	5.409

mal	0.063	0.313	5.347
gens	0.062	0.361	5.284
harems	0.056	0.428	4.757
rentrais	0.055	0.302	4.685

5.3.6 **Classe 'Midhat-pacha', 72 passages, numéro 6.**

C'est le grand vizir Midhat-pacha qui dépose le sultan Abd-Ul-Aziz et amène Abd-Ul-Hamid II au pouvoir. L'agitation politique de cette période est bien décrite par le premier passage, cependant le véritable thème de cette classe est l'agitation et le bruit en général.

> ...bruit se fit entendre, bruit de pas et de voix humaines ; une bande de softas entrait par les portiques du centre, portant des lanternes et des bannières ; ils criaient : « Vive le sultan ! vive Midhat-pacha ! vive la constitution ! vive la guerre ! » Ces hommes étaient comme enivrés de se croire libres ; et, seuls, quelques vieux Turcs qui se souvenaient du passé haussaient les épaules en regardant courir ces foules exaltées. Allons saluer Midhat-pacha, s'écrièrent les softas...
>
> ...vieux Phanar, même au milieu de ce bruit, ne pouvait s'empêcher d'avoir l'air sinistre...
>
> ...bruit de la fête se perdit dans la brume, et nous retombâmes dans le silence et l'obscurité...
>
> ...bruit de sa chute dans le silence profond indiquait lequel de nous deux avait deviné...
>
> ...Samuel se terminent en ate ; tout ce qui fait du bruit se dit : fate boum (faire boum...
>
> ...orage est passé et le temps est radieux ; on n'entend que le bruit lointain des chiens errants qui jappent dans le silence du soir...
>
> ...bruit s'était renouvelé, plus distinct et moins terrible, si caractéristique même qu'il ne laissait plus d'équivoque : Setchan ! (Les souris !) dit-elle en riant...
>
> ...premières nuits qu'elle passa dans cette case isolée d'Eyoub, un bruit rapproché, dans l'escalier même du vieux logis, nous fit tous deux frémir. Tous deux nous crûmes entendre à notre porte une troupe de djinns, ou des hommes à turban, rampant sur les marches vermoulues, avec des poignards et des yatagans dégainés. Nous avions tout à craindre, quand nous étions réunis, et il nous était permis de trembler...
>
> ...bruit et du sang aux douleurs turques...
>
> ...porte, lourde et ferrée ; deux petites esclaves circassiennes viennent sans bruit vous ouvrir...

Les facteurs parlent à peu prés d'eux-mêmes à part le Facteur 'À' qui introduit des expressions comme « À onze heure... », « À mes pieds... », etc. et que l'on retrouve ici du fait du passage :

> **À** une heure, un tapage inattendu dans le silence de cette nuit : des harpes et des voix de femmes ; on nous crie gare, et à peine avons-nous le temps de nous garer. Un canot de la

> Maria Pia passe grand train près de notre barque ; il est rempli d'officiers italiens en partie fine, ivres pour la plupart ; il avait failli passer sur nous et nous couler
>
> À onze heures, un léger bruit d'avirons sur la mer calme ; un point lointain s'approche en glissant comme une ombre. C'est la barque de Samuel. Les factionnaires le couchent en joue et le hèlent. Samuel ne répond rien, et cependant les fusils s'abaissent ; les factionnaires ont une consigne secrète qui concerne lui seul, et le voilà le long du bord

Le facteur 'arrêté' arrive un peu par la bande, en effet à l'agitation et au bruit s'ajoute souvent le sang or il est un épisode dramatique où Aziyadé se blesse en brisant une tasse à café dans sa main et le passage correspondant se trouve dans cette classe :

> Cependant la tache s'élargissait par terre, et un liquide sombre tombait toujours de sa main fermée, goutte à goutte d'abord, ensuite en mince filet noir. Une lanterne éclairait misérablement cette chambre. Je m'approchai pour regarder : il y avait près d'elle une mare de sang. La porcelaine brisée avait entaillé cruellement sa chair, et l'os seulement avait **arrêté** cette coupure profonde.

Qui trouve un écho dans cet autre passage :

> C'est la vieille qui avait un jour **arrêté** le sang de sa main qui la soigne ; elle est sa confidente et je crois qu'elle l'a dénoncée pour de l'argent.

Le facteur 'coutume' peu intense ni très spécifique s'explique certainement par le passage :

> ...commença à s'endormir tout doucement ; le jour se mit à poindre, et je la laissai, comme de coutume avant le soleil, dormant d'un bon sommeil tranquille...

Qui lui même ne se rattache que très faiblement à cette classe.

Nous n'avons pas trouvé d'explications satisfaisantes pour le facteur 'commença'

Facteur	Intensité	Spécificité	Contribution
Midhat-pacha	0.138	0.599	9.953
vive	0.116	0.478	8.379
softas	0.115	0.582	8.287
bruit	0.101	0.68	7.29
cri	0.069	0.476	4.989
À	0.059	0.48	4.218
arrêté	0.049	0.495	3.559
voix	0.045	0.265	3.26
coutume	0.042	0.466	3.047
commença	0.041	0.66	2.939

5.3.7 **Classe 'yeux', 69 passages, numéro 2.**

Cette classe des yeux est celle des beaux yeux verts d'Aziyadé et plus généralement de ceux des femmes turques dont le reste du visage est dissimulé au regard et qui fût plus souvent associé à Shéhérazade qu'à Aziyadé[106]. C'est donc aussi la classe des bijoux, des toilettes et de la bague offerte à Loti par sa maîtresse qui tient une certaine importance dans le roman. La conjonction d'une manche, d'une référence à sa petite tête et du gland de soie de son fez fait apparaître un passage relatif au pauvre Achmet dans cette classe si féminine.

> ...bras agitait avec colère sa large manche blanche ; sa petite tête faisait danser furieusement le gland de soie de son fez...
>
> ...jeune femme qui avait ces yeux se leva, et montra jusqu'à la ceinture sa taille enveloppée d'un camail à la turque (féredjé) aux plis longs et rigides. Le camail était de soie verte, orné de broderies d'argent. Un voile blanc enveloppait soigneusement la tête, n'en laissant paraître que le front et les grands yeux. Les prunelles étaient bien vertes, de cette teinte vert de mer d'autrefois chantée par les poètes d'Orient...
>
> ...hauteurs d'Eyoub s'étalait la masse mouvante des dames turques. Tous ces corps de femmes, enveloppés chacun jusqu'aux pieds de pièces de soie de couleurs éclatantes, toutes ces têtes blanches cachées sous les plis des yachmaks d'où sortaient des yeux noirs, se confondaient sous les cyprès avec les pierres peintes et historiées des tombes. Cela était si coloré et si bizarre, qu'on eût dit moins une réalité qu'une composition fantastique...
>
> ...yachmak, très épais, était ramené sur ses yeux jusqu'à dérober tout son front ; à peine voyait-on, par l'ouverture du voile, rouler ses prunelles, si limpides et si mobiles ; son féredjé d'emprunt était d'une couleur foncée, d'une coupe sévère, que n'adoptent point d'ordinaire les femmes élégantes et jeunes. Et le vieil Abeddin lui-même ne l'eût point...
>
> ...Samuel reçoit dans sa barque les deux premiers de ces personnages, et s'éloigne sans mot dire. Je suis resté seul avec la femme au voile, aussi muette et immobile qu'un fantôme blanc ; j'ai pris les rames, et, en sens inverse, nous nous éloignons aussi dans la direction du large. Les yeux fixés sur elle, j'attends avec anxiété qu'elle fasse un mouvement ou un signe...
>
> ...soirée une toilette qui la rendait étrangement belle ; la richesse orientale de son costume contrastait maintenant avec l'aspect de notre demeure, redevenue sombre et misérable. Elle portait une de ces vestes à longues basques dont les femmes turques d'aujourd'hui ont presque perdu le modèle, une veste de soie violette semée de roses d'or. Un pantalon de soie jaune descendait jusqu'à ses chevilles, jusqu'à ses petits pieds chaussés de pantoufles dorées. Sa chemise en gaze de Brousse lamée d'argent, laissait échapper ses bras ronds, d'une teinte mate et ambrée, frottés d'essence de roses. Ses cheveux bruns étaient divisés en huit nattes, si épaisses, que deux d'entre elles auraient suffi au bonheur d'une merveilleuse de Paris ; ils s'étalaient à côté d'elle sur le divan, noués au bout par des rubans jaunes, et mêlés de fils d'or, à la manière des femmes arméniennes. Une masse d'autres petits cheveux plus courts et plus rebelles formaient nimbe autour de ses joues rondes, d'une pâleur chaude et dorée. Des teintes d'un ambre plus foncé entouraient ses paupières ; et ses sourcils, très rapprochés d'ordinaire, se rejoignaient ce soir-là avec une expression de profonde douleur...

[106]Et donc à la Perse plus qu'à la Turquie

> ...canots des escadres étaient partis quand je revins sur le quai ; les yeux verts m'avaient légèrement captivé, bien que le visage exquis caché par le voile blanc me fût encore inconnu ; j'étais repassé trois fois devant la mosquée aux cigognes, et l'heure s'en était allée sans que j'en eusse conscience...
>
> ...vint m'éveiller pour le quart ; il était minuit. Le timonier alluma une bougie dans ma chambre : je vis briller les dorures et les fleurs de soie...
>
> ...vision, et il semble qu'elle illumine les lieux par lesquels elle passe. On cherche des rayons autour de sa tête enfantine et sérieuse, et on en trouve en effet, quand la lumière tombe sur certains petits cheveux impalpables, rebelles à toutes les coiffures, qui entourent délicieusement ses joues et son front...
>
> ...habitais, tout au fond du Prince-of-Wales, un réduit blindé confinant avec la soute aux poudres. J'avais meublé d'une manière originale ce caveau, où ne pénétrait pas la lumière du soleil : sur les murailles de fer, une épaisse soie rouge à fleurs bizarres ; des faïences, des vieilleries redorées, des armes, brillant sur ce fond sombre...

Il n'y a pas de commentaires particuliers à faire sur ces facteurs si évidents.

Facteur	Intensité	Spécificité	Contribution
yeux	0.1	0.225	6.903
bague	0.071	0.311	4.879
roses	0.065	0.381	4.46
expression	0.055	0.328	3.819
bijoux	0.055	0.338	3.805
argent	0.054	0.443	3.748
femmes	0.054	0.218	3.741
camail	0.054	0.336	3.709
voile	0.047	0.553	3.209
cheveux	0.045	0.24	3.127

5.3.8 **Classe 'horreur', 54 passages, numéro 8.**

Cette classe est bien décrite par le très long passage qui arrive en première position où s'expriment les convictions royalistes du jeune Loti. Le thème est une fois encore centré autour des considérations morales, religieuses et politiques du héros :

> ...récit, circonstancié et agrémenté de descriptions, d'une amourette à la turque. Nous vous suivons, Georges et moi, à travers les méandres fantasmagoriques d'une grande fourmilière orientale. Nous restons la bouche béante en face des tableaux que vous nous tracez ; je songe à vos trois poignards, comme je songeais au bouclier d'Achille, si minutieusement chanté par Homère ! Et puis enfin, peut-être parce que vous avez reçu un grain de poussière dans l'oeil, peut-être parce que votre lampe s'est mise à fumer comme vous acheviez votre lettre, peut-être pour moins que cela, vous terminez en nous lançant la série des lieux communs édités au siècle dernier ! je crois vraiment que les lieux communs des frères ignorantins valent encore mieux que ceux du matérialisme, dont le résultat sera l'anéantissement de tout ce qui existe. On les acceptait au XVIIIe siècle, ces idées matérialistes : Dieu était un préjugé ; la morale était devenue l'intérêt bien entendu, la société un vaste champ d'exploitation pour l'homme habile. Tout cela séduisait beaucoup de gens par sa nouveauté et par la sanction qu'en recevaient les actes les plus immoraux. Heureuse époque où aucun frein ne vous retenait ; où l'on pouvait tout faire ; l'on pouvait rire de tout, même des choses les moins drôles, jusqu'au moment où tant de têtes tombèrent sous le couteau de la Révolution, que ceux qui conservèrent la leur commencèrent à réfléchir. Ensuite vint une époque de transition, où l'on vit apparaître une génération atteinte de phtisie morale, affligée de sensiblerie constitutionnelle, regrettant le passé qu'elle ne connaissait pas, maudissant le présent qu'elle ne comprenait pas, doutant de l'avenir qu'elle ne devinait pas. Une génération de romantiques, une génération de petits jeunes gens passant leur vie à rire, à pleurer, à prier, à blasphémer, modulant sur tous les tons leur insipide complainte pour en venir un beau jour à se faire sauter la cervelle...
>
> ...Samuel a peur des kédis (des chats). Le jour, les kédis lui inspirent des idées drôles ; il ne peut les regarder sans rire. La nuit, il devient très respectueux, et s'en tient à distance...
>
> ...grande dame, en passant le seuil de ma demeure, eut un mauvais rire qui me fit monter la colère au visage, et je ne fus pas loin de saisir son bras rond pour la retenir...
>
> ...Loti, j'étais presque une petite fille. Quand pour la première fois je t'ai vu, il n'y avait pas dix lunes que j'étais dans le harem d'Abeddin, et je ne m'ennuyais pas encore. Je me tenais dans mon appartement, assise sur mon divan, à fumer des cigarettes, ou du hachisch, à jouer aux cartes avec ma servante Emineh, ou à écouter des histoires très drôles du pays des hommes noirs, que Kadidja sait raconter parfaitement...
>
> ...ferme mes rideaux, j'allume ma lampe et mon feu : le décor change et mes idées aussi. Je continue ma lettre devant une flamme joyeuse, enveloppé dans un manteau de fourrure, les pieds sur un épais tapis de Turquie. Un instant je me prends pour un derviche, et cela m'amuse...
>
> ...pense aller bientôt à Jérusalem, où je tâcherai de ressaisir quelques bribes de foi. Pour l'instant, mes croyances religieuses et philosophiques, mes principes de morale, mes théories sociales, etc., sont représentés par cette grande personnalité...
>
> ...sorte de révélation semble alors se faire ; on dirait qu'on vient de naître une seconde fois, car dès lors on vit davantage, on fonctionne tout entier ; tout ce qu'il y a en nous d'idées, de sentiments, se réveille et s'avive comme la flamme du punch que l'on agite. (Littérature de l'avenir...

> ...point de visiteurs inattendus ou déplaisants. Si quelques Turcs me visitent discrètement quand je les y invite, mes amis ignorent absolument le chemin de ma demeure, et des treillages de frêne gardent si fidèlement mes fenêtres qu'à aucun moment du jour un regard curieux n'y saurait pénétrer...
>
> ...Dieu, il n'y a pas de morale, rien n'existe de tout ce qu'on nous a enseigné à respecter ; il y a une vie qui passe, à laquelle il est logique de demander le plus de jouissances possible, en attendant l'épouvante finale qui est la mort...
>
> ...imprudences, toutes les maladresses, entassées jour par jour pendant un mois, dans le but d'arriver à un résultat par lui-même impossible...

Ces facteurs sont ni spécifiques ni intenses et sont principalement agrégés autour du premier passage qui aurait mérité d'être découpé.

Facteur	Intensité	Spécificité	Contribution
horreur	0.072	0.204	3.894
génération	0.065	0.22	3.534
obligé	0.064	0.388	3.464
rideaux	0.06	0.304	3.227
rire	0.059	0.31	3.206
soldats	0.059	0.116	3.18
communs	0.058	0.302	3.144
époque	0.057	0.318	3.058
siècle	0.056	0.233	3.042
fenêtres	0.054	0.282	2.934

5.3.9 Classe 'LOTI', 18 passages, numéro 1.

On termine par une classe de péritexte épistolaire sans intérêt qui ne sera pas pris en considération pour l'indexation.

> ...WILLIAM BROWN, LIEUTENANT...
>
> ...LOTI A PLUMKETT...
>
> ...PLUMKETT A LOTI...
>
> ...LOTI, DE SA SOEUR...
>
> ...LOTI A SA SOEUR...
>
> ...LOTI A PLUMKETT...
>
> ...LOTI, DE SA SOEUR...
>
> ...LOTI A PLUMKETT...
>
> ...LOTI A SA SOEUR...
>
> ...LOTI A WILLIAM BROWN...

Facteur	Intensité	Spécificité	Contribution
LOTI	0.994	1	17.889
caïque	0.015	0.012	0.271
factionnaires	0.015	0.019	0.27
Phanar	0.012	0.014	0.209
personne	0.006	0.004	0.112
aimées	0.004	0.007	0.073
nuit	0.003	0.002	0.055
minarets	0	0	0.003
trouvaient	0	0	0
cours	0	0	0

5.4 Classification par isotopies

Nous voici arrivé à la dernière partie de cette analyse, notre but est maintenant de s'assurer que les isotopies suggérées par les classes de passages précédemment calculées permettent d'interpréter les documents initiaux, à savoir les chapitres du roman. Comme nous avons tenté de l'expliquer dans les sections 2 et 4, nous allons une fois encore avoir recours à la classification, l'idée étant que les documents regroupés sur la base de ces isotopies doivent pouvoir aisément s'interpréter.

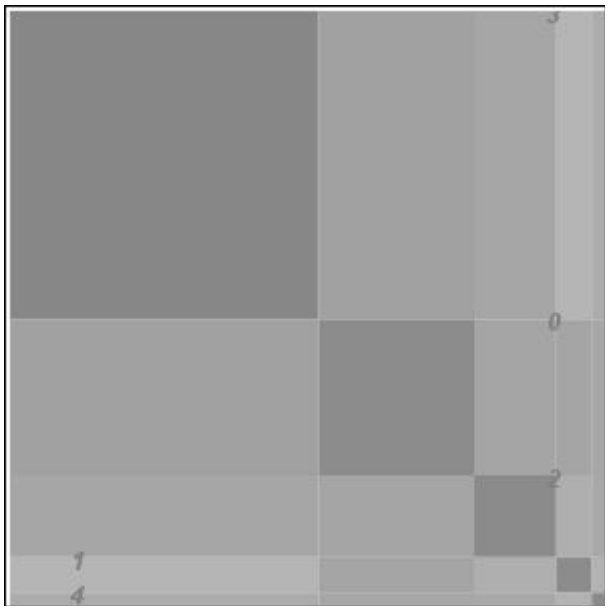

Figure 20 : Matrice de densité de la classification des chapitres en fonction des isotopies

La figure 20 montre la matrice de densité de la classification des chapitres en fonctions des isotopies suggérées par la classification précédente classification des passages. Le seuil de densité est de 0,75 car les isotopies[107] tendent à plus rapprocher les documents que les facteurs les termes[108]. Six classes ont été ainsi obtenues dont une réduite à un élément qui n'apparaît pas sur le graphique. Comme nous l'avions fait précédemment nous avons utilisé l'application BIRT pour générer automatiquement un rapport de classification : pour chaque classe nous présentons les cinq passages de meilleure contribution des cinq

[107] Nous les appellerons dorénavant isotopies pour ne pas trop alourdir le texte, nous sommes cependant conscient qu'il ne s'agit pas là d'isotopies mais leur manifestation au travers du prisme de nos traitements des chaînes de cractères

[108]Ce qui explique d'ailleurs l'aspect plus foncé de la matrice de densité.

documents de meilleure contribution ainsi que les cinq isotopies de meilleure contribution. De façon générale les classes obtenues correspondent aux principales classes de passages et portent d'ailleurs les mêmes noms[109], ce qui est assez normal d'ailleurs si l'on considère que les isotopies se déduisent naturellement des facteurs principaux des classes de passages. L'objet de cette dernière phase de l'analyse est principalement de vérifier la qualité des regroupements effectués, c'est pourquoi nous multiplions l'affichage des passages dans le rapport.

5.4.1 **Classe 'Loti', taille 79, numéro 3**

Comme on pouvait s'y attendre la classe la plus grande se constitue autour des personnages principaux du roman.

Document EYOUB À DEUX Chapitre XXVIII

Ce chapitre est consacré à la vielle amie d'Aziyadé, Béhidjé.

> Aziyadé m'exprime quelque idée neuve, quelque notion nette et profonde sur des choses qu'elle semblerait devoir ignorer absolument, et que je lui demande : « Qui t'a appris cela, ma chérie ? » Aziyadé répond : « C'est ma mère Béhidjé »
>
> Béhidjé-hanum passe ses journées à cet observatoire, étendue sur un fauteuil, et Aziyadé est souvent à ses pieds, Aziyadé attentive au moindre signe de sa vieille amie, et dévorant ses paroles
>
> belle encore, affirme Aziyadé, malgré ses quatre-vingts ans, « belle comme les beaux soirs »
>
> Aziyadé arrive le soir, l'imagination plus surexcitée que de coutume, je puis en toute sûreté lui dire : Tu as passé ta journée, ma chère petite amie, aux pieds de ta mère Béhidjé
>
> Béhidjé-hanum n'est point une mère pour Aziyadé. Tout au moins est-ce une mère imprudente, qui ne craint pas d'exalter terriblement la jeune imagination de son enfant

Document 1 SALONIQUE JOURNAL DE LOTI Chapitre XVII

Le premier entretien entre Aziyadé et Loti, Loti ne parlant pas encore le turc Samuel est l'interprète.

> heure pour Aziyadé de repartir, et, l'instant d'après, elle nous quitta
>
> Aziyadé avait dit à Samuel qu'il resterait cette nuit-là auprès de nous. Je la regardais faire avec étonnement : elle m'avait prié de m'asseoir entre elle et lui, et commençait à lui parler en langue turque

[109] Le nom d'une classe est donnée par l'étiquette de l'isotopie de meilleure contribution.

Aziyadé comprit, elle passa ses bras en tremblant autour de mon cou ; et nous nous penchâmes tous deux sur l'eau

traduisait Samuel, que son Dieu n'est pas le même que le tien, et qu'elle n'est pas bien sûre, d'après le Koran, que les femmes aient une âme comme les hommes ; elle pense que, quand tu seras parti, vous ne vous verrez jamais, même après que vous serez morts, et c'est pour cela qu'elle pleure. Maintenant, dit Samuel en riant, elle demande si tu veux te jeter dans la mer avec elle tout de suite ; et vous vous laisserez couler au fond en vous tenant serrés tous les deux ... Et moi, ensuite, je ramènerai la barque, et je dirai que je ne vous ai pas vus

pourvu qu'elle ne pleure plus ; partons tout de suite, ce sera fini

Document MANÉ, THÉCEL, PHARÈS Chapitre XXIX

Lettre déchirante d'Achmet à Loti.

lettre de Mytilène à Aziyadé par la vieille effendi ; elle l'a serrée dans sa robe, et n'a pas pu se la faire lire encore, parce qu'elle n'est pas sortie depuis ton départ

turc, écrit sous la dictée d'Achmet par un écrivain public de la place d'Emin-Ounou à Stamboul, et adressé à Loti, à Brightbury

On ne m'a pas encore appelé pour la guerre, à cause de mon père, qui est très vieux ; cependant je pense qu'on m'appellera

Aziyadé te fait dire qu'elle ne vit pas sans toi ; qu'elle ne voit pas le moment de ton retour à Constantinople ; qu'elle ne croit pas qu'elle puisse jamais voir tes yeux face à face et qu'il lui semble qu'il n'y a plus de soleil

vieux Abeddin a soupçonné et tout deviné, car nous avions été sans prudence pendant les derniers jours. Il ne lui a pas fait de reproches, a dit effendi, et ne l'a pas chassée, parce qu'il l'aimait beaucoup. Seulement, il n'entre plus dans son appartement ; il ne prend plus garde à elle et il ne lui parle plus. Les autres femmes aussi du harem l'ont abandonnée, excepté Fenzilé-hanum, qui est allée pour elle consulter

Document MANÉ, THÉCEL, PHARÈS Chapitre XV

La séparation, Loti doit rembarquer sur le Deerhound.

Aziyadé a voulu venir me conduire ; elle a juré d'être sage ; elle est à cette dernière heure d'un calme inattendu

voiture est là qui stationne, commandée par Achmet, pour ramener Aziyadé dans sa demeure

fini sans retour ! si je reviens jamais comme je l'ai juré, les années auront secoué sur tout cela leur cendre, ou bien j'aurai creusé l'abîme

Foundoucli est encore un coin de la vieille Turquie, qui semble détaché du fond de Stamboul : petite place dallée, au bord de la mer, antique mosquée à croissant d'or, entourée de tombes de derviches, et de sombres retraites d'oulémas

canot est rendu : elle et Achmet se retirent dans un angle obscur de la mosquée ; je pars, et je les perds de vue

Document EYOUB À DEUX Chapitre XXII

Loti à Eyoub, intégré aux habitants du quartier rêvant de devenir pour toujours Arif auprès d'Aziyadé.

> idée, qui est d'Aziyadé, se présente à mon esprit par instants sous des aspects étrangement
>
> possible, après tout, et je serais là moins malheureux qu'ailleurs. Je te jure, Aziyadé, dis-je, que je laisserais tout sans regret, ma position, mon nom et mon pays. Mes amis ... je n'en ai pas et je m'en moque ! Mais, vois-tu, j'ai une vieille mère
>
> Aziyadé ne dit plus rien pour me retenir, bien qu'elle ait compris peut-être que cela ne serait pas tout à fait impossible ; mais elle sent par intuition ce que cela doit être qu'une vieille mère, elle, la pauvre petite qui n'en a jamais eu ; et les idées qu'elle a sur la générosité et le sacrifice ont plus de prix chez elle que chez d'autres, parce qu'elles lui sont venues toutes seules, et que personne ne s'est inquiété de les lui donner
>
> café turc, chez le cafedji Suleïman, on élargit le cercle autour du feu, quand j'arrive le soir, avec Samuel et Achmet. Je donne la main à tous les assistants, et je m'assieds pour écouter le conteur des veillées d'hiver (les longues histoires qui durent huit jours, et où figurent les djinns et les génies). Les heures passent là sans fatigue et sans remords ; je me trouve à l'aise au milieu d'eux, et nullement
>
> Être batelier en veste dorée, quelque part au sud de la Turquie, là où le ciel est toujours pur et le soleil toujours chaud

Isotopie	Discrimination	Intensité	Spécificité	Contribution
Loti	1765.5	0.439	0.773	34.704
Achmet	1336.356	0.388	0.683	30.66
Allah	1213.777	0.287	0.505	22.688
Aziyadé	605.832	0.162	0.285	12.774
Samuel	879.305	0.125	0.22	9.87

5.4.2 Classe 'minarets', taille 40, numéro 0

Paysages et impressions d'orient.

Document SOLITUDE Chapitre V

Première installation de Loti à Stamboul dans le quartier du Taxim, longues descriptions de Stamboul, c'est dans se chapitre où se promenant dans un cimetière il manque de se faire violer par un veilleur de nuit.

> maison était située en un point retiré de Péra, dominant de haut la Corne d'or et le panorama lointain de la ville turque ; la splendeur de l'été donnait du charme à cette habitation. En travaillant la langue de l'islam devant ma grande fenêtre ouverte, je

> planais sur le vieux Stamboul baigné de soleil. Tout au fond, dans un bois de cyprès, apparaissait Eyoub, où il eût été doux d'aller avec elle cacher son existence, point mystérieux et ignoré où notre vie eût trouvé un cadre étrange et charmant
>
> choeur passa, et se perdit dans l'éloignement. Par ma fenêtre grande ouverte, on ne voyait que la vapeur du matin, le vide immense du ciel ; et puis, tout en haut, quelque chose se dessina en rose, un dôme et des minarets ; la silhouette de la ville turque s'esquissa peu à peu, comme suspendue dans l'air ... Alors, je me rappelai que j'étais à Stamboul, et qu'elle avait juré d'y venir
>
> ciel blanchissait à l'orient quand je regagnai ma chambre. La pâle débauche me retenait souvent par les rues jusqu'à ces heures matinales. À peine étais-je endormi, qu'une suave musique vint m'éveiller ; une vieille aubade d'autrefois, une mélodie gaie et orientale, fraîche comme l'aube du jour, des voix humaines accompagnées de harpes et de guitares
>
> mois, je demeurai à Péra, songeant aux moyens d'exécuter ce projet impossible, aller habiter avec elle sur l'autre rive de la Corne d'or, vivre de la vie musulmane qui était sa vie, la posséder des jours entiers, comprendre et pénétrer ses pensées, lire au fond de son coeur des choses fraîches et sauvages à peine soupçonnées dans nos nuits de Salonique
>
> souvent parcouru la nuit ces cimetières, et j'y ai fait plus d'une fâcheuse rencontre

Document EYOUB À DEUX Chapitre XVII

Surpris par un orage Loti se réfugie dans un café turc. C'est précisément au moment de l'annonce de la constitution ; bousculade dans la rue, superbe description de la promulgation à la Sublime Porte sous la pluie et la grêle.

> vieux turbans dans ce café, et de vieilles barbes blanches. Des vieillards (des hadj-baba) étaient assis, occupés à lire les feuilles publiques, ou à regarder à travers les vitres enfumées les passants qui couraient sous la pluie. Des dames turques, surprises par l'ondée, fuyaient de toute la vitesse que leur permettaient leurs babouches et leurs socques à patins. C'était dans la rue une grande confusion et dans le public, une grande bousculade ; l'eau tombait à torrents
>
> ciel était noir et tourmenté ; pluie et grêle tombaient abondamment et inondaient tout ce monde. Sous ces cataractes, on donnait au peuple lecture de la charte, et les vieilles murailles crénelées du sérail, qui fermaient le tableau, semblaient s'étonner beaucoup d'entendre proférer en plein Stamboul ces paroles
>
> À la même heure, à l'autre bout de Constantinople, au palais de l'Amirauté, s'étaient réunis les membres de la conférence internationale
>
> entré, pour laisser passer une averse, dans un café turc près de la mosquée
>
> voï, voï, Allah !... et nos femmes ne couraient point en voile de gaze ; et les croyants disaient plus régulièrement leurs prières

Document SOLITUDE Chapitre VII

Première apparition d'Achmet dans le roman, longues promenades à deux dans Eyoub occasions de belles descriptions.

À mes pieds, les vieilles cases arméniennes sont obscures et endormies ; j'ai vue sur un très profond ravin, au bas duquel un bois de cyprès séculaires forme une masse absolument noire ; ces arbres tristes ombragent d'antiques sépultures de musulmans ; ils exhalent dans la nuit des parfums balsamiques. L'immense horizon est tranquille et pur ; je domine de haut tout ce pays. Au-dessus des cyprès, une nappe brillante, c'est la Corne d'or ; au-dessus encore, tout en haut, la silhouette d'une ville orientale, c'est Stamboul. Les minarets, les hautes coupoles des mosquées se découpent sur un ciel très étoilé où un mince croissant de lune est suspendu ; l'horizon est tout frangé de tours et minarets, légèrement dessinés en silhouettes bleuâtres sur la teinte pâle de la nuit. Les grands dômes superposés des mosquées montent en teintes vagues jusqu'à la lune, et produisent sur l'imagination l'impression du gigantesque

croissant s'abaisse lentement derrière Stamboul, derrière les dômes de la Suleïmanieh. Dans cette grande ville, je suis étranger et inconnu. Mon pauvre Samuel était le seul qui y sût mon nom et mon existence, et sincèrement je commençais à l'aimer

palais là-bas, le Seraskierat, il se passe à l'heure qu'il est une sombre comédie ; les grands pachas y sont réunis pour déposer le sultan Mourad ; demain, c'est Abd-Ul-Hamid qui l'aura remplacé. Ce sultan pour l'avènement duquel nous avons fait si grande fête, il y a trois mois, et qu'on servait aujourd'hui encore comme un dieu, on l'étrangle peut-être cette nuit dans quelque coin du sérail

comptais que mon pauvre Samuel serait auprès de moi ce soir, et sans doute je ne le reverrai jamais. J'en ai le coeur serré et ma solitude me pèse. Il y a huit jours, je l'avais laissé partir pour gagner quelque argent, sur un navire qui s'en allait à Salonique. Les trois bateaux qui pouvaient me le ramener sont revenus sans lui, le dernier ce soir, et personne à bord n'en avait entendu parler

chantent dans les cyprès, avec la même voix que celles de mon pays ; j'aime ce bruit d'été qui me ramène aux bois du Yorkshire, aux beaux soirs de mon enfance, passée sous les arbres, là-bas, dans le jardin de Brightbury

Document EYOUB À DEUX Chapitre XLII

La rencontre de trois femmes aguicheuses et la cavalcade dans les rues de Stamboul derrière leur voiture est l'occasion de superbes descriptions de la ville.

passâmes Dolma-Bagtché, Sali-Bazar, Top-Hané, le bruyant quartier de Galata, et puis le pont de Stamboul, le triste Phanar et le noir Balate. A Eyoub enfin, dans une vieille rue turque, devant un Conak antique, à la mine opulente et sombre, les trois femmes s'arrêtèrent

certaine après-midi de janvier, le ciel sur Constantinople était uniformément sombre ; un vent froid chassait une fine pluie d'hiver, et le jour était pâle comme un jour britannique

tenaient fort mal, à la façon de toutes les hanums de grande maison qui ne craignent guère d'adresser aux Européens dans les rues les regards

soldats et des eunuques noirs gardaient ces entrées défendues. Les styles de ces portiques semblait indiquer lui-même que le seuil en était dangereux à franchir ; les colonnes et les frises de marbre, fouillées à jour dans le goût arabe, étaient couvertes de dessins étranges et d'enroulements mystérieux

suivais à cheval une longue et large route, bordée d'interminables murailles de trente pieds de haut, droites, polies, inaccessibles comme des murailles de prison

Document 1 SALONIQUE JOURNAL DE LOTI Chapitre XII

L'emménagement du harem de Aziyadé dans un yali[110] sur le chemin de Monastir est l'occasion de descriptions des paysages avoisinants.

> journées à errer sur ce chemin de Monastir. C'était une campagne nue et triste, où l'oeil s'étendait à perte de vue sur des cimetières antiques ; des tombes de marbre en ruine, dont le lichen rongeait les inscriptions mystérieuses ; des champs plantés de menhirs de granit ; des sépultures grecques, byzantines, musulmanes, couvraient ce vieux sol de Macédoine où les grands peuples du passé ont laissé leur poussière. De loin en loin, la silhouette aiguë d'un cyprès, ou un platane immense, abritant des bergers albanais et des chèvres ; sur la terre aride, de larges fleurs lilas pâle, répandant une douce odeur de chèvrefeuille, sous un soleil déjà brûlant. Les moindres détails de ce pays sont restés dans ma mémoire
>
> nuit, c'était un calme tiède, inaltérable, un silence mêlé de bruits de cigales, un air pur rempli de parfums d'été ; la mer immobile, le ciel aussi brillant qu'autrefois dans mes nuits des tropiques
>
> jour je descendais en armes. Par grosse mer, toujours, un canot me jetait sur les quais, au milieu de la foule des bateliers et des pêcheurs ; et Samuel, placé comme par hasard sur mon passage, recevait par signes mes ordres pour la nuit
>
> certaine heure du matin, avant le jour, je pouvais, avec mille dangers, rejoindre ma corvette par un moyen convenu avec les officiers de garde
>
> venue habiter avec les trois autres femmes de son maître un yali de campagne, dans un bois, sur le chemin de Monastir

Isotopie	Discrimination	Intensité	Spécificité	Contribution
minarets	1177.828	0.294	0.262	11.773
pays	1168.555	0.219	0.195	8.769
mer	736.484	0.182	0.162	7.281
Stamboul	758.344	0.181	0.161	7.232
Loti	1765.5	0.162	0.145	6.488

5.4.3 Classe 'sens', taille 21, numéro 2

Les sens, la débauche, la posture cynique.

Document EYOUB À DEUX Chapitre XL

Lettre de Plumckett à Loti, pleines de considérations sur l'amour, l'amitié. Passage vraiment difficile à lire, confus et brouillon, référence à « l'odalisque Aziyadé et autre cocottes turques ».

[110] Belle demeure au bord de mer.

Femme ? Bien souvent cela tient uniquement à ce que la courbe de son nez, l'arc de ses sourcils, l'ovale de son visage, que sais-je ? ont ce je ne sais quoi auquel correspond en vous un autre je ne sais quoi qui fait le diable à quatre dans votre imagination. Ne vous récriez pas ! la moitié du temps, votre amour

femme un charme moral, une délicatesse de sentiment, une élévation de caractère qui sont la vraie cause de votre amour ... Hélas ! gardez-vous bien de confondre ce qui est en elle et ce qui est en vous. Toutes nos illusions viennent de là : attribuer ce qui est en nous et nulle part ailleurs à ce qui nous plaît. Faire une châsse à la femme que l'on aime et prendre son ami pour un homme de génie

pénétrer de cette idée en laquelle j'ai foi : il n'y a pas de douleur morale qui n'ait son remède. C'est à notre raison de le trouver et de l'appliquer suivant la nature du mal et le tempérament du sujet

heureux, et tout ce qui est antérieur à ce bonheur disparaît dans une sorte de nuit. Il semble qu'on était dans les limbes ; on vivait, relativement à la vie actuelle, comme l'enfant en bas âge par rapport au jeune homme. Les sentiments par lesquels on passe lorsque l'on est amoureux, on ne peut les décrire qu'au moment même où on les éprouve, et certes, je ne ressens rien de pareil en ce moment-ci. Et pourtant, tenez, sapristi ! je m'emballe en remuant toutes ces idées-là, je m'exalte, je perds la tête, je ne sais plus où j'en suis !... Quelle bonne chose d'aimer et d'être aimé ! savoir qu'une nature d'élite a compris la vôtre ; que quelqu'un rapporte toutes ses pensées, tous ses actes à vous ; que vous êtes un centre, un but, en vue duquel une organisation aussi délicatement compliquée que la vôtre, vit, pense et agit ! Voilà qui nous rend forts ; voilà qui peut faire des hommes de génie

amitié, qui est un sentiment plus sévère, plus solidement assis, puisqu'il repose sur tout ce qu'il y a de plus élevé en nous, la partie purement intellectuelle de nous-même. Quel bonheur de pouvoir dire tout ce que l'on sent à quelqu'un qui vous comprend jusqu'au bout et non pas seulement jusqu'à un certain point, à quelqu'un qui achève votre pensée avec le même mot qui était sur vos lèvres, dont la réplique fait jaillir de chez vous un torrent de conceptions, un flot d'idées. Un demi-mot de votre ami vous en dit plus que bien des phrases, car vous êtes habitué à penser avec lui. Vous comprenez tous les sentiments qui l'animent et il le sait. Vous êtes deux intelligences qui s'ajoutent et se complètent

Document EYOUB À DEUX Chapitre XLVI

Trahison avortée de Loti avec l'une des trois belles aguicheuses rencontrées au chapitre XLII, dans Eyoub à deux, Séniha.

ardente volupté se pâmait dans le sourire de cette bouche, dans le mouvement lent de ces yeux noirs, à moitié cachés sous la frange de leurs cils. J'en avais rarement vu de plus belle, là, près de moi, attendant mon bon plaisir, dans la tiède solitude d'une chambre parfumée ; et cependant il se livrait en moi-même une lutte inattendue ; mes sens se débattaient contre ce quelque chose de moins défini qu'on est convenu d'appeler l'âme, et l'âme se débattait contre les sens. À ce moment, j'adorais la chère petite que j'avais chassée ; mon coeur débordait pour elle de tendresse et de remords. La belle créature assise près de moi m'inspirait plus de dégoût que d'amour ; je l'avais désirée, elle était venue ; il ne tenait plus qu'à moi de l'avoir ; je n'en demandais pas davantage et sa présence m'était odieuse

lendemain soir, ma case était parée et parfumée, pour recevoir la grande dame qui avait désiré faire, en tout bien tout honneur, une visite à mon logis solitaire. La belle Séniha arriva très mystérieusement sur le coup de huit heures, heure indue pour Stamboul

femme à précautions : un aimable eunuque, habitué sans doute aux escapades de sa maîtresse, se tenait, à toute éventualité, près de la porte de ma maison

enleva son voile et le féredjé de laine grise qui, par prudence, la couvrait comme une femme du peuple, et laissa tomber la traîne d'une toilette française dont la vue ne me charma pas. Cette toilette, d'un goût douteux, plus coûteuse que moderne, allait mal à Séniha, qui s'en aperçut. Ayant manqué son effet, elle s'assit cependant avec aisance et parla avec volubilité. Sa voix était sans charme et ses yeux se promenaient avec curiosité sur ma chambre, dont elle louait très fort le bon air et l'originalité. Elle insistait surtout sur l'étrangeté de ma vie, et me posait sans réserve une foule de questions auxquelles j'évitais de répondre

grande dame, en passant le seuil de ma demeure, eut un mauvais rire qui me fit monter la colère au visage, et je ne fus pas loin de saisir son bras rond pour la retenir

Document SOLITUDE Chapitre XXVI

Lettre de sa soeur à Loti, pleine de lamentations et de reproches.

prie à toute heure, bien-aimé ; jamais ta pensée ne m'avait tant rempli le coeur ... Ne serait-ce que dans dix ans, dans vingt ans, je sais que tu croiras un jour. Peut-être ne le saurai-je jamais, peut-être mourrai-je bientôt, mais j'espérerai et je prierai

chair. Tu le ferais donc, puisque tu le dis ; tu le ferais d'un visage froid, d'un coeur sec, puisque tu te persuades suivre un fil fatal et maudit, puisque je ne suis plus rien dans ton existence ... Ta vie est ma vie, il y a un recoin de moi-même où personne n'est ... c'est ta place à toi, et quand tu me quitteras, elle sera vide

perdu mon frère, je suis prévenue affaire de temps, de quelques mois peut-être, il est perdu pour le temps, et l'éternité, déjà mort de mille morts. Et tout s'effondre, et tout se brise. Le voilà, l'enfant chéri qui plonge dans un abîme sans fond, l'abîme des abîmes ! Il souffre, l'air lui manque, la lumière, le soleil ; mais il est sans force ; ses yeux restent attachés au fond, à ses pieds ; il ne relève plus sa tête, il ne peut plus, le prince des ténèbres le lui défend ... Quelquefois pourtant il veut résister. Il entend une voix lointaine, celle qui a bercé son enfance ; mais le prince lui dit : « Mensonge, vanité, folie ! » et le pauvre enfant, lié, garrotté, au fond de son abîme, sanglant, éperdu, ayant appris de son maître à appeler le bien mal, et le mal bien, que fait-il ?... il sourit

malheureux, troublé, ballotté, confiant, plus je t'aime. Ah ! mon bien-aimé frère, mon chéri, si tu voulais revenir à la vie ! si Dieu voulait ! si tu voyais la désolation de mon coeur, si tu sentais la chaleur de mes prières

pauvre âme travaillée et chargée, même pas le sourire moqueur de Satan

Document EYOUB À DEUX Chapitre XI

Autre lettre de sa soeur à Loti, un peu le même genre que la précédente.

Cher frère, tu es à moi, tu es à Dieu, tu es à nous. Je le sens, un jour, bientôt peut-être, tu reprendras courage, confiance et espoir. Tu verras combien cette erreur est douce et délicieuse, précieuse et bienfaisante. Oh ! mensonge mille fois béni, que celui qui me fait vivre et me fera mourir, sans regrets, et sans frayeur ! qui mène le monde depuis des siècles, qui a fait les martyrs, qui fait les grands peuples, qui change le deuil en allégresse, qui crie partout : « Amour, liberté et charité »

lettre ! C'est tout ce que je puis demander pour le moment, et je puis dire comme la Sunamite voyant son fils mort :

pauvre coeur est plein de contradictions, ainsi que tous les cœurs troublés qui flottent sans boussole. Tu jettes des cris de désespoir, tu dis que tout t'échappe, tu en appelles passionnément à ma tendresse, et, quand je t'en assure moi-même, avec passion, je trouve que tu oublies les absents, et que tu es si heureux dans ce coin de l'Orient que tu voudrais toujours voir durer cet Éden. Mais voilà, moi, c'est permanent, immuable ; tu le retrouveras, quand ces douces folies seront oubliées pour faire place à d'autres, et peut-être en feras-tu plus tard plus de cas que tu ne penses

Chère frère

Brightbury, décembre 1876

Document EYOUB À DEUX Chapitre V

Lettre de Loti à sa soeur, promesse de ne pas devenir « un vieux débauché ».

parlé d'Aziyadé, je puis bien te dire qu'elle est arrivée. Elle m'aime de toute son âme, et ne pense pas que je puisse me décider à la quitter jamais. Samuel est revenu aussi ; tous deux m'entourent de tant d'amour, que j'oublie le passé et les ingrats, un peu aussi les absents

dur et ingrat de ne pas t'écrire plus tôt. Je t'ai fait beaucoup de mal, tu le dis, et je le crois. Malheureusement, tout ce que j'ai écrit, je le pensais, et je le pense encore ; je ne puis rien maintenant contre ce mal que je t'ai fait ; j'ai eu tort seulement de te laisser voir au fond de mon coeur, mais tu l'avais voulu

retour, je ferai un suprême effort. Quand je serai au milieu de vous, mes idées changeront ; si vous me choisissez une jeune fille que vous aimiez, je tâcherai de l'aimer, et de me fixer, pour l'amour

avis, je ne connais pas de chose plus repoussante qu'un vieux débauché qui s'en va de fatigue et d'usure, et qu'on abandonne. Mais je ne serai point cet objet-là : quand je ne serai plus bien portant, ni jeune, ni aimé, c'est alors que je disparaîtrai

intéresser à quelque chose, dis-tu ? à quelque chose de bon et d'honnête, et le prendre à coeur. Mais j'ai ma pauvre chère vieille mère ; elle est aujourd'hui un but dans ma vie, le but que je me suis donné à moi-même. Pour elle, je me compose une certaine gaieté, un certain courage : pour elle, je maintiens le côté positif et raisonnable de mon existence, je reste Loti, officier de marine

Isotopie	Discrimination	Intensité	Spécificité	Contribution
sens	404.766	0.187	0.088	3.931
âme	348.566	0.18	0.084	3.787
femme	436.043	0.178	0.083	3.739
temps	335.117	0.17	0.079	3.562
abîme	763.043	0.158	0.074	3.324

5.4.4 **Classe 'Eyoub', taille 9, numéro 1**

L'opposition de l'occident et de l'orient, Eyoub le sombre et terrible coeur musulman de l'empire Ottoman[111], Brightbury la sage, la froide, l'ennuyeuse et puissante Angleterre.

Document MANÉ, THÉCEL, PHARÈS Chapitre XXXII

Un ratage du système du au « site funèbre » qui rappelle Eyoub, en fait ce chapitre réduit à un paragraphe clôt la description poignante et lugubre de l'enterrement d'une enfant grecque durant lequel chante joyeusement une mésange.

> oiseau était drôle de se trouver si heureux de vivre, et d'être si gai au milieu de ce site funèbre

Document Azraël Chapitre V

Finalement sa bien aimée et son cher ami Achmet morts durant de son séjour à Brightbury, Loti rongé par le remord s'engage dans l'armée turque et meurt sous son nom turc d'Arif lors de la seconde bataille de Kars qui scelle la victoire à venir de la Sainte Russie sur l'empire Ottoman, la montée en puissance des jeunes turcs, qui se terminera cinquante plus tard par le démembrement de l'empire. Le véritable Loti est resté fidèle jusqu'à son dernier souffle et malgré ses convictions royalistes de jeune aspirant de marine il recevra dans ses vieux jours une délégation officielle de la jeune république turque reconnaissante.

> lit dans le Djerideï-havadis, journal de Stamboul : « Parmi les morts de la dernière bataille de Kars, on a retrouvé le corps d'un jeune officier de la marine anglaise, récemment engagé au service de la Turquie sous le nom de Arif-Ussam effendi. » Il a été inhumé parmi les braves défenseurs de l'islam (que Mahomet protège !), aux pieds du Kizil-Tépé, dans les plaines

Document 1 SALONIQUE JOURNAL DE LOTI Chapitre II

Tout début du roman, les puissances occidentales impose le châtiment des responsables de l'assassinat de consuls européens.

> exécution terminée, les soldats se retirèrent et les morts restèrent jusqu'à la tombée du jour exposés aux yeux du peuple. Les six cadavres, debout sur leurs pieds, firent, jusqu'au soir, la hideuse grimace de la mort au beau soleil de Turquie, au milieu de promeneurs indifférents et de groupes silencieux de jeunes femmes

[111] C'est en tous cas la vison de Loti.

Document 1 SALONIQUE JOURNAL DE LOTI Chapitre III

Toujours ce thème du début du roman de l'intervention occidentale.

> européennes avaient envoyé sur rade de Salonique d'imposants cuirassés. L'Angleterre s'y était une des premières fait représenter, et c'est ainsi que j'y étais venu moi-même, sur l'une des corvettes de Sa Majesté
>
> France et d'Allemagne avaient exigé ces exécutions d'ensemble, comme réparation de ce massacre des consuls qui fit du bruit en Europe au début de la crise orientale

Document 1 SALONIQUE JOURNAL DE LOTI Chapitre XVI

Découverte de l'orient...

> Salonique, juin 1876
>
> soir, c'était pour les yeux un enchantement d'un autre genre : tout était rose ou doré. L'Olympe avait des teintes de braise ou de métal en fusion, et se réfléchissait dans une mer unie comme une glace. Aucune vapeur dans l'air : il semblait qu'il n'y avait plus d'atmosphère et que les montagnes se découpaient dans le vide, tant leurs arêtes les plus lointaines étaient nettes
>
> bonheur de faire à Salonique ces corvées matinales qui vous mettaient à terre avant le lever du soleil. L'air était si léger, la fraîcheur si délicieuse, qu'on n'avait aucune peine à vivre ; on était comme pénétré de bien-être. Quelques Turcs commençaient à circuler, vêtus de robes rouges, vertes ou orange, sous les rues voûtées des bazars, à peine éclairées encore d'une demi-lueur transparente
>
> souvent assis le soir sur les quais où se portait la foule, devant cette baie tranquille. Les orgues de Barbarie d'Orient y jouaient leurs airs bizarres, accompagnés de clochettes et de chapeaux chinois ; les cafedjis encombraient la voie publique de leurs petites tables toujours garnies, et ne suffisaient plus à servir les narguilhés, les skiros, le lokoum et le raki
>
> Samuel était heureux et fier quand nous l'invitions à notre table. Il rôdait alentour, pour me transmettre par signes convenus quelque rendez-vous d'Aziyadé, et je tremblais d'impatience en songeant à la nuit qui allait venir

Isotopie	Discrimination	Intensité	Spécificité	Contribution
Eyoub	1179.098	0.444	0.089	3.996
Salonique	401.848	0.269	0.054	2.42
Brightbury	76.051	0.265	0.053	2.384
caïque	270.445	0.246	0.049	2.215
nuit	1032.484	0.153	0.031	1.374

5.4.5 **Classe 'vieille', taille 4, numéro 4**

Nous avons vu que la classe de passages correspondante s'analysait difficilement. Ici les choses sont plus claires peut-être.

Document SOLITUDE Chapitre XXVII

Ce chapitre est celui de l'épisode du vieux Kaïroullah...

> Marketo, dit-il, ayez pitié de moi ! Je demeure très loin et on croit que j'ai de l'or. Mieux vaudrait me tuer de votre main que me mettre à la porte à pareille heure. Laissez-moi dormir dans un coin de votre maison, et, avant le jour, je vous jure de partir
>
> courage pour mettre dehors ce vieillard, qui y fût mort de froid et de peur, en admettant qu'on ne l'eût point assassiné. Je me contentai de lui assigner un coin de ma maison, où il resta accroupi toute une nuit glaciale, pelotonné comme un vieux cloporte dans sa pelisse râpée. Je l'entendais trembler ; une toux profonde sortait de sa poitrine comme un râle ; et j'en eus tant de pitié, que je me levai encore pour lui jeter un tapis qui lui servît de couverture
>
> eus chassé tout ce monde comme une troupe de bêtes galeuses, je vis de nouveau paraître la tête allongée du vieux Kaïroullah, soulevant sans bruit la draperie de ma porte
>
> heures, une nuit d'hiver, et le quartier d'Eyoub était aussi noir et silencieux qu'un tombeau
>
> vieux Kaïroullah était assis devant moi par terre. Il était ramassé sur lui-même, comme un insecte malfaisant et immonde ; son crâne chauve et pointu luisait à la lueur de ma lampe

Document SOLITUDE Chapitre XIX

Installation à Eyoub, les débuts sont difficiles dans la vieille case...

> présence de cette case vide, de ces murailles nues, de ces fenêtres disjointes et de ces portes sans serrures. C'était si loin d'ailleurs, si loin du Deerhound, et si peu pratique

Document SOLITUDE Chapitre XXIV

Loti à soeur et le thème de la vielle mère.

> désiré me marier, je te l'avais dit ; je t'avais confié le soin de chercher une jeune fille qui fût digne de notre toit de famille et de notre vieille mère. Je te prie de n'y plus songer : je rendrais malheureuse la femme que j'épouserais, je préfère continuer une vie de plaisirs
>
> Tant que je conserverai ma chère vieille mère, je resterai en apparence ce que je suis aujourd'hui. Quand elle n'y sera plus, j'irai te dire adieu, et puis je disparaîtrai sans laisser trace de moi-même

écris dans ma triste case d'Eyoub ; à part un petit garçon nommé Yousouf, que même j'habitue à obéir par signes pour m'épargner l'ennui de parler, je passe chez moi de longues heures sans adresser la parole à âme qui vive

croyais à l'affection de personne ; cela est vrai. J'ai quelques amis qui m'en témoignent beaucoup, mais je n'y crois pas. Samuel, qui vient de me quitter, est peut-être encore de tous celui qui tient le plus à moi. Je ne me fais pas d'illusion cependant : c'est de sa part un grand enthousiasme d'enfant. Un beau jour, tout s'en ira en fumée, et je me retrouverai

ouvrir mon coeur devient de plus en plus difficile, parce que chaque jour ton point de vue et le mien s'éloignent davantage. L'idée chrétienne était restée longtemps flottante dans mon imagination alors même que je ne croyais plus ; elle avait un charme vague et consolant. Aujourd'hui, ce prestige est absolument tombé ; je ne connais rien de si vain, de si mensonger, de si inadmissible

Document EYOUB À DEUX Chapitre LXVII

Ce chapitre débute par une vieille poésie orientale, sur le thème de la joie éphémère et du temps qui passe.

saison de la joie et du plaisir : la saison vernale est arrivée. « Ne fais pas de prière avec moi, ô prêtre ; cela a son propre temps »

printemps, les amandiers fleurissent, et moi, je vois avec terreur, chaque saison qui m'entraîne plus avant dans la nuit, chaque année qui m'approche du gouffre ... Où vais-je, mon Dieu ?... Qu'y a-t-il après ? et qui sera près de moi quand il faudra boire la sombre coupe

Qui sait, quand la belle saison finira, lequel de nous sera encore envie ? « Soyez gais, soyez pleins de joie, car la saison du printemps passe vite, elle ne durera pas. » Écoutez la chanson du rossignol : la saison vernale s'approche. « Le printemps a déployé un berceau de joie dans chaque bosquet. » Où l'amandier répand ses fleurs argentées. « Soyez gais, soyez pleins de joie, car la saison du printemps passe vite, elle ne durera pas » (Extrait d'une vieille poésie orientale)

Isotopie	Discrimination	Intensité	Spécificité	Contribution
vieille	383.254	0.256	0.023	1.026
tête	392.004	0.21	0.019	0.841
Arif	167.684	0.208	0.019	0.831
mère	323.906	0.189	0.017	0.755
porte	176.188	0.176	0.016	0.705

5.4.6 Classe 'dévouement', taille 1, numéro 5

Un autre ratage du système dans cette classe marginale où l'isotopie 'dévouement' est déclenchée par le mot 'capables' mais ce n'est pas de dévouement que ces grandes dames sont capables.

Document EYOUB À DEUX Chapitre XLIII

> femmes turques, les grandes dames surtout, font très bon marché de la fidélité qu'elles doivent à leurs époux. Les farouches surveillances de certains hommes, et la terreur du châtiment sont indispensables pour les retenir. Toujours oisives, dévorées d'ennui, physiquement obsédées de la solitude des harems, elles sont capables de se livrer au premier venu, au domestique qui leur tombe sous la patte, ou au batelier qui les promène, s'il est beau et s'il leur plaît. Toutes sont fort curieuses des jeunes gens européens, et ceux-ci en profiteraient quelquefois s'ils les avaient, s'ils l'osaient, ou si plutôt ils étaient placés dans des conditions favorables pour le tenter. Ma position à Stamboul, ma connaissance de la langue et des usages turcs, ma porte isolée tournant sans bruit sur ses vieilles ferrures, étaient choses fort propices à ces sortes d'entreprises ; et ma maison eût pu devenir sans doute, si je l'avais désiré, le rendez-vous des belles désoeuvrées des harems

Isotopie	Discrimination	Intensité	Spécificité	Contribution
dévouement	166.602	0.285	0.006	0.285
ami	278.16	0.271	0.006	0.271
pauvres	80.523	0.247	0.006	0.247
capables	115.82	0.237	0.005	0.237
monde	123.41	0.209	0.005	0.209

5.5 Discussion

Nous savons que l'art d'interpréter est trompeur, et l'on pourra peut-être nous reprocher d'avoir décelé dans ces rapprochements un sens qui n'y était pas. Pourtant, connaissant le roman, il nous semble avoir mis en évidence dans un minimum de temps et sans tricher avec les règles que nous nous étions imposées au travers de la méthode et du logiciel, les thèmes fondamentaux du roman et produit par les isotopies principales de la dernière classification une indexation de qualité malgré quelques erreurs mineures.

Notre pari était de montrer qu'il n'est pas besoin d'utiliser de ressources linguistiques sophistiquées pour produire des propositions d'indexation pertinentes. Nous pensons en avoir fourni la preuve dans cet exemple que tout un chacun peut vérifier. Comme nous l'avons dit : il est facile de se faire une idée claire du contenu d'un roman et nous mettrons très prochainement notre logiciel qui se compose d'un ensemble de *plugins* Eclipse à disposition sous une licence *open source* comme nous le verrons dans le chapitre suivant.

Il faut par ailleurs noter qu'il reste de vastes marges d'amélioration, en effet notre découpage des passages est on ne peut plus grossier nous n'avons utilisé aucun algorithme

de reconnaissance d'entités nommées par exemples ou d'autres méthodes d'analyse de surface qui de notre point de vue pourraient être fort utiles. Notre système repose entièrement sur une analyse sémantique latente améliorée guidée par la théorie de la sémantique interprétative.

6 Pour une plateforme de philologie numérique

C'est à dessein que nous reprenons le terme de « philologie numérique » introduit par François Rastier dans un chapitre de son ouvrage *Arts et sciences du texte* [Rastier 2001a] et que nous l'accolons à celui de plateforme. La dimension informatique que sous-tend l'idée même de philologie numérique nous rappelle les multiples difficultés auxquelles se heurte quiconque aborde le traitement des données textuelles digitales. Ces difficultés sont de toutes natures : techniques, architecturales, économiques, stratégiques, théoriques. Les quelques pages de cette thèse qui fut une aventure de quatre ans n'est que la partie émergée de l'iceberg, la partie immergée est le programme qui les supporte. Ce fut une contrainte forte de cette thèse de ne pas seulement tester des idées mais aussi de les mettre en œuvre de la façon la plus industrielle possible, c'est à dire, prenant autant que faire ce peut des contraintes de facilité de déploiement, de maintenance, d'évolutivité, d'intégration, de coûts, d'adoption de standards. Mais reprenons la nature des difficultés que nous évoquions dans l'ordre inverse de leur énumération.

6.1 Théorie

Pour qui aborde le domaine du traitement automatique de la langue aussi bien que pour celui qui le connaît mieux, il est difficile de faire la part des choses tant les théories abondent, se contredisent, se superposent. Elles proviennent très souvent de disciplines qui n'ont pour ainsi dire rien en commun, elles sont assujetties aux modes et aux chapelles. Elles nécessitent l'apprentissage de vocabulaires spécialisés, la découverte d'algorithmes complexes, l'utilisation de logiciels épars reflets de la disparité des idées, des approches et convenons-en de leur peu d'efficacité dans le contexte marchand : l'ordinateur avec lequel on converse que l'on imaginait il y a cinquante ans n'a pas vu le jour, les interfaces graphiques se sont depuis imposées. C'est en fait l'objectif que l'on se fixe qui détermine la ou les théories auxquelles l'on tente de se rattacher. Ce n'est pas tant que l'on aille au marché des théories remplir son panier mais il faut bien, ne serait-ce que pour les mettre à l'épreuve se fixer à un moment. Dans notre cas, l'objectif est assez clair : c'est l'indexation de textes. Il ne faut pas croire pour autant qu'il existe une théorie de l'indexation automatique dont on pourrait se revendiquer et à laquelle on pourrait apporter sa contribution. A nos yeux, le premier des mérites de François Rastier a été une volonté

opiniâtre depuis plusieurs années de rendre compte de ces difficultés et de proposer une nouvelle perspective à la linguistique centrée autour du texte et de son interprétation et non pas autour de la phrase et de la syntaxe, autour du sens et non pas autour de la signification. La théorie de l'apprentissage statistique de Vladimir Vapnik se situe quant à elle dans une tout autre dimension, très éloignée des préoccupations purement linguistiques en généralisant la théorie de Glivenko-Cantelli-Kolmogorov pour construire une théorie de la Minimisation du Risque Empirique. Cette théorie a suscité un énorme engouement à partir des années 90, qui c'est traduit par le développement des Machines à Vastes Marges et la recherche de fonctions à noyaux utilisables par ces machines. Des fonctions à noyaux ont été utilisées dans toute sortes de domaines allant du traitement des séries temporelles à celui des chaînes d'ADN. Ces travaux en ont relancé d'autres plus anciens sur les la sémantique latente et les comparaisons de chaînes de caractères plus proche de nos préoccupations. D'un point de vue théorique notre ambition a donc été d'amorcer un rapprochement entre deux courants de pensée fort éloignés en nous fondant sur l'idée que s'il est difficile d'imaginer des machines interprétantes, il est plus facile d'imaginer des machines reproduisant des interprétations. Notre contribution aura de ce point de vue été de proposer d'encadrer l'interprétation de textes en la ramenant à l'interprétation de classes de contextes et de classes de textes.

6.2 Stratégie

Les difficultés stratégiques relèvent de la difficulté d'évaluer la pertinence d'objectifs sur le long terme, de la difficulté de choisir les théories et les moyens nécessaires à leur réalisation, de l'ampleur de la vision qu'ils sous-tendent. Les études du vol des oiseaux et des chauves-souris de Léonard de Vinci étaient-elles les plus appropriées pour construire des machines volantes ? Steve Jobs compris l'importance de la souris et de l'interface graphique lors de sa visite à Xerox PARC, il su mettre en œuvre ce que d'autres avaient inventé. Stratégiquement, la philologie numérique répond à un besoin dont l'importance n'est pas toujours bien perçu, celui des moyens nécessaires à l'interprétation des textes numériques et au delà des documents numériques. Nous avons tenté dans la première section, de montrer la nécessité de prendre explicitement en compte la dimension herméneutique de l'analyse des données ce qui suppose le développement de démarches et d'outils spécifiques. Lorsque ces données sont des textes ces exigences nouvelles sont celles de la philologie numériques, ce sont ces considérations qui nous ont amenés á

envisager non pas l'écriture d'un nouveau logiciel mais plutôt l'intégration de programmes dans un ensemble plus vaste que nous désignons par l'expression *plateforme de philologie numérique*. Il est en effet difficile de penser qu'un programme, aussi bien conçu soit-il, réponde sur le long terme à des demandes que l'on n'imagine pas aujourd'hui ; par ailleurs beaucoup de programmes existent déjà et il serait impossible et de toutes façons trop coûteux de les récrire. Ces contraintes existent pour de nombreuses autres applications et un vaste mouvement commencé avec ARPANET[112] dans les années 60 puis continué dans les années 70 par Richard Stallman avec GNU[113] arrive aujourd'hui à maturité pour y répondre sous le nom générique d'*open source*[114]. GNU et OSI[115] sont d'abord des licences qui permettent l'utilisation libre de logiciels tout en les garantissant contre leur appropriation abusive. *Open source* est une redéfinition de la licence GNU par Bruce Perens et Eric Raymond ouvrant la porte aux industriels, en leur permettant notamment d'y insérer certaines clauses relatives au nom commercial [Leclercq 1999]. Dans les faits l'*open source* est un ensemble de règles et de pratiques promouvant l'accès, la conception, et la production de logiciels mais au delà de biens et de connaissances en général (voir par exemple le projet Gutenberg). Son principe fondamental, comme son nom, l'indique est la transparence du code source, son libre accès et sa modification à condition que soit inséré la licence permettant son utilisation. L'ouverture vers les industriels et le positionnement stratégique de certains poids lourds comme IBM et Sun ont considérablement changé le paysage du logiciel[116], et de vastes communautés d'industriels, de développeurs indépendants et d'utilisateurs se sont regroupées autour de projets communs. Le travail collaboratif permis par Internet et l'*open source* ouvre la voie à ce que nous appelons peut-être improprement plateforme. Une plateforme est un ensemble intégré de logiciels *open source* utilisés et entretenus par une communauté dans le but de traiter un problème commun. Notre thèse se veut donc un manifeste pour une plateforme de philologie numérique.

Les difficultés économiques sont souvent de peu d'intérêt dans ce genre d'exercice académique, mais n'oublions pas que l'ENST Bretagne est une école d'ingénieurs et que les ingénieurs doivent considérer les contraintes économiques qui s'imposent à leurs

[112] Advanced Research Projects Agency Network

[113] GNU pour GNU is Not Unix

[114] Nous renonçons à utiliser une traduction française tant le terme s'est imposé en tant que tel

[115] Open Source Initiative

[116] Cette thèse, par exemple, est écrite en utilisant l'éditeur de texte *open source* de Sun, Open Office.

réalisations techniques. Pour pouvoir prétendre à une véritable pérennité un projet *open source* doit pouvoir attirer des industriels, doit donc pouvoir offrir des moyens de faire des bénéfices commerciaux. Donc deux questions se posent : de façon générale comment gagne-t-on de l'argent avec une application *open source* ? De façon particulière quel est l'attrait commercial d'une plateforme de philologie numérique ?

6.3 Économie

Comment les compagnies *open source* font-elles des bénéfices ? Tout d'abord en déplaçant le centre de gravité de leur stratégie commerciale du flux de revenu généré par le nombre de licences vendues aux flux de revenu généré par les services proposés autour des produits qu'elles offrent gratuitement[117]. Concrètement, la plupart des compagnies *open source* procèdent de la manière suivante : à partir d'une idée[118] chercher les programmes *open source* existant[119] proposant des solutions proches ou liées, les intégrer, ajouter les fonctionnalités manquantes et conditionner le tout en un nouveau produit mis gratuitement à disposition sous une licence *open source*. Dans un souci de courtoisie, mais aussi pour s'intégrer dans des communautés existantes il est d'usage de contacter les autres compagnies, organisations et individus dont on utilise les produits *open source*, ce qui nous amène à la seconde phase : construire une communauté d'utilisateurs autour du nouveau produit. Bien évidemment, il faut mettre gratuitement son produit sur Internet, en utilisant son propre site ou plus préférablement en inscrivant son projet sur des sites spécialisés comme FreashMeat ou SourceForge qui *de facto* assureront une certaine audience au projet naissant.[120] Les arguments marketing auprès de futures clients sont essentiellement : la transparence, le code est disponible et aucune épée de Damoclès n'est suspendue au dessus de leur tête par des clauses spécifiques les empêchant de continuer à développer leurs applications ; la pérennité : si pour une raison ou une autre l'équipe ayant développé le code disparaît[121] le code reste à disposition et la communauté continue à le faire vivre ; des considérations régionales peuvent aussi être de puissants arguments commerciaux par

[117]Ce n'est pas exactement l'idée qui nous intéresse ici mais les licences *open source* permettent d'utiliser les programmes qui en relèvent dans des programmes propriétaires

[118] La première idée que nous proposons d'implémenter au sein de la plateforme de philologie numérique est le développement d'un programme d'indexation.

[119] En visitant les sites FreshMeat.net ou SourceForge.net par exemple

[120] Une bonne documentation et un scrupuleux respect des règles usuelles de clarté de programmation est on s'en doute un atout significatif.

[121] Retraite anticipée pour cause de succès foudroyant

l'assurance que les services fournis respectent les lois locales et que les clients de ces services bénéficient des garanties auxquelles ils s'attendent habituellement, que le support fourni correspond aux heures travaillées du pays où le produit est distribué et qu'il est délivré dans la langue adéquate. Ce qui nous amène à la troisième phase, comment faire entrer l'argent nécessaire au fonctionnement de l'organisation ? La réponse est simple, il faut que la communauté d'utilisateurs atteigne une masse critique. Le support qui était jusqu'alors gratuit peut se décliner sur des modes commerciaux différents, offrant par exemple des supports dédiés se facturant entre 100 et 200 € de l'heure[122], mais beaucoup d'autres formes de service peuvent être proposées : formations, séminaires, conseil, la vente d'encarts publicitaires, etc...[123] La clef de voûte est donc l'utilité et l'efficacité du produit proposé, qu'en est-il d'une plateforme de philologie numérique ? Tout d'abord, rappelons le, il ne s'agit pas d'un unique produit mais de ce que l'on a coutume d'appeler une suite, l'application d'indexation par isotopie décrite dans cette dissertation n'en est qu'un élément[124]. Qu'est-ce qu'une plateforme de philologie numérique peut apporter et à qui ? Quelles sont les applications *open source* existantes ayant des objectifs similaires ? Il existe trois grands projets *open source* ou GNU : GATE, OpenNLP et UIMA dont la vocation est voisine de celle la philologie numérique. Ces trois projets sont comparés dans « Étude des frameworks UIMA[125], GATE[126] et OpenNLP » [Bond 2006] ; ils ont pour ambition de proposer des cadres unificateurs pour l'intégration d'outils de traitement automatique de la langue. Nous verrons plus loin que nous utilisons UIMA dans notre architecture et nous nous en expliquerons, ce qui nous intéresse pour lors ce sont les aspects différenciateurs. Suivant la stratégie commerciale *open source* esquissée plus haut, une fois les applications *open source* existantes sur lesquelles on veut s'appuyer identifiées, il faut écrire les parties particulières à l'application que l'on veut développer : ce qui de façon générale n'existe pas ou peu sont les outils interprétatifs. La plupart des outils proposés dans GATE sont des outils traditionnels de traitement de la phrase qui ne correspondent pas vraiment aux objectifs herméneutiques de la philologie numérique. Sur la question plus particulière de l'indexation nous n'avons pas trouvé d'applications *open*

[122] Il faut évidemment que le support soit à la hauteur du prix demandé

[123] On pourra utilement consulter le site http://www.wiredjournal.com/ pour un exemple de modèle commercial *open source.*

[124] C'est en tous cas notre vœu le plus cher,

[125] *Unstructured Information Management Application*

[126] *General Architecture For Text Engineering*

source proches de nos préoccupations, les projets *open source* d'indexation étant le plus souvent des variantes des systèmes Saltoniens *The Latent Semantic Indexing Projec*t mis à part. Il n'existe pas à notre connaissance d'outils *open source* d'indexation interprétative telle que nous la concevons et dont nous avons développés les principes dans les sections précédentes. Or il existe un réel besoin de ce type d'indexation prenant tout à la fois en compte la dimension herméneutique de toute indexation de qualité, les contraintes économiques qui supposent la mécanisation de certaines tâches et la reproductibilité automatique des interprétations. Nous avons déjà souligné le déficit herméneutique de l'analyse des données en général ; ce déficit se retrouve dans le traitement et l'analyse des données textuelles et à notre sens une approche y remédiant, fournissant des outils simples et des méthodes claires est promise à un avenir dépassant le cadre universitaire, cette approche est celle de la philologie numérique dont François Rastier trace les grandes lignes dans [Rastier 2001a] et dont notre application d'indexation s'inspire.

6.4 Architecture

Le chapitre précédent montre l'importance de savoir intégrer des projets *open source* de provenance diverses dont la complexité et la qualité peuvent varier considérablement. Nous avons donc décidé d'utiliser Eclipse comme plateforme d'intégration en raison de ses incomparables qualités. Eclipse a été créé par IBM en Novembre 2001. Le fondation Eclipse créée en Janvier 2004 est une organisation indépendante à but non lucratif représentant la communauté qui s'est formée autour de ce qui n'était au départ qu'un environnement de développement Java[127]. Eclipse est donc avant tout une vaste communauté *open source*, regroupée autour d'une plateforme qui intègre des centaines d'applications[128]. Voici comment Eclipse se présente sur son site :

> « Eclipse is an open source community whose projects are focused on building an open development platform comprised of extensible frameworks, tools and runtimes for building, deploying and managing software across the lifecycle. A large and vibrant ecosystem of major technology vendors, innovative start-ups, universities, research institutions and individuals extend, complement and support the Eclipse platform. »

Ce fut donc naturel de choisir Eclipse qui est au cœur d'applications aussi ambitieuses que Maestro, le projet de la NASA, qui développe les outils Java de pilotage à distance de

[127] On trouvera une description d'Eclipse à http://www.eclipse.org/org/

[128] Au moment où nous écrivons ces lignes le site www.eclipse*plug-in*central.com/ recense 942 *plug-ins* Eclipse.

véhicules et d'expériences lors des missions spatiales comme « Mars Polar Lander » dont on voit une image écran à la figure 21. Nous reviendrons plus en détail sur la structure d'Eclipse et ce qui en fait son succès, mais l'importance stratégique de cette plateforme explique que nous avons subordonné tout autre application *open source* à la facilité de l'intégrer dans Eclipse et donc pour partie notre choix d'UIMA comme application de base de traitement textuelle[129]. Dans la section 6 consacrée à un exemple d'application nous avons utilisé l'application Eclipse BIRT[130] pour en créer la structure de base à laquelle nous avons seulement ajouté des commentaires ; ce n'est qu'un exemple parmi d'autres d'utilisation d'applications Eclipse existantes dans le développement d'une plateforme de philologie numérique.

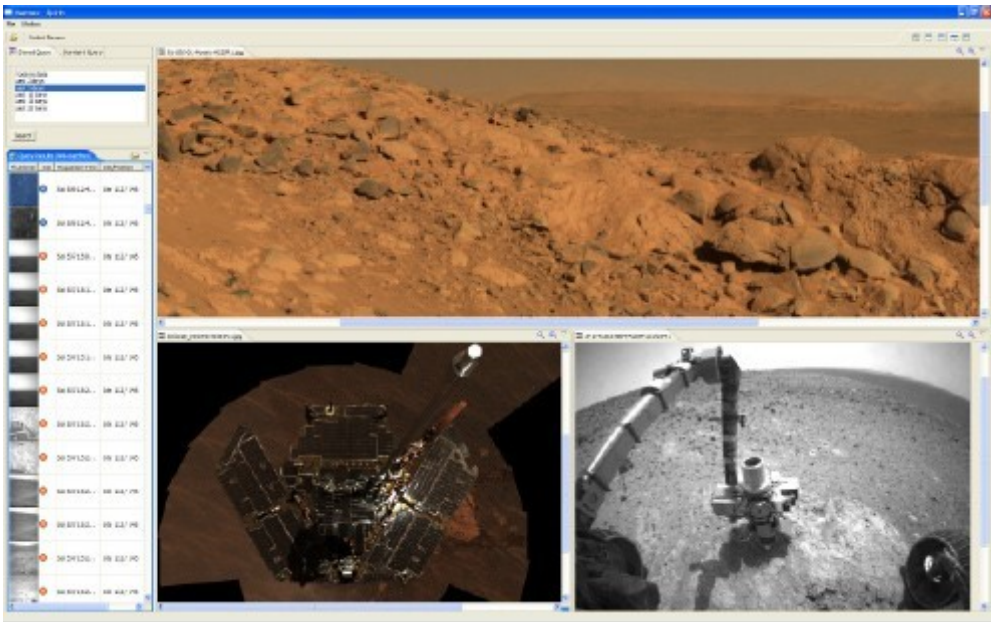

Figure 21: Maestro

6.4.1 Eclipse comme plateforme de philologie numérique

GATE que nous avons déjà mentionné se présente comme « the Eclipse of Natural Language Engineering, the Lucene of Information Extraction, a leading toolkit for Text Mining ». Ce n'est malheureusement pas pour ses promoteurs une application Eclipse, mais qu'est-ce qu'une application Eclipse ? Les auteurs de [McAffer 2006] décrivent Eclipse comme un « système de composants logiciels »[131]. L'unité de base de ce système est le

[129] Nous verrons que stratégiquement UIMA possède des qualités intrinsèques autres qui justifie notre choix.

[130] Business Intelligence Reporting Tool

[131] « We described the essence of Eclipse as its role as a component framework »

plug-in[132], tout est *plug-in* dans Eclipse, les outils de développements JDT et PDE sont entre autres des *plug-in* comme on le voit à la figure 22, une nouvelle application est aussi un *plug-in*. Ces *plug-in* se fichent dans la plateforme Eclipse qui fournit les composantes essentielles que sont l'interface graphique de base, (le *workbench*) et l'accès aux ressources (projets et répertoires regroupés dans un *workspace*).

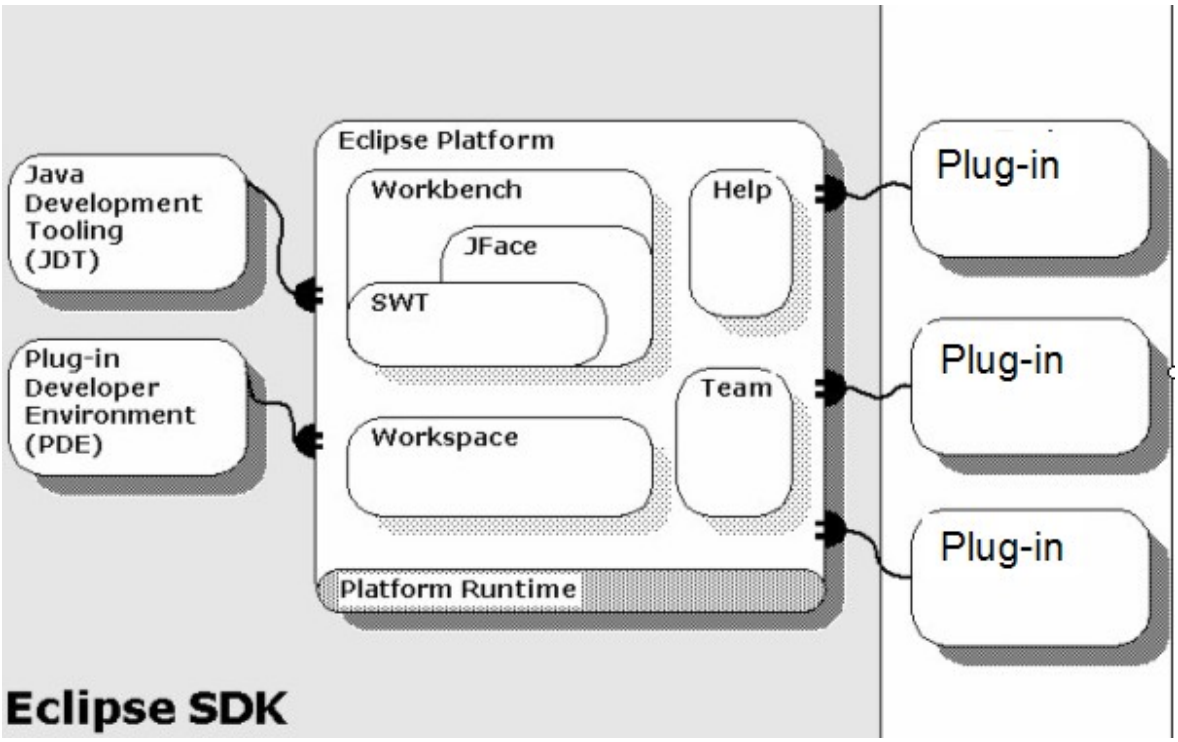

Figure 22: Architecture d'Eclipse

On voit que la plateforme par elle-même peut supporter un ensemble de *plug-ins* sans qu'il soit nécessaire d'y inclure les *plug-ins* de développement c'est cette constatation qui donné jour à la notion de *Rich Client Plateform* (RCP) qui permet d'utiliser la plateforme comme support d'applications indépendantes ; par exemple, Maestro, le produit de la NASA précédemment évoqué, est une application RCP. Ainsi de simple environnement de développement Java Eclipse est devenu une plateforme générale d'intégration de *plug-ins*, d'intégration d'applications et par restriction une plateforme de philologie numérique, il suffit de créer de nouveaux *plug-ins* et de les intégrer à des *plug-ins* déjà existant.

[132] Que nous renonçons aussi à traduire

6.4.2 Anatomie d'un *plug-in*

La figure 23 montre la structure sur disque du *plug-in* utilisé pour implémenter la boite à outils d'automates à états finis utilisée dans notre application d'indexation. Le répertoire 'src' contient le code, le répertoire 'bin' les exécutables résultants qui seront compilés en une archive Java (JAR). Le manifeste de cette archive se trouve dans le fichier MANIFEST.MF dont le contenu est montré à la figure 24. Le fichier plug-in.xml décrit la façon dont le *plug-in* s'intègre dans Eclipse et comment d'autres peuvent s'y intégrer, c'est à dire comment ce *plug-in* utilise les points d'extension d'Eclipse et quels sont les points d'extension qu'il offre.

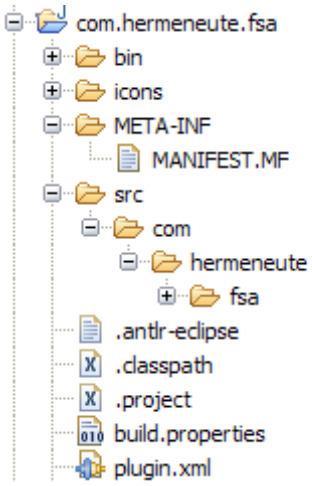

Figure 23: Structure d'un plug-in

Le manifeste de ce plug-in est montré à la figure 24 les sept premières lignes fournissent des informations sur la version du *plug-in* son nom, qui en est l'auteur ; mais surtout il introduit l'importante notion de *bundle*[133] qui rappelle qu'avant tout un *plug-in* est un *bundle* OSGi[134]. Sur le site http://france.osgiusers.org/Main/Objectif on trouve la définition suivante :

> L'OSGi Alliance a été fondée en Mars 1999 par une quinzaine d'entreprises. L'intention originelle était de définir une spécification pour développer et déployer des télé-services dans des passerelles domotiques (aussi par exemple des set-top-box, des modems ADSL, etc.). La première version du standard (élaborée par les quinze membres fondateurs) était strictement orientée par ce marché de niche. Mais ces premières spécifications se sont révélées très intéressantes par les nombreuses implémentations qu'elles ont encouragées, et beaucoup d'autres sociétés se sont jointes pour l'élaboration de leur deuxième version.

[133] Lot, paquet.

[134] *Open Service Gateway Initiative*

Les spécifications OSGi sont un cadre pour la définition, la composition et l'exécution de composants logiciels Java, les *bundles*. Il n'y a pas à proprement parler de différences entre un *bundle* OSGi et un *plug-in* Eclipse. OSGi gère les *bundles* et leur code en fournissant les moyens de retrouver et charger dynamiquement leurs classes. La façon de retrouver dynamiquement ces classes[135] est décrite dans le manifeste.

```
Manifest-Version: 1.0
Bundle-ManifestVersion: 2
Bundle-Name: FSA Plug-in
Bundle-SymbolicName: com.hermeneute.fsa; singleton:=true
Bundle-Version: 1.0.0
Bundle-Activator: com.hermeneute.fsa.Activator
Bundle-Vendor: Hermeneute
Bundle-Localization: plugin
Require-Bundle: org.eclipse.ui,
 org.eclipse.core.runtime,
 org.eclipse.jface.text,
 org.eclipse.ui.editors,
 org.eclipse.ui.workbench.texteditor,
 org.antlr,
 org.antlr.eclipse.core,
 org.eclipse.platform,
 org.eclipse.core.filesystem,
 org.eclipse.core.resources,
 org.eclipse.core.expressions,
 org.eclipse.ui.ide,
 org.eclipse.ui.views,
 org.eclipse.ui.forms
Eclipse-LazyStart: true
Export-Package: com.hermeneute.fsa,
 com.hermeneute.fsa.editors,
 com.hermeneute.fsa.editors.contentassist
```

Figure 24: Manifeste d'un plug-in

[135] On parle de CLASSPATH du *bundle* dans le jargon Java

```
<extension-point
        id="editorActions"          name="%ExtPoint.editorActions"
        schema="schema/editorActions.exsd"/>
<extension-point
        id="editors"                name="%ExtPoint.editors"
        schema="schema/editors.exsd"/>
```

Figure 25: Points d'extensions du plug-in org.eclipse.ui pour les éditeurs

Par exemple dans le manifeste montré à la figure 24, le nom du *bundle*, son nom symbolique et sa version permettent de l'identifier sans équivoque, la section 'Require-*Bundle*:' indique quels sont les *bundles* qu'il utilise lors de son exécution. On voit ainsi qu'en plus de nombreux *plug-ins* Eclipse comme celui qui gère l'interface graphique org.eclipse,core.runtime, ce *plug-in* utilise deux *plug-in open source* ANTLR[136] permettant de définir la syntaxe du langage d'expressions rationnelles que l'on utilise. Enfin la section 'Export-Package:' indique quelles sont les archives JAR qui seront exportées pour mettre ce *bundle* en œuvre. Le cadre OSGi permet aux *plug-ins* d'interagir et de collaborer grâce à ces descriptions d'eux-mêmes fournies par leur manifeste. En plus de la gestion de l'interaction des *plug-ins* par OSGi lors de leur exécution, Eclipse fournit un mécanisme d'enregistrement de leurs extensions. Un *plug-in* peut ainsi s'ouvrir à d'autres *plug-ins* en indiquant les informations dont il a besoin pour accomplir certaines tâches.

[136] ANother Tool for Language Recognition.

```
<extension
      point="org.eclipse.ui.editors">
    <editor
        name="FSA Editor"
        extensions="fsa"
        icon="icons/sample.gif"
        contributorClass="org.eclipse.ui.texteditor.BasicTextEditorActionContri
butor"
        class="com.hermeneute.fsa.editors.FSAEditor"
        id="com.hermeneute.fsa.editors.FSAEditor">
    </editor>
    <editor
        name="HTML editor"
        icon="icons/scribe.jpg"
        extensions="html"
        contributorClass="com.hermeneute.fsa.editors.ResultEditorContributor"
        class="com.hermeneute.fsa.editors.ResultEditor"
        id="com.hermeneute.fsa.editors.ResultEditor" />
  </extension>
```

Figure 26: Contributions du plug-in com.hermeneute.fsa au plug-in org.eclipse.ui pour la création d'éditeurs de texte spécialisés

Pour ce faire, il utilise ce que l'on appelle un point d'extension, par exemple à la figure 25 on voit la déclaration des points d'extension du *plug-in* org.eclipse.ui qui permettent de créer des éditeurs dans le Workbench. Les schémas XML editorActions.exsd et editors.exsd indiquent quelles sont les informations nécessaires à ce *plug-in* pour créer de nouveaux éditeurs. La figure 26 montrent comment le *plug-in* com.hermeneute.fsa utilise ces points d'extensions pour créer des éditeurs spécialisés, ces déclarations XML se trouvent dans le fichier *plug-in*.xml en bas de la figure 23. La première déclaration décrit un nouvel éditeur nommé FSA Editor qui permet d'éditer les fichiers suffixés par '.fsa'

reconnaissables par l'icône 'sample.gif'. Cet éditeur utilise les menus, barres d'outils et ligne d'état standard d'un éditeur de texte standard assuré par la classe :

```
org.eclipse.ui.texteditor.BasicTextEditorActionContributor
```

Les fonctions spécialisée de cet éditeur, vérification syntaxique du langage, coloriage des parties remarquables, compilation au moment de la sauvegarde, etc. sont assurées par la classe :

```
com.hermeneute.fsa.editors.FSAEditor
```

propre à l'application.

Ce sont ces mécanismes de gestion des *plug-ins* par les spécifications OSGi et le registre des extensions qui donnent à Eclipse son incroyable flexibilité. Ce sont eux qui sont utilisés pour ajouter à la plateforme de nouvelles fonctionnalités allant de l'ajout contextuel de menus déroulants, à la gestion des accès aux ressources en passant par la création d'éditeurs spécialisés. Ces mécanismes sont déclaratifs et utilisés parcimonieusement, c'est-à-dire que les composants ne sont chargés que lorsqu'ils sont nécessaires.

6.4.3 SWT et JFace

Pour compléter le tableau, le *workbench* utilise deux librairies, SWT[137] et JFace, permettant d'implémenter un puissant modèle d'interface graphique générique basé sur les notions de fenêtres, de perspectives, de vues, d'éditeurs et d'actions. SWT est une librairie graphique de bas niveau implémentant les objets de base d'une interface graphique que sont les boutons, les listes, les arbres, les fontes, les couleurs..., toutes choses que le système sous-jacent offre de façon standard. SWT présente donc le système de fenêtrage du système d'exploitation existant au travers d'une API qui ne dépend pas de lui. SWT est implémenté sur une large variété de systèmes, c'est lui qui donne aux applications Eclipse la même apparence que des applications natives. JFace se présente comme une boite à outils pour construire des interfaces utilisateurs indépendamment du système de fenêtrage, selon le modèle MVC[138] introduit par Steve Burbeck dans les années 80 avec Smalltalk [Burbeck 1992]. JFace utilise SWT mais n'en masque pas l'accès. La conjugaison du système de *plug-in* et de ces deux librairies permet de définir le Workbench Eclipse.

[137] Standard Widget Toolkit

[138] Model View Controller

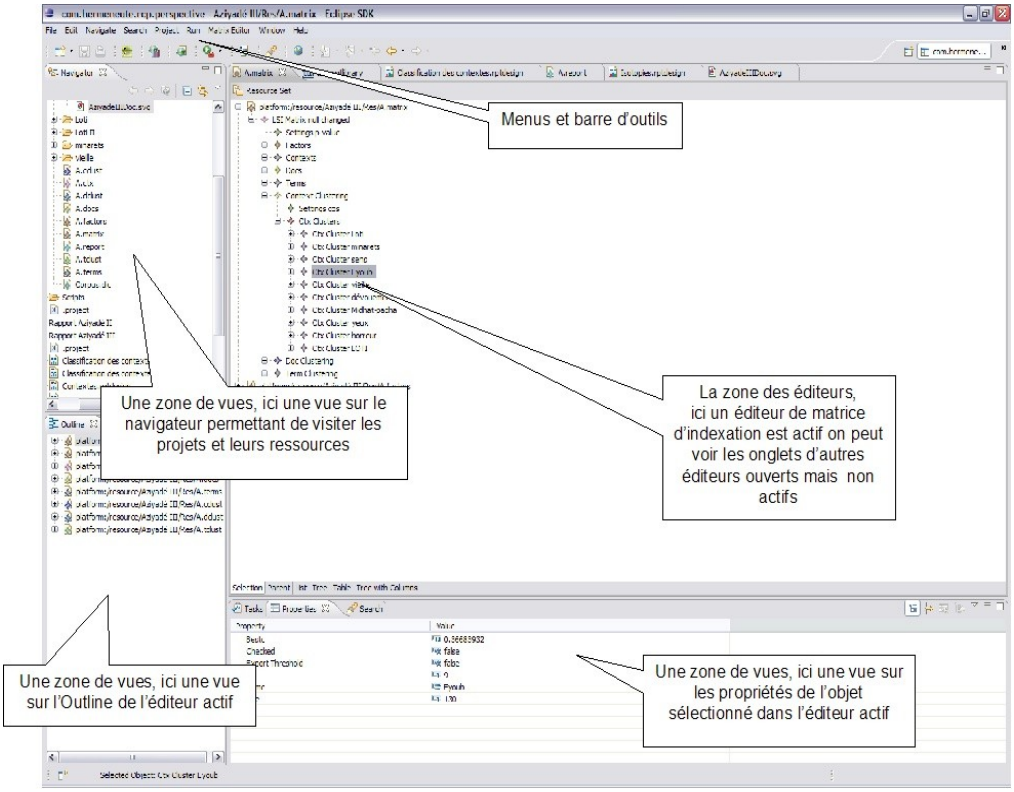

Figure 27: Une perspective sur le Workbench

Ainsi à un coût minimum, toute application peut s'intégrer dans une interface graphique de qualité, robuste, intuitive s'intégrant naturellement dans le système de fenêtrage sous-jacent, le Workbench. Comme on le voit à la figure 160 il apparaît comme un ensemble de fenêtres, ces fenêtres sont de deux types : les éditeurs et les vues s'organisent par zones, une seule pour les éditeurs et plusieurs pour les vues. Dans ces zones les vues et les éditeurs peuvent être empilées, les parties non actives dans chaque zone peuvent être activées grâce à des onglets. Un agencement particuliers de ces zones est une perspective à la figure 28 une perspective BIRT sur le Workbench.

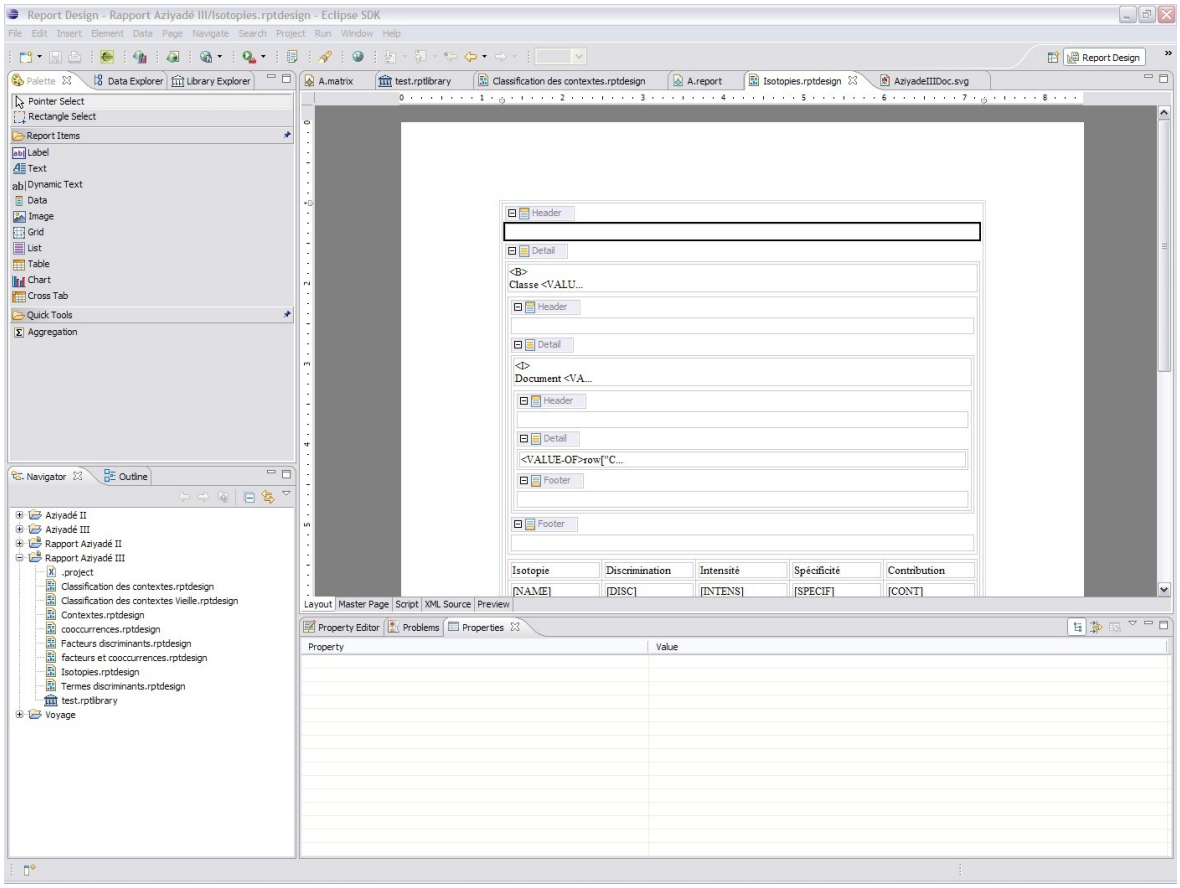

Figure 28: Perspective BIRT sur le Workbench

La différence entre vue et éditeur réside essentiellement dans la rémanence des données. Les données d'un éditeur peuvent être sauvées d'une session sur l'autre, une vue ne contient pas de données rémanentes. Très souvent les vues servent à fournir des informations complémentaires sur l'éditeur actif comme c'est le cas des vues 'Properties' et 'Outline' à la figure 27, de palette comme à la figure 28 ou de vue sur des ressources comme c'est le cas des navigateurs de ces deux figures, des explorateurs de bases de données, de package Java ou de *plug-ins*. Des mécanismes d'écoute d'évènements et de registres permettent aux éditeurs de communiquer avec les vues, les notions de 'drag and drop' sont très simples à mettre en œuvre. Au total, pour un effort de programmation très minime, il est assez simple d'intégrer une application dans le *workbench* ce qui fait d'Eclipse une plateforme d'intégration incomparable, d'autant plus qu'elle s'intègre naturellement dans la plupart des systèmes d'exploitation existants[139]. Mais les raisons d'utiliser Eclipse comme plateforme de philologie numérique ne s'arrêtent pas là.

[139] En tous cas Linux, Mac OS X et Windows.

6.4.4 EMF

EMF[140] est un environnement de modélisation et permettant de générer trois *plug-ins* à partir d'un modèle décrit par UML, un schéma XML ou bien encore un ensemble d'interfaces Java annotées. Ces trois *plug-ins* implémentent respectivement le modèle sous forme de classes Java, un ensemble d'adaptateurs[141] permettant l'édition et la rémanence du modèle sur la base d'un système de commandes et de notifications, un éditeur Eclipse de base comme cela est présenté à la figure 29 empruntée à [Butler 2005]. D'un point de vue pratique, EMF permet de gagner un temps considérable dans la conception d'une application, trois principaux éditeurs de notre application sur quatre sont en fait des adaptations, souvent très superficielles d'éditeurs générés par EMF.

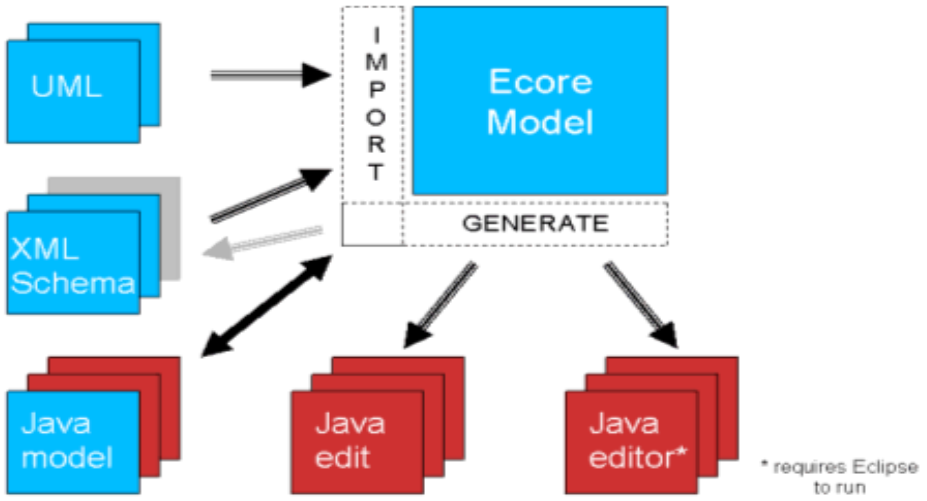

Figure 29: Le fonctionnement d'EMF

Le principe que nous avons employé constamment pour accélérer le développement de notre application a donc été de définir nos modèles de données au travers de schémas XML, de générer les éditeurs correspondants puis en utilisant les points d'extensions permettant d'ajouter des menus déroulant contextuels permettant de lancer les actions correspondant à nos traitements : recherche de facteurs, exportations d'automates,

[140] *Eclipse Modeling Framework*

[141] *Adapters*, ce sont des écouteurs (*listeners*) Java d'un certain type massivement utilisés dans EMF, permettant d'étendre les capacités des objets auxquels ils sont attachés.

classifications, etc. Les données de chaque éditeur sont sauvées sous le format XML correspondant aux schémas ayant permis de générer leur génération[142].

6.4.5 UIMA

Deux cents personnes travaillent dans la division de recherche d'IBM sur des sujets liés au traitement automatique de la langue. Il y a quelques années, une étude interne montra qu'il était crucial pour ces chercheurs de pouvoir découvrir rapidement les résultats des travaux de chacun et de pouvoir utiliser des composants logiciels existants mis au point par d'autres plutôt que de les réécrire [Ferrucci 2004]. C'est de cette nécessité de combiner des composants logiciels existants, écrits dans des langages différents par des équipes différentes qu'est né UIMA. Après avoir été un succès interne, UIMA est maintenant un projet de la fondation Apache disponible sous une licence *open source*.

Les principes architecturaux d'UIMA sont décrits par la figure 30 provenant de la documentation UIMA [UIMA 2006][143]. Nous y voyons apparaître les concepts de base de l'architecture : descripteur de composants[144], annotateur, contextes UIMA et structures d'analyse commune[145].

[142] Il faut noter que c'est la solution la plus simple ; rien n'empêche de rendre ces données persistantes en modifiant le mécanismes standard de sauvegarde, des applications spécialisées comme JPOX ou Hibernate, Elver, www.elver.org propose des solutions *open source* pour ce faire, dans le projet Eclipse Teneo.

[143] Disponible sous licence Apache http://www.apache.org/licenses/LICENSE-2.0

[144] *Component descriptor*

[145] *Common Architecture Structure* (CAS)

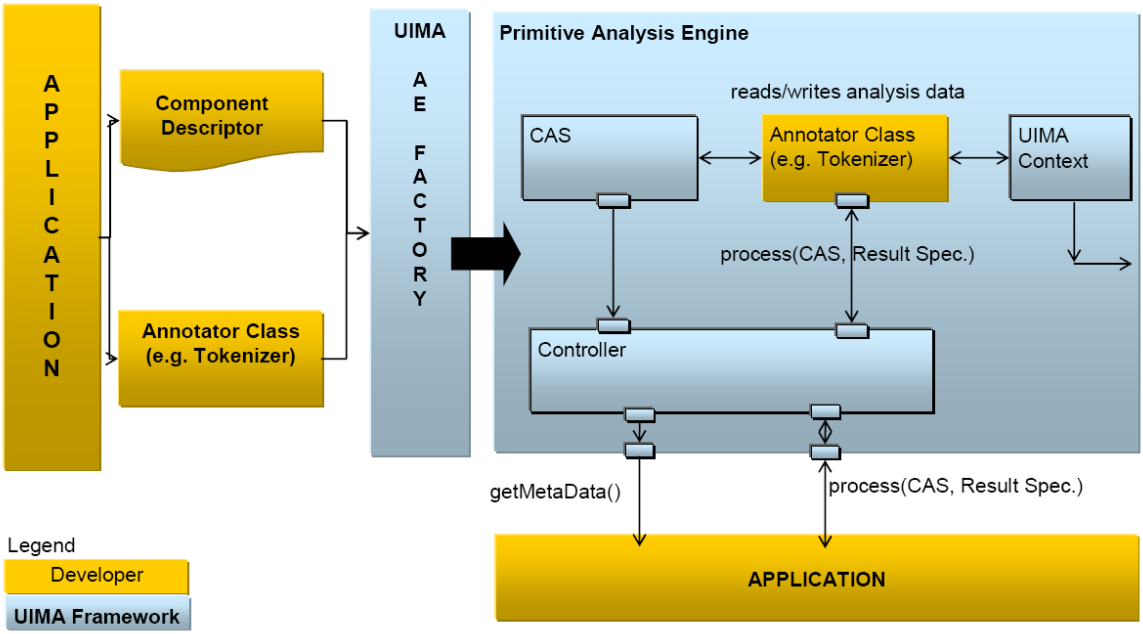

Figure 30: Architecture UIMA

Un descripteur de composant est un fichier XML décrivant les caractéristiques[146] d'un annotateur ou plus généralement d'un moteur d'analyse. Un annotateur reçoit un CAS et y écrit des annotations sur le document qu'il contient, en tenant peut-être compte des annotations qui s'y trouvent déjà, un CAS peut ainsi passer de moteur d'analyse en moteur d'analyse s'enrichissant au fur et à mesure de sa progression dans la chaîne de traitements. Une chaîne de traitements est mise en œuvre par un moteur de traitement de collections[147], lui-même décrit par un fichier XML, un descripteur[148]. Outre les moteurs d'analyse évoqués plus haut, un CPE utilise deux types de programmes eux-mêmes décrits par des descripteurs XML, un lecteur de collection et des consommateurs de CAS[149].

[146]Comme par exemple le langage dans lequel il est écrit, la façon de trouver l'exécutable, les types d'annotations qu'il s'attend à trouver, les types d'annotations qu'il va écrire et des paramètres de fonctionnement.

[147]*Collection Processing Engine* (CPE)

[148]*CPE descriptor*

[149]*Collection reader* et *CAS consumer*

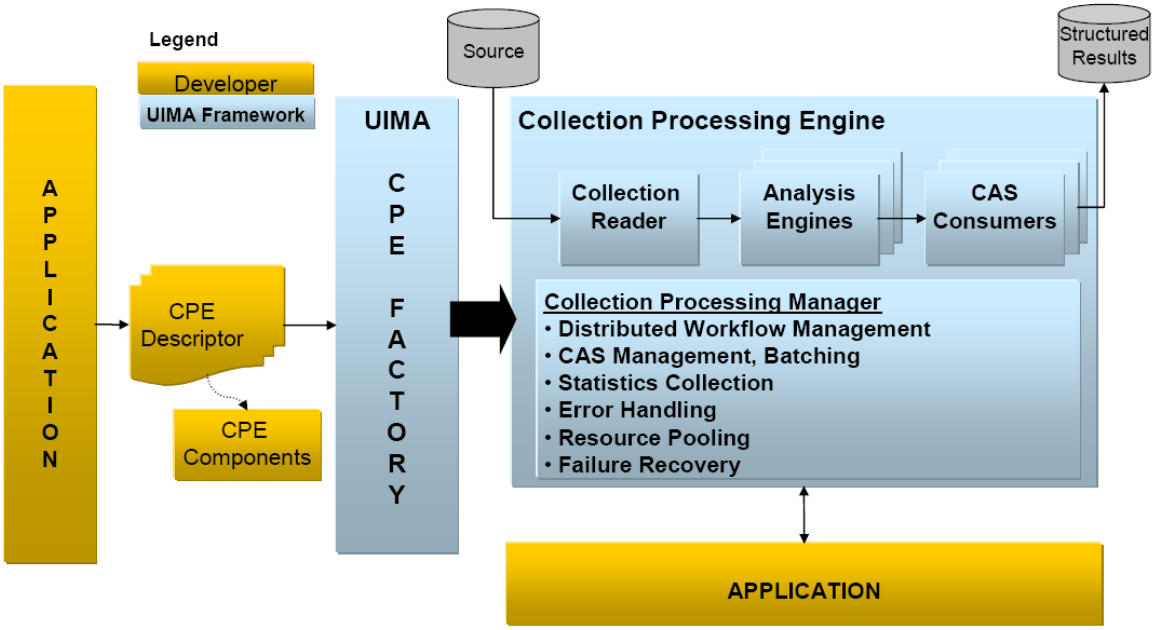

Figure 31: Moteur de traitement de collections

Un lecteur de collection peut-être aussi bien un simple programmes de lectures de fichiers regroupés dans un répertoire qu'un *crawler* Web ou une requête sur une base de données. Les consommateurs de CAS arrivent en fin de chaîne de traitements pour alimenter des applications avales en données structurées. Un gestionnaire de moteurs de collections gère les CAS, leurs flux, le traitement par lots et la gestion des erreurs.

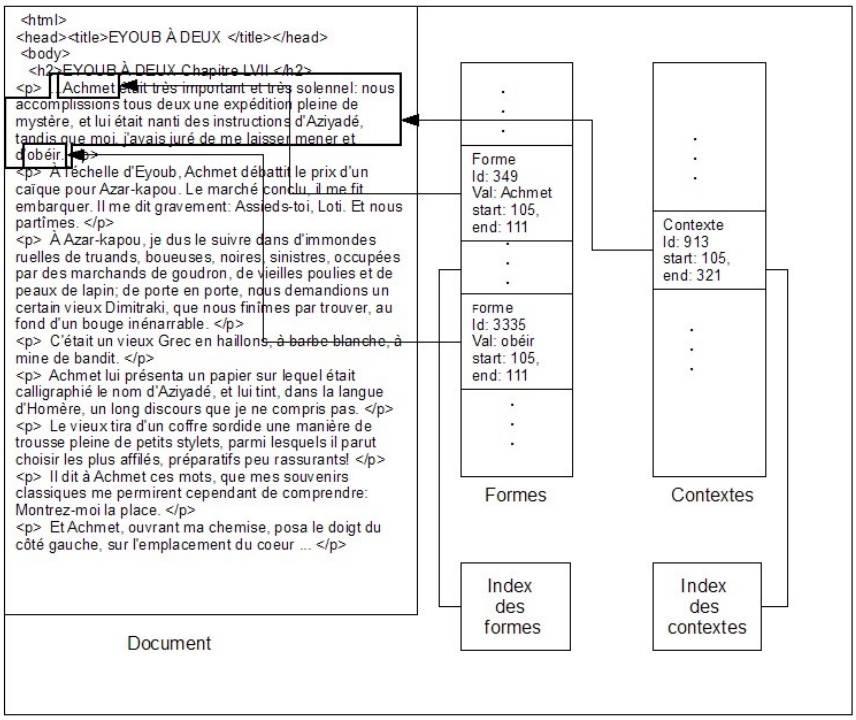

Figure 32: Schématisation d'un CAS

Donc l'objet au cœur de ce dispositif est la structure d'analyse, le CAS[150]. C'est en effet cette structure qui porte les résultats d'analyse, c'est elle qui est échangée par les moteurs. Un schéma XML décrit le CAS et tout moteur d'analyse doit être capable, pour pouvoir s'intégrer dans UIMA, de lire et de produire des fichiers respectant ce schéma et donc de disposer d'une interface CAS dans son langage. Deux langages possèdent une telle interface, Java et C++. L'interface C++ est utilisée par les langages interprétés comme Perl et Python pour pouvoir s'intégrer dans un cadre UIMA. De façon simplifiée, un CAS est composé d'un document, d'un ensemble d'annotations typées et d'un ensemble d'index sur ces annotations permettant de les parcourir efficacement comme le schématise la figure 31. Dans cette figure le document représenté est le chapitre LVII d'*Aziyadé*. Dans notre application, un contexte est une séquence de formes fréquentes[151], situées à l'intérieur d'un

[150] On devrait dire la CAS ou plutôt même la SAC mais dans les deux cas on y trouve des connotations dysphoriques, quoiqu'un casse ne soit pas non plus réjouissant, les acronymes anglais dominent le jargon informatique français, le masculin utilisé ici est une forme d'emploi neutre souvent utilisé dans ce jargon, nous sacrifions donc à cet usage.

[151] Nous utilisons le terme de *forme* plutôt que celui de *mot* pour éviter les débats byzantins sur ce qu'est un mot ; de plus nous pensons dans le future utiliser des tables de suffixes (*Suffix Array*) pour mettre en évidence des récurrences de chaînes de caractères plus fines que les mots afin d'améliorer la pertinence des facteurs : le terme de *forme* nous semble plus approprié dans ce cas de figure.

paragraphe[152]. UIMA offre la possibilité de sous-typer sa structure d'annotation de base et un ensemble de types d'objets ne dépendant pas d'un langage de programmation en particulier[153]. Nous avons ainsi défini deux types d'annotations : forme et contexte. Une forme est définie par un identifiant unique de type entier, sa représentation par une chaîne de caractères[154] et par l'héritage du type d'annotation standard d'UIMA représentant les positions de départ et d'arrivée de l'occurrence repérée dans le texte. De la même façon, un contexte est décrit par un identifiant unique et les positions de départ et d'arrivée[155]. Un CAS peut contenir plusieurs vues sur le document et ses annotations, par exemple pour une page HTML le texte nettoyé de la plupart de ses balises et autres scripts, accompagné de ses annotations propres. Un CAS peut donc contenir plusieurs vues sur les données désignées du joli nom de SofA[156]. Voici donc succinctement présenté l'objet central d'UIMA[157] ; passons donc à ce qui en fait sa spécificité et son intérêt. David Ferrucci et Adam Lally [Ferrucci 2004] définissent les principes qui ont présidé à la création d'UIMA : d'une part, encourager et permettre la réutilisation de composants logiciels, d'autre part, séparer des rôles dans le développement de ces composants. Deux caractéristiques d'une architecture basée autour de l'idée de structure d'analyse commune favorisent l'atteinte du premier objectif :

1. La récursivité de la structure,
2. son centrage sur les données,

Le couple CAS annotateur est récursif en ce sens que plusieurs annotateurs agissant sur des CAS peuvent être agrégés pour produire un CAS unique. Il est ainsi possible d'utiliser des descripteurs particuliers appelés agrégateurs qui ne nécessitent pas l'écriture de moteurs particuliers mais seulement la réutilisation des descripteurs déjà existants, leur combinaison permettant de produire de nouveaux CAS, ainsi des moteurs écrits dans des langages différents. Java, C++, Perl peuvent être utilisés conjointement et remplacés aussi longtemps qu'ils procèdent aux mêmes traitements sur les CAS qui leurs sont soumis.

[152]Une chaîne de caractères commençant par la balise et se terminant par la balise .

[153] Par exemple uima.cas,String pour les chaînes de caractères ou uima.cas.Integer pour les entiers.

[154] Ici le mot lui-même.

[155]la position de départ de la première forme repérée dans le paragraphe considéré et la position finale de la dernière forme repérée dans ce même paragraphe.

[156]*Subject of Analysis*

[157]Pour une description plus détaillée on pourra avantageusement consulter [Götz 2004] et[UIMA 2006].

Qu'il décrive le comportement d'un agrégateur ou d'un moteur d'analyse, un descripteur est défini par le type des données qui lui sont fournies. Ainsi, aussi longtemps qu'un composant reçoit des données correspondant aux types de données spécifiées par son descripteur, il peut être utilisé dans une chaîne de traitement. Un descripteur ne décrit pas uniquement les données que son composant attend en entrée mais aussi la nature des données qu'il ajoute au CAS qu'il produit ; en ce sens un composant est entièrement décrit par les données qu'il reçoit en entrée et qu'il produit en sortie.

Toujours dans [Ferrucci 2004] David Ferrucci et Adam Lally identifient trois rôles dans le développement d'une application traitant des données non structurées.

1. Le développeur d'annotateurs,
2. l'assembleur de moteurs d'analyse,
3. le déployeur de moteurs d'analyse.

Sans entrer dans les détails, les intitulés parlent d'eux-mêmes, dans le cas d'une plateforme de philologie numérique, des gens dont les compétences sont très différentes seront amenés à l'utiliser. Une personne ayant d'abord un profil littéraire n'a, en général, cure des subtilités algorithmiques d'un *parser* : elle sera plus intéressée par son utilisation pour tester des hypothèses ou pour remplacer dans une chaîne de traitements un module par un autre plus performant. Le succès de Perl auprès de nombreux linguistes s'explique par sa facilité d'apprentissage et ses grandes capacités à traiter les chaînes de caractères. Pouvoir intégrer des composants Perl dans une telle plateforme est à notre avis essentiel et à notre connaissance UIMA est la seule architecture permettant l'intégration de composants Perl ou Python dont le succès ne se dément pas auprès d'une vaste population venue à l'informatique par nécessité pratique. À l'opposé, il est déterminant que des virtuoses de Java et C++ puissent construire des modules extrêmement complexes et puissants nécessitant des connaissances informatiques approfondies. Enfin, n'oublions pas qu'en dernier ressort ces applications sont appelées à être déployées ; sur des serveurs Web par exemple, par des gens qui n'ont pour leur part que des soucis de performance machines qui, pour être d'un intérêt crucial, ne seront pas abordées ici sinon pour dire que cet aspect aussi a été considéré dans UIMA. Dans le contexte de la philologie numérique, il est aussi important de noter qu'UIMA permet d'annoter autre chose que des chaînes de caractères, en effet il est par exemple possible de sous-typer *uima.cas.AnnotationBase* et d'utiliser deux paires de coordonnées (x,y) de types uima.cas.double pour définir une zone sur une image,

ce qui est particulièrement important lorsqu'il s'agit d'annoter des documents manuscrits, des textes dont on ne dispose que des versions photos difficilement transcodables par reconnaissance optique de caractères ou encore des pages Web produites par des logiciels de composition graphique. Enfin, comme nous l'avions déjà annoncé, UIMA s'intègre bien dans Eclipse, à tel point qu'il existe cinq *plug-ins* UIMA dans Eclipse, offrant la possibilité d'éditer les descripteurs de composants, de générer automatiquement les classes Java correspondant aux types nouveaux types définis par les utilisateurs. Cette intégration est imparfaite, il n'existe pas d'éditeurs de descripteurs de CPE, mais elle existe et ne demande qu'à être étendue[158].

En résumé, UIMA est une structure d'accueil pensée pour des applications devant traiter des données non structurées et absolument neutre d'un point de vue théorique. Son ambition première est de faire communiquer efficacement des composants logiciels hétérogènes tout en assurant une efficacité maximum. Elle a prouvé cette efficacité par sa mise en œuvre par IBM dans des projets extrêmement complexes dans l'industrie automobile[159], les services de renseignement[160] ou des études de satisfaction des clientèles[161].

6.4.6 Une première expérimentation

Nous ne prétendons pas être les seuls ou les premiers à avoir développé une application liée au traitement automatique de la langue sous Eclipse en utilisant EMF et UIMA, nous n'avons par contre pas trouvé de projets promouvant ce type d'architecture comme accélérateur de développement. Beaucoup d'entreprises vantent leur utilisation d'UIMA[162] et le monde universitaire commence à s'y intéresser comme en témoigne le workshop UIMA à Tübingen cette année dans le cadre de la conférence de printemps GLDV[163]. Les *plug-ins* Eclipse témoignent plus du fait qu'Eclipse est une version *open source* de l'environnement de développement IBM Websphere Studio, que d'une réelle volonté d'utiliser Eclipse comme une plateforme d'intégration pour les application traitant des

[158]Ce qui pourrait être un projet *open source* en soit.

[159] *Quality Early Warning for the Automotive Industry*

[160]*Advanced Intelligence for Anti-Terrorism and Law Enforcement*

[161]*Customer Service and Problem Detection for Durable Goods and High-Tech Manufacturers*

[162]Arisem , Digimind ou Temis en France, Nstein au Québec

[163] *Gesellschaft für linguistische Datenverarbeitung*

données non structurées. Lorsque nous avons commencé ce projet d'indexation par isotopie, nous avions pour objectif de minimiser l'énorme charge de travail que représente aujourd'hui l'interface utilisateur dans tout projet informatique. Les parties propres à notre projet se réduisaient à cette époque à l'utilisation d'une boite à outils de machines à état finis et de programmes de classifications. Comme beaucoup, nous avons réalisé que la plupart des mécanismes nécessaires à l'interface de qualité que nous recherchions pour notre application se trouvaient déjà dans l'environnement de développement Java que nous utilisions : Eclipse, c'est d'ailleurs ainsi qu'est né le concept RCP d'Eclipse. Nous avons réellement pu mesurer au cours de ce projet l'énorme gain de temps offert par Eclipse en tant que plateforme d'intégration. Par contre, UIMA n'étant pas disponible en tant que produit *open source*, nous avons perdu un temps considérable à récrire en Java notre boite à outils de machines à états finis qui se présentait initialement sous forme de programmes C++ et Perl. Nous n'aurions pas non plus hésité à utiliser nombre de produits *open source* si nous avions utilisé UIMA plus tôt et donc négativement cette fois nous avons pu mesurer le temps que nous aurions pu gagner en l'adoptant plus tôt. Il y a bien sûr une phase d'apprentissage de l'utilisation d'Eclipse, EMF et UIMA mais tous comptes faits nous estimons que si nous avions disposé de cette plateforme dès le début accompagnée de tutoriels, de documentation, d'un Wiki, d'une liste de diffusion et de quelques améliorations simples nous aurions économisé les trois quarts du temps que nous avons consacré au développement informatique, c'est à dire un bon tiers du temps de travail que représente cette thèse. Mais ce n'est pas tout, la promotion de la sémantique interprétative passe par la pratique et l'expérimentation, par le développement et le partage d'outils et de données dédiés à la démarche interprétative. En d'autres termes, sans nier l'intérêt de ce qui existe déjà, force est de constater qu'il manque encore un support informatique ouvert accessible à tous et sans exclusives[164], à la philologie numérique que François Rastier appelle de ses vœux.

[164] Il est plus important d'inciter les gens à s'enrichir grâce aux technologies qui pourraient naître d'une telle initiative plutôt que de les en dissuader : ce qui est important c'est la diffusion et l'échange des idées.

7 Conclusion

Google, ce doit être un record dans une thèse consacrée à l'indexation, n'a été cité que deux fois jusqu'à présent, non pas que nous le déconsidérions, bien au contraire, mais parce que ce qui fait la force de Google est son système de classement des résultats par ordre de pertinence. Ce système qui reste secret est, d'après leurs inventeurs Sergey Brin et Larry Page, essentiellement basé sur la topologie du Web : une page est importante si elle est référencée par d'autres pages et d'autant plus si ces autre pages sont importantes. L'efficacité de ce système explique la formidable puissance financière de cette entreprise, en effet les encarts publicitaires qu'elle vend sont extrêmement pertinents et peuvent rapporter énormément d'argent. Le système est extrêmement difficile à tromper et pour une entreprise, se trouver dans les cinq premières réponses peut être la garantie d'un considérable accroissement de clientèle, à tel point que des gourous indépendants de la Silicon Valley vendent leurs services pour améliorer le score de leurs clients pour certaines requêtes types. Il existe bien sûr une dimension textuelle dans ce moteur de recherche, certainement plus sophistiquée que l'on ne croit d'ailleurs mais elle est elle aussi secrète. On peut donc admirer la réussite de Google mais on ne peut s'y comparer, à quoi bon, dès lors gloser sur ce sujet. Mais, plus généralement, cela pose la question de savoir comment comparer deux systèmes d'indexation. La réponse peut être commerciale comme dans le cas de la bataille entre Google et Yahoo : le meilleur système est celui qui rapporte le plus de dividendes mais on conviendra que par sa démesure une telle comparaison n'aurait aucun sens ici. On peut aussi demander à des échantillons représentatifs d'utilisateurs de noter à l'aveugle la qualité des résultats obtenus mais là encore, en dehors du coût, que représente ce genre d'opération, il n'est pas sûr que dans tous les cas de figure le gagnant soit le meilleur, en effet, tout dépend des critères qui auront guidé le choix des juges et de leur compétence à émettre des jugements de qualité. L'évaluation de la qualité d'une indexation est on le voit un problème difficile car il suppose que l'on sache par avance quels sont les réponses pertinentes à une requête donnée, les indicateurs de rappel et de précision abondamment utilisés dans la littérature sur le sujet sont toujours plus ou moins contestables de fait et les protocoles d'expérimentation extrêmement coûteux. Pour ces raisons : secret des procédures d'indexation, coût des procédures de comparaison et variabilité des interprétations, nous ne pouvons fournir une comparaison de notre approche avec des approches existantes et nous le regrettons.

Par ailleurs est-il pertinent de comparer des systèmes de natures différentes? Google se présente comme un système de filtrage collaboratif passif, mettant à profit l'information implicite fournie par les personnes ayant créé des liens entre différents sites. À l'opposé le système que nous présentons demande à une personne son avis sur la pertinence de regroupements effectués sur la base de leur contenu textuel, le regroupement des opinions de plusieurs personnes jugeant de tels regroupements sur des sujets divers formerait un système de filtrage collaboratif actif. De tels systèmes sont peu utilisés car ils demandent à des personnes de fournir une effort spécifique, mais lorsque l'on achète un ouvrage sur Amazon considère-t-on de la même manière les comptes rendus et le fait même qu'il y en ait (filtrage actif) avec les informations du type « les personnes ayant acheté ce livre ont aussi acheté ceux-ci » ? On en revient alors à l'ultime critère pour juger de la pertinence d'une approche : correspond-elle à un besoin actuel ou potentiel ? Concernant la nôtre, nous nous basons sur plusieurs constats. La demande croissante d'analyses textuelles spécialisées qui se manifestent au travers de la course aux ontologies, avatars modernes des anciens thésaurus, la survie de petites entreprises spécialisées sur ce sujet dans un contexte économique difficile, les investissements en recherche et développement de très grandes entreprises montrent que le besoin existe. L'échec des approches traditionnelles met en lumière la nécessité d'aborder le texte dans son ensemble et la nécessité de prendre en compte la dimension interprétative de leur analyse.

Il existe donc bien un besoin d'indexation interprétative, la question suivante est celle de sa compétitivité. En fait quand on considère les sommes englouties dans le développement d'ontologies dont on ne sait pas très bien comment elles seront utilisées sinon comme l'étaient les thésaurus, par des armées d'indexeurs, ce type d'indexation est particulièrement économique dans la mesure où il est naturellement mécanisable[165]. Nous n'avons pas abordé ce point dans notre exposé[166], mais différentes classifications par isotopies peuvent être combinées pour faire valoir des interprétations complémentaires. Nous sommes certains de pouvoir, par cette approche, indexer de vastes ensembles de textes, en effet, une fois interprétée et validée, une classification peut être appliquée à de très grands ensembles de documents en des temps extrêmement brefs et comme nous l'avons déjà souligné elles

[165]On peut en effet nous opposer que les mêmes techniques d'apprentissage statistique peuvent être utilisées pour faire apprendre à des machines comment caractériser des textes en fonctions des termes d'une ontologie d'après une indexation humaine. Pourquoi donc avoir perdu du temps et de l'argent à construire les dites ontologies quand ont peut directement obtenir ces catégories par une analyse interprétative?

[166]Car il justifie une étude complète en lui-même

peuvent être combinées avec des techniques utilisant des liens hypertextes si ceux-ci existent. Une approche interprétative n'est pas une approche lente, c'est une approche incrémentale qui peut progressivement améliorer un système existant moins élaboré.

Quels sont les apports de cette thèse? Ils sont de trois ordres : technique, méthodologique et architectural. D'un point de vue technique, nous avons proposé une amélioration de la sémantique latente en filtrant les cooccurrences de la matrice de projection par le test exact de Fisher et en utilisant une approximation d'analyse en composantes principales pour trouver les facteurs principaux de cette matrice. Nous avons aussi proposé une nouvelle méthode de classification basée sur la notion de densité d'une fonction à noyaux Cf. [Ho 2007]. D'un point de vue méthodologique, nous avons montré que la sémantique interprétative permet d'interpréter en termes de molécules sémiques et d'isotopies les regroupements de passages et de documents obtenus grâce aux techniques précédentes et d'en corriger les imperfections par une méthode simple d'interprétation de ces regroupements, ouvrant ainsi la voie à une méthode d'indexation interprétative. D'un point de vue architectural, nous avons montré que la combinaison d'Eclipse et d'UIMA offre les fondations d'une plateforme ouverte de philologie numérique qui pourrait accueillir et combiner d'autres outils d'aide à l'interprétation des textes et à l'apprentissage automatique d'interprétations.

La réalisation effective d'une telle plateforme plus peut-être que le reste de cette thèse retient aujourd'hui notre attention. Comme nous le savons, les temps de développement associés à la création d'outils d'analyse textuelle sont très souvent grevés par des aspects absolument étrangers au sujet principal : interface graphique, réutilisation et combinaison de composants, etc. Par ailleurs, il est à peu près impossible à qui n'est pas informaticien de combiner à sa guise différents modules pour tester de nouvelles idées ou vérifier des hypothèses. Les applications se présentent sous une forme monolithique, disposant de leurs propres interfaces graphiques, n'ayant pas de systèmes de communication homogènes. Si, Eclipse et UIMA offrent des solutions à ces problèmes, elles se situent à des niveaux différents et beaucoup reste à faire pour obtenir une véritable plateforme. Tout d'abord les *plug-ins* Eclipse d'UIMA n'offrent qu'un niveau d'intégration minimum, on notera en particulier l'absence d'interfaces graphiques intuitives de création de CPE. Nous avons utilisé dans notre application le concept de *wizard* Eclipse pour collecter les données nécessaires à la génération des descripteurs de moteurs d'analyse et de CPE mais cette

solution reste encore bien peu souple. L'application Eclipse GEF[167] devrait permettre d'écrire des éditeurs graphiques de descripteurs rendant l'utilisation d'UIMA plus intuitive. Mais l'enjeu essentiel est de proposer une gamme d'outils prêts à l'emploi, les outils OpenNLP[168] sont *open source* et facilement intégrables dans UIMA de même que les outils GATE[169] et ceux du Standford NLP Group[170] sous licence GNU[171] mais ces outils ne s'inscrivent pas dans une démarche interprétative mais peuvent bien sûr en être des 'briques' importantes.

La question à laquelle nous voudrions tenter de répondre est de définir ce que nous entendons par démarche interprétative du point de vue informatique. Nous savons que l'intelligence artificielle est une impasse ; cependant une part importante des travaux entrepris sous son égide ont abouti à des résultats qu'il serait dommage d'ignorer ceux concernant en particulier l'apprentissage statistique. Ces travaux permettent de mettre au point des outils qui dans un certain cadre et souvent imparfaitement permettent de reproduire des activités humaines comme la reconnaissance de la parole, la reconnaissance de caractères optique, la segmentation d'image. Le point faible de ces outils est qu'il n'est pas toujours facile de leur fournir les exemples nécessaires à leur apprentissage et qu'ils sont extrêmement sensibles aux variations de contextes: changement de locuteurs en reconnaissance de la parole, qualité de la reproduction en OCR, changement d'environnement en segmentation d'image, toutes choses naturelles et aisées pour une personne. La réflexion sur les interfaces permettant de piloter les outils d'apprentissage statistique ne s'est pas développée car l'on a longtemps cru[172] que l'amélioration de ces outils permettrait à terme de faire disparaître ces imperfections. Nous croyons au contraire que ces imperfections ne pourront jamais être totalement surmontées et ce plus particulièrement dans la compréhension des textes[173], nous pensons au contraire que l'avenir est dans le pilotage de ces outils. Nous nous sommes plus particulièrement

[167] Graphical Editing Framework

[168] http://opennlp.sourceforge.net/

[169] http://gate.ac.uk/

[170] http://www-nlp.stanford.edu/index.shtml

[171] La licence GNU n'est pas open source elle autorise la redistribution et la modification du code à la condition que le produit résultant ne puisse pas être utilisé sous une forme commerciale. On trouvera ici les arguments de Richard Stallman contre l'open source http://www.gnu.org/philosophy/free-software-for-freedom.fr.html, arguments que nous ne partageons pas.

[172] Et l'on croit encore

[173] Ce qui nous paraît même être d'une parfaite incongruité

intéressé dans cette thèse à cette approche dans le cadre de l'indexation, la classification et la recherche des facteurs latents étant les outils à piloter pour obtenir une indexation de qualité, mais nous pensons que la sémantique interprétative suggère de nombreuses autres outils interprétatifs, en traduction automatique, en recherche d'information ciblée, en aide à la rédaction, etc. Il n'est donc pas pour nous question de remettre en cause les nombreux travaux effectués aussi bien en traitement automatique de la langue qu'en apprentissage statistique nous prônons plutôt une approche « pragmatique » centrée sur le sens et donc sur l'interprétation.

Mais surtout une plateforme dépend avant tout d'une communauté, communauté regroupant des utilisateurs et des développeurs, qu'ils soient des individus, des institutions académiques ou des entreprises. Notre conclusion est donc un plaidoyer pour la constitution de cette communauté autour de standards reconnus sur le marché pour fédérer les démarches favorisant « l'essor de nouvelles pratiques interprétatives » [Rastier 2001a]

8 Annexe A

Facteur Loti

Cooccurence	Contribution	Discrimination
Allah	0.46	102.566
Aziyadé	0.399	843.621
Achmet	0.322	476.016
fini	0.261	300.523
disait-elle	0.199	78.793
jours	0.165	443.684
commençait	0.153	15.457
reviendras	0.123	-1.77
volonté	0.123	14.652
Arif	0.123	122.629
attendre	0.123	17.992
morte	0.123	60.789
dit-il	0.123	44.91
ensuite	0.123	77.445
versen	0.107	-2.562
protège	0.107	5.965
revoir	0.107	64.781
sérieux	0.107	28.375
Deerhound	0.107	120.762

Facteur Allah

Cooccurence	Contribution	Discrimination
Aziyadé	0.367	843.621
commençait	0.228	15.457
voir	0.228	159.211
jours	0.226	443.684
hommes	0.221	779.961
Dieu	0.199	227.809
versen	0.171	-2.562
protège	0.171	5.965
esprit	0.17	159.641
aperçoit	0.142	0.121
revoir	0.142	64.781
attendre	0.142	17.992
disait-elle	0.142	78.793
calme	0.129	218.711
haut	0.116	282.645

ouléma	0.114	4.473
sélamet	0.114	-0.871
répéterai	0.114	-0.871
consistance	0.114	-0.871
reviens	0.114	-3.387
amères	0.114	11.23
an	0.114	1.25
Pourtant	0.114	2.57
pleuré	0.114	0.453
reviendras	0.114	-1.77
Béhidjé	0.114	12.637
fanfares	0.114	27.754
soldats	0.113	24.43
Taxim	0.113	15.777
promener	0.113	70.289
volonté	0.113	14.652
rester	0.113	21.66
morte	0.113	60.789
sérieux	0.112	28.375
visite	0.111	91.043
retour	0.111	102.875
rendait	0.11	11.613
ensuite	0.108	77.445
impossible	0.107	46.383
idée	0.106	146.77
Deerhound	0.1	120.762

Facteur pays

Cooccurence	Contribution	Discrimination
mer	0.475	337.555
grande	0.362	897.734
souvent	0.294	538.176
loin	0.292	251.059
nouvelles	0.181	119.371
soleil	0.171	291.496
ciel	0.164	335.969
pluie	0.158	31.105
immense	0.147	-0.328
serais	0.136	10.48
va-et-vient	0.136	12.66
pareille	0.132	104.543
bois	0.121	32.656
appelez	0.113	13.703
Grecs	0.113	9.434
atteindre	0.113	16.637

nord	0.113	0.305
restais	0.113	26.672
faudra	0.113	7.414
arriver	0.113	46.23
traversé	0.113	5.574
Espagne	0.113	9.77
bateaux	0.113	13.898
nuages	0.113	-6.961
Levant	0.113	29.879
pointe	0.113	-4.633
étendue	0.113	18.484
étoile	0.113	16.035
Marmara	0.113	17.742
idée	0.105	146.77
bleue	0.101	57.445

Facteur Samuel

Cooccurence	Contribution	Discrimination
madame	0.471	193.875
barque	0.36	205.539
Salonique	0.332	298.316
dormir	0.249	62.219
kédi	0.222	9.629
pauvre	0.215	-40.395
couverture	0.194	10.355
pense	0.159	275.156
revoir	0.138	64.781
sommeil	0.136	16.703
mer	0.126	337.555
figure	0.119	100.809
boum	0.111	12.953
factionnaires	0.111	3.047
dormait	0.111	11.566
apporté	0.11	8.629
avirons	0.11	-5.898
immobile	0.107	13.941
bord	0.107	162.352
heure	0.105	528.129
servir	0.102	29.969

Facteur génération

Cooccurence	Contribution	Discrimination
communs	0.224	-0.406
siècle	0.224	3.203

époque	0.224	8.629
lieux	0.224	24.988
morale	0.224	14.641
rire	0.224	59.141
gens	0.224	26.379
blasphémer	0.112	9.43
réfléchir	0.112	2.699
résultat	0.111	9.457
récit	0.111	4.492
insipide	0.111	6.922
champ	0.111	7.461
suivons	0.111	8.609
frein	0.111	-0.641
frères	0.111	0.82
complainte	0.111	5.051
béante	0.111	0.844
anéantissement	0.111	5.301
fumer	0.111	10.984
chanté	0.111	-1.762
connaissait	0.111	11.734
Homère	0.111	10.762
existe	0.111	9.707
poignards	0.111	8.32
devinait	0.111	0.746
tableaux	0.111	3.512
drôles	0.111	-3.414
prier	0.111	10.598
restons	0.111	4.348
retenait	0.111	3.375
devenue	0.111	27.086
songeais	0.111	9.297
reçu	0.11	19.086
cervelle	0.11	4.117
pleurer	0.109	13.555
présent	0.109	28.023
intérêt	0.109	12.469
nouveauté	0.109	11.336
têtes	0.108	27.395
vit	0.108	13.465
vaste	0.107	4.168
poussière	0.106	9.816
venir	0.106	49.719
lampe	0.106	45.84
actes	0.104	8.867
vint	0.104	-14.402
travers	0.104	29.004

passant	0.102	25.562
avenir	0.101	69.086

Facteur aime

Cooccurence	Contribution	Discrimination
charme	0.632	9.836
affection	0.221	37.133
âme	0.22	84.555
détails	0.158	6.988
foyer	0.158	19.211
patrie	0.157	16.574
bois	0.153	32.656
heureuses	0.126	-1.172
aiment	0.126	-1.172
au-delà	0.126	13.879
entoure	0.126	0.543
déserté	0.126	1.301
douces	0.126	-2
bergers	0.126	0.023
premières	0.126	20.754
famille	0.126	34.246
adore	0.126	11.52
difficile	0.125	26.883
anciens	0.125	14.34
années	0.123	79.434
rendait	0.123	11.613
présence	0.116	-21.59
langue	0.108	34.766
autrefois	0.104	63.418

Facteur loin

Cooccurence	Contribution	Discrimination
nuit	0.35	421.645
tombes	0.296	27.363
portes	0.263	92.332
cyprès	0.244	211.277
mer	0.241	337.555
parti	0.229	218.215
pareille	0.229	104.543
lampe	0.196	45.84
antiques	0.194	8.406
lampes	0.164	-1.691
éclaire	0.164	-1.691
rares	0.164	-0.59

funéraires	0.163	-13.742
campagne	0.163	181.699
murs	0.16	28.672
bornes	0.131	-1.484
montagne	0.131	12.312
lichen	0.128	-5.055
turbans	0.128	6.027
inscriptions	0.128	-5.625
ouvrent	0.124	2.625
mystérieuses	0.118	7.195
herbe	0.109	40.805
odeur	0.108	-12.078

Facteur gens

Cooccurence	Contribution	Discrimination
pauvres	0.346	11.719
capables	0.283	-2.344
charité	0.252	-0.984
seraient	0.252	-0.586
fort	0.215	74.172
braves	0.189	13.949
actes	0.189	8.867
Dieu	0.186	227.809
fortune	0.157	-2.27
chrétienne	0.157	17.254
intime	0.157	6.754
affection	0.156	37.133
sûrs	0.126	2.508
confier	0.126	8.938
sauver	0.126	-0.449
satisfaction	0.126	2.043
empêche	0.126	5.848
passions	0.126	4.473
dépit	0.126	2.031
impunité	0.126	7.359
pousser	0.126	5.16
retirent	0.126	6.652
pourraient	0.126	3.465
illimités	0.126	-5.586
exemple	0.125	19.324
honnêtes	0.125	1.148
contrôle	0.125	5.613
communs	0.124	-0.406
dévouement	0.124	16.621
siècle	0.122	3.203

époque	0.122	8.629
aimer	0.122	30.07
conscience	0.121	22.262
lieux	0.119	24.988
chercher	0.116	80.195
bons	0.114	103.805
morale	0.111	14.641
amis	0.111	53.727
moitié	0.103	87.648

Facteur Achmet

Cooccurence	Contribution	Discrimination
Aziyadé	0.603	843.621
madame	0.324	193.875
place	0.316	287.023
assis	0.2	194.965
fini	0.161	300.523
parti	0.157	218.215
orchestre	0.147	25.043
lendemain	0.143	30.938
situation	0.137	21.531
guerre	0.133	123.957
caïque	0.121	264.117
embrouillait	0.117	3.031
voisins	0.114	3.719
Allons	0.114	6.855
Ibrahim	0.112	11.324
audace	0.112	20.68
rôle	0.111	22.668
père	0.105	14.875

Facteur marbre

Cooccurence	Contribution	Discrimination
blancheur	0.325	0.891
blanc	0.285	79.023
fontaine	0.244	14.527
sombres	0.244	13.277
palais	0.203	15.664
sol	0.203	12.406
cimetières	0.203	-22.148
cyprès	0.171	211.277
dalles	0.163	-22.719
chemins	0.163	4.625
nuances	0.163	-5.203

cours	0.162	1.016
froide	0.161	-30.848
gris	0.161	-26.566
bleu	0.159	21.504
neige	0.159	-12.676
partie	0.151	16.309
front	0.146	111.887
crues	0.122	-1.633
colonnes	0.122	-2.211
gardes	0.122	-0.668
musiciens	0.122	4.168
chamarrés	0.122	0.645
laquais	0.12	0.137
Dolma-Bagtché	0.119	-11.34
vêtus	0.118	-23.402
quai	0.117	-1.742
pont	0.113	5.477
invraisemblable	0.11	21.668
marches	0.108	17.824

Facteur madame

Cooccurence	Contribution	Discrimination
Galata	0.346	65.539
dormir	0.311	62.219
café	0.277	90.965
vieille	0.276	173.148
coquine	0.242	-4.938
tenait	0.207	9.496
langues	0.207	3.508
parlait	0.207	2.258
métiers	0.207	-0.469
kédi	0.207	9.629
couru	0.207	29.516
Europe	0.207	7.297
quartier	0.207	317.234
porte	0.203	88.949
revint	0.104	1.449
lisez	0.104	3.547
frapper	0.104	-1.5
bir	0.104	-1.797
Bir	0.104	-6.203
chatte	0.104	3.625
bal	0.104	2.902
disait-il	0.104	1.953
apporté	0.103	8.629

impasse	0.103	8.387
ambassade	0.103	13.129
quartiers	0.103	-7.934

Facteur saison

Cooccurence	Contribution	Discrimination
joie	0.429	-16.289
printemps	0.399	6.898
durera	0.245	-0.328
Soyez	0.245	-0.617
gais	0.245	5.547
pleins	0.245	12.574
passe	0.245	66.473
vernale	0.184	-4.676
rossignol	0.153	3.602
approche	0.153	41.332
fleurs	0.153	77.484
argentées	0.123	-0.484
répand	0.123	-0.484
amandier	0.123	-0.484
bosquet	0.123	-0.484
berceau	0.123	-0.484
déployé	0.123	-0.484
Écoutez	0.123	-0.484
finira	0.123	-0.484
"Qui	0.123	-0.484
poésie	0.123	-5.695
Extrait	0.123	5.941
envie	0.123	12.23
chanson	0.123	69.301
orientale	0.121	48.262
belle	0.119	187.098

Facteur printemps

Cooccurence	Contribution	Discrimination
joie	0.385	-16.289
durera	0.256	-0.328
Soyez	0.256	-0.617
gais	0.256	5.547
pleins	0.256	12.574
passe	0.256	66.473
fleurs	0.211	77.484
matinée	0.171	0.102
approche	0.171	41.332

bois	0.165	32.656
argentées	0.128	-0.484
répand	0.128	-0.484
amandier	0.128	-0.484
bosquet	0.128	-0.484
berceau	0.128	-0.484
déployé	0.128	-0.484
Écoutez	0.128	-0.484
finira	0.128	-0.484
"Qui	0.128	-0.484
poésie	0.128	-5.695
rossignol	0.128	3.602
vernale	0.128	-4.676
Extrait	0.128	5.941
cigognes	0.127	7.297
envie	0.127	12.23
chanson	0.117	69.301

Facteur charme

Cooccurence	Contribution	Discrimination
affection	0.273	37.133
coin	0.272	39.805
campagne	0.271	181.699
difficile	0.227	26.883
années	0.227	79.434
bois	0.216	32.656
heureuses	0.182	-1.172
aiment	0.182	-1.172
entoure	0.182	0.543
déserté	0.182	1.301
douces	0.182	-2
Marguerite	0.182	5.973
bergers	0.182	0.023
premières	0.182	20.754
poème	0.182	-4.445
famille	0.181	34.246
foyer	0.181	19.211
anciens	0.179	14.34
rendait	0.175	11.613
patrie	0.166	16.574
amour	0.161	155.879
jeunesse	0.152	49.32
autrefois	0.131	63.418
pousse	0.128	-0.438
oriental	0.105	15.66

Facteur mer

Cooccurence	Contribution	Discrimination
grande	0.486	897.734
souvent	0.298	538.176
ciel	0.242	335.969
Marmara	0.224	17.742
Levant	0.187	29.879
bleue	0.186	57.445
pareille	0.185	104.543
calme	0.171	218.711
serais	0.15	10.48
appelez	0.15	13.703
Grecs	0.15	9.434
atteindre	0.149	16.637
nord	0.149	0.305
restais	0.149	26.672
faudra	0.149	7.414
arriver	0.149	46.23
va-et-vient	0.149	12.66
traversé	0.149	5.574
Espagne	0.149	9.77
bateaux	0.149	13.898
nuages	0.149	-6.961
pointe	0.149	-4.633
étendue	0.149	18.484
étoile	0.149	16.035
nouvelles	0.146	119.371
pluie	0.143	31.105
idée	0.141	146.77

Facteur nuit

Cooccurence	Contribution	Discrimination
heure	0.472	528.129
cyprès	0.355	211.277
campagne	0.269	181.699
Corne	0.263	202.332
souvent	0.212	538.176
obscures	0.203	-2.027
séculaires	0.2	-11.73
Phanar	0.192	158.703
étoilée	0.162	8.766
belle	0.147	187.098
route	0.145	51.77
viendra	0.143	9.906

platanes	0.143	-14.473
silence	0.141	-74.73
souviens	0.136	-2.074
herbe	0.123	40.805
veilleurs	0.114	6.934
larmes	0.113	110.238
suivit	0.111	34.922
sèche	0.11	-4.363
blancs	0.108	11.883
bornes	0.102	-1.484
montagne	0.101	12.312
échelle	0.1	34.93

Facteur fond

Cooccurence	Contribution	Discrimination
abîme	0.383	6.812
mal	0.341	254.574
temps	0.277	463.066
prince	0.256	-1.109
perdu	0.255	133.355
enfant	0.254	435.84
morts	0.213	48.895
lumière	0.207	147.477
Corne	0.201	202.332
maître	0.155	157.395
mois	0.148	119.023
dormait	0.127	11.566
attachés	0.126	6.484
sanglant	0.126	12.066
plonge	0.126	11.578
vanité	0.125	26.492
manque	0.125	17.613
éternité	0.123	57.523
appeler	0.12	16.051
entend	0.116	37.816
appris	0.115	11.047
lointaine	0.112	10.039
brise	0.108	8.656
souffre	0.107	24.117

Facteur jaloux

Cooccurence	Contribution	Discrimination
moindres	0.289	1.637
figure	0.289	100.809

mains	0.289	92.656
vue	0.289	191.148
avenir	0.192	69.086
âges	0.144	4.012
caresses	0.144	-0.316
pressent	0.144	0.754
touchent	0.144	9.309
fondre	0.144	2.234
aient	0.144	2.152
Canlidja	0.144	-2.039
touchée	0.144	-0.523
toucher	0.144	4.285
suffit	0.144	22.461
aimée	0.144	7.242
reprendre	0.144	1.277
école	0.144	29.387
paroles	0.144	8.699
chérie	0.144	70.656
points	0.144	18.941
présente	0.144	21.555
lèvres	0.144	4.895
donné	0.144	53.719
histoires	0.143	21.199
bien-aimée	0.143	83.168
pouvoir	0.142	60.199
sentiments	0.142	26.258
enfance	0.142	94.48
maître	0.141	157.395
fille	0.141	18.016
bouche	0.14	31.215
enfant	0.132	435.84
sens	0.126	45.879
petites	0.125	65.219
vieux	0.113	671.559

Facteur Eyoub

Cooccurence	Contribution	Discrimination
soir	0.498	521.648
mosquée	0.346	466.617
Corne	0.307	202.332
quartier	0.302	317.234
1876	0.231	202.488
échelle	0.231	34.93
logis	0.23	24.195
hiver	0.211	139.543

Phanar	0.189	158.703
assez	0.166	104.031
saint	0.152	33.719
sainte	0.149	53.07
caïque	0.141	264.117
case	0.116	452.188
sultans	0.114	-0.613
tombeau	0.114	-4.863
sacrés	0.106	4.68
située	0.106	1.309

Facteur bague

Cooccurence	Contribution	Discrimination
bijoux	0.305	-0.527
gravé	0.204	-2.641
nom	0.203	145.902
brodée	0.153	-0.938
relatif	0.153	-0.938
emporterais	0.153	-0.938
rêvait	0.153	-0.938
propres	0.153	4.562
énorme	0.153	-0.004
discrétion	0.153	2.973
servante	0.153	0.184
Emineh	0.153	0.184
bazar	0.153	-1.613
payer	0.153	0.125
apporter	0.153	0.777
différents	0.153	0.941
vendre	0.153	0.812
apportait	0.153	3.004
songé	0.153	2.629
pièces	0.153	2.203
vivait	0.153	-1.48
aisance	0.153	0.375
contrôle	0.153	5.613
luxe	0.152	0.098
donnait	0.152	8.488
difficile	0.152	26.883
Scutari	0.152	21.598
envoyer	0.152	13.094
coussins	0.152	10.082
possible	0.152	24.77
passait	0.152	45.418
donner	0.151	22.297

large	0.151	52.973
objets	0.151	-0.043
peur	0.15	24.379
argent	0.145	29.793
petites	0.142	65.219
soie	0.142	67
pauvre	0.14	-40.395

Facteur rapports

Cooccurence	Contribution	Discrimination
sons	0.427	4.348
suite	0.384	24.859
chiffres	0.256	-1.531
plaît	0.256	4.172
phénomène	0.256	11.23
exprimés	0.171	-1.523
vous-même	0.171	15.098
différentes	0.171	8.062
certaines	0.171	12.965
compose	0.128	-2.527
vibrations	0.128	4.293
sympathies	0.128	2.762
intervalles	0.128	7.211
entendez	0.128	-0.684
impressionné	0.128	-6.723
bizarre	0.128	2.578
simplement	0.128	19.441
sonore	0.128	0.168
sensation	0.128	1.074
sympathie	0.128	1.969
occasion	0.128	19.805
sympathiques	0.128	9.973
heureuse	0.128	10.762
phrase	0.128	22.844
entendre	0.128	-2.176
cervelle	0.128	4.117
affaire	0.128	62.777
impression	0.127	51.855
mouvement	0.127	100.305
foule	0.122	179.746
corps	0.112	123.219

Facteur abîme

Cooccurence	Contribution	Discrimination

prince	0.292	-1.109
perdu	0.292	133.355
enfant	0.292	435.84
mal	0.292	254.574
temps	0.292	463.066
attachés	0.146	6.484
sanglant	0.146	12.066
plonge	0.146	11.578
vanité	0.146	26.492
manque	0.146	17.613
éternité	0.146	57.523
appeler	0.146	16.051
entend	0.146	37.816
appris	0.146	11.047
lointaine	0.146	10.039
brise	0.146	8.656
souffre	0.146	24.117
sourit	0.145	21.746
affaire	0.145	62.777
chéri	0.145	50.609
morts	0.145	48.895
force	0.145	78.668
mille	0.145	35.027
enfance	0.144	94.48
maître	0.144	157.395
mois	0.143	119.023
frère	0.143	110.758
pourtant	0.143	99.07
lumière	0.141	147.477
pieds	0.133	88.043
pauvre	0.131	-40.395
voix	0.127	238.742
mort	0.127	365.863

Facteur garçon

Cooccurence	Contribution	Discrimination
monde	0.319	489.047
ans	0.273	77.434
jeune	0.248	500.793
viens	0.226	16.312
invraisemblable	0.226	21.668
existence	0.213	96.797
Péra	0.213	70.426
vivre	0.208	166.449
triste	0.196	66.77

ensemble	0.173	117.324
déterminé	0.169	-3.688
intelligent	0.169	-3.688
vingt-deux	0.168	-2.164
Plumkett	0.167	96.41
connu	0.167	16.957
moyen	0.165	47.926
vingt	0.156	13.57
âge	0.119	20.945
"Tout	0.11	3.035
Derrière	0.106	2.012
uniforme	0.105	-1.719
marchand	0.104	-6.215
prévu	0.104	-2.039
paletot	0.103	4.832
manière	0.102	37.316
lois	0.102	-3.766
chapeau	0.101	-13.398
cœur	0.1	-2.023

Facteur peur

Cooccurence	Contribution	Discrimination
derrière	0.361	72.555
Jésus	0.241	-0.086
conversion	0.241	0.887
ennui	0.24	13.539
marchais	0.18	-5.523
haine	0.18	-1.203
ridicule	0.18	-3.84
sentais	0.18	-1.617
perdre	0.179	1.879
créature	0.178	19.207
grilles	0.177	12.285
mourir	0.17	58.285
maîtresse	0.169	69.594
chercher	0.164	80.195
venu	0.156	76.625
pense	0.149	275.156
maison	0.144	-165.43
kédis	0.119	6.148
pleurera	0.115	-0.625
brûles	0.115	-0.625
auras	0.115	-0.625
séduire	0.115	-0.625
Chanaan	0.115	-0.625

patois	0.115	-0.625
austérité	0.115	-0.625
méthodisme	0.115	-0.625
absurdes	0.115	-0.625
ennuyeux	0.115	-0.625
sermons	0.115	-0.625
ignoble	0.115	-0.625
blafardes	0.115	-0.625
vois-tu	0.112	2.688
première	0.11	121.191
souffres	0.103	0.199
radieux	0.103	16.133

Facteur sons

Cooccurence	Contribution	Discrimination
suite	0.425	24.859
chiffres	0.283	-1.531
plaît	0.283	4.172
phénomène	0.283	11.23
exprimés	0.189	-1.523
vous-même	0.189	15.098
différentes	0.189	8.062
certaines	0.189	12.965
compose	0.142	-2.527
vibrations	0.142	4.293
sympathies	0.142	2.762
intervalles	0.142	7.211
entendez	0.142	-0.684
impressionné	0.142	-6.723
bizarre	0.142	2.578
simplement	0.142	19.441
sonore	0.142	0.168
sensation	0.142	1.074
sympathie	0.142	1.969
occasion	0.142	19.805
sympathiques	0.142	9.973
heureuse	0.142	10.762
phrase	0.141	22.844
entendre	0.141	-2.176
cervelle	0.141	4.117
affaire	0.141	62.777
impression	0.141	51.855
mouvement	0.14	100.305
foule	0.135	179.746
corps	0.122	123.219

Facteur Aziyadé

Cooccurence	Contribution	Discrimination
jours	0.476	443.684
larmes	0.394	110.238
mère	0.339	122.84
jeune	0.262	500.793
barque	0.243	205.539
tapis	0.23	44.492
commençait	0.225	15.457
maîtresse	0.204	69.594
Béhidjé	0.194	12.637
idée	0.187	146.77
morte	0.167	60.789
Kadidja	0.15	4.75
yeux	0.122	143.543
contraire	0.119	29.496
parti	0.112	218.215
robe	0.108	48.613

Facteur minarets

Cooccurence	Contribution	Discrimination
mosquées	0.389	15.324
Stamboul	0.379	496.504
ciel	0.345	335.969
dômes	0.26	-25.473
horizon	0.26	-1.426
immense	0.259	-0.328
haut	0.258	282.645
cyprès	0.256	211.277
air	0.198	409.594
pur	0.144	79.77
ville	0.136	40.52
lune	0.136	81.527
foule	0.136	179.746
Corne	0.135	202.332
pointes	0.129	15.586
Au-dessus	0.128	-2.465
montent	0.127	12.926
séculaires	0.118	-11.73

Facteur vie

Cooccurence	Contribution	Discrimination
Salonique	0.369	298.316

vivre	0.307	166.449
mienne	0.245	49.203
but	0.222	29.734
aimer	0.22	30.07
kodja	0.21	-1.793
hélas	0.209	22.371
humaine	0.209	30.953
signe	0.207	10.824
maîtresse	0.199	69.594
décrire	0.198	0.453
souffert	0.197	23.543
horreur	0.182	9.867
existence	0.166	96.797
vieillesse	0.165	-0.543
vîmes	0.154	12.797
éternellement	0.151	-4.043
vision	0.141	-1.703
souci	0.139	-1.727
simple	0.136	-4.082
visages	0.134	1.855
genre	0.134	20.242
donné	0.107	53.719
habiter	0.106	9.812

Facteur ans

Cooccurence	Contribution	Discrimination
vingt	0.442	13.57
vieux	0.35	671.559
mort	0.228	365.863
compte	0.221	24.566
aimés	0.221	6.293
années	0.216	79.434
milliers	0.214	27.98
âge	0.204	20.945
suivant	0.201	78.832
vingt-sept	0.165	12.184
seize	0.163	-2.344
vingt-deux	0.161	-2.164
jolie	0.148	31.434
tour	0.138	47.234
vieillard	0.13	11.297
prierai	0.108	-1.996
espérerai	0.108	-1.996
mourrai-je	0.108	-1.996
saurai-je	0.108	-1.996

croiras	0.108	-1.996
serait-ce	0.108	-1.996
poupée	0.102	3.438
immatérielle	0.1	-3.242
prie	0.1	-0.504

Facteur yeux

Cooccurence	Contribution	Discrimination
femme	0.403	284.289
tête	0.359	231.547
voile	0.325	105.488
front	0.325	111.887
regard	0.275	66.84
prunelles	0.232	-0.461
camail	0.185	0.551
féredjé	0.184	-9.215
plis	0.183	7.543
verts	0.179	18.953
blanc	0.179	79.023
blanche	0.167	65.977
jeune	0.146	500.793
limpides	0.137	-1.18
miens	0.134	-19.77
semblait	0.132	31.711
yachmak	0.131	1.902
fixés	0.127	-11.918
poètes	0.123	-3.324
taille	0.111	0.766

Facteur suite

Cooccurence	Contribution	Discrimination
chiffres	0.324	-1.531
phénomène	0.324	11.23
plaît	0.324	4.172
entendez	0.216	-0.684
simplement	0.216	19.441
exprimés	0.162	-1.523
compose	0.162	-2.527
vibrations	0.162	4.293
sympathies	0.162	2.762
intervalles	0.162	7.211
impressionné	0.162	-6.723
bizarre	0.162	2.578
sonore	0.162	0.168

sensation	0.162	1.074
sympathie	0.162	1.969
occasion	0.161	19.805
sympathiques	0.161	9.973
heureuse	0.161	10.762
vous-même	0.161	15.098
différentes	0.161	8.062
phrase	0.16	22.844
impressions	0.16	1.211
entendre	0.16	-2.176
cervelle	0.16	4.117
affaire	0.157	62.777
impression	0.156	51.855
mouvement	0.149	100.305
foule	0.111	179.746
pleure	0.101	-7.781

Facteur patrie

Cooccurence	Contribution	Discrimination
face	0.286	6.234
histoire	0.286	19.312
tour	0.283	47.234
corps	0.268	123.219
siècles	0.205	-2.977
ancêtres	0.132	0.949
consolider	0.127	-0.957
vivifier	0.127	-0.957
octroyée	0.127	-0.957
Naguère	0.127	-0.957
glorieusement	0.127	-0.957
figurera	0.127	-0.957
renonçons	0.127	-0.957
meurtrière	0.127	-0.957
glorieuse	0.127	-0.957
honteusement	0.127	-0.957
éteindre	0.127	-0.957
arrêt	0.127	-0.957
incliner	0.127	-0.957
terme	0.127	-0.957
Providence	0.127	-0.957
décrets	0.127	-0.957
États	0.127	-0.957
pères	0.125	4.406
paru	0.123	5.832
héritage	0.123	4.879

balle	0.122	2.16
périr	0.121	2.816
fixé	0.121	2.016
montre	0.121	3.293
ministres	0.118	-1.496
Majesté	0.117	6.98
Sublime	0.113	0.43
reposent	0.112	-4.223
dos	0.111	1.246
loi	0.111	5.305
chrétien	0.11	-9.66
cendres	0.108	0.281
attend	0.108	17.059
inerte	0.108	-2.277
Porte	0.107	4.48
défendre	0.103	-3.609
charte	0.103	7.645

Facteur siècle

Cooccurence	Contribution	Discrimination
communs	0.274	-0.406
époque	0.274	8.629
lieux	0.274	24.988
morale	0.273	14.641
rire	0.273	59.141
reçu	0.204	19.086
vint	0.201	-14.402
lettre	0.195	93.359
idées	0.175	119.398
blasphémer	0.127	9.43
réfléchir	0.125	2.699
résultat	0.125	9.457
récit	0.124	4.492
insipide	0.123	6.922
champ	0.123	7.461
suivons	0.123	8.609
frein	0.123	-0.641
frères	0.123	0.82
complainte	0.121	5.051
béante	0.12	0.844
anéantissement	0.12	5.301
fumer	0.12	10.984
chanté	0.119	-1.762
connaissait	0.118	11.734
Homère	0.118	10.762

existe	0.117	9.707
poignards	0.116	8.32
devinait	0.116	0.746
tableaux	0.116	3.512
drôles	0.113	-3.414
prier	0.113	10.598
restons	0.113	4.348
retenait	0.112	3.375
devenue	0.112	27.086
songeais	0.111	9.297
choses	0.107	256.164
cervelle	0.105	4.117

Facteur Stamboul

Cooccurence	Contribution	Discrimination
mosquées	0.5	15.324
grandes	0.332	129.301
haut	0.294	282.645
rue	0.279	158.309
dômes	0.265	-25.473
grande	0.261	897.734
vieux	0.238	671.559
saint	0.226	33.719
Scutari	0.212	21.598
silhouette	0.178	18.605
Péra	0.142	70.426
pont	0.141	5.477
lieu	0.13	84.609
bagages	0.125	0.922
connaissons	0.124	6.133
faubourg	0.121	3.734
nombre	0.106	-3.906
retrouve	0.102	13.031

Facteur circonstances

Cooccurence	Contribution	Discrimination
souffert	0.348	23.543
quelles	0.232	8.051
air	0.215	409.594
connaître	0.174	1.484
formé	0.174	-3.461
né	0.174	-2.738
jugement	0.174	-2.203
voyez-vous	0.174	-1.852

exprime	0.174	1.512
dépens	0.174	3.375
malheur	0.174	9.648
produit	0.174	15.684
touchée	0.174	-0.523
comprendre	0.174	16.367
type	0.174	2.273
individu	0.174	2.418
porter	0.174	-0.828
pourraient	0.174	3.465
école	0.174	29.387
appris	0.174	11.047
servir	0.173	29.969
raisonnable	0.173	-6.43
parler	0.173	50.133
connaissance	0.173	12.543
scène	0.172	45.551
drôle	0.172	33.34
expression	0.172	7.66
manière	0.17	37.316
garde	0.169	49.285

Facteur joie

Cooccurence	Contribution	Discrimination
pleins	0.406	12.574
durera	0.304	-0.328
Soyez	0.304	-0.617
gais	0.304	5.547
passe	0.304	66.473
fleurs	0.251	77.484
vernale	0.203	-4.676
argentées	0.152	-0.484
répand	0.152	-0.484
amandier	0.152	-0.484
bosquet	0.152	-0.484
berceau	0.152	-0.484
déployé	0.152	-0.484
Écoutez	0.152	-0.484
finira	0.152	-0.484
"Qui	0.152	-0.484
poésie	0.152	-5.695
rossignol	0.152	3.602
Extrait	0.152	5.941
envie	0.151	12.23
approche	0.146	41.332

chanson	0.141	69.301

Facteur face

Cooccurence	Contribution	Discrimination
histoire	0.301	19.312
tour	0.295	47.234
corps	0.269	123.219
intérêt	0.222	12.469
vit	0.219	13.465
"Aziyadé	0.151	-2.254
croit	0.147	1.09
moment	0.138	59.707
puisse	0.136	0.918
consolider	0.129	-0.957
vivifier	0.129	-0.957
octroyée	0.129	-0.957
Naguère	0.129	-0.957
glorieusement	0.129	-0.957
figurera	0.129	-0.957
renonçons	0.129	-0.957
meurtrière	0.129	-0.957
glorieuse	0.129	-0.957
honteusement	0.129	-0.957
éteindre	0.129	-0.957
arrêt	0.129	-0.957
incliner	0.129	-0.957
terme	0.129	-0.957
Providence	0.129	-0.957
décrets	0.129	-0.957
États	0.129	-0.957
pères	0.125	4.406
paru	0.123	5.832
héritage	0.123	4.879
balle	0.122	2.16
périr	0.121	2.816
fixé	0.121	2.016
montre	0.121	3.293
ministres	0.116	-1.496
Majesté	0.115	6.98
Sublime	0.11	0.43
reposent	0.108	-4.223
dos	0.106	1.246
loi	0.106	5.305
chrétien	0.106	-9.66
cendres	0.102	0.281

inerte	0.102	-2.277
Porte	0.101	4.48

Facteur ciel

Cooccurence	Contribution	Discrimination
pur	0.469	79.77
bleu	0.312	21.504
noirs	0.289	101.387
immense	0.28	-0.328
nuages	0.261	-6.961
pâle	0.237	52.387
pavés	0.201	-3.125
toits	0.193	3.16
aspect	0.189	55.719
noir	0.181	-9.199
horizon	0.176	-1.426
cases	0.174	5.844
pluie	0.17	31.105
lune	0.169	81.527
Portia	0.156	-24.754
déluge	0.136	-0.715
clair	0.12	-12.52
beau	0.119	127.891

Facteur communs

Cooccurence	Contribution	Discrimination
époque	0.291	8.629
lieux	0.291	24.988
morale	0.291	14.641
rire	0.291	59.141
blasphémer	0.139	9.43
réfléchir	0.138	2.699
résultat	0.137	9.457
récit	0.137	4.492
insipide	0.136	6.922
champ	0.136	7.461
suivons	0.136	8.609
frein	0.136	-0.641
frères	0.136	0.82
complainte	0.135	5.051
béante	0.134	0.844
anéantissement	0.134	5.301
fumer	0.134	10.984
chanté	0.133	-1.762

connaissait	0.133	11.734
Homère	0.133	10.762
existe	0.132	9.707
poignards	0.131	8.32
devinait	0.131	0.746
tableaux	0.131	3.512
drôles	0.129	-3.414
prier	0.129	10.598
restons	0.129	4.348
retenait	0.128	3.375
devenue	0.128	27.086
songeais	0.128	9.297
reçu	0.125	19.086
cervelle	0.123	4.117
pleurer	0.118	13.555
présent	0.117	28.023
intérêt	0.117	12.469
nouveauté	0.116	11.336
têtes	0.112	27.395
vit	0.108	13.465
vaste	0.106	4.168
poussière	0.101	9.816

Facteur sultan

Cooccurence	Contribution	Discrimination
grande	0.419	897.734
palais	0.317	15.664
Abd-Ul-Hamid	0.254	19.094
sérail	0.25	-11.867
sacre	0.19	-9.809
Mourad	0.19	-5.621
trône	0.19	-6.855
Leurs	0.19	2.641
laquais	0.189	0.137
vive	0.185	7.203
caïques	0.184	3.078
pouvoir	0.167	60.199
envier	0.126	4.289
couverte	0.123	-2.062
livrée	0.123	-2.062
éperon	0.123	-2.062
magnificence	0.123	-2.062
vingt-six	0.123	-2.062
conduits	0.123	-2.062
avènement	0.122	-15.574

portent	0.121	2.566
rameurs	0.12	1.23
ciselés	0.119	-1.359
comédie	0.119	1.777
élégance	0.119	0.848
constitutionnel	0.117	-0.242
vieilli	0.115	-0.672
fatigué	0.115	2.84
déposer	0.113	-2.008
placé	0.106	0.367

Facteur air

Cooccurence	Contribution	Discrimination
vif	0.35	-1.113
vivre	0.32	166.449
mine	0.289	9.059
posées	0.233	-2.34
respirer	0.231	-5.5
pont	0.223	5.477
piteux	0.172	-2.73
douche	0.172	-2.73
vices	0.172	-2.73
sale	0.172	-2.73
piètre	0.172	-2.73
tas	0.17	-0.922
ternes	0.169	-0.492
glacée	0.169	-5.207
fleurissent	0.167	-5.531
monté	0.167	-0.555
retombe	0.167	-0.77
saisit	0.166	-1.426
kédi	0.163	9.629
retrouve	0.162	13.031
bougies	0.157	-1.699
Sodome	0.157	-2.352
vide	0.157	39.648
au-dedans	0.149	-2.996
pur	0.143	79.77

Facteur cyprès

Cooccurence	Contribution	Discrimination
tombes	0.387	27.363
campagne	0.387	181.699
cimetières	0.332	-22.148

séculaires	0.276	-11.73
bois	0.267	32.656
haut	0.236	282.645
sépultures	0.219	-10.453
horizon	0.214	-1.426
arbres	0.212	-13.566
antiques	0.206	8.406
gigantesque	0.159	-4.469
obscures	0.155	-2.027
cimetière	0.15	-19.094
platanes	0.145	-14.473
vagues	0.129	-20.832
lune	0.117	81.527
mosquées	0.109	15.324

Facteur infini

Cooccurence	Contribution	Discrimination
songez	0.344	-3.25
espace	0.344	0.137
vue	0.344	191.148
terre	0.287	251.008
embrasse	0.172	50.176
éternels	0.172	-2.387
éternité	0.172	57.523
tournant	0.172	4.859
essence	0.172	12.723
particulier	0.172	20.598
inconnue	0.172	25.852
raison	0.172	62.801
points	0.172	18.941
centre	0.172	58.977
passent	0.172	82.57
instant	0.171	155.816
suivant	0.171	78.832
pensée	0.17	100.891
esprit	0.17	159.641
moi-même	0.167	98.027
autour	0.163	82.418
choses	0.155	256.164
temps	0.13	463.066
point	0.123	239.785
soleil	0.101	291.496

Facteur impressions

Cooccurence	Contribution	Discrimination
monde	0.316	489.047
écrit	0.253	70.34
causera	0.19	6.504
moindre	0.19	0.492
rappeler	0.19	-2.504
mélange	0.19	-2.336
prononcer	0.19	-1.934
meilleur	0.19	-1.754
réalité	0.19	0.629
entendez	0.19	-0.684
simplement	0.19	19.441
aimons	0.19	5.488
image	0.19	3.578
produit	0.19	15.684
mystérieuse	0.19	5.332
objet	0.19	13.762
papier	0.19	22.273
bien-aimée	0.189	83.168
mille	0.188	35.027
imagination	0.188	87.051
absolument	0.188	-44.199
esprit	0.185	159.641
pur	0.183	79.77
idées	0.18	119.398
nom	0.179	145.902
femme	0.141	284.289

Facteur vive

Cooccurence	Contribution	Discrimination
Midhat-pacha	0.353	1.16
softas	0.353	9.754
bruit	0.353	24.926
saluer	0.177	8.52
libres	0.177	3.973
enivrés	0.177	8.531
criaient	0.177	0.688
Allons	0.177	6.855
portant	0.177	7.637
épaules	0.177	4.219
constitution	0.177	5.367
bande	0.177	23.941
foules	0.176	16.059
courir	0.176	36.566
entendre	0.176	-2.176

centre	0.176	58.977
croire	0.176	30.926
lanternes	0.176	-30.023
portiques	0.176	15.289
humaines	0.176	33.707
Turcs	0.174	189.539
guerre	0.173	123.957
voix	0.159	238.742

Facteur café

Cooccurence	Contribution	Discrimination
turc	0.531	95.984
impasse	0.265	8.387
blancs	0.198	11.883
tasses	0.197	7.762
servait	0.196	1.766
matelots	0.196	4.855
Suleïman	0.195	16.941
narguilhé	0.184	-66.289
rue	0.178	158.309
profond	0.173	46.949
Galata	0.151	65.539
passants	0.15	20.633
revolver	0.124	-2.766
prudent	0.124	-2.766
marchés	0.124	-2.766
contrebande	0.124	-2.766
suspects	0.124	-2.766
maltais	0.124	-2.766
commerce	0.124	-2.766
débouché	0.124	-2.766
famée	0.124	-2.766
bruyante	0.124	-2.766
cercle	0.121	-8.543
intimidée	0.12	3.043
tasse	0.12	-10.152
main	0.12	116.672
cigarette	0.118	2.359
mastic	0.116	1.906
traitait	0.111	-1.906
italiens	0.105	-0.934
apporte	0.104	8.766

Facteur planche

Cooccurence	Contribution	Discrimination
laisser	0.319	15.809
derrière	0.317	72.555
clef	0.154	-1.484
retirait	0.154	-1.484
bonsoir	0.154	-1.484
mouillés	0.154	-1.484
chaussures	0.154	-1.484
four	0.154	-1.484
cuisine	0.154	-1.484
rez-de-chaussée	0.154	-1.484
fermer	0.153	5.578
appartements	0.152	2.965
serrures	0.152	2.949
aveugles	0.151	-3.41
vingt-quatre	0.151	2.367
boue	0.15	4.766
brasero	0.149	-13.852
tourner	0.149	11.516
refermer	0.148	-2.73
traverser	0.148	0.371
verrous	0.146	-12.359
nus	0.146	-14.715
chaleur	0.145	2.906
vase	0.144	11.922
arabe	0.143	11.402
vêtements	0.142	10.547
escalier	0.139	3.172
intérieur	0.137	21.809
monter	0.136	14.086
pleines	0.133	3.875
portière	0.132	10.781
vagues	0.132	-20.832
déserte	0.129	1.285
visite	0.128	91.043
asseoir	0.126	22.57
échelle	0.125	34.93
tomber	0.118	6.738
nattes	0.104	-3.27
épais	0.102	-10.762

Facteur temps

Cooccurence	Contribution	Discrimination
amour	0.467	155.879
viendra	0.416	9.906

mal	0.269	254.574
rêve	0.254	73.023
femme	0.249	284.289
perdu	0.224	133.355
noms	0.208	2.348
restera	0.208	-10.246
nous-mêmes	0.208	0.879
prince	0.206	-1.109
aimez	0.2	8.027
vue	0.174	191.148
enfant	0.148	435.84
enfance	0.139	94.48
éternité	0.136	57.523
orage	0.124	11.383
entend	0.105	37.816

Facteur histoire

Cooccurence	Contribution	Discrimination
tour	0.323	47.234
corps	0.283	123.219
venue	0.236	13.824
scène	0.233	45.551
début	0.161	-36.543
aimées	0.141	4.992
décors	0.141	-15.055
changements	0.141	-3.668
consolider	0.138	-0.957
vivifier	0.138	-0.957
octroyée	0.138	-0.957
Naguère	0.138	-0.957
glorieusement	0.138	-0.957
figurera	0.138	-0.957
renonçons	0.138	-0.957
meurtrière	0.138	-0.957
glorieuse	0.138	-0.957
honteusement	0.138	-0.957
éteindre	0.138	-0.957
arrêt	0.138	-0.957
incliner	0.138	-0.957
terme	0.138	-0.957
Providence	0.138	-0.957
décrets	0.138	-0.957
États	0.138	-0.957
pères	0.133	4.406
paru	0.129	5.832

héritage	0.129	4.879
nature	0.129	62.906
balle	0.128	2.16
périr	0.127	2.816
fixé	0.127	2.016
montre	0.127	3.293
ministres	0.121	-1.496
Majesté	0.119	6.98
Sublime	0.112	0.43
reposent	0.11	-4.223
dos	0.108	1.246
loi	0.108	5.305
chrétien	0.107	-9.66
cendres	0.102	0.281
inerte	0.102	-2.277
Porte	0.1	4.48

Facteur tête

Cooccurence	Contribution	Discrimination
vieille	0.524	173.148
front	0.455	111.887
eau	0.399	72.035
fontaine	0.252	14.527
mésange	0.202	18.016
souvenirs	0.168	35.008
froide	0.165	-30.848
Kadidja	0.163	4.75
danse	0.142	4.914
sérieuse	0.135	5.805
sommet	0.134	5.59
appuyé	0.132	-2.551
compliquée	0.119	-1.645
voyais	0.103	-20.926
plis	0.102	7.543

9 Index des figures

9 Index des figures

10 Bibliographie

Abeillé et al 1994: Anna Abeillé, Marc Cavazza, François Rastier, *Sémantique pour l'analyse*, 1994, Dunod

Alché-Buc, Siolas 2000: Georges Siolas , Florence d'Alché-Buc, *Support vector machines based on semantic kernel for text categorization*, 2000, IEEE-INNS-ENNS

Arendt 1968a: Hannah Arendt, *La crise de la culture, le concept d'histoire*, 1972, Gallimard

Arendt 1968b: Hannah Arendt, *La crise de la culture, la tradition et l'age moderne*, 1972, Gallimard

Arrow 1951: Kenneth J. Arrow, *Social Choice and Individual Values*, 1951,

Besançon et al. 2000: Martin Rajman, Romaric Besançon et Jean-Cédric Chappelier., *Le modèle DSIR*, 2000

Bond 2006: Stéphane Bond, Mustapha Es-Salihe, *Étude des frameworks UIMA, GATE et OpenNLP*, 2006

Borges 1944: Jorge Luis Borges, *Ficciones*, 1944, http://www.fcsh.unl.pt/borgesjorgeluis

Bouroche, Saporta 1980: J.M. Bouroche, G. Saporta, *L'analyse des données*, 1980, PUF

Briet 1951: Suzanne Briet, *Qu'est-ce que la documentation?*, 1951, Revue de la documentation

Burbeck 1992: Steve Burbeck, *How to use Model-View-Controller (MVC)by*, 1992

Butler 2005: John Butler, Ravi Hubbly, Walcelio Melo, *An MOF-based repository for enterprise architecture models*, 2005

Church, Gale 1995: Kenneth W. Church, William A. Gale, *Inverse document frequency (IDF): a measure of deviations from Poisson*, 1995,

Cornuéjols 2005: Antoine Cornuéjols, *Méthodes à noyaux et SVMs*, 2005

Crépel 1990: Pierre Crépel, *Le dernier mot de Condorcet sur les élections*, 1990, Mathématiques et Sciences Humaines

Cristianini et al 2001: Nello Cristianini, John Shawe-Taylor, Huma Lodhi, *Latent Semantic Kernel*, 2001, ICML 2001

Cristianini et al 2004: Nello Cristianini, John Shawe-Taylor, *Kernel Methods for Pattern Analysis*, 2004, Cambridge University Press

Cristianini et al. 2004: Nello Cristianini, John Shawe-Taylor, *Kernel Methods for Pattern Analysis*, 2004, Cambridge University Press

Deerwester et al 1988: Deerwester, Dumais, Furnas, Harshman, Landauer, *Using Latent Semantic Analysis to improve access to textual information*, 1988, CHI's 88

Descartes 1999: René Descartes, *Discours de la méthode*, 1999, descartes.free.fr

Dunning 1994: Ted Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence*, 1994

Dunning 1994: Ted Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence*, 1994, Computational Linguistics

Engel 1880: Friedrich Engels, *Socialisme utopique et socialisme scientifique*, 2005, Aden Editions

Favier 1998: Laurence Favier, *Recherche et application d'une méthodologie d'analyse de l'information...*, 1998,Université Lyon II

Ferrucci 2004: David Ferrucci, Adam Lally, *UIMA: an architectural approach to unstructured information processing ...*, , Natural Language Engineering

Funk et al. 1983: M E Funk and C A Reid, *Indexing Consistency in Medline*, 1983

Gadamer 1960: Hans Georg Gadamer, *Vérité et Méthode*, 1976, Seuil

Gérard 2004: Christophe Gérard, *CONTRIBUTION À UNE SÉMANTIQUE INTERPRÉTATIVE DES STYLES*, 2004,Université de Toulouse Le Mirail

Gonick et al 1994: Larry Gonick, Woollcott Smith, *Cartoon Guide to Statistics*, 1994

Götz 2004: T. Götz and O. Suhre, *Design and implementation of the UIMA Common Analysis System*, 2004, IBM System Journal

Greimas 1983: Algirdas Julien Greimas, *Du Sens*, 1983, Seuil

Heidegger 1927: Martin Heidegger, *Être et temps*, 1927, Emmanuel Martineau

Ho 2007: Christian Mauceri, Diem Ho, *Clustering by Kernel Density*, 2007, Computational Economics

ISO 1985: ISO Technical Committee, *Methods for examining documents, determining their subjects and select ...*, 1985

Leclercq 1999: Nicolas Leclerq, *Logiciel libre : une volonté de transparence*, 1991,

Lengaigne et al. 2004: Benoît Lengaigne, Nicolas Postel, *Arrow et l'impossibilité : une démonstration par l'absurde*, 2004

Loti 1879: Pierre Loti, *Aziyadé*, 1999,

Loti 1991: Pierre Loti, *Aziyadé suivi de Fantôme d'Orient*, 1991, Gallimard

Mai 2000: Jens-Erik Mai, *The subject indexing process: an investigation of knowlege representation*, 2000,University of Texas

Mallery, Hurwitz 1987: John C. Mallery and R. Hurwitz and G. Duffy, *Hermeneutics: from textual explication to computer understanding*, 1987

Marcotorchino, Michaud 1979: F. Marcotorchino, P. Michaud, *Modèles d'optimisation en analyse des données relationnelles*, 1979, Mathématiques et Sciences Humaines

Marcotorchino, Michaud 1983: F. Marcotorchino, P. Michaud, *Agrégation de similarités en classification automatique*, 1982, Revue de Statistique Appliquée

McAffer 2006: Jeff McAffer, Jean-Michel Lemieux, *Eclipse rich client platform*, 2006,

Michaud 1987: Pierre Michaud, *Condorcet, a man of avant-garde*, 1987

Missire 2005: Régis Missire, *SÉMANTIQUE DES TEXTES ET MODÈLE MORPHOSÉMANTIQUE DE L'INTERPRÉTATION*, 2005,Université Toulouse II Le Mirail

Moore 2004: Robert C. Moore, *On Log-Likelihood-Ratios and the Significance of Rare Events*, 2004, Conference on Empirical Methods in NLP

Petit 1983: J.L. Petit, *Généralisation de la méthode des partitions centrales*, 1993, Revue de Statistique Appliquée

Pincemin 1999a: Bénédicte Pincemin, *Sémantique interprétative et analyses automatiques de textes : que devie...*, 1999, Sémiotiques

Pincemin 1999b: Bénédicte Bommier-Pincemin, *Diffusion ciblée automatique d'informations*, 1999,Université de Paris IV

Pugh 2004: Glendon Ralph Pugh, *An Analysis of the Lanczos Gamma Approximation*, 2004,University of British Columbia

Rastier 1987: François Rastier, *Sémantique Interprétative*, , PUF

Rastier 1989: François Rastier, *Sens et Textualité*, 1989, Hachette

Rastier 1996: François Rastier, *Sémantique Interprétative*, 1996, PUF

Rastier 1999: François Rastier, *DE LA SIGNIFICATION AU SENS - POUR UNE SÉMIOTIQUE SANS ONTOLOGIE*, 1999

Rastier 2001a: François Rastier, *Arts et Sciences du texte*, 2001, Formes Sémiotiques

Rastier 2003: François Rastier, *Herméneutique et Linguistique : dépasser la méconnaissance*, 2003, Revue Texto

Rastier 2006: François Rastier, *Formes sémantiques et textualité*, 2006, Langages

Rastier 2006b: François Rastier, *FORMES SÉMANTIQUES ET TEXTUALITÉ*, 2006, Langages

Régnier 1983: Simon Régnier, *Sur quelques aspects mathématiques des problèmes de classification automati*, 1983, Mathématiques et Sciences Humaines

Rossignol 2005: Mathias Rossignol, *Acquisition sur corpus d'informations lexicales fondées sur la sémantiqu...*, 2005,

Salton, Yang 1975: Gerard Salton, C.S. Yang, *A vector space model for Automatic indexing*, 1975, Communicationof the ACM

Saussure 1916: Ferdinand de Saussure, *Cours de linguistique générale*, 1995, Payot

Schelling 1799: Friedrich Wilhelm Joseph von Schelling, *Introduction à l'ssquisse d'un système de philosophie de la nature*, 2001, Livre de poche

Shawn 2006: Shawn Martin, *An Approximate Version of Kernel PCA*, 2006, Conference on Machine Learning and Applications

Sokal 1996: Alan Sokal, *A Physicist Experiments With Cultural Studies*, 1996, Lingua Franca

Tanguy 1997: Ludovic Tanguy, *Une approche linguistique du TALN : la sémantique interprétative*, 1997,Université de Rennes I

UIMA 2006: The Apache UIMA Development Community, *UIMA: Overview and setup*, 2006

Vapnik 1998: Vladimir N. Vapnik, *Statistical Learning Theory*, 1998, Wiley, Inter-Science

Wittgenstein 1921: Ludwig Wittgenstein, *Tractatus logico-philosophicus*, 1993, Gallimard

Wittgenstein 1953: Ludwig Wittgenstein, *Recherches philosophiques*, 2005, Gallimard

Yarowski 1995: David Yarowsky, *Unsupervised word sense disambiguation rivaling supervised methods*, 1995, Association for Computational Linguistics

Zarader 2005: Marlène Zarader, *Compréhension et interprétation dans l'herméneutique de Gadamer*, 2005

: , ,