UNIVERSITÉ PARIS DESCARTES
UFR DE MATHÉMATIQUES ET INFORMATIQUE

THÈSE DE DOCTORAT

présentée par

Zoulikha BELLIA HEDDADJI

pour obtenir le grade de docteur de l'Université Paris Descartes

Spécialité informatique

Modélisation et classification de textes. Application aux plaintes liées à des situations de pollution de l'air intérieur

Le jury est composé de :

Rapporteurs :

Christine LARGERON	Professeur, Université Jean Monnet (Saint-Étienne)
Djamel Abdelkader ZIGHED	Professeur, Université Lumière (Lyon)

Examinateurs :

Denis CHARPIN	Praticien Hospitalier, Hôpital Nord de Marseille
Séverine KIRCHNER	Responsable du Pôle Air Intérieur (CSTB)
Alain NICOLAS	Sami de Liège
Georges STAMON	Professeur, Université Paris Descartes
Gilles VENTURINI	Professeur, Université François Rabelais (Tours)
Nicole VINCENT	Professeur, Université Paris Descartes

Table des matières

Introduction

Pendant longtemps, la propagation des allergies et d'autres problèmes de santé a été imputée essentiellement à la qualité de l'air extérieur sans trop s'inquiéter des conséquences sanitaires que pouvait avoir la qualité de l'air respiré dans les lieux fermés, à usage d'habitation notamment, où nous passons la majorité de notre temps. Les premières actions menées dans le domaine de l'air au sein des lieux de vie, sont récentes et pour la plupart encore en cours[1], l'expertise est par conséquent nouvelle. Il existe des experts spécialistes dans le domaine de la microbiologie, d'autres sont ingénieurs en ventilation ou ingénieurs chimistes. L'expertise dans le domaine « air intérieur » proprement dit n'existe pas encore en tant que domaine de spécialité à part entière. Par ailleurs, les plaintes liées aux malaises ressentis à l'intérieur des ouvrages de construction (locaux de travail, écoles, domiciles, etc.) où nous passons près de 80% de notre temps parviennent en nombre important aux organismes en charge de les traiter et sont le plus souvent laissées sans réponse.

Dans cette thèse, nous nous proposons d'étudier le niveau de faisabilité d'une approche automatique que nous allons définir pour apporter des solutions au traitement des plaintes écrites en langue naturelle. Nous évaluons le niveau de faisabilité de notre approche automatique, constituée d'une application informatique, en confrontant les résultats du système que nous proposons, aux points de vue des experts du domaine, en nous basant sur un corpus de plaintes. Par « solution » nous entendons une précision de la nature du problème de pollution domestique exprimé dans le texte de la plainte ainsi qu'un ensemble d'options de gestion et d'actions correctives permettant de réduire les effets du problème sanitaire cité.

La démarche que nous adoptons dans cette étude est fondée sur une hypothèse d'homogénéité des motifs des plaintes du domaine de la pollution intérieure. Cette hypothèse, conforme à l'observation de notre corpus, doit être vérifiée auprès des spécialistes afin d'organiser l'ensemble des plaintes au sein de scénarios. En effet, les experts que nous avons rencontrés dans le cadre de cette étude, n'étaient pas en mesure d'affirmer l'existence d'une régularité des thèmes de pollution domestique. Une fois cette hypothèse vérifiée, des solutions génériques réalisées par les experts du domaine, sont rattachées aux scénarios correspondant au contexte de la solution en question. Le principe de notre approche est de mettre en évidence le scénario auquel appartient

[1] En France, la première campagne logement est celle de l'Observatoire de la Qualité de l'Air Intérieur (OQAI). Lancée en 2003, la mission de l'OQAI consiste à mieux cerner la pollution intérieure et ses effets à travers les mesures réalisées dans 600 logements.

la plainte nouvelle à traiter et d'assigner à cette dernière la solution attribuée au scénario désigné.

Les plaintes considérées sont écrites en langue naturelle. Le domaine scientifique apportant le plus d'éléments de réponse à ce problème général est le domaine du traitement automatique de la langue et plus précisément les systèmes de recherche d'information. Ce travail est donc une démarche de modélisation de l'information et de recherche d'information appliquée au domaine particulier de la pollution domestique représenté par des plaintes liées à des situations de crise sanitaire.

Ce mémoire se compose de six chapitres et respecte le plan suivant :

- Le chapitre 1 expose le contexte actuel en matière de pollution domestique dans lequel notre système évolue. Les polluants existant dans nos environnements intérieurs connus à ce jour, leurs sources potentielles, leurs impacts sur la qualité de l'atmosphère et sur la santé des individus sont exposés. Ce chapitre présente également des études qui se sont intéressées à l'analyse des plaintes des particuliers (spécifiques[2] et générales) ainsi qu'à leur suivi au niveau des services sollicités. À la fin de ce chapitre, nous citons des réalisations, inscrites dans le domaine de la pollution de l'air dans les logements, intervenant à différents niveaux dans le cadre du processus de réponse aux plaintes.

- Le chapitre 2 expose le principe des systèmes de recherche d'information, notamment ceux agissant sur les textes. On y présente les étapes classiques du processus de traitement automatique de la langue. Les systèmes de recherche implémentant les modèles d'appariement inter-textuel sont abordés dans ce chapitre également.

- Dans le chapitre 3, nous présentons le langage XML et son niveau d'importance pour mettre en évidence l'intérêt des initiatives d'adaptation des systèmes de recherche aux corpus structurés et semi-structurés. Quelques études portant sur l'adaptation des modèles de recherche classiques, à la sémantique gérée au moyen de ressources externes sont exposées dans ce chapitre.

- Un panorama des ressources sémantiques les plus utilisées dans le domaine du traitement automatique de la langue est exposé dans le chapitre 4. Les mesures de similarité correspondant aux différentes ressources étudiées sont également présentées.

- Nous présentons ensuite dans le chapitre 5, le principe de notre approche pour la résolution automatique des plaintes écrites. Nous réalisons une étude du corpus des plaintes, ainsi qu'une analyse du raisonnement des experts en charge des diagnostics de l'air dans les ouvrages de construction. Dans ce chapitre nous exposons nos contributions dans le cadre des systèmes de recherche, notamment par rapport à : l'application de la sémantique, l'adaptation à la structure ainsi que la définition d'un nouveau modèle de recherche fondé sur

[2] Liées à la qualité de l'air intérieur.

la théorie du signal. Nous proposons aussi une méthode de construction automatique des scénarios nécessaires à la réalisation de l'assignation des solutions.

- Le chapitre 6 et le dernier, présente les résultats des expérimentations menées à partir d'un échantillon représentatif des plaintes. Ces expérimentations concernent les comparaisons des différents modèles de recherche implémentés par rapport notamment à leur applications sémantiques. Nous mettons en évidence le niveau d'accord entre les scénarios automatiques et les observations des experts établies à partir d'un corpus. Pour l'ensemble des systèmes implémentés nous exposons les taux de réussite des assignations de notre prototype relativement aux propositions des experts. Enfin, nous présentons les résultats de l'analyse statistique témoignant du niveau de corrélation entre la taille des requêtes et l'efficacité de notre système selon la méthodologie de recherche appliquée.

Chapitre 1

Plainte « air intérieur »

1.1 Introduction

Dans ce chapitre, pour présenter la notion de « plainte air intérieur », nous apportons d'abord une définition de la thématique « air intérieur » (section 1.2.1). Ensuite, nous faisons le lien entre la plainte au sens large et les réclamations possibles suite à des situations de pollution de l'air dans les lieux de vie (section 1.2.2). Pour présenter plus précisément la « pollution intérieure » nous citons subséquemment les principaux polluants existant dans les ambiances des sites clos, leurs éventuelles sources constituant le bâti (section 1.3) ainsi que leurs conséquences sur le bien-être et la santé des occupants (section 1.4). Après avoir défini ces différents éléments en lien avec la « plainte air intérieur », des études concernant l'analyse des caractéristiques des plaintes et de leurs circuits de gestion au sein des organismes dans la pratique sont présentées (section 1.5). Les caractéristiques d'une plainte peuvent être notamment leur nombre, leurs motifs (ventilation, animaux parasites, bruit, etc.) et leurs aspects possibles (sanitaire, juridique, etc.). Les circuits de gestion des plaintes concernent les pratiques mises en oeuvre pour y répondre. De manière concrète, ces dernières sont disparates d'une structure à l'autre. Nous présentons néanmoins dans ce chapitre une norme internationale dédiée à la gestion des réclamations ainsi qu'un guide de bonnes pratique pour la gestion administrative des plaintes air intérieur (section 1.6). Après que l'on s'est intéressé aux caractéristiques des plaintes et à leurs circuits de réception nous présentons dans la section 1.7 des réalisations permettant de répondre aux plaintes. Un standard Européen, une norme et un système expert tous consacrés à la réalisation des diagnostics de la qualité de l'air intérieur sont décrits à la fin de ce chapitre.

1.2 Définitions

Les exemples récents dans la presse témoignent d'une importante recrudescence des plaintes en lien avec l'air des lieux de vie (écoles, bureaux, hôpitaux, logements, etc.). Dans le contexte de la thèse, on s'intéresse exclusivement aux logements. Notre choix s'est porté sur les ouvrages de construction à usage d'habitation et cela par rapport au temps important passé par la population

au sein de ces lieux de vie, et également par rapport à la nécessité de la prise en compte des personnes sanitairement sensibles (enfants, personnes âgées, etc.).

1.2.1 Notion de « air intérieur »

La notion d'« air » signifie le milieu gazeux que nous respirons. Dans le cadre de notre thèse, et comme nous venons de le spécifier, le terme « intérieur » est utilisé pour désigner les locaux non industriels à usage d'habitation (maison individuelle et appartement). L'association de ces deux termes « air intérieur » (abrégé souvent en AI) exprime dans le contexte de notre travail l'air que nous respirons au sein de nos logements.

1.2.2 Précisions au sujet de la notion de « plainte air intérieur »

De manière générale, une plainte dénonce une nuisance qui est un ensemble de facteurs d'origine technique ou sociale qui rendent la vie pénible ou malsaine [46]. La plainte « air intérieur » (ou bien plainte « habitat ») est la dénonciation d'une gêne d'ordre visuel, olfactif ou sanitaire exacerbée en milieu intérieur. Par conséquent, les demandes de renseignement abondantes concernant les effets éventuels du bâtiment, de ses équipements ou des activités de vie quotidienne sur la qualité de l'air et la santé ne sont pas prises en compte dans le cadre de cette étude.

Dans le contexte précis de cette thèse, nous entendons par le mot « plainte » le courrier manuscrit ou électronique adressé par un particulier aux institutions en charge d'enquêtes sur la qualité de l'air dans les logements suite aux intolérances sus-citées. La plainte désigne pour nous également une reprise écrite d'un entretien téléphonique avec un occupant recueilli par un interlocuteur d'un organisme en charge d'accueillir les plaintes et complété par un constat des lieux, établi par le diagnostiqueur sur site. Par le mot « plaignant » nous entendons le particulier occupant le logement, concerné par le problème de la pollution de l'air et émettant la plainte sous sa forme écrite ou orale. Dans la partie 1.5 nous réalisons une analyse détaillée sur les caractéristiques de la plainte air intérieur dans la pratique et des différents niveaux de sa prise en compte par divers services de l'état à travers un panorama d'études réalisées dans le domaine.

L'altération du bien-être dénoncée au moyen de ces plaintes est due essentiellement à des facteurs de risque[1], présents dans les habitations, et que nous présentons dans la section suivante.

1.3 Les polluants de l'air intérieur et leurs sources potentielles

Les populations sont exposées à la pollution intérieure aux travers des différentes voies d'expositions : *ingestion*, par *contact cutané* et par *inhalation*. L'exposition à la pollution de l'air des logements est associée notamment à l'état de l'atmosphère extérieure, aux différentes sources de

[1] Le facteur ou acteur de risque est une source de risque. Nous utilisons cette expression pour désigner les polluants intérieurs ainsi que les activités entraînant la pollution des lieux de vies (tabagisme, utilisation des parfums d'intérieurs, etc.) .

contamination intérieures sans oublier les occupants eux-mêmes, ainsi qu'au niveaux d'aérations des locaux. Les sources de pollution se regroupent en deux catégories d'éléments polluants : les « facteurs structurels » réunissant tous les éléments constituant le bâti ainsi que les « facteurs comportementaux » relatifs aux activités de l'occupant [79]. Dans la littérature spécialisée en matière de AI la classification des polluants se fait traditionnellement en fonction de son origine : chimique, biologique et physique. Nous suivons cette démarche pour présenter les polluants les plus fréquemment relevés dans l'air des espaces fermés d'habitation ainsi que leurs provenances. Pour présenter les contaminants des trois classes de polluants nous nous basons sur un standard d'harmonisation à échelle européenne de l'investigation de l'habitat en matière de pollution intérieure (section 1.7.2) [6]. En effet, ce standard élaboré en 2005 désigne l'ensemble des paramètres nécessaires aux prélèvements dans le cadre des investigations des logements.

1.3.1 Les polluants chimiques et leurs sources

Les différents paramètres chimiques nécessaires à l'analyse dans le cadre d'investigations des locaux à usage d'habitation établis par le guide concernent les composés organiques. Ces derniers peuvent être volatils (COV), semi-volatils (SCOV) ou non volatils. Dans le guide considéré, le monoxyde de carbone n'est pas pris en compte. Nous devons néanmoins en tenir compte par rapport à la fréquence des intoxications oxycarbonées enregistrées en France. En effet, une étude récente dans le cadre du plan national santé-environnement (PNSE)[2] 2004-2008, a rapporté une moyenne de 6000 cas d'intoxication par an avec une moyenne de 300 décès à déplorer.

Le CO

- Définition : le monoxyde de carbone (ou oxyde de carbone) est un composé chimique de formule CO. C'est un gaz difficile à détecter par l'occupant car incolore et inodore.

- Ses sources : le monoxyde de carbone est produit lors de la combustion incomplète des combustibles fossiles (pétrole, gaz, charbon) dans les chaudières, les chauffe-eau ou les moteurs automobiles par exemple. Dans les milieux intérieurs, il est aussi généré par la fumée de tabac[3]. Il est produit essentiellement en cas d'installations défectueuses ou mal entretenues de chauffages [2], en cas d'usage d'équipements non-prévus pour l'intérieur (brûleurs, équipement de chantiers, etc.) le tout associé à une mauvaise ventialtion.

Les COV

- Définition : les « composés organiques volatils » désignent des substances d'origines et de propriétés très diverses, naturelles ou non, dont le point commun est la capacité de s'évaporer à la température ambiante et se répandre ainsi dans l'air [2].

[2] www.sante.gouv.fr/htm/dossiers/pnse/sommaire.htm
[3] fr.ekopedia.org

- Les sources des COV : tout produit émanant une odeur (notamment les « odeurs de neuf ») dégagent des COV. Les COV ne sont pas toujours odorants. Les principales sources des COV sont : les panneaux de particules, les bois agglomérés ou contre-plaqués, certains textiles d'ameublement ainsi que des revêtements synthétiques (moquette, planchers, faux plafonds, etc.). Les COV sont émis par les peintures, les vernis[4] et les produits d'entretien des bois. En matière d'activités, le nettoyage, le bricolage, la cuisine des aliments et toutes les combustions (surtout le tabagisme) produisent des COV.

- Ses variétés : dans le bâtiment, plusieurs centaines de COV peuvent exister. Le formaldéhyde est très présent notamment dans les bois agglomérés, dans les textiles, dans les mousses isolantes, les produits cosmétiques, etc. Les composés organiques volatiles considérés par le guide de standardisation comprennent : les aldéhydes, les composés organiques volatils (COV) aromatiques et aliphatiques, les cycloalcanes, les alcools, les amines, les esters, les ethers, les cétones, les dérivés glycoliques et les terpènes [2, 6].

Les familles des composés organiques non ou semi-volatils existantes désignent : les biocides, les pesticides, les insecticides, les fongicides, les pyréthrinoides, les retardateurs de flamme, les biphenyles polychlorés (ou PCB), les composés aromatiques polycycliques, les dioxines, les phthalates, les métaux lourds ainsi que les fibres et particules. L'amiante et les fibres minérales artificielles sont deux formes de pollution domestique ayant des conséquences sur la santé.

Les fibres

- Définition : une fibre désigne toute particule d'aspect filamenteux[5] allongée aux cotés parallèles et dont le rapport longueur/diamètre est supérieur ou égal à 3.

- Variétés : il existe plusieurs genres de fibres dans la nature : d'origine végétale non comestible (cellulose, coton, lin, coco, bois, liège, etc.), d'origine animale comme la laine de mouton et d'origine minérale (issues des roches) comme l'amiante. Par ailleurs, il existe d'innombrables sortes de fibres artificielles telles que : les fibres d'origine métallique (laine d'acier, etc.), d'origine non métallique (fibres minérales vitreuses, plastiques, polyester, nylon, etc.) [2].

- Leurs sources : dans le bâtiment, les fibres se trouvent essentiellement dans les matériaux d'isolation (thermique ou acoustique), les matériaux de couvertures (des toits notamment), les revêtements (sols, murs), les matières d'aménagement (cloisons, panneaux, etc. (fibres de bois)).

[4] Les peintures « à l'eau » dégagent moins de COV.
[5] Se présentant sous forme de faisceaux.

Les fibres d'amiante

L'amiante est un minéral fibreux. Depuis 1997, il est interdit de vendre des produits contenant des fibres d'amiante. Néanmoins, on en rencontre encore souvent dans des bâtiments construits avant cette date.

1.3.2 Les polluants biologiques et leurs sources

La classe des polluants biologiques intègre : les moisissures (spores et mycotoxines), les levures, les bactéries, les allergènes (d'acariens, de chats, etc.).

Les moisissures

- Définition : les moisissures sont une variété particulière de champignons microscopiques (donc invisible à l'œil nu). Les moisissures que l'on retrouve le plus souvent dans les habitations sont : Cladosporium, l'Aspergilus, le Penicillium et l'Alternaria.

- Leurs sources : pour que les moisissures se développent, la présence d'humidité dans des atmosphères douces est nécessaire. Elle pénètrent par les ouvertures (portes, fenêtres, bouches d'aération, etc.), elles peuvent être transportée par les occupants (vêtements, cheveux, etc.) ou par les animaux et les insectes. Pour proliférer, les moisissures se nourrissent en absorbant de l'eau, de l'amidon contenus dans la matière organique, et même de petits débris alimentaires.

- Développement en milieu intérieur : dans le bâtiment, la condensation est principalement à l'origine de la multiplication des moisissures. Ce phénomène est particulièrement important dans les pièces d'eau (salle de bain (figure 1.1), cuisines, toilettes), les logements surpeuplés, etc. Par ailleurs, certaines matières au contact de l'humidité constituent un support favorable au développement des moisissures. Nous pouvons citer : certains papiers peints, certaines peintures, le papier, le carton, certains joints de silicone[6], le plâtre, les textiles mal entretenus, etc.

Les spores sont de minuscules particules vivantes d'origine sexuée et/ou asexuée qui permettent aux champignons microscopiques, les moisissures notamment, de se multiplier. Elles sont produites en très grand nombre, par exemple le champignon mérule peut produire en moyenne 3000 spores par mm^2. Les spores peuvent survivre très longtemps, plusieurs mois à plusieurs années. C'est sous cette forme que les moisissures se dispersent puis se déposent sur des supports nouveaux. Les mycotoxines sont des toxines provenant de diverses espèces de champignons microscopiques tel que les moisissures. Le terme de mycotoxine est utilisé pour décrire des métabolites présentant une action toxique à faible dose sur les animaux, par opposition aux termes

[6] Par rapport à la manière avec laquelle le joint a été posé.

FIG. 1.1 – Moisissure noire dans une salle de bain

de phytotoxine ou antibiotique utilisés pour décrire des métabolites qui présentent une action toxique à faible dose sur les plantes et les bactéries respectivement.

Les acariens

- Définition[7] : les acariens appartiennent à la classe des Arachnida, c'est à dire qu'ils sont de la même classe que les araignées et les scorpions. Les tiques, ces petites bêtes qui prolifèrent dans les bois en été, sont les plus grands des acariens. Les dermatophagoides ou les « acariens des maisons » (figure 1.2) sont les plus répandus des acariens, ils vivent dans les habitations et mesurent entre 285 μm[8] et 350 μm[9].

- Leurs sources : les « acariens des maisons » sont liés à la présence de l'homme.

- Développement : les poussières et principalement les squames humaines[10] permettent aux acariens de prospérer. Les acariens se logent essentiellement dans les matelas, les oreillers, les supports en fibres ou en tissus, etc. Par ailleurs, il existe un lien direct entre la température et le cycle biologique des acariens. En effet, plus il fait doux, plus les acariens éclosent, parviennent à maturité et se reproduisent rapidement[11].

- Leurs allergènes : un allergène est une substance, une particule ou un corps organique (atome, molécule, protéine). Les allergènes d'acariens sont issus de leurs déchets qui se présentent sous forme de débris de carapace ou de déjections. Ainsi, les acariens morts sont encore sources d'allergies.

À l'exemple des moisissures, les levures sont des espèces connues de la famille des champignons très présentes dans les milieux de vie mais résistantes (et très souvent inaccessible) aux modes de nettoyage domestique. La principale différence entre ces deux espèces de champignons est que les levures sont des organismes unicellulaires, alors que les moisissures sont des organismes

[7] Le site www.acarien.net établi la présentation la plus complète des questions au sujet des acariens.
[8] Poids de la femelle adulte.
[9] Poids du mâle adulte.
[10] Lamelles d'épiderme qui se détachent de la peau.
[11] www.omafra.gov.on.ca/

FIG. 1.2 – Acarien des maisons

pluricellulaires.

Les bactéries sont des organismes vivants unicellulaires caractérisées par une absence de noyau et d'organites[12]. Ils sont également pris en compte dans le guide en tant que contaminant d'origine biologique. En effet, les bactéries sont présentes dans les milieux de vie notamment dans les tapis, la terre des plantes d'intérieurs, etc.

1.3.3 Les polluants physiques et leurs sources

Dans cette catégorie de polluants, le standard Européen distingue les paramètres suivants :

- Radioactivité (radon, gamma)
- Champs électriques de basses fréquences
- Champs électromagnétiques de hautes fréquences
- Champs électrostatiques
- Champs magnétostatiques
- Champs magnétiques terrestres
- Champs acoustiques (bruit, vibrations)

Le radon

Un des contaminants physiques le plus souvent analysé dans le domaine de la pollution intérieure est le radon. Il est recensé dans la rubrique radioactivité des paramètres physiques pris en compte par le standard. Le radon est un gaz radioactif qui provient de la dégradation de l'uranium et du radium présent naturellement dans la croûte terrestre [80]. Il est diffusé à partir des sols et de l'eau, et se trouve par effets de confinement à des concentrations plus élevées à l'intérieur des bâtiments qu'à l'extérieur.

Les champs électromagnétiques

Un champ électrique est un champ créé par des particules électriquement chargées. Un champ magnétique est un champ de force invisible, mesurable et orientée créé par les deux pôles, nord

[12] Le terme organite (ou organelle) désigne différentes structures spécialisées contenues dans le cytoplasme des cellules des êtres vivants et délimitées par une membrane.

ou sud, d'un aimant naturel ou artificiel. Dans notre vie quotidienne les sources artificielles dominantes des champs électromagnétiques sont les lignes de transport au courant, les installations et équipements électriques, installations et appareils de communication (télévision, téléphone, etc.), etc [2].

Le bruit

Ce facteur est associé aux ambiances bruyantes comme c'est le cas des nuisances sonores provoquées par le passage des avions, par la présence de gros chantiers à proximité des habitations, par certaines enseignes publicitaires, etc.

Il existe un lien très fort entre l'environnement dans lequel nous vivons et notre état de santé. Après avoir cité les familles de contaminants les plus répandus dans les ambiances des ouvrages de construction à usage d'habitation ainsi que leurs provenances, nous présentons dans la suite les altérations que la santé des habitants pourrait bien subir en présence de ces facteurs de risque.

1.4 Les effets des polluants de l'air intérieur

Les espaces fermés dans lesquels nous passons la plus grande partie de notre temps sont très souvent à l'origine de la croissance de troubles divers chez l'homme allant du simple inconfort à des pathologies voire à des intoxications. Avant d'annoncer les conséquences sanitaires connues à ce jour et dues à certains polluants domestiques, nous commençons d'abord dans cette section en faisant une réflexion sur le sujet. Dans la suite, nous mettons en évidence les différents aspects des nuisances à caractère environnementale.

1.4.1 La nature des nuisances dues aux polluants domestiques

Longtemps, la question des nuisances environnementales a eu un caractère ambigu : doivent-elles être associées à la santé uniquement ou bien à la qualité de vie également. En effet, la frontière entre la santé et la qualité de vie des habitants a des contours assez flous. L'Organisation Mondiale de la Santé (OMS) définit la santé comme « un état de bien-être physique, mental et social total et non simplement comme une absence de maladie ou d'infirmité ». Par ailleurs, et abstraction faite de certains polluants toxiques à effets systémiques, la majorité des polluants auxquels les occupants sont exposés (le plus souvent a faible dose) dans les logements n'entraînent pas de symptômes spécifiques. Non seulement leurs effets sont diffus et combinés mais en plus ils touchent des organes différents et agissent à long terme. La réaction des personnes exposées à un certain type de polluants dépend également de la typologie génétique, la vulnérabilité constitutionnelle ainsi que l'adéquation du système de défenses de chacun (personnes âgées, nourrissons, malades chroniques, etc.). De plus, des polluants différents de l'air intérieur peuvent avoir des effets semblables. En effet, il a été constaté que d'un point de vue médical, les relations cause à effet ne peuvent pas encore être entièrement prouvées [20]. Par ailleurs, les personnes exposées, incommodées par une source intérieure, exhalant une odeur désagréable par exemple,

présentent des mécanismes physiopathologiques [13] qui ne s'expliquent pas par l'approche toxicologique classique[14] (c'est à dire en utilisant les connaissances résultant d'expériences humaines (ou animales)). Toutefois, dans les sections suivantes nous mentionnons quelques problèmes de santé et/ou de gêne qui ont tendance à s'exacerber dans le logement et à diminuer ailleurs.

1.4.2 Les problèmes sanitaires dus aux contaminants chimiques

Effets sanitaires du CO

L'oxyde de carbone pénètre aisément dans les alvéoles pulmonaires et se combine à l'hémoglobine beaucoup plus rapidement que ne le fait l'oxygène. Le système nerveux est particulièrement affecté, car il est très sensible à la privation d'oxygène. L'inhalation (par exposition chronique) de CO entraîne des maux de tête, des vertiges et des troubles sensoriels, particulièrement une diminution de l'acuité visuelle. Un autre effet du CO, dû à des modifications bio-chimiques, est le dépôt de cholestérol sur les artères, produisant des troubles cardio-vasculaires. La mort survient, suite à une exposition aiguë, par anoxie (manque d'oxygène) lorsque l'attachement du CO aux sang atteint 65 % de l'hémoglobine initiale[15]. À plus faible dose, le CO est à l'origine d'irritations des yeux, d'irritations des voies respiratoires, de maux de tête et de l'augmentation des crises chez les personnes sujettes à l'asthme.

Effets sanitaires des COV

Compte tenu de leur nombre il est difficile de savoir si les COV présentent un risque pour la santé. Certains n'ont pas ou très peu de conséquences, alors que d'autres peuvent provoquer des affections graves. Leurs effets sont diffus et imprécis par rapport à leurs interactions inévitables et leur présence et concentrations variables dans la pratique. Néanmoins, des intoxications aiguës ont été constatées en pratique, dues à des expositions courtes à des quantités importantes de COV. À la suite d'expositions prolongées à de faibles quantités, le benzène peut induire des leucémies et le formaldéhyde a été classé comme « cancérigène certain » par le centre international de recherche sur le cancer (CIRC). Il peut également entraîner des irritations des muqueuses (nez, oreille, bouche, etc.) et des yeux. Les COV peut également être à l'origine d'irritations de la peau. Ils peuvent être à l'origine de malaises, de nausées, d'états de fatigue et de somnolence. Ces états peuvent conduire à une hypersensibilité aux odeurs.

Effets sanitaires des fibres

Par rapport à leur taille, les fibres pénètrent dans les voies respiratoires. Les fibres les plus fines peuvent aller jusqu'au poumon profond, les autres selon leurs dimensions, s'arrêtent dans

[13] La physiopathologie est la discipline biologique qui traite des dérèglements de la physiologie. Cette dernière étudie notamment les interactions entre un organisme et son environnement.

[14] La toxicologie est la science étudiant les substances toxiques (ou poisons), leur étiologie (origine) ainsi que les circonstances de leur contact avec l'organisme.

[15] www.futura-sciences.com/fr/doc/t/developpement-durable/

le nez, la gorge, la trachée ou les bronches. Leurs effets sur la santé les plus répandus sont dus au contact avec la peau. Ces effets cutanés se présentent sous forme de démangeaisons et de rougeurs. Les yeux, le nez et la gorge peuvent également être irrités suite à l'exposition aux fibres. En effet, les fibres peuvent entraîner conjonctivites, écoulements nasaux, toux sèches favorisant ainsi les infections et les allergies [2, 29].

Effets sanitaires de l'amiante

Les maladies liées à l'amiante sont provoquées essentiellement par l'inhalation des fibres. Toutes les variétés d'amiante sont classées comme substances cancérigènes[16] avérées chez l'homme. L'amiante est à l'origine des cancers qui peuvent atteindre soit la plèvre[17] (mésothéliomes), soit les bronches et/ou les poumons (cancers broncho-pulmonaires). D'autres pathologies non cancéreuses, peuvent être causées également par l'amiante, notamment les épanchements pleuraux[18].

1.4.3 Les problèmes sanitaires dus aux contaminants biologiques

Effets sanitaires des moisissures

Les agents allergènes sont à l'origine de pathologies telles que les rhinites, les allergies, les infections respiratoires et pulmonaires, en particulier chez les personnes prédisposées. Les personnes prédisposées sont notamment les personnes fragilisées sur le plan immunitaire, les enfants et les personnes âgées. Les moisissures produisent des allergènes (des spores provocant des réactions allergiques ou irritantes) et dans certains cas des substances toxiques (les mycotoxines) [89]. Les allergies aux moisissures sont fréquentes, leurs symptômes sont variables et peuvent être sévères : irritation des yeux, du nez, de la gorge, toux et difficultés respiratoires, exacerbation de l'asthme chez les personnes déjà atteintes. Les mycotoxines peuvent provoquer également des intoxications alimentaires. Les mycotoxines ne sont pas toutes nocives pour l'homme. Il s'avère néanmoins que certaines sont cancérigènes et mutagènes[19] alors que d'autres peuvent être dommageables pour le foie, les reins ou le système nerveux [61]. Plus récemment, des études ont suggéré la possibilité d'autres effets des moisissures sur la santé tels que maux de tête, irritabilité, fatigue et symptômes gastro-intestinaux. La recherche sur les effets des moisissures sur la santé se poursuit [31, 89].

Effets sanitaires des allergènes d'acariens

À l'exemple des allergènes des moisissures, les allergènes des acariens issus de leur déjections principalement et présents dans les poussières sont à l'origine d'irritations du nez (rhinite), de la gorge et des yeux (conjonctivites). La manifestation la plus inquiétante est l'asthme (qui peut

[16] Classement effectué par le Centre international de Recherche sur le Cancer (CIRC) en 1976.

[17] La plèvre est une membrane lisse des cavités corporelles (thorax et abdomen).

[18] Présence de liquide dans la plèvre.

[19] Les mutations, en dehors de celles qui affectent les cellules reproductives, ne sont pas inoffensives. Si elles n'induisent pas toutes des cancers, elles constituent la première étape nécessaire vers la cancérisation.

être grave ou mortel). Un adulte sur cent présente un asthme allergique aux acariens, et 6% des 13-14 ans sont concernés. De plus, la présence d'allergènes d'acariens peut participer à certaines allergies cutanées (les poussées d'eczéma notamment).

1.4.4 Les problèmes sanitaires dus aux contaminants physiques

Effets sanitaires des champs électromagnétiques

Les effets biologiques avérés des champs magnétiques et électriques surviennent suite à des expositions intenses. Dans ce cas, les champs magnétiques peuvent induire des courants électriques circulant dans le corps humain. À ce jour, il n'existe pas de preuves convaincantes de risques sanitaires dus à de faibles niveaux de champs. En effet, les approches épidémiologiques (statistiques) ne suffisent pas pour établir un lien de causalité (présence de champs magnétique basse fréquence alors leucémie de l'enfant), puisque d'autres facteurs de risques associés à la présence des installations électriques pourraient bien expliquer la statistique.

Effets sanitaires du bruit

Le bruit peut provoquer plusieurs effets sur le développement de l'enfant. Il peut entraîner une diminution de l'intelligibilité de la parole, pour la raison que la distortion des sons par le bruit provoque une confusion des consonances [80]. Par ailleurs, les ambiances bruyantes peuvent induire des troubles du comportement. Cela peut se manifester sous la forme d'agressivité, d'irritabilité, de stress, etc.

1.4.5 Les odeurs et leurs effets

Longtemps, des nuisances dues aux odeurs ont été dénoncées par les citadins. Aujourd'hui des travaux en psychologie de l'environnement ont établi la relation entre la gêne et le stress qui a de mauvaises conséquences sur la santé. Les odeurs délétères sont un phénomène récurrent dans les plaintes. Non seulement ces odeurs sont difficiles à supporter et éprouvantes mais elles sont censées avoir un lien avec la santé. Certaines maladies sont imputées aux odeurs telles que la fièvre typhoïde [45]. Très souvent, les habitants abordent le problème des odeurs en tant que « émanations » de sources polluantes (telles que les moisissures). Les troubles dus aux odeurs ont tendance à se juxtaposer aux autres effets des polluants. Hormis le stress, les odeurs environnementales peuvent avoir d'autres effets psychologiques sur les personnes exposées, telles que : l'anxiété, les troubles du sommeil, etc.

1.4.6 Le syndrome des bâtiments malsains

Le « Syndrome des Bâtiments Malsains » (SBM) ou « Sick Building Syndrome » (SBS) en anglais, est un ensemble de symptômes qui touchent de façon permanente plusieurs personnes fréquentant un même immeuble. Le terme syndrome est utilisé pour désigner les symptômes et

cela par rapport à leur subjectivité (bien que leur réalité ne soit plus contestée[20]). Ces symptômes sont mal définis [126], comme états de fatigue, irritations des muqueuses (gorge sèche, nez sec, picotements oculaires), céphalées, divers symptômes oculaires (prurit, irritation, écoulement), sensation de peau sèche, gêne respiratoire, sifflement, congestion des sinus, difficulté à porter des lentilles de contact, goût inhabituel dans la bouche, langue et lèvres sèches, gorge serrée, nausée, difficulté à se concentrer avec mémoire diminuée, nez bouché, poitrine oppressée, éternuements, engourdissements, étourdissements, sensations d'éblouissement, etc.

Dans son article [17], De blay présente une étude détaillée sur les effets des polluants de l'habitat sur les sujets allergiques et asthmatiques. Par rapport à la prise de conscience croissante des populations aux facteurs de risque sus-cités et à l'augmentation et aux changements importants des problèmes de santé enregistré ces dernières décennies, le nombre des plaintes ayant pour origine les nuisances de l'air intérieur ne cesse de s'accroître. Il existe aujourd'hui un grand nombre de demandes de renseignements et d'investigations témoignant de l'étendue du phénomène à travers le pays. Ces demandes d'intervention sont reçues par les autorités et se présentent sous différents aspects. Très souvent, les organismes en charge de compiler ce type de doléances réceptionnent des appels téléphoniques de personnes qui présentent une intolérance à l'impureté de l'air au sein de leurs lieux de vie. Ces protestations doivent être traitées comme tout autre type de plaintes, car si elles ne sont pas traitées rapidement, les occupants prendront l'affaire très à cœur et continueront à se plaindre. Nous présentons dans la section suivante des études ayant porté sur l'analyse de différents aspects des plaintes air intérieur dans la pratique.

1.5 Présentation des plaintes dans la pratique

Dans un but d'étudier les plaintes « air intérieur » et les environnements dans lesquels elles évoluent, quelques études ont eu lieu ces dernières années. Nous présentons dans cette section une étude ayant pour but de connaître le niveau de connaissance des professionnels des domaines connexes à la pollution intérieure en matière des conséquences de l'air intérieur sur la santé. Ensuite, nous exposons l'enquête de Merlo [79], qui s'est intéressée à analyser les plaintes « air intérieur » et à apporter des renseignements sur le suivi pratique organisé au sein de différents organismes. Une étude de terrain réalisée à Toulouse [96], et exposée dans la section 1.5.3, s'est quant à elle penchée sur les éléments entraînant la plainte.

1.5.1 Les acteurs de la prise en compte des conséquences sanitaires de la qualité de l'air intérieur

La qualité de l'air intérieur est aujourd'hui un problème de santé environnementale reconnu. Le volume des problèmes associés à la qualité de l'air intérieur a augmenté avec l'utilisation de plus en plus fréquente de matériaux de construction synthétiques et l'accentuation de la pollution extérieure. L'apparition des mesures d'économie des énergies et de l'isolation acoustique

[20] www.sante.gouv.fr

privilégie la construction de logements de plus en plus hermétiques entraînant la stagnation et la re-circulation de l'air confiné.

Des professionnels issus majoritairement du milieu de la santé, de l'hygiène, des affaires sociales ou de l'urbanisme, rencontrent dans leur pratique quotidienne des situations en lien avec les nuisances dues à l'habitat. Ces acteurs sont peu nombreux à être organisés concrètement face à la demande du public en matière de conseil et de prévention. Une enquête descriptive sur la prise en compte des relations habitat-santé par les médecins, les architectes et les travailleurs sociaux [26] a été établie en 1996 sur un échantillon de professionnels du Languedoc Roussillon pour « évaluer leur niveau de connaissance, de pratique, et d'implication » dans le domaine. Cette enquête est un sondage à base de questionnaires réalisé auprès de 512 médecins qui sont à l'écoute des symptômes et du mal être des individus, de 88 travailleurs sociaux et de 110 architectes. Les questionnaires ont été adressés aux travailleurs sociaux en fonction de leur implication dans la vie des populations les moins favorisées. Les architectes ont été impliqués dans cet exercice compte tenu de leur connaissance dans le domaine de la conception des constructions et des matériaux.

L'étude menée par l'équipe de Olivo [26] et qui avait pour but de connaître « comment les relations entre habitat et santé sont-elles prises en compte par les architectes, médecins et travailleurs sociaux »[21] a rapporté que 62% des médecins, 47,2% des architectes et 42,5% des travailleurs sociaux déclarent n'avoir pas du tout été informés des conséquences sanitaires de l'air intérieur au cours de leur formation. Le taux des professionnels formés et qui sensibilisent leurs patients sur l'effet éventuel de l'habitat sur la santé est deux fois plus important que le nombre des acteurs non sensibilisés pendant leur formation initiale signalant ce phénomène à leur public (chez les médecins : 64,5% contre 35,5%). Ce travail a fait constater également que les acteurs parmi les professionnels interrogés qui connaissent le mieux le lien entre la pollution de l'air intérieur et la santé sont les travailleurs sociaux.

1.5.2 Études des caractéristiques des plaintes et de leur suivi

Différents services déconcentrés de l'état sont sollicités par les particuliers préoccupés par la dégradation de leur confort, dégradation attribuée à leur environnement intérieur. Une circulaire d'enquête datant de l'année 2005, établie par la Direction Générale de la Santé (DGS), auprès des Directions des Affaires Sanitaires et Sociales (DDASS) et également auprès des Services Communaux d'Hygiène et de Santé (SCHS) a procuré des résultats intéressants [79, 120]. Les conclusions de cette enquête à base de questionnaires ont permis de faire le point au niveau national sur la **caractérisation** des plaintes relatives à la qualité de l'air intérieur ainsi que sur la traçabilité ou le **suivi** de leur traitement [120].

Caractérisation des plaintes : le nombre d'organismes ayant répondu à cette enquête est de 70 DDASS et de 122 SCHS. Leurs réponses sont relativement disparates. Leurs conceptions de la notion d'« *air intérieur* » sont différentes.

[21] C'est le titre de l'étude.

	Pourcentage des plaintes habitat-air-intérieur				
	[0-20[	[20-40[	[40-60[	[60-80[	[80-100[
Nombre de DDASS	11	10	16	15	8
Nombre de SCHS	37	39	23	25	5

TAB. 1.1 – Pourcentage des plaintes habitat liées à l'air intérieur

Environnement intérieur		Jamais	Rarement	Souvent
Humidité	SCHS	3%	4%	93%
	DDASS	0%	3%	97%
Système de combustion	SCHS	9%	59%	33%
	DDASS	6%	51%	43%
Ventilation	SCHS	7%	18%	75%
	DDASS	7%	41%	52%
Matériaux de construction	SCHS	54%	40%	6%
	DDASS	45%	53%	2%
Produits chimiques	SCHS	44%	52%	4%
	DDASS	49%	50%	1%
Animaux parasites	SCHS	12%	20%	68%
	DDASS	0%	63%	37%
Animaux domestiques	SCHS	17%	34%	49%
	DDASS	16%	57%	27%

TAB. 1.2 – Pourcentage des motifs des plaintes pour les DDASS et SCHS

– *Volume des plaintes* : la quasi-totalité des organismes enquêtés reçoivent des plaintes « air intérieur » (seules 3 DDASS et 5 SCHS déclarent ne jamais en recevoir). Ces services ont pu donner des estimations sur le pourcentage des plaintes dans le domaine de l'habitat relatives à la qualité de l'air intérieur (tableau 1.1).

– *Motif des plaintes* : dans le domaine de l'environnement intérieur, l'humidité et les problèmes de ventilation ressortent comme étant les motifs les plus invoqués par les plaintes « air intérieur ». Le tableau 1.2 indique les motifs de plaintes les plus répétitifs selon les structures interrogées.

– *Aspects sanitaires* : les estimations des acteurs enquêtés du pourcentage des plaintes où les occupants évoquent des problèmes de santé sont estimées dans le tableau 1.3.

Suivi des plaintes :

– Système d'enregistrement des plaintes : 74% des DDASS et 50% des SCHS déclarent mettre

	Pourcentage des plaintes liées à la santé				
Fourchette de réponse (%)	[0-20[	[20-40[	[40-60[	[60-80[	[80-100[
Nombre de DDASS	5	17	14	7	8
Nombre de SCHS	26	26	22	7	10

TAB. 1.3 – Pourcentage des plaintes « air intérieur » mentionnant des problèmes de santé évoquées par les DDASS et les SCHS

en œuvre un registre informatique de recueil des plaintes. Par contre, 26% des DDASS et 28% des SCHS conservent les plaintes sous format papier. Le reste des services a déclaré procéder aux deux méthodes de sauvegarde.

- La réponse aux plaintes : la majorité des services concernés par cette enquête réserve une part de leur activité pour répondre aux plaintes (95% environ des SCHS et des DDASS). L'étude révèle également que les conseils par téléphone sont fréquents comme moyen de réponse. L'envoi de fascicules de conseils comme moyen de réponse aux plaignants est moins fréquent. Les SCHS effectuent plus souvent des visites à domicile que les DDASS.

- Circuit de réponse aux plaintes : la majorité des DDASS et des SCHSS s'associent pour répondre aux plaintes et agissent en partenariat avec d'autres services (80% des DDASS et 78% des SCHS) quand la plainte n'est pas de leur ressort. Leurs principaux partenaires cités par l'enquête et vers qui ils orientent les plaignants sont : les mairies, les associations (de consommateurs, de locataires, etc.) ainsi que les professionnels du domaine médical.

1.5.3 Enquête et indice d'évaluation de l'aspect physique et psychosocial de la nuisance reportée dans les plaintes

L'enquête menée par l'équipe de Rajon dans le cadre d'une étude sur les aspects physiques et psychosociaux des nuisances liées à l'habitat de 1987 à 1988 à Toulouse a retenu deux facteurs essentiels motivant la plainte [96]. Cette étude a analysé des plaintes en insalubrité concernant l'habitat et son environnement immédiat (voisinage). Cette étude a souligné que toute plainte en insalubrité est sous-tendue par des facteurs objectifs et subjectifs. Par exemple, l'âge du plaignant est en rapport avec les nuisances dénoncées. Bien évidemment, la tolérance à une nuisance varie en fonction de l'âge des résidants. Cette étude rapporte également que les nuisances sont fortement ressenties et dénoncées, surtout si le logement abritait des enfants en bas âge ou des personnes âgées.

Pour réaliser son analyse des aspects physiques et psychosociaux des nuisances recensées en région toulousaine et reportées dans des plaintes, l'application de l'équipe de Boussin [46] utilise

l'Indicateur de Santé Perceptuelle de Nottingham l'(ISPN)[22]. Cet indicateur d'auto-évaluation de la santé est utilisé pour connaître le niveau de perception négative de la santé chez le plaignant. Le questionnaire créé à cet effet, regroupe 38 items auxquels on peut répondre par « oui » ou par « non ». Ces 38 questions appartiennent à 6 dimensions : *tonus, douleur, réactions émotionnelles, isolement social, sommeil et mobilité physique*. Dans chaque dimension les questions positives sont pondérées par un cœfficient ; des scores correspondant à chaque dimension sont élaborés à la fin de chaque questionnaire. Plus un score est élevé, plus l'auto-perception de l'état de santé de l'occupant dans une dimension donnée est mauvaise.

1.5.4 Résultats de l'enquête menée dans le cadre de l'ISPN

D'autres études révèlent que certaines périodes de vie suscitent une vision alarmiste de l'état de santé de l'individu (litige entre le locataire et le propriétaire, attente de déménagement, divorce, chômage, etc.). L'analyse de l'équipe de Boussin [46] révèle quant à elle une corrélation significative entre l'âge et la perception de l'état de santé. En revanche, elle déclare que le sexe du sujet enquêté n'a pas d'ascendant sur la perception générale de son état de santé. Cette analyse a également établi un lien entre l'auto-perception de l'état de santé et le motif de la plainte d'une part et avec la justification de la plainte d'autre part.

- *Perception de la santé selon le motif de la plainte* : L'étude constate que les personnes qui se plaignent du bruit ont tendance à avoir une perception négative de leur état de santé dans les domaines de l'émotion et du sommeil relativement aux autres motifs.

- *Perception de la santé selon la justification de la plainte* : Les plaignants pour causes non justifiées ont une perception plus négative de leur état de santé dans les dimensions de l'isolement, de l'émotion et du tonus.

Quelque soit la motivation des plaignants, il est encore difficile aujourd'hui pour les particuliers d'identifier les structures compétentes en matière de prise en charges des nuisances liées à l'air dans les logements. De plus, le circuit de réception des doléances des habitants incommodés par l'air intérieur et leur traçabilité pour l'archivage est complexe. Dans la section suivante nous présentons des guides adaptés à l'harmonisation des modes de réception et de suivi des plaintes.

1.6 Circuit de réception

L'étude de l'équipe de Frère [45], réalisée en 2005, a rapporté que la majorité des dossiers de plaintes ne suivent pas un seul circuit. Au cours de cette étude, 7 à 8 circuits différents ont pu être

[22] Traduction française de « Nottingham Health Profile ».

identifiés avec une complexité variable. Cette étude a rapporté par ailleurs que pour traiter une plainte, entre 3 et 5 interlocuteurs peuvent intervenir pour la maîtrise du problème. Cet échange d'informations et de dossiers entre des institutions éparpillées augmente le temps de traitement. D'autre part, les institutions confrontées aux plaintes recevant des appels téléphoniques, souvent, n'en gardent pas de traces écrites. Il devient donc difficile d'avoir une représentation réelle des plaintes et de leur mode de réception.

Par conséquent, nous avons cherché à savoir si il existait des réglementations dédiées à la normalisation du processus de gestion des plaintes de manière générale. Nous avons ainsi pris connaissance de la norme ISO 10002 [3] que nous décrivons dans la partie suivante.

1.6.1 Norme ISO 10002 pour la gestion des réclamations

La norme internationale sur la gestion des réclamations ISO 10002[23] définit des référentiels pour la gestion des plaintes sous forme de lignes directrices pour le traitement des réclamations dans les organismes. Cette norme fournit des conseils en matière de mise en œuvre du processus de traitement des réclamations pour tous types d'activités, commerciales ou non. La tendance croissante des achats en ligne de services ou de produits est à l'origine de l'instruction de cette norme. Rien ne dit que cette norme sur l'instruction des plaintes ne peut s'appliquer à d'autres domaines que ceux prévus, comme notamment le domaine de la santé publique lié aux environnements intérieurs.

Pour l'harmonisation des processus de gestion des réclamations, la norme 10002 comporte un certain nombre de principes de base :

- *La visibilité* : le public doit savoir où adresser sa réclamation. Il n'est pas toujours évident de savoir où formuler sa réclamation surtout quand l'organisme en charge de traiter les réclamations possède plusieurs sites.

- *L'accessibilité* : les organismes doivent faciliter le dépôt des réclamations au moyen de formulaires et de fiches de notations des demandes de leurs clients. Ce principe indique également qu'il convient d'informer le public sur la formulation des réclamations, et ce, dans toutes les langues et sous différents formats, notamment les impressions en gros caractères, le braille ou la bande audio pour ne pénaliser aucun réclamant.

- *La réceptivité* : il est nécessaire d'accuser réception immédiatement de chaque réclamation et de la traiter promptement.

- *Frais et équité* : les frais de traitement ne doivent pas incomber aux réclamants. Il convient de juger les réclamations objectivement en respectant les parties concernées.

[23] La norme ISO 10018 a été annulée et remplacée par la ISO 10002, qui elle est identique à la norme FD X50-187.

- *Une approche centrée sur le client* : il convient que l'organisme encourage en paroles et en actions les retours d'informations, y compris les réclamations.

- *La responsabilité* : les responsabilités et les délégations doivent être établies clairement dans l'organisme en charge de traiter les réclamations.

- *Une amélioration continue* : l'amélioration continue du processus doit être un objectif constant dans l'organisme.

1.6.2 Guide de bonnes pratiques pour la gestion administrative des plaintes

Outre la norme ISO 10002, le guide de bonnes pratiques pour la gestion administrative des plaintes a été mis en œuvre par l'Inspection Générale des Affaires Sociales IGAS [34]. Il a été établi à partir des pratiques en vigueur dans certaines DDASS ou DRASS pour mettre en évidence les principes majeurs pour la gestion administrative des plaintes. Le guide comprend quatre fiches méthodologiques sur les étapes de la gestion administrative des plaintes.

- *Fiche n°1 : L'enregistrement de la plainte*
  Selon le guide de l' IGAS, il convient d'identifier un agent « centralisateur » des plaintes dans chaque service déconcentré pour ne pas perdre trace de l'information (tableau 1.4).

- *Fiche n°2 : L'accusé de réception*
  Le guide détermine trois degrés de responsabilité pour l'instruction de la plainte reçue par la DDASS ou la DRASS : de compétence exclusive DDASS ou DRASS, de compétence partagée avec d'autres organismes et un degré traduisant que la plainte est en dehors des compétences de la DDASS ou de la DRASS (tableau 1.5).

- *Fiche n°3 : Le traitement de la plainte par le service instructeur*
  Dans la fiche n°3, le guide indique les tâches à effectuer par l'instructeur et les moyens qu'il doit mettre en œuve dans le cas où une inspection sur place doit être réalisée (tableau 1.6).

- *Fiche n°4 : L'archivage du dossier de la plainte*
  La fiche n°4 du guide de l'IGAS préconise une série d'actions pour retrouver les informations relatives au processus d'instruction des plaintes une fois le traitement des dossiers achevé (tableau 1.7).

- *Les annexes*
  Les annexes du guide proposent des modèles de documents usuellement utilisés par les différents services de l'état tels que : les accusés de réception du réclamant suivant les différents degrés de compétence, les transmissions des différentes pièces aux institutions

1- S'assurer que les plaintes sont adressées au centralisateur
Faire un point régulier avec les services internes et les institutions externes susceptibles d'être destinataires de plaintes
2- Enregistrer la plainte
Assurer une traçabilité de chaque plainte en tenant un registre informatisé qui peut comporter : - Un code d'enregistrement pour chaque plainte - Les dates d'envoi et de réception de la plainte - L'identification de la structure concernée - La provenance de la plainte - Le motif de la plainte - La date d'envoi de l'accusé de réception
3- Enregistrer les étapes du traitement de la plainte
Les différentes étapes du traitement de la plainte sont également consignées dans le registre informatisé. Ce suivi nécessite : - D'organiser une transmission systématique des informations entre le service instructeur et le centralisateur - Si besoin, de relancer périodiquement le service instructeur

TAB. 1.4 – Tâches à effectuer par le centralisateur

compétentes, les courriers d'information au plaignant, les courriers de demande d'explication au responsable de la structure de traitement de la plainte ayant apporté une réponse au plaignant.

1.7 La réponse aux plaintes

Comme nous l'avons cité précédemment, les conditions de vie récentes sont à l'origine de symptomatologies nouvelles (section 1.4). Au cours de ces dernières années, de nombreux efforts sont réalisés dans ce domaine comme en témoignent la panoplie des programmes lancés récemment, notamment, l'initiative « Science, Children, Awareness, Legal instrument, Evaluation » (SCALE)[24], le projet « THADE »[25], le plan « Environment & Health Action Plan » (EHAP)[26] de l'Union européenne, les directives de l'OMS ou encore les nombreuses initiatives nationales comme le programme d'audit (sur la qualité de l'air intérieur des bâtiments) de l'Observatoire de la Qualité de l'Air Intérieur en France (OQAI). Dans l'ensemble des pays Européens, des structures sont mises en places pour analyser les habitations pour identifier les charges pol-

[24] L'initiative SCALE vise à approfondir les connaissances sur l'interaction complexe entre l'environnement et la santé.

[25] Le projet THADE a été conçu pour étudier l'impact de la qualité de l'air intérieur sur la santé et pour trouver de nouvelles mesures pouvant améliorer les conditions de vie.

[26] Le plan EHAP 2004-2010 de l'Union Européenne.

1- Identification de l'autorité compétente	
- Examiner la plainte pour identifier l'autorité compétente	
2- Accusé de réception au plaignant	
- Adresser au plaignant un accusé de réception de sa demande - Adapter le contenu du courrier comme suit :	
Plainte exclusive DDASS ou DRASS	- Date de réception de la plainte - Code d'enregistrement de la plainte - Coordonnées du service instructeur
Plainte partagée	- Date de réception de la plainte - Code d'enregistrement de la plainte - Coordonnées du service instructeur
Plainte hors DDASS ou DRASS	- Informer le plaignant du transfert
3- Transmission de la plainte	

TAB. 1.5 – Réalisation de l'accusé de réception de la plainte

1- Transmission de la plainte à la structure
- Adresser un courrier signé au directeur de la structure - Joindre une copie de la plainte - Demander une copie de la réponse établie au plaignant
2- Réponse au plaignant
-Informer par courrier le plaignant de la décision de l'autorité compétente
3- À l'issue de l'instruction de la plainte
- Analyser la réponse faite au plaignant par la structure - Décider des suites éventuelles à donner - Communiquer les éléments du dossier au centralisateur - Archiver le dossier

TAB. 1.6 – Tâches à effectuer par l'instructeur

1- Conservation du dossier de plainte
- Conserver les pièces originales du dossier dans un lieu sécurisé
2- Composition d'un dossier type de plainte
- Le courrier du plaignant - L'accusé de réception - Les courriers de demandes d'explication au directeur de la structure - La réponse écrite du directeur de la structure - Le courrier de réponse au plaignant - Toutes pièces concernant l'instruction

TAB. 1.7 – Tâches à effectuer pour l'archivage des dossiers de plaintes

luantes[27]. Leur mode de fonctionnement varie : paramètres analysés, méthodologies et appareils de mesures, procédures d'analyses des laboratoires chimiques ou mycologiques, interprétation des résultats, valeurs limites à appliquer, etc. Nous allons citer dans la suite des normes spécifiques standardisant le déroulement des enquêtes au sein des logements.

1.7.1 La norme expérimentale XP X 43-403

Cette norme est adoptée pour la mise en œuvre des audits de la qualité de l'air dans les bâtiments à usage d'habitation. Les audits de qualité de l'air effectués dans d'autres sites figurent dans d'autres normes, par exemple, la norme expérimentale codifiée « XP X 43-401 » par l'AFNOR concerne les bâtiments non industriels à usage de bureaux et locaux similaires. Le principe de la norme XP X 43-403 s'articule autour des points suivants :

- *Le domaine d'application* : Cette norme s'applique aux habitats individuels ou collectifs dans la totalité des pièces y compris les pièces à pollution spécifique (garages, cave, etc.). L'audit doit prendre en compte les matériaux de construction, les équipements intérieurs du logement et extérieurs, les produits d'entretien et de bricolage rangés à l'intérieur du bâtiment.

- *Les paramètres pouvant avoir un intérêt* : Cette norme tient compte de paramètres physiques, chimiques et micro-biologiques. En l'absence de source évidente de contaminants, la norme préconise de s'en tenir à un ensemble de paramètres d'approche simple tels que les teneurs en CO et CO_2. Par contre, l'existence d'une source potentielle de contamination dans le logement tel que la présence de revêtement des sols, de matériaux d'isolation, de panneaux de particules encollés, des taches de moisissures, etc., justifiera le choix d'indicateurs plus spécifiques.

- *L'enquête in-situ* : Elle est définie par la norme en tant qu'étape décisive. Elle permet de sélectionner les agents physiques, chimiques et biologiques pouvant être responsables des troubles et leurs sources. La norme exige la participation de plusieurs acteurs pour recueillir efficacement les informations. Ces acteurs sont : l'occupant, le propriétaire ou son représentant syndical et le responsable technique du bâtiment. Pour bien examiner les problèmes de l'air intérieur cette norme préconise de documenter, pour chaque local, les points suivants : localisation du bâtiment, description du local, description des équipements, nombre de personnes y séjournant habituellement.

- *La stratégie d'échantillonnage* : Dans le cas où des agents physiques, chimiques ou biologiques potentiellement responsables des troubles ont été identifiés, les choix des appareils et leur mise en œuvre pour les mesurages de l'audit font appel chacun à des normes spécifiques. La norme X35-202 concerne les appareils et méthodes de mesure des grandeurs physiques. La norme expérimentale XP X43-402 s'adresse à la stratégie d'échantillonnage

[27] Ces structures sont appelées communément « ambulances de l'environnement ou ambulances vertes ».

des polluants chimiques de l'atmosphère intérieure des locaux. Les prélèvements aériens des allergènes sont réalisés selon la norme X43-404.

1.7.2 Standard Européen pour l'investigation de l'air intérieur dans l'habitat

Dans le cadre de la Présidence luxembourgeoise de l'Union Européenne, un guide a été établi en 2005 dans le cadre de la standardisation et l'harmonisation à échelle européenne de l'investigation de l'habitat en matière de pollution intérieure [6]. Ce guide, élaboré par des experts en matière d'investigation de l'habitat disposant d'expérience pratique sur le terrain, concerne différents points :

1. Mode d'investigation de l'habitation : les données à récolter, coopération avec des médecins de l'environnement, etc.
2. Paramètres analysés : les substances chimiques, les analyses mycologiques et les facteurs physiques.
3. Techniques et appareils de mesure : conditions indispensables pour comparer les résultats des analyses.
4. Conditions de mesure : ventilation, température, humidité de l'air ambiant, etc.
5. Les analyses des laboratoires.
6. Valeurs d'orientation ou seuils limites : établir des valeurs guides des polluants dans les lieux clos. Ces valeurs d'orientation doivent être issues de données statistiques tout en tenant compte des connaissances toxicologiques. Ces mesures guides font défaut pour de nombreux paramètres.
7. Assainissements (solutions) à proposer : mesures d'assainissement, catalogue des matériaux problématiques. Homogénéiser la réponse aux situations problématiques.
8. Les contrôles à effectuer : ces contrôles sont primordiaux pour d'une part vérifier le diagnostic proposé par l'expert et d'autre part contribuer à l'acquisition des éléments indispensables à la mise en évidence d'une relation de cause à effet entre les polluants de l'habitation et les symptômes en question. Le standard recommande de réaliser les contrôles six ou douze mois après que les mesures de minimisation (ou amélioration) ont été réalisées.
9. Gestion et centralisation des données par pays : mise en place des structures nécessaires à la centralisation et au partage des données.
10. Formation et formation continue : cela consiste notamment à rendre possible l'accès aux séminaires organisés régulièrement dans les différents pays à tous les partenaires oeuvrant sur le secteur de la pollution intérieure.

La mise en œuvre d'un tel standard est complexe notamment par rapport aux différentes structures spécifiques, existantes dans chaque pays. Par ailleurs, les connaissances en matière de polluants (disparition, apparition, nouvelles sources, nouvelles formes, nouveaux comportements,

etc.) évoluent constamment. Les standards devront par conséquent être adapté continuellement par rapport aux nouvelles exigences. Par ailleurs, malgré la mise en place de procédures réglementaires normées pour la gestion des plaintes, les réponses institutionnelles ont besoin d'être outillées afin d'améliorer concrètement leurs interventions et satisfaire leur public. La revue de la littérature effectuée dans la thématique des environnements intérieurs, rapporte quasi-unanimement, la nécessité de mettre en place une approche pluridisciplinaire pour la compréhension des interactions entre l'habitat et la santé. Une telle alliance des compétences permettrait de donner une dynamique plus forte à la prise en compte des préoccupations des habitants en réduisant les coûts des interventions en matière de temps et de ressources. En effet, automatiser le processus de diagnostic à l'aide d'outils informatiques respectant le cadre normatif classique d'enquête est une perspective nouvelle dans le domaine. Dans la section suivante nous allons citer une réalisation qui s'inscrit dans cette démarche.

1.7.3 Système expert dédié au diagnostic de la qualité de l'air intérieur

Le Centre d'Études Techniques de l'Équipement (CETE) est un service déconcentré du ministère de l'équipement. Il s'occupe de prestations variées, comme la réalisation d'expertises et de conseils sur des domaines d'interventions multiples. Il agit notamment sur la ville et l'aménagement du territoire, les transports et la gestion des risques naturels[28]. Toujours dans un objectif de mise au point d'une méthode de diagnostic de la qualité de l'air intérieur, le CETE Nord Picardie a développé un système expert conduisant à une série de préconisations sur les gestes et attitudes à suivre pour le logement concerné [75]. La description de la situation de vie est réalisée en sélectionnant les caractéristiques adaptées à la situation courante à partir d'un ensemble de listes de choix. Rappelons que nous nous intéressons aux méthodes et aux outils permettant d'apporter des réponses aux plaintes écrites. Cependant, apparier la démarche de notre système à l'approche définie par le CETE en focalisant sur leurs principales différences nous semble être pertinent dans le contexte de notre recherche.

Les données à recueillir

Comme tout autre système expert, l'utilisateur de l'applicatif du CETE a besoin de recueillir un certain nombre d'informations à travers une série de questions. Les renseignements à acquérir se présentent sous forme de 11 familles de données :

- 1ère famille : Le contexte
  Dans ce volet on trouve des questions portant sur des informations d'ordre général comme : la date de la demande, la date de la réalisation du diagnostic, le Numéro du dossier, le département, le nombre d'occupants ainsi que des informations sur le diagnostiqueur (nom, organisme, etc.).

[28] Le CETE Nord Picardie : www.cete-nord-picardie.equipement.gouv.fr

- 2ème famille : L'environnement extérieur
  Cette rubrique comporte des informations sur le voisinage susceptible de causer une gêne, notamment la présence de routes, de végétations allergisantes, d'aéroports, etc. Le diagnostiqueur est également soumis à réaliser des prélèvements de certains paramètres extérieurs. Les paramètres devant être mesurés sont la température, l'humidité relative, le CO et le CO_2. De manière générale, les mesures physiques considérées par le système sont : les débits de ventilation, la teneur en CO et en CO_2, la température et l'humidité relative. Les autres polluants ne sont pas pris en compte en raison du coût que cela nécessiterait, et du poids de l'appareillage à transporter et à sa fiabilité.

- 3ème famille : Des généralités sur le logement
  Cette catégorie de données recense l'année de construction, le type du logement, le nombre de pièces, la présence d'annexes mitoyennes (caves, garages, greniers, etc.), les diagnostics réglementaires réalisés (plomb ou amiante), etc. Des informations générales (type, puissance, configuration, etc.) sur le système de chauffage et de ventilation sont également indispensables pour renseigner ce volet d'informations.

- 4ème famille : L'entrée du logement
  Cette section concerne les dimensions de cette partie du logement, le type de fenêtres, le système de chauffage et de ventilation, l'état des revêtements, le type de mobilier, la présence de plantes, les pratiques courantes (séchage de linge, usage d'humidificateurs, etc.) ainsi que les pathologies observées (problèmes d'humidité, de moisissures, etc.) et les taux de prélèvement des 4 paramètres (T, HR, CO, CO_2).

- 5ème à 8ème famille : Les pièces principales
  Les 5ème (pièces principales à vivre), 6ème (cuisine), 7ème (salle d'eau) et 8ème (WC) familles de données concernent les mêmes paramètres descriptifs que les éléments à renseigner pour la description de l'entrée du logement (famille N°4).

- 9ème famille : Des généralités sur les occupants
  Ce volet doit répondre à des questions relatives principalement au comportement des occupants. Comme notamment, le tabagisme, les habitudes de ventilation, la nature et la fréquence des activités domestiques « polluantes » (bricolage, peinture, etc.), la présence d'animaux ainsi que les conditions d'occupation.

- 10ème famille : Le questionnaire « occupant »
  C'est un questionnaire très détaillé concernant le profil de l'occupant, ses pathologies et ses antécédents de santé et les ressentis qu'il suspecte être en lien avec la qualité de l'air intérieur.

- 11ème famille : Les travaux
  Une liste exhaustive de questions sur les travaux pouvant être réalisés dans un logement

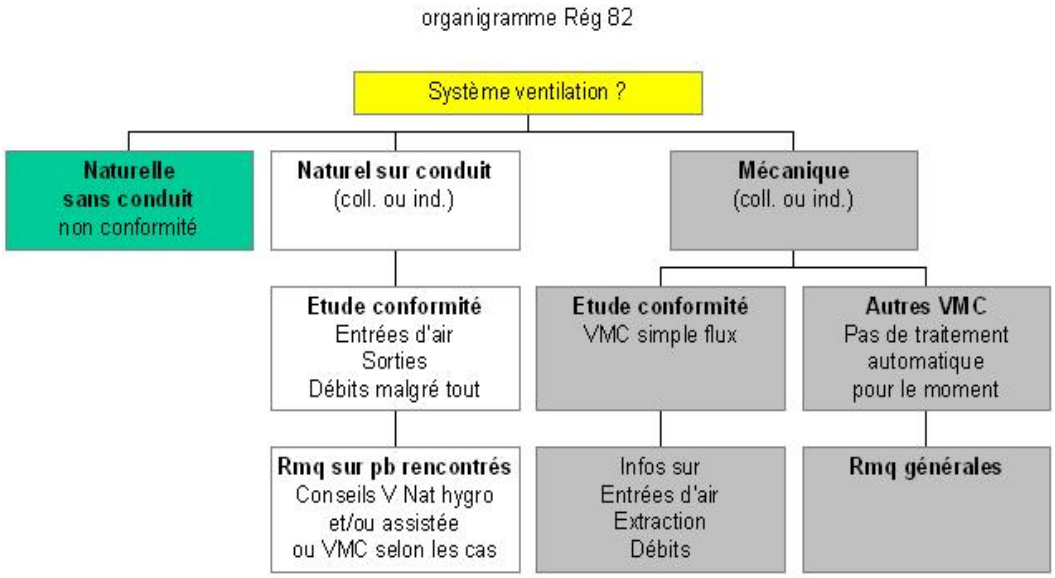


FIG. 1.3 – Organigramme réglementaire N° 82

quelconque est à compléter par le diagnostiqueur. Ce formulaire concerne les travaux d'isolation, le chauffage, la ventilation, etc.

Les principes d'analyse

La base de règles du système expert réalisée manuellement par le CETE et qui est à la base du raisonnement de l'application est axée principalement sur les exigences réglementaires applicables sur les composants du logement. Le système réalise des couplages de réglementations en cas de présence de plusieurs dispositifs à la fois. La figure 1.3 présente un extrait succinct de la base de règles administrant les situations de conformité et de non conformité des installations de ventilation.

Le protocole d'enquête

Les interventions réalisées par le CETE dans le cadre des enquêtes de diagnostic de la qualité de l'air dans les logements se font en trois temps :

- Contact préalable
  Par anticipation à l'enquête in situ, le diagnostiqueur vérifie les principales caractéristiques du logement et de la ventilation. Dans un souci de gain de temps, un questionnaire santé est envoyé à l'occupant préalablement à la démarche sur place de l'enquêteur. Ce dernier doit préparer et vérifier le matériel nécessaire à l'enquête.

- Arrivée sur site
  Le diagnostiqueur prend note de la situation du logement par rapport à son environnement

externe, il repère également les conduits d'aération et mesure les paramètres extérieurs.

- Collecte des informations dans le logement
  Les mesures physiques sont prélevées à l'aide de sondes dans les différentes pièces à vivre. Un croquis d'organisation du logement est également établi. Le logement est re-visité en entier pour une étude plus approfondie. Les pièces sont diagnostiquées séparément en remplissant les questionnaires dédiés à chaque pièce en fonction de son usage.

En moyenne le coût d'une intervention est aux alentours de 750 et dure entre 2 heures et un quart et 3 heures, ce qui correspond à une demi-journée de charge de travail. D'un point de vue complexité, le questionnaire comporte 250 questions incluant le questionnaire sanitaire renseigné par l'occupant, ce qui est considérable. L'outil de diagnostic du CETE est passé récemment au stade d'outil de pré-diagnostic. Ceci est dû entre autres aux différentes interrogations au sujet de l'exhaustivité des questionnaires et principalement aux questionnaires sanitaires et environnementaux et à la déduction automatique inhérente du système. Il a été néanmoins remarqué lors de la phase de tests établis à partir de 60 enquêtes sur site, que le système analysait mieux les situations dues à des anomalies d'ordre technique. Cependant, lorsqu'il s'agit de cas atypiques issus également de conditions techniques incohérentes, le système ne permet pas la mise en exergue du problème et sa solution. Par conséquent, le CETE a pour objectif d'alléger l'outil expert et de le dédier aux diagnostics des équipements de ventilation.

1.8 Conclusion

Au terme de ce chapitre, rappelons que nous nous sommes intéressés ici à la présentation du domaine de la pollution de l'air dans les lieux de vie. Nous avons exposé les différents polluants, leurs sources et leurs effets sur la santé de la population. Nous avons également exposé une panoplie d'études s'étant intéressées aux caractéristiques des plaintes « air intérieur » et au niveau de leur prise en charge. Et avant de conclure, nous avons présenté une réalisation informatique du CETE s'inscrivant dans le même contexte que le notre. Néanmoins, l'objet de notre travail est dans une perspective plus large que celle du système expert du CETE. De manière générale, nous souhaitons concevoir un système informatique permettant à ses utilisateurs de comprendre les problèmes à l'origine des situations de pollution intérieure et cela en accédant aux connaissances du domaine qui se présentent sous forme de problèmes et de solutions à ces problèmes. Le principe de l'approche est de simuler l'activité de jugement des diagnostiqueurs en optimisant leur modèle de raisonnement, en développant un système de recherche d'information qui s'articule autour d'une base de faits. Ces faits correspondent à un historique d'actions valides résolues sur le terrain et ayant fait l'objet d'enquêtes approfondies et d'un suivi de la part de l'organisme en charge du diagnostic. Par ailleurs, l'avantage supplémentaire de la méthode que nous proposons de mettre en place par rapport au système expert sus-cité est que la

base archive de faits s'enrichit au fur et à mesure que de nouvelles situations sont analysées et validées. À l'exemple du diagnostiqueur qui acquiert plus d'expérience à travers les nouveaux cas rencontrés, plus le recueil de faits est considérable et cohérent plus le système est performant.

Le système que nous proposons doit permettre aux usagers de s'exprimer en langue naturelle comme ils le feraient dans le cadre d'une lettre classique. Cette exigence du sujet de la thèse donne une importance très grande à la conception et à la réalisation du système. En outre la simplicité d'utilisation du système accordée par l'expression écrite naturelle et son exhaustivité (d'un point de vue utilisateur), l'intérêt de la contrainte exigée vis à vis de l'usage de la langue naturelle est multiple. Étant donné que le domaine de la pollution des environnements clos est un domaine récent et qu'il reste encore des facteurs de risques peu ou mal connus, la description des cas ne peut se limiter aux éléments de questionnaires fermés. Et pour que le système puisse résoudre des cas issus de conditions nouvelles et par conséquent profiter de son insertion en mémoire archive évolutive et réutilisable il est nécessaire que l'interface usager permette une saisie libre des observables. Par conséquent, le système souhaité s'inscrit plus exactement dans la lignée des systèmes de recherche d'information. Le principe des systèmes de recherche d'information est détaillé dans le chapitre suivant.

Chapitre 2

Les systèmes de recherche d'information

Nous abordons dans ce chapitre la définition et l'organisation d'un système de recherche d'information « SRI ». Nous commençons cette partie par une définition de la discipline de la recherche d'information (section 2.1). Les paliers de traitement automatique des textes écrits en langue naturelle et qui sont à la base des techniques pouvant améliorer les performances des systèmes de recherche d'information sont ensuite détaillés (section 2.2). L'élaboration d'index est la première étape nécessaire au SRI. Nous étudions par conséquent le principe de l'indexation dans la partie qui suit (section 2.3). L'automatisation du traitement de l'information a permis de développer des outils pour la représentation des documents et des requêtes à l'aide d'index et ensuite de comparer par appariement leurs représentations. Ainsi, les modèles de recherche implémentant les mesures d'appariement inter-textuel sont abordées dans les sections suivantes (section 2.4).

2.1 La recherche d'information

Les techniques de la recherche d'information (abrégée en RI ou IR en anglais pour « Information retrieval ») sont directement issues des sciences de l'information et plus précisément de la bibliothéconomie. La problématique majeure de cette discipline qui est ancienne et antérieure à l'apparition des ordinateurs a toujours été de permettre un accès rapide aux documents. Ces accès nécessitent plusieurs intermédiaires et de gros moyens. Il faut en effet établir un classement des documents existants et sélectionner pour chaque document un jeu de mots clés représentatif. Cette description synthétique par mots clés appelés, « index », suppose du documentaliste une connaissance suffisante de chaque ouvrage pour pouvoir en traduire le contenu.

On peut aujourd'hui dire que la recherche d'information s'est développée et est devenue un champ transdisciplinaire. La recherche d'information multimédia, par exemple, qui combine des données de différents types (texte, image, audio, vidéo) présente un panorama des différentes modélisations et interrogations possibles des documents multimédia [43]. La notion de document

dans le contexte de la RI a connu également une extension au domaine des bases de données dédiées à un accès local, ou bien mises en réseau reliées par des liens hypertextes comme sur la toile du Web.

Cependant, lorsque la recherche d'information agit sur des textes ses frontières avec le traitement automatique du langage naturel (TALN) ou le traitement automatique des langues (TAL) deviennent perméables. En effet, le TAL qui se situe à la frontière de la linguistique et de l'informatique concerne l'application de techniques informatiques à toutes les formes du langage humain parlé et écrit.

2.2 Le traitement automatique de la langue

Parmi les champs d'action les plus classiques du TAL on peut citer : la morphologie, la syntaxe, la sémantique et la pragmatique. Ces domaines constituent un système d'analyse des textes par couches où chaque niveau apporte une analyse supplémentaire par rapport au niveau précédent. Nous citons subséquemment ces 4 paliers de traitement. Mais avant cela, nous présentons ici la définition de quelques concepts premiers nécessaires à la mise en évidence de l'intérêt de chaque niveau.

2.2.1 Notions premières

- *La polysémie*
  C'est la propriété d'un mot (ou d'une expression) qui a plus d'un sens. Le sens propre d'un mot polysémique correspond à son premier sens étymologique (d'origine) alors que le sens figuré correspond à une image abstraite symbolique (métaphore) : exemple « café » graines du fruit du caféier (sens propre) et « café » le lieu de consommation (sens figuré).

- *L*'homonymie
  C'est la relation entre des homonymes, qui sont des mots d'une langue donnée qui ont la même forme écrite ou orale mais des sens différents. Deux homonymes ont deux sens différents totalement disjoints, par rapport à la polysémie où il est question d'un sens propre et d'un sens figuré : exemple « avocat » le fruit et « avocat » le professionnel du droit.

- *L*a synonymie
  La synonymie est une relation de proximité sémantique entre des mots d'une même langue.

- *L*es anaphores
  Une anaphore grammaticale est un mot (anaphore par reprise) ou un pronom (anaphore pronominale) qui assure une reprise sémantique d'un précédent élément appelé « antécédent » afin d'éviter une répétition lexicale dans un énoncé : exemple d'une anaphore par reprise : « Anna a acheté des roses. Ces fleurs sont ses préférées ». Exemple d'une anaphore pronominale : « Marie est en Chine. Elle est ravie de connaître une nouvelle culture ».

– *L*es ellipses
  En linguistique, une ellipse est un raccourci ou une suppression de mots nécessaires dans la structure syntaxique des phrases. Elle est intéressante puisqu'elle permet d'éviter la lourdeur dans un discours. Par exemple « Je ne lui dit rien. Trop bavard. Faut s'en méfier. »

– *L*e morphème
  Le morphème est la plus petite unité lexicale ayant un sens spécifique, c'est-à-dire que chaque morphème est indivisible tout en ayant un sens particulier. Les morphèmes sont isolés par segmentation : « inutilement » est constitué par le préfixe « in- », la racine « utile » (adjectif), et le suffixe adverbial « -ment ». Tous les mots simples du français ne peuvent être segmentés ; les éléments isolés n'auraient pas de sens. Par exemple : on peut pas segmenter « machine », « vapeur », etc.

– *L*es allomorphes
  Un allomorphe est une variante fléchie d'un morphème (pluriel, féminin, terminaisons verbales). Par exemple, « utiles » pour le mot « utile ».

2.2.2 L'étude morphologique

La première étape de l'étude morphologique est la segmentation. D'un point de vue analyse automatique, un texte est une chaîne de caractères. Les chaînes de caractères sont segmentées en unités minimales plus connues par les développeurs sous leur appellation anglaise « tokens ». Les séparateurs par défaut sont les espaces, d'autres choix peuvent être plus adaptés en fonction de la représentation des textes. Tous les caractères non alphabétiques peuvent être paramétrés en tant que séparateurs comme les tirets, les apostrophes, etc.

L'analyse lexicale est la deuxième étape. Elle consiste à attribuer aux mots (ou aux tokens d'un point de vue machine) une forme de base et nécessite un lexique répertoriant la liste des mots de la langue. Elle est constituée plus précisément de l'analyse de la morphologie flexionnelle et de l'étude de la morphologie lexicale. Cette dernière consiste à étudier la forme des mots construits (composés de plusieurs morphèmes). Elle est complémentaire à l'étude de la morphologie flexionnelle, qui étudie la variation (ou la flexion) des mots suite aux conjugaisons, déclinaisons (en fonction du Genre, le Nombre, etc.).

De manière générale, la morphologie étudie la composition des mots. La composition des mots se fait à partir de plus petites entités appelées morphèmes. Il existe deux aspects permettant de distinguer le morphème. Un aspect flexionnel : les morphèmes sont distingués en fonction des caractéristiques de conjugaison, du genre et du nombre des mots. Le mot « voisins » comporte par exemple deux morphèmes : voisin et le pluriel (morphème grammatical). Un aspect dérivationnel : permet de mettre en évidence les morphèmes en fonction du premier mot duquel le mot concerné a été dérivé. (par exemple fleurette et fleurir sont dérivés du mot fleur).

La morphologie en français notamment est suffisamment maîtrisée pour avoir permis la conception de logiciels qui produisent des racines ou des lemmes comme Flemm [85]. Ces analyses consistent à extraire de la forme initiale du mot une forme de base unique. Deux techniques existent pour l'obtention de la forme de base d'un mot : la racinisation et la lemmatisation.

La racinisation ou le « stemming »
C'est la méthode la plus utilisée par les moteurs de recherche qui ont pour but, le plus souvent, d'élargir le champ de la requête. Cette technique ramène chaque terme à une racine privée de ses suffixes (ou « stem »). L'affectation d'un mot à une racine peut être établie par des postulats adaptés aux lexiques considérés, ce que nous verrons par la suite dans la partie (section 5.9). Une autre classe de méthodes réalise la racinisation à l'aide d'un ensemble de règles et d'exceptions. Chaque suffixe connu est traité par une règle, tandis que les formes contenant le suffixe traité par la règle et pour lesquelles la règle ne s'applique pas sont recensées en tant qu'exceptions à cette règle.

La lemmatisation
Parallèlement à la racinisation, elle consiste à reconnaître la forme canonique d'un mot à partir d'une de ses formes dérivées (verbe conjugué, forme au féminin, forme au pluriel). Par exemple : vitraux sera reconnu comme le pluriel de vitrail.

Des études ont rapporté que la racinisation et la lemmatisation sont deux méthodes quasi-équivalentes pour les langues de morphologie simple comme l'anglais, mais la lemmatisation est significativement plus efficace pour les langues de morphologie complexe telles que le français. Il a été également constaté que pour certaines langues qui ont une morphologie très riche, la lemmatisation est une méthode difficile à appliquer.

A ce premier niveau, plusieurs problématiques doivent être considérées. En premier lieu, concernant la phase de segmentation, l'inconvénient majeur de l'utilisation des tokens-mots réside dans le fait que ces éléments ne correspondent pas toujours à des mots. Les unités obtenues par segmentation sont recomposées pour identifier différents nouveaux mots composés possibles. Cependant, certaines de ces re-compositions peuvent être ambiguës : par exemple dans l'expression "pomme de terre cuite » doit-on mettre en évidence le mot composé « terre-cuite » ? Ensuite, au sujet de la morphologie lexicale qui s'intéresse à attacher une forme au lexique de base, il n'est pas (encore) possible de réunir tous les mots possibles d'une langue dans un lexique. Les noms propres (de personnes, de villes, etc) constituent un inventaire ouvert[1]. De plus, en fonction des besoins, des mots nouveaux sont créés régulièrement par abréviation, siglaison, etc..

Par ailleurs, l'ambiguïté purement lexicale se situe à un niveau local (au niveau du mot). Le

[1] Le site Internet de l'Institut national de la statistique et des études économiques (Insee) possède des fichiers mis à jour chaque année recensant certaines entités. On y trouve la liste des communes, des cantons, des régions de la métropole et des départements outre-mer ainsi que la liste des pays et territoires étrangers.

mot « souris » peut désigner un animal ou une forme du verbe « sourire », « couvent » peut être un verbe ou un nom également. Connaître la catégorie grammaticale des mots à l'aide d'une analyse syntaxique permettrait dans ce cas de lever les ambiguïtés résiduelles.

2.2.3 L'étude syntaxique

Les relations syntaxiques entre les mots dans les phrases représentent une information supplémentaire que les systèmes de RI ont tout intérêt à prendre en considération. L'analyse syntaxique permet de reconnaître les catégories grammaticales des termes (Nom, Verbe, Adjectif, etc.). Grâce à cette information supplémentaire un certain nombre d'ambiguïtés de type homonymie peuvent être levées, et ce, en considérant le contexte syntaxique général des phrases (cheminée participe passé ou nom ?).

L'analyse syntaxique s'intéresse également à la reconnaissance des groupes syntaxiques, tels que les syntagmes nominaux[2] ou syntagmes verbaux[3]. Ces unités linguistiques plus complexes permettent de lever certaines ambiguïtés. Par exemple dans la phrase « pommes-de-terre cuites » l'interrogation au sujet du mot composé terre-cuite n'a plus lieu d'être. De plus, ces entités sont moins polysémiques que les mots simples.

Les méthodes d'identification des index (ou index composites) selon des patrons spécifiques ou par co-occurrences statistiques sont très coûteuses à mettre en oeuvre ce qui est défavorable au succès d'une application de recherche d'information. De plus, une validation humaine des résultats de la segmentation automatique est nécessaire ce qui ne fait qu'alourdir l'analyse. En plus des relations grammaticales qui apparaissent directement dans un texte, il en existe d'autres plus implicites et non pas moins pertinentes à l'extraction. C'est le cas des anaphores, qui peuvent être très éloignées de leurs antécédents, ce qui rend difficile la mise en relation automatique. La résolution d'anaphores mobilise davantage de sémantique et de pragmatique. D'autres ambiguïtés résistent à l'analyse syntaxique. Par exemple de la phrase « la petite porte le voile » deux structures sont envisageables : dans la première, petite est un nom féminin, porte est un verbe, le est un article, voile est un nom masculin et dans la deuxième, petite est un adjectif, porte est un nom, le est un pronom et voile est un verbe. Connaître la signification des mots (niveau sémantique) ou la situation dans laquelle le message est émis (niveau pragmatique) permet éventuellement de lever ce genre d'ambiguïtés.

2.2.4 L'étude sémantique et pragmatique

L'analyse sémantique prend en entrée les résultats de l'analyse syntaxique. De ce fait, une bonne analyse lexicale et syntaxique diminue l'effort dans le cadre de l'étude sémantique. En effet, une bonne désambiguïsation syntaxique peut aider : on en sait davantage sur le sens de « parallèle » si l'on connaît son genre (Sur Terre, « un parallèle » est un cercle imaginaire reliant

[2] groupe syntaxique dont l'élément principal,ou noyau est un nom.
[3] groupe syntaxique dont l'élément principal,ou noyau est un verbe.

tous les lieux situés sur une même latitude, « une parallèle » est une droite.).

L'étiquetage sémantique (partie importante de l'analyse) consiste à attribuer une étiquette sémantique à un mot. Cette étiquette peut porter une information sémantique d'ordre général (agent, artefact, espace, etc.) ou être une information sémantique précise en portant le sens du mot étiqueté (par exemple « baie » : « fenêtre », « baie » : « crique », « baie » :« fruit ») ou bien (« cousin » et les étiquettes « parent » et « insecte »).

L'analyse sémantique nécessite un traitement assez fin, comme notamment, la résolution d'anaphores. Les anaphores dites « anaphores par lien sémantique » doivent être traitées à ce dernier échelon. L'anaphore par lien sémantique reprend un terme sémantiquement lié à son antécédent. L'anaphore peut être plus générique ou plus spécifique que l'antécédent en question « Marie aime le dessin. Elle s'est inscrite à cette matière au début de l'année ». L'ambiguïté traitée au palier sémantique se situe au niveau du sens des mots également. Elle est due à la polysémie ou à l'homonymie : par exemple : « il a bu un verre ».

Alors que la sémantique analyse le sens d'un énoncé en dehors de son contexte, l'analyse pragmatique (ou analyse au niveau discours), elle, ouvre une perspective plus profonde en se basant sur des connaissances extra-linguistiques. L'analyse pragmatique prend en entrée les résultats de l'analyse sémantique et vise à analyser le sens de l'énoncé dans le cadre de son contexte, « Le professeur a envoyé l'élève chez le proviseur. Parce qu'il le trouvait insupportable (« il » est le professeur). Parce qu'il lançait des boulettes au plafond (« il » est l'élève). Parce qu'il voulait le voir (« il » est le proviseur). ». L'identification du discours permet de lever l'ambiguïté inhérente. Les ellipses qui, par définition, sont dépourvues de structure grammaticale, nécessitent une analyse pragmatique afin de gérer les ambiguïtés qui y sont associées.

Les différents aspects du problème de traitement automatique de la langue naturelle cités dans le cadre des paliers d'analyse sont traités dans des réalisations séparées. Une analyse qui se prétend robuste a nécessairement besoin d'une validation humaine, ce que tous les systèmes de recherche d'information ne peuvent se permettre notamment ceux basés sur des corpus à l'échelle du Web. Les systèmes d'analyse linguistique les plus utilisés à l'heure actuelle intègrent les modules d'analyse morphologique et syntaxique.

Pour que le système que nous souhaitons mettre en place apporte une réponse à une plainte écrite il est nécessaire qu'il analyse son texte. L'idéal serait bien entendu d'appliquer une chaîne de traitement correspondant aux paliers du TAL présentés précédemment. Pour l'analyse morphologique et syntaxique nous utilisons l'outil TreeTagger adapté au français et que nous présentons dans la suite de cette section. Concernant l'aspect sémantique de la langue, nous avons choisi d'utiliser un dictionnaire en tant que ressource externe pour l'extraction du sens, ce que nous détaillons dans le chapitre 5 du mémoire. La dimension pragmatique qui se réfère au contexte, et qui est très implicite, nécessite forcément une intervention humaine par rapport aux autres niveaux de la langue pris en compte par notre système.

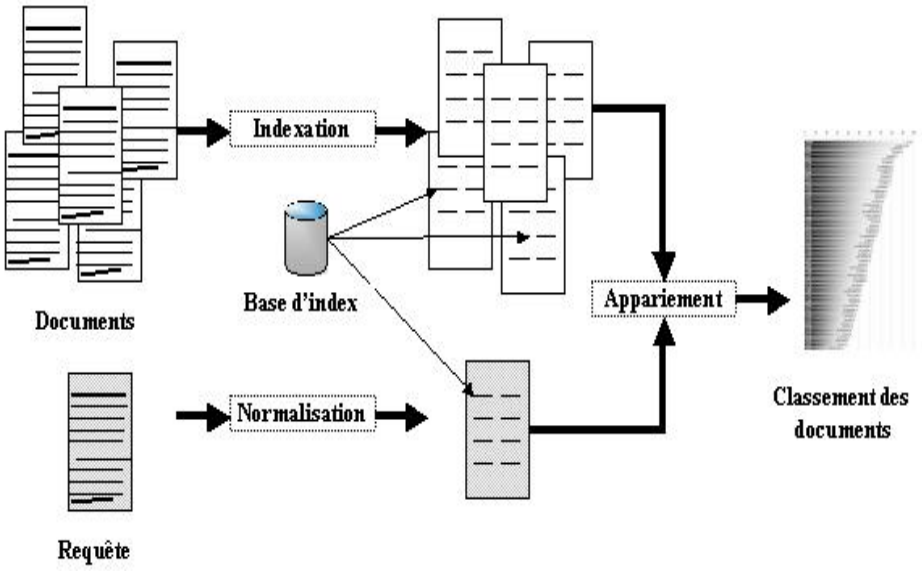


FIG. 2.1 – Exemple d'architecture d'un SRI

Au sens large, les SRI s'appuient sur deux actions principales : l'indexation des corpus (y compris les requêtes) (section suivante 2.3) et l'interrogation du fond documentaire (section 2.4) comme le montre la figure 2.1. Ces deux aspects sont au coeur des problèmes abordés par la RI.

2.3 L'indexation

La pertinence de la démarche d'indexation des documents est une première étape dans la mise au point d'un système de recherche d'information de qualité. Les normes ISO 5963, 1985[4] et NF Z 47-102, 1993[5] exposent les principes généraux de l'indexation. L'examen de quelques définitions données de cette opération fait ressortir ses principales fonctionnalités, tout en laissant place aux divers aspects qu'elle peut avoir.

Selon la norme ISO 5963, 1985
Cette norme est dédiée à l'indexation manuelle. Elle incite les services d'indexation et d'autres centres de documentation à unifier leurs pratiques d'indexation humaine. La norme présente des recommandations pour l'analyse des documents afin de repérer les identifiants significatifs, apporte des recommandations en matière de sélection des concepts fondamentaux.

Selon la norme NF Z 47-102, 1993
L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage do-

[4] Méthodes pour l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation.

[5] Principes généraux pour l'indexation des documents.

cumentaire les concepts après les avoir extraits du document par une analyse.

D'un point de vue recherche documentaire[6]
L'indexation est le fait de dresser un répertoire ou une liste, généralement alphabétique, des sujets traités, des noms cités dans un ouvrage, suivis des références aux pages, aux paragraphes, etc.

D'un point de vue lexicologique[7]
L'indexation d'un corpus est le fait d'établir un relevé complet du vocabulaire, c'est-à-dire un inventaire de toutes les formes ou des unités lexicales qui figurent dans un énoncé ou un ensemble d'énoncés déterminés. Le premier problème documentaire qui se pose pour le traitement des titres est celui de la confection du « dictionnaire des mots vides » ou « dictionnaire négatif » ou « dictionnaire d'arrêt », c'est-à-dire de la liste des mots qui ne seront pas retenus par l'ordinateur pour l'indexation.

D'un point de vue économique[8]
L'indexation est l'action consistant à lier la valeur d'un capital ou d'un revenu à l'évolution d'une variable de référence (prix, production, productivité, par exemple). Exemple : indexation d'un emprunt sur le prix de l'or, indexation des traitements sur le coût de la vie.

En considérant l'indexation dans une perspective de recherche d'information, son rôle est de distinguer les unités descriptives (les index) attribuées à la représentation des documents. Salton et Mc Gill [108] discernent deux aspects essentiels aux index. Ces unités doivent être descriptives (représentatives du contenu du document) et discriminantes (mettant en évidence ce qui distingue le contenu du document de celui d'un autre dans le corpus). Pour déterminer les index, deux références (ou langages) d'indexation existent. L'opération d'indexation libre utilise librement tous les mots d'une langue naturelle donnée, voire même des groupes de mots ou des groupes de caractères (N-grammes) [50]. Dans ce cas le langage d'indexation consiste en une liste de tous les « tokens-mots » (section 2.2.2). Par conséquent, cette liste n'est pas connue a priori. A contrario, l'indexation contrôlée utilise les entités répertoriées dans une liste de référence prédéfinie (vocabulaire ou terminologie).

2.3.1 Indexation contrôlée Versus indexation libre

Une étude de la société américaine des indexeurs ASI [9] a rapporté que la qualité de l'indexation dépendait plus du degré de représentativité des documents par les index sélectionnés plutôt que de la nature de la base d'indexation établie (libre ou contrôlée). Les résultats de Salton [105, 106] concluent à une différence non significative entre les résultats de recherche d'informa-

[6] Définition du portail lexical du site centre national de ressources textuelles et lexicales www.cnrtl.fr/portail/
[7] Définition de « Trésor de Langue Française » électronique : atilf.atilf.fr/tlf.htm
[8] Définition du portail lexical du site centre national de ressources textuelles et lexicales www.cnrtl.fr/portail/
[9] American Society for Indexers.

tion établie à partir des deux langages d'indexation. L'étude de Leonard [68] va dans le même sens. Plus précisément, Leininger [67] estime que le choix du langage contrôlé favoriserait la précision alors que l'indexation libre favoriserait le rappel. Leininger, tout comme Sparck-Jones [115] précise que l'utilisation d'un vocabulaire contrôlé est conditionnée par un thésaurus adapté à la base documentaire considérée.

2.3.2 Indexation manuelle, automatique ou assistée

Dans cette partie l'indexation est caractérisée en fonction des moyens mis en œuvre pour l'extraction des descripteurs.

- L'indexation manuelle est l'attribution d'index à un document, par un humain. Elle est précise mais requiert beaucoup de temps et une formation spécifique de l'agent indexeur (ou expert). Un consensus générique concernant la méthodologie d'indexation humaine est établi [4] mettant en évidence les étapes suivantes :

  - Analyse du texte
  - Interprétation dans le vocabulaire contrôlé
  - Relecture et révision. Cette étape est décisive, puisque c'est à ce moment que l'agent décide de maintenir ou de rejeter un index.

  Bien que les indexeurs suivent une même démarche, les résultats de l'indexation peuvent varier d'un agent à un autre.

- L'indexation automatique est l'un des premiers débouchés de l'analyse syntaxique. Elle dénote l'assignation d'index à un document par un ensemble de programmes implémentés sur machine informatique sans implication humaine (mise à part la phase de conception ou de mise en oeuvre).

- L'indexation assistée ou semi-automatique est un compromis des deux méthodes qui précèdent. C'est une indexation automatique ponctuée par des interventions d'agents humains. Ces interventions peuvent intervenir à n'importe quel moment de l'indexation automatique et consistent en des corrections ou des enrichissements de la base des descripteurs.

D'une manière générale, l'indexation est une étape coûteuse et souvent irréversible. Elle doit être évaluée avant d'être intégrée dans le SRI.

2.3.3 Éléments d'évaluation de l'indexation

Parmi les mesures qualitatives permettant d'évaluer un système d'indexation nous pouvons citer :

- L'objectivité
  L'objectivité d'une opération caractérise sa validité indépendamment de la connaissance de son réalisateur. Dans un contexte de SRI, ce critère est écarté en tant que caractéristique du système d'indexation puisque la tâche d'indexation et le fonctionnement du système de recherche d'information en général sont orientés « utilisateur ». Mais puisque la régularité de l'indexation est visée sa consistance est étudiée.

- La consistance
  Cette notion désigne le degré d'accord entre des indexations établies pour un même document par deux indexeurs ou deux méthodes d'indexation différentes, ou par un indexeur et une méthode (consistance inter-indexeurs). Et lorsqu'il s'agit de comparer des indexations établies sur un même document par un même indexeur on parle de la consistance intra-indexeur.

Un certain nombre de facteurs sont susceptibles d'influencer la variabilité de l'indexation entre indexeurs :

- Facteurs externes : l'utilisation d'un vocabulaire contrôlé ou d'un vocabulaire libre, par exemple. La consistance inter-indexeurs semble meilleure dans le cas où il s'agit d'indexation contrôlée par opposition à une indexation libre [86]. La consistance dépend également des outils utilisés (logiciels, fascicules, Internet, etc), de l'environnement (l'ordre des documents à indexer influence considérablement les résultats) et d'éventuelles contraintes temporelles exigées (dans le cadre d'un projet par exemple).

- Facteurs internes : les connaissances propres à l'indexeur (sa formation initiale, technicien ou expert, etc.).

Les mesures de consistance inter-indexeurs les plus utilisées sont la mesure de Hooper [51] et la mesure de Rolling [99]. Pour définir les principes de chaque mesure, nous désignons par A et B deux indexeurs et par AD et BD le nombre de descripteurs proposés par respectivement A exclusivement et par B exclusivement. Le paramètre ABD est le nombre de termes communs retenus par l'indexeur A et l'indexeur B (figure 2.1). La mesure de Hooper (formule 2.1) est très simple. Elle évalue l'accord entre deux descriptions en fonction de la proportion de descripteurs proposés par les deux indexeurs à la fois, sur l'ensemble des descripteurs proposés.

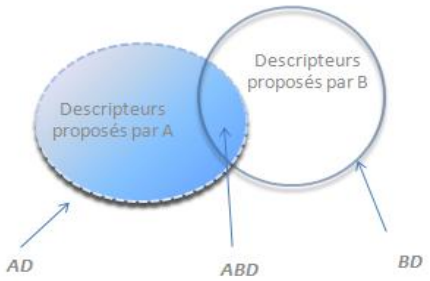


FIG. 2.2 – Répartition de descripteurs attribués par A et B pour un même document

$$Accord_{Hooper} = \frac{100 \times ABD}{AD + BD + ABD} \tag{2.1}$$

La mesure de Rolling est une variante de la mesure de Hooper. Elle attribue un poids double aux descripteurs communs par rapport aux descripteurs issus d'une divergence d'opinion entre les deux indexeurs.

$$Accord_{Rolling} = \frac{100 \times 2 \times ABD}{AD + BD + 2 \times ABD} \tag{2.2}$$

Qualité de l'indexation
L'inconvénient majeur dans l'évaluation de l'indexation est qu'il n'existe pas d'indexation de « référence » absolue appelée aussi « gold standard » pour confronter l'indexation à évaluer. Néanmoins, il existe deux méthodes d'évaluation dans la littérature :

- Méthode a priori : elle consiste à comparer l'indexation à une indexation référentielle particulière établie par un indexeur expert.

- Méthode a posteriori : c'est une validation de l'indexation par un expert.

Aujourd'hui, dans la plupart des SRI informatisés, des systèmes automatiques d'indexation linguistique sont utilisés pour identifier automatiquement les indexes des textes à partir d'analyses morphologiques et syntaxiques. Dans la section suivante, nous présentons les outils les plus fréquemment employés pour l'indexation linguistique.

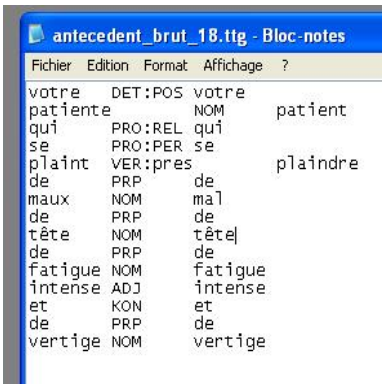


FIG. 2.3 – Exemple d'un résultat d'étiquettage réalisé par TreeTagger

2.3.4 Les étiqueteurs et les lemmatiseurs

Les étiqueteurs désignent des outils d'ingénierie linguistique dont la fonction est d'associer à chaque mot une catégorie grammaticale. En général, les étiqueteurs sont aussi des lemmatiseurs. Ces derniers sont des outils qui fournissent les lemmes des mots. Parmi les analyseurs réalisant la fonction d'étiqueteur et de lemmatiseur les plus répandus nous pouvons citer : TreeTagger et Brill. Les logiciels Cordial analyseur et l'analyseur flexionnel Flemm [85] (qui lui-même utilise les étiquettes de TreeTagger ou de Brill) sont dédiés à la lemmatisation uniquement.

*T*reeTagger :

Développé par l'institut de linguistique informatique de l'Université de Stuttgart[10], TreeTagger [109] nécessite une mise en forme spécifique des fichiers texte en entrée (un mot par ligne) (figure 2.3). Initialement dédié au traitement des textes allemands, les nouvelles versions adaptées au traitement des textes écrits dans d'autres langues (anglais, français, italien, danois, espagnol, bulgare, russe, grec, portugais, chinois et ancien français) existent maintenant[11]. En réalité, TreeTagger est un étiqueteur qui n'est pas dédié à une langue particulière. Son principal avantage est qu'il sépare les données des programmes. En effet, il se compose d'un programme principal fondé sur le principe des arbres de décision binaires et de quelques fichiers de paramètres qui dépendent de la langue à analyser. En plus de sa disponibilité gratuite, un autre avantage de TreeTagger est qu'il soit utilisable en ligne de commande, par conséquent une application donnée peut l'utiliser de manière automatique et complètement transparente.

*B*rill :

L'outil a été créé dans le cadre d'une thèse[12] à l'université de Pennsylvanie. Brill infère des règles d'étiquetage à partir d'un corpus annoté manuellement. Initialement, c'est le corpus « Wall Street Journal » qui a servi de base pour l'étiquetage et la lemmatisation des textes en anglais. Il est

[10] Institute for Computational Linguistics de l'Université de Stuttgart.
[11] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[12] Thèse de Eric Brill (d'où le nom de l'étiqueteur).

possible d'entraîner Brill sur tout type de corpus étiqueté. Par exemple, concernant le français, c'est l'INaLF[13] qui a adapté Brill à la langue en effectuant un apprentissage sur la base textuelle Frantext[14]. Une convention est nécessaire pour l'obtention des droits d'utilisation de Brill fonctionnant à partir des règles inférées de cet apprentissage.

*C*ordial :

Cordial 8 de la société Synapse-Développement est un correcteur orthographique et grammatical du français. La version « Cordial Analyseur », qu'il est possible d'acquérir à des fins de recherche, intègre entre autres fonctionnalités un étiqueteur morphosyntaxique. Cordial n'étant pas libre de droit, et non utilisable en ligne de commande, il est par conséquent moins exploité.

Après que l'on ait détaillé les principes de l'indexation, nous présentons dans la section qui suit les modèles de recherche les plus fréquemment implémentés pour la mise en évidence des rapprochements entre textes indexés de manière à interroger le fond documentaire.

2.4 Les modèles de recherche

Cette partie a pour objet de décrire les systèmes de recherche d'information les plus classiques. Selon la philosophie adoptée par les SRI pour le formalisme des documents et le mode de recherche, nous pouvons les classer dans 4 groupes :

- Ceux qui sont fondés sur l'algèbre vectorielle ;
- Ceux qui fonctionnent sur le principe de la théorie des ensembles ;
- Ceux qui reposent sur le principe des classements probabilistes ;
- Ceux qui utilisent les réseaux de neurones.

Dans la littérature, on parle également parfois d'une cinquième classe qui regroupe les modèles hybrides. Ces modèles fusionnent le principe de au-moins deux modèles issus des quatre classes principales sus-citées. Dans cette partie nous décrivons les principes de fonctionnement des quatre modèles principaux ainsi que les techniques qui s'y sont greffées pour les adapter et les compléter. Pour définir un SRI, nous allons décrire de manière précise ses composantes essentielles, dont la représentation des documents et des requêtes ainsi que leur mode d'appariement.

2.4.1 Les modèles fondés sur l'algèbre vectorielle

Nous allons citer subséquemment les modèles de recherche à base vectorielle. Nous commercerons par détailler le modèle vectoriel qui est un des formalismes de représentation les plus

[13] INaLF : Institut National de la Langue Française.
[14] atilf.atilf.fr/frantext.htm

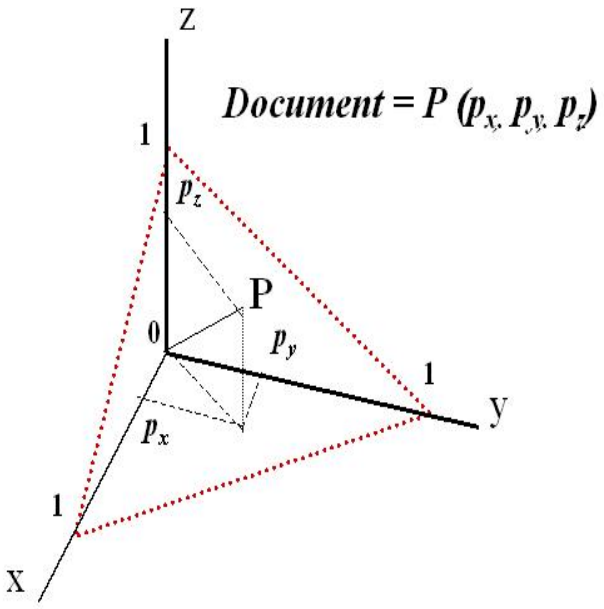


FIG. 2.4 – Exemple d'une projection 3D d'un texte

simples et les plus applicables [88]. Ensuite nous énonçons le principe du modèle « Distributional Semantics based Information Retrieval » (DSIR) et le modèle « Latent Semantic Indexing » (LSI) qui utilisent le formalisme vectoriel pour mettre en évidence la sémantique inhérente de la co-occurence des termes.

Le modèle vectoriel

Pour la représentation des documents du corpus, l'approche vectorielle utilise une représentation d'inspiration géométrique. L'idée principale est de considérer les caractéristiques décrivant l'ensemble des documents en tant que dimensions dans un espace d'information multidimensionnel. Dans la plupart des approches, les caractéristiques sont les termes d'indexation et les coordonnées des vecteurs sont les poids de ces derniers. Ainsi, les requêtes et les documents sont représentés par des vecteurs dans un même espace, et la pertinence d'un document par rapport à une requête est calculée en fonction des positions du document et de la requête dans l'espace de représentation. L'estimation de cette pertinence correspond à une distance établie à partir de son sens géométrique.

Le poids des termes peut être attribué de façons différentes. La mesure la plus connue et la plus utilisée pour les documents non structurés est le poids TF-IDF défini par Salton [104]. Le « term frequency » (TF) signifie le nombre d'occurrences d'un terme dans un document. Le « inverted document frequency » (IDF) désigne la valeur inverse du nombre de documents dans lesquels le terme considéré est présent. Pour estimer la spécificité d'un terme par rapport à un document donné, le nombre de ses occurrences ne suffit pas. Un terme commun à de nombreux

documents est moins utile qu'un terme commun à peu d'entre eux. Le score TF-IDF combine le nombre d'occurrences du terme dans le document TF (pondération locale) et sa distribution dans l'ensemble du corpus pour évaluer sa pertinence (son utilité ou sa rareté relative) IDF (pondération globale). Plusieurs formes de pondération des termes ont été proposées, et elles sont plus ou moins équivalentes. Nous présentons ici la mesure classique la plus utilisée par les modèles de recherche implémentant le principe vectoriel.

$$TF - IDF_{t,d} = TF_{t,d} \times \left(\log(\frac{|D|}{DF_t}) + 1 \right) \tag{2.3}$$

$TF_{t,d}$ désigne la fréquence d'apparition du terme t dans le document d, |D| est le nombre de documents dans le corpus et DF_t représente le nombre de documents comprenant le terme t.

En plus de la représentation des documents, la finalité des modèles vectoriels est la recherche des documents pertinents par rapport à une requête. La mesure de similarité documentaire est calculée à l'aide de la distance angulaire correspondant au cosinus de l'angle formé par le vecteur représentant la requête de l'utilisateur et le vecteur du document de la base. Intuitivement, cette fonction constitue une évaluation de la proximité thématique entre deux documents. Pour une distance angulaire inférieure à $\frac{\pi}{4}$ les deux documents sont thématiquement proches, pour une distance supérieure à $\frac{\pi}{4}$ la proximité thématique est considérée comme faible et pour une distance aux alentours de $\frac{\pi}{2}$ les deux documents sont totalement différents.

$$cosinus(r,d) = \frac{\sum_{t \in r \cap d} TF - IDF_{t,d} \times TF - IDF_{t,r}}{\sqrt{\sum_{t \in r}(TF - IDF_{t,r})^2} \times \sqrt{\sum_{t \in d}(TF - IDF_{t,d})^2}} \tag{2.4}$$

Dans le cas où un nouveau document est introduit dans le système, il devient nécessaire de recalculer tous les scores. Il s'avère néanmoins, que, lorsque le nombre de documents dans la base est élevé, la mémorisation d'un nouvel élément n'a pas un impact important sur les scores. Ils peuvent être recalculés séparément lors d'une étape spéciale dédiée à la mise à jour des archives ultérieurement. Nous verrons par la suite (dans les sections 2.4.1, 2.4.1 et 3.4), que l'avantage des représentations vectorielles, c'est qu'elles permettent l'utilisation de techniques simples, telles des extensions ou des agrégations indispensables pour la gestion et la mise en évidence de la structure des documents ou de la sémantique notamment.

Distributional Semantics based Information Retrieval

Ce modèle dont l'appellation abrégée en *DSIR* et qui signifie en français recherche documentaire à base de sémantique distributionnelle, est un modèle dérivé du modèle vectoriel classique. Il vise à faire valoir le contenu de la matrice des co-occurrences des termes, issue des corpus de référence [95]. Plus précisément, le but de ce modèle est de tenir compte des dépendances sémantiques entre mots, exprimées par la fréquence de co-occurrence des mots[15].

[15] De manière générale, la sémantique distributionnelle *SD* peut être résumée par la phrase suivante : « deux unités linguistiques sont sémantiquement similaires si leurs contextes textuels sont similaires ».

Dans le cadre du modèle DSIR les termes d'indexation (unités linguistiques) de l'ensemble T et notés u_i sont représentés par un vecteur $C_i = (C_i^1, C_i^2, ..., C_i^{|T|})$, où la composante C_i^j correspond à la fréquence de co-occurrence de l'unité linguistique u_i avec le terme t_j. Ce vecteur est appelé *profil de co-occurrence*. Un document est représenté par le vecteur $d = (d_1, ..., d_{|T|})$, sachant que d_j correspond à la somme pondérée des fréquences de co-occurrences des unités linguistiques qu'il comporte relativement au terme j.

$$d_j = \sum_{u_i \in d} p_i \; C_i^j \tag{2.5}$$

Le poids p_i correspond à la pondération du modèle vectoriel classique. Suivant cette définition, une unité linguistique n'a de poids au sein d'un document que par le biais de ses co-occurrences et plus à travers de ses occurrences. Le modèle *DSIR hybride* [102, 15] redéfinit la mesure de poids d'une unité linguistique dans un document en prenant en compte à la fois ses occurrences et ses co-occurrences à l'aide de la formule suivante :

$$d_j = \alpha \; p_j + (1 - \alpha) \sum_{u_i \in d} p_i \; C_i^j \tag{2.6}$$

Le paramètre $\alpha \quad (0 \leq \alpha \leq 1)$ correspond au degré d'hybridation qui permet d'ajuster l'importance relative dans le modèle augmenté hybride du modèle DSIR par rapport au modèle vectoriel standard. Concernant les mesures d'appariement des documents implémentées dans le cadre d'une modélisation DSIR ou DSIR hybride, la méthode respecte le système des modèles algébriques standards.

Latent Semantic Indexing

De manière générale, les modèles de recherche utilisent un ensemble de mots clés pour représenter les documents. Ces mots clés bien qu'ils appartiennent tous aux documents, ils ne sont pas tous très pertinents pour une description idéale des textes qu'ils caractérisent, ils sont néanmoins tous des descripteurs potentiels. Partant du principe qu'une représentation géométrique traditionnelle basée uniquement sur les mots clés contient trop de bruit, le modèle LSI [32] propose de sélectionner de nouveaux descripteurs en transformant la représentation vectorielle classique en une représentation qui vise à favoriser l'association des documents sémantiquement proches[16] et cela en mettant en évidence les concepts de discussion plutôt que les occurrences des termes (*« You shall know a word by the company it keeps. »* J.R. Firth, Linguiste anglais, 1957). Ainsi un document peut être considéré pertinent par rapport à une requête même si celle ci ne contient aucun terme en commun avec le document (problème de synonymie). Et inversement, la technique a également pour ambition de pouvoir mettre plus en évidence la différence thématique

[16] Ce qui est initialement *latent* ou sous-jacent.

entre deux documents utilisant un certain nombre de termes écrits de manière identique et ayant des sens différents (problème d'homonymie ou de polysémie).

Le principe de la méthode LSI est de redéfinir une nouvelle procédure d'indexation en réalisant une décomposition en valeurs singulières de la matrice globale X des poids des termes au sein du corpus (matrice termes-documents). Le modèle LSI propose ainsi de créer un nouvel espace « sémantique » qui représente le mieux les corrélations entre les termes[17]. Cette technique peut être considérée comme une généralisation de l'extraction des vecteurs propres dans le cadre d'une analyse en composantes principales (ACP) qui permet en effet de compresser un ensemble de variables aléatoires en un ensemble de critères de meilleur choix. La principale idée de la méthode LSI est de projeter sur une même dimension les termes apparaissant de manière co-occurrente dans les textes du corpus, et pouvoir ainsi passer d'une description à base de termes à une description établie à partir d'une combinaison de ces termes.

De manière plus formelle la procédure d'indexation est réalisée à travers les étapes suivantes :

- Création de la matrice initiale et globale des termes X ;

- Décomposition de la matrice X ;
  Le modèle de décomposition en valeurs singulières [18] permet à toute matrice rectangulaire $t \times d$ d'être décomposée à l'aide du produit de trois autres matrices. La factorisation de X est réalisée comme suit :

$$X_{t \times d} = T_{t \times n}\, S_{n \times n}\, (D_{d \times n})^T \tag{2.7}$$

  Les matrices T et D sont respectivement les matrices des vecteurs singuliers à gauche et à droite pour S associés à X. S est la matrice diagonale des valeurs singulières, elles sont toutes positives et par convention triées par ordre décroissant. n est le **rang** de X, formellement, il correspond au nombre de valeurs singulières non-nulles de X ($n \leq min(t, d)$).

- Choix du facteur de réduction de dimension ;
  De la matrice S, seulement les k plus grandes valeurs singulières sont retenues. La matrice $\hat{X}$ est déduite de cette réduction, elle est par conséquent de rang plus petit que la matrice initiale X. Le choix de la dimension finale k de la matrice $\hat{X}$ est très important. Les résultats obtenus sont liés en grande partie à la valeur de cette donnée. Se limiter à quelques dimensions (accorder une trop petite valeur à k) reviendrait à réduire énormément le nombre de concepts (thématiques) de la structure sémantique (parce que les relations entre termes

[17] D'autres méthodes statistiques sont utilisées pour la réduction de dimension à l'aide de la sélection de nouveaux descripteurs [69] [130].

[18] Ou SVD, de l'anglais : Singular Value Decomposition.

sont trop complexes pour être ramenées à une dizaine de concepts). Et inversement, attribuer une trop grande valeur à k reviendrait à autoriser, après transformations, la présence de détails non pertinents (le tout est de savoir s'arrêter là où commence le bruit).

- Reconstitution de la matrice finale $\hat{X}$.

$$\hat{X}_{t\times d} = T_{t\times k}\ S_{k\times k}\ (D_{d\times k})^T \tag{2.8}$$

Les termes et les documents sont alors représentés dans un nouvel espace dont la dimension correspond au rang de la matrice reconstituée. Chaque terme et chaque document a donc ses coordonnées dans cet espace vectoriel caractérisé par k facteurs orthogonaux. La similarité entre deux termes, entre deux documents ou entre un terme et un document est alors obtenue, à l'instar des modèles algébriques, en calculant le cosinus de l'angle formé par les deux vecteurs correspondants ou leur produit scalaire. Afin d'apparier une nouvelle requête aux documents du corpus, le modèle LSI utilise le vecteur des poids des termes initial de la requête, X_q, pour calculer sa représentation approchée dans le nouvel espace, D_q.

$$D_q = (X_q)^T\ T\ S^{-1} \tag{2.9}$$

Ensuite, la similarité entre la requête et les documents est calculée en mesurant la distance angulaire entre leurs représentations vectorielles dans l'espace à k dimensions. La valeur de k est un point critique de la méthode, en effet, elle doit être différente en fonction des contextes et des objectifs poursuivis. Par conséquent, il n'y a que des expérimentations qui peuvent déterminer une valeur optimale pour k. Par ailleurs, la signification théorique des nouveaux axes et les valeurs au sein de la matrice réduite sont loin d'être clairs. En effet, la représentation à l'aide de la méthode LSI est bien une représentation de la sémantique des informations du corpus, mais cette sémantique ne peut être utilisée que dans le cadre d'un système de recherche d'information.

2.4.2 Les modèles fondés sur la théorie des ensembles

Nous citons dans cette section les modèles de recherche dont le formalisme interprète les documents et les requête en tant qu'ensembles en conservant simplement la forme canonique des termes. Les opérateurs logiques sont employés pour évaluer la pertinence d'un document. Le premier modèle cité est le modèle booléen de base. Dans la suite des modèles étendant le principe booléen sont présentés. La première contribution citée est celle du modèle booléen pondéré, qui comme son nom l'indique attribue un poids aux termes par rapport au premier modèle. Le modèle de proximité floue qui prend en compte la densité des termes de la requête au sein d'un document

pour évaluer sa pertinence, est présenté ensuite. Dans la dernière section, une dernière extension du modèle booléen de base est annoncée. Le modèle p-norme est un modèle paramétré, et dont le principe varie selon la valeur du facteur p.

Le modèle booléen

Pour représenter une requête, le modèle est fondé sur le principe de l'algèbre de BOOLE. En effet, une requête r est une expression logique constituée d'une combinaison des termes identifiant à l'aide des opérateurs ET, OU et NON. L'appariement entre une requête et la base de documents est établi à partir d'une correspondance pleine. C'est à dire, qu'un document doit s'accorder parfaitement au besoin exprimé par la requête. Exemple : en réponse à la requête « pollution ET habitat » le système proposera les documents qui comprennent les deux termes. Par le système booléen la requête en entrée « humidité OU moisissure » permettra de retrouver les documents qui contiennent au moins un des deux termes. L'avantage principal de ce système réside dans la facilité de sa mise en œuvre. En effet, les applications de recherche d'information fondées sur ce principe comme celles opérant sur les bases de données bibliographiques ou les moteurs de recherche utilisent la technique des fichiers inversés. Cette méthode est connue comme étant un dispositif permettant d'améliorer la performance d'un système de recherche en réduisant les coûts liés au temps d'accès à l'information. Son principe est fondé sur la création d'un fichier qui dresse pour chaque terme retenu pour l'indexation, la liste des documents où ce terme est présent.

$$Sim(t_i, d) = 1 \; si \; t_i \in d \; , \; 0 \; sinon. \tag{2.10}$$

$$Sim(r_1 \vee r_2, d) = 0 \; ssi \; Sim(r_1, d) = Sim(r_2, d) = 0 \; , \; 1 \; sinon. \tag{2.11}$$

$$Sim(r_1 \wedge r_2, d) = 1 \; ssi \; Sim(r_1, d) = Sim(r_2, d) = 1 \; , \; 0 \; sinon. \tag{2.12}$$

Cependant, les limites du système booléen sont bien connues :

- L'inconvénient majeur du modèle booléen est dû à son évaluation strictement binaire des taux de pertinence. Il répond à une requête de manière dichotomique. En effet, un document répond soit précisément au besoin d'une requête, dans quel cas le document retourne un score de 1, ou bien il ne correspond pas à la requête et son score est nul. Aucune autre pondération n'est réalisable, et ainsi il n'est pas possible de classer les documents pertinents et leur nombre est difficile à contrôler. En effet, pour une requête formulée sous forme d'une longue conjonction

de mots clés, un document qui satisfait la majorité de ces mots clés est aussi impertinent qu'un document qui ne satisfait aucun mot clé, et inversement. Pour une requête qui est une longue disjonction, un document qui contient un seul mot clé est aussi bon qu'un document qui satisfait la totalité des mots clés.

- Aucune pondération des termes n'est possible, les termes ont tous la même importance.

- Les résultats de la recherche dépendent du degré de maîtrise des opérateurs booléens alors qu'il n'est pas toujours évident de traduire un besoin exprimé en langue naturelle à l'aide des opérateurs logiques cités.

Compte tenu de ces limites, il semble indispensable de donner un peu de souplesse à ce modèle. Les adaptations proposées dans le cadre des modèles booléens étendus proposent de considérer les opérations booléennes en termes notamment de distances.

Le modèle booléen pondéré

Afin de réaliser un ordonnancement des réponses du système booléen, la version pondérée de Salton [107] associe aux termes de la requête et à ceux des documents, un poids. Pour définir une requête dans le cadre d'un système booléen pondéré, l'utilisateur associe préalablement des poids aux termes selon l'intérêt qu'il accorde à chaque descripteur. Il est tout à fait possible d'utiliser des pondérations négatives, celles-ci indiquent que l'usager est à la recherche de documents ne parlant pas du concept associé.

Le modèle de proximité floue

L'approche de Mercier [10, 11] repose sur l'hypothèse que plus les occurrences des termes d'une requête se retrouvent proches dans un document de la base, plus ce document est pertinent par rapport à cette requête.

L'approche de Mercier [10, 11] développe la notion de proximité en *« flouïfiant »* l'opérateur de proximité binaire NEAR. Les systèmes booléens qui implémentent l'opérateur NEAR ont besoin que l'on précise une distance maximale entre deux termes de la requête. Par exemple, pour A NEAR 6 B, la proximité binaire évalue un document à 1 si le terme A est à, au maximum, 6 mots du terme B. L'opérateur de proximité a la même fonctionnalité qu'un opérateur AND mais avec une contrainte supplémentaire sur la distance maximale tolérée entre les mots clés concernés. Pour flouïfier cette notion de proximité, Mercier représente un document d par une suite finie de longueur l de termes t. La fonction de Mercier qui calcule la proximité floue entre deux termes A et B est modélisée par $\mu_{NEAR(A,B)}(d)$.

$$\mu_{NEAR(A,B)}(d) = \max_{\substack{i \in d^{-1}(A) \\ j \in d^{-1}(B)}} \left(\max(\frac{k - |j - i|}{k}, 0) \right) \tag{2.13}$$

La notation $d^{-1}(t)$ correspond à l'ensemble des positions prises par le terme t dans d. Le paramètre k est une constante qui caractérise le degré d'influence d'une occurrence. Une valeur aux alentours de 5 permet d'évaluer la proximité entre deux termes dans le cadre d'une expression, une valeur de k entre 15 et 30 est utilisée pour mesurer le degré de proximité au niveau de la phrase et une valeur de 100 estime la proximité dans un contexte paragraphe, etc.

La notion de pertinence locale

La fonction $\mu_{NEAR(A,B)}(d)$ attribue un degré de proximité floue entre deux termes d'un texte. La motivation des travaux de Mercier est d'établir une distance entre une requête r et un document d. Pour ce faire, elle calcule par la fonction μ_t^d un degré de pertinence pour chaque terme t de r dans l'ensemble des positions possibles x dans d. Les positions x sont définies par des entiers aussi bien positifs que négatifs puisque l'influence d'un terme s'étale de part et d'autre les positions de ses occurrences et des fois même déborde avant le commencement du document ou après sa fin.

$$\mu_t^d(x) = \max_{i \in d^{-1}(t)} \left(\max(\frac{k - |x - i|}{k}, 0) \right) \tag{2.14}$$

La pertinence des termes de la requête est calculée afin d'évaluer la pertinence de la requête en fonction du document. L'approche étudiée repose l'hypothèse que les requêtes obéissent au modèle booléen classique, c'est à dire qu'une requête est une série de conjonctions et/ou de disjonctions de mots clés. Par conséquent la pertinence de la requête $r = A\ AND\ B$ à une position x est définie comme suit :

$$\mu_r^d(x) = min\left(\mu_A^d(x), \mu_B^d(x)\right) \tag{2.15}$$

De même pour l'opérateur OR :

$$\mu_r^d(x) = max\left(\mu_A^d(x), \mu_B^d(x)\right) \tag{2.16}$$

De la même manière, pour les requêtes composées de plus de 2 mot-clés, leur pertinence locale est évaluée en appliquant les formules logiques correspondantes. Dans le tableau 2.1 nous illustrons l'exemple de Mercier où nous constatons parfaitement que les positions dans un texte représentent un ensemble flou et la pertinence locale relative à la requête μ_r^d qui retourne ses valeurs dans l'intervalle [0, 1] détermine leur degré d'appartenance. Par analogie avec l'ancienne

x	0	1	2	3	4	5	6	7	8	9	10
d		A		B			C		A	B	C
μ_A^d	0.9	1	0.9	0.8	0.7	0.7	0.8	0.9	1	0.9	0.8
μ_B^d	0.7	0.8	0.9	1	0.9	0.8	0.7	0.8	0.9	1	0.9
μ_C^d	0.4	0.5	0.6	0.7	0.8	0.9	1	0.9	0.8	0.9	1
μ_{AETB}^d	0.7	0.8	0.9	0.8	0.7	0.7	0.7	0.8	0.9	0.9	0.8
$\mu_{(AETB)OUC}^d$	0.7	0.8	0.9	0.8	0.8	0.9	1	0.9	0.9	0.9	1

TAB. 2.1 – Les valeurs de proximité locale dans l'exemple de Mercier

mesure de similarité en RI, *Le niveau de coordination*[19], Mercier modélise le score d'une requête par rapport à un document de la manière suivante :

$$Score(r,d) = \sum_{x \in d^{-1}} \mu_r^d(x) \tag{2.17}$$

Reposant sur l'hypothèse que le degré de similarité entre une requête et un document est compris dans l'intervalle [0, 1], le score Score(r,d) est normalisé en le divisant par le cardinal de l'ensemble flou d^{-1}.

$$Sim(r,d) = \frac{\sum_{x \in d^{-1}} \mu_r^d(x)}{|d^{-1}|} \tag{2.18}$$

Pour le modèle de Mercier il n'est pas nécessaire d'avoir un lexique de base pour reconnaître les termes (pour leur attribuer un poids dans le cadre des configurations algébriques par exemple). En effet, la reconnaissance des termes est réalisée simplement à partir d'une correspondance suivant la forme des mots clés (ou token-mots) sur les lemmes du document.

Le modèle p-norme

Cette approche est également une extension du modèle booléen classique. La réflexion principale qui est à l'origine de cette contribution a été initiée à partir de la table de vérité 2.2 correspondant au principe de fonctionnement du modèle booléen standard.

Le principe est de voir que dans le cadre d'une conjonction, pour maximiser la pertinence d'une requête il faut se rapprocher le plus possible du cas où A est vrai « 1 » et B vrai « 1 ». Et dans

[19] En anglais « The Coordination Level » compte le nombre d'occurrence des termes de la requête dans le document pour évaluer sa pertinence.

A	B	$A \vee B$	$A \wedge B$
1	1	1	1
1	0	1	0
0	1	1	0
0	0	0	0

TAB. 2.2 – Table de vérité pour l'évaluation du modèle booléen standard

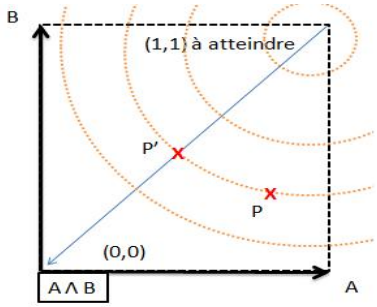


FIG. 2.5 – Projection de la requête conjonctive

le cadre d'une disjonction, A mis à faux « 0 » et B mis à faux « 0 » est le cas de figure à éviter. Pour flouïfier l'évaluation strictement binaire proposée par le modèle booléen, Salton [107] propose de calculer une distance entre la requête et les deux situations (à éviter ou à atteindre). Les deux figures suivantes représentent une projection des résultats de pertinence d'une requête établie à partir d'une conjonction (figure 2.5) et d'une disjonction (figure 2.6) de mots clés dans l'espace de ces derniers (ici A et B). Dans le cas d'une conjonction, plus un point est éloigné du point (1, 1) moins il est pertinent. Par conséquent, la pertinence (Pert) d'une requête représentée par le point P est calculée par le complément de la distance entre le point (1, 1) et le point P'.

Le complément de la distance est considéré pour évaluer la pertinence parce que cette dernière est d'autant plus grande que la distance du point à atteindre est petite. La distance entre le point

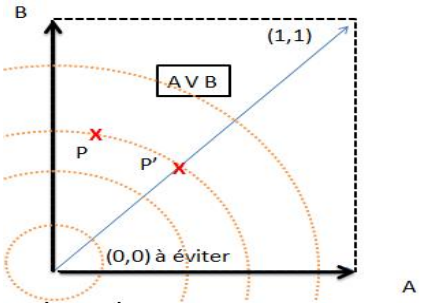


FIG. 2.6 – Projection de la requête disjonctive

à atteindre et P est égale à la distance entre le point (1, 1) et le reste des points du cercle (P' inclus) dont le centre de gravité est (1,1). Les distances sur le plan sont calculées en appliquant le principe de la distance Euclidienne. Cette extension du modèle booléen, admet également une pondération des termes dans les documents (p_i est le poids de t_i dans d).

$$Pert(t_i, d) = p_i \tag{2.19}$$

$$Pert(r_1 \wedge r_2, d) = 1 - \left(((1 - Pert(r_1, d))^2 + (1 - Pert(r_2, d))^2)/2\right)^{1/2} \tag{2.20}$$

$$Pert(r_1 \vee r_2, d) = \left((Pert(r_1, d)^2 + Pert(r_2, d)^2)/2\right)^{1/2} \tag{2.21}$$

À partir de ces formules, plusieurs généralisations ont été réalisées. Une généralisation parmi celles étudiées, consiste à attribuer un poids p aux opérateurs. La pertinence des requêtes établies à partir d'opérateurs pondérés ($\vee^p$ et $\wedge^p$) est calculée comme suit :

$$Pert(r_1 \wedge^p r_2, d) = 1 - (((1 - Pert(r_1, d))^p + (1 - Pert(r_2, d))^p)/p)^{1/p} \tag{2.22}$$

$$Pert(r_1 \vee^p r_2, d) = ((Pert(r_1, d)^p + Pert(r_2, d)^p)/p)^{1/p} \tag{2.23}$$

Ceci revient à remplacer les distances Euclidiennes par les p-distances, donc à utiliser la « p-norme » [20]. Lorsque la valeur de p est égale à 1, la formule de pertinence correspond à une variante du mode de calcul du taux de pertinence à l'aide du modèle vectoriel. Lorsque la valeur de p tend vers l'infini la p-norme correspond au modèle flou (formules 2.24 et 2.25).

$$P(r_1 \wedge^\infty r_2, d) = \min\left(P(r_1, d), P(r_2, d)\right) \tag{2.24}$$

$$P(r_1 \vee^\infty r_2, d) = \max\left(P(r_1, d), P(r_2, d)\right) \tag{2.25}$$

[20] $p \in [1, +\infty[$

2.4.3 Les modèles fondés sur les classements probabilistes

La famille des systèmes probabilistes se distingue rigoureusement des autres approches. Au lieu de calculer un coefficient de pertinence pour un document, c'est le degré de probabilité d'une hypothèse qui est estimé. Cette hypothèse est que le document D_i est pertinent et sa probabilité est notée par $P(rel/D_i)$, sachant que rel et nrel présentent respectivement la pertinence et la non-pertinence. Nous citons dans la section suivante le principe des modèles probabilistes.

Le principe probabiliste

P(rel) est la probabilité de pertinence, c'est-à-dire, la chance de tomber sur un document pertinent si on choisit un document au hasard dans le corpus. Et inversement, P(nrel) est la probabilité d'avoir un document non pertinent si on réalise un tirage au hasard. P(D) est la probabilité d'extraire le document D du corpus.

De façon générale, le principe des modèles probabilistes est de déterminer les probabilités P(rel|D) et P(nrel|D) d'un document donné par rapport à une requête donnée. Le modèle le plus simple pour calculer la probabilité de pertinence $P(rel/D_i)$ est de compter le nombre de mots pertinents (appartenant à la requête) dans le document. Par exemple, si on fait une recherche avec les mots « maison voiture radon » et que seul le mot maison est présent dans le document, on pourrait dire que $P(rel|D_i) = 1/3$. Cette façon de faire n'est pas très convaincante, étant donné qu'elle ne prend pas en considération le pouvoir discriminatoire des mots. Un document ne contenant que le mot (plus rare et par conséquent plus pertinent) « radon » se verra aussi attribuer une probabilité de $P(rel|D_i) = 1/3$ (ce qui est indiqué par le coefficient idf dans le modèle vectoriel). Il devient alors nécessaire de trouver une autre définition du taux de pertinence d'un document suivant le principe de probabilité.

Selon le théorème de Bayes :

$$P(rel|D_i) = \frac{P(D_i|rel)P(rel)}{P(D_i)} \tag{2.26}$$

$$P(nrel|D_i) = \frac{P(D_i|nrel)P(nrel)}{P(D_i)} \tag{2.27}$$

Le but est de mettre en évidence les documents maximisant $\frac{P(rel|D_i)}{P(nrel|D_i)}$ et qui sont, selon le théorème de Bayes, les mêmes que ceux maximisant la valeur de $\frac{P(D_i|rel)}{P(D_i|nrel)}$ (puisque P(rel), P(nrel) et $P(D_i)$ ne dépendent pas à la fois du document et de la requête, elles peuvent être considérées

ici comme constantes). Un modèle plus sophistiqué que le précédent calcule $P(D_i|rel)$ à partir des termes apparaissant dans D_i.

$$P(D_i|rel) = \prod_{t_j \in D_i} P(t_j|rel) \prod_{t_j \in D_i} (1 - P(t_j|rel)) \tag{2.28}$$

$$P(D_i|nrel) = \prod_{t_j \in D_i} P(t_j|nrel) \prod_{t_j \in D_i} (1 - P(t_j|nrel)) \tag{2.29}$$

La valeur de $P(D_i|rel)$ est estimée comme étant le produit des probabilités associées à chaque terme dans le document, multipliées par le produit des probabilités que les termes absents n'apparaissent pas dans un document pertinent (ce qui est calculé par $\prod_{t_j \in D_i}(1 - P(t_i|rel))$).

$$\begin{aligned} \log \frac{P(D_i|rel)}{P(D_i|nrel)} = & \sum_{t_j \in D_i} \log P(t_j|rel) + \sum_{t_j \in D_i} \log (1 - P(t_j|rel)) \\ & - \sum_{t_j \in D_i} \log P(t_j|nrel) - \sum_{t_j \in D_i} \log (1 - P(t_j|nrel)) \end{aligned} \tag{2.30}$$

La valeur $\log \frac{P(D_i|rel)}{P(D_i|nrel)}$ est appelée « valeur de statut de recherche » et le but est de trouver les documents qui la maximisent. Le problème est maintenant d'estimer $P(t_j|rel)$ et $P(t_j|nrel)$. Il existe deux manières d'estimer $P(t_j|rel)$ et $P(t_j|nrel)$. L'estimation a priori de ces facteurs est une technique qui suggère de donner une valeur fixe à $P(t_j|rel)$ et de calculer $P(t_j|nrel)$ en fonction de sa distribution dans l'ensemble des documents. Plus un terme t_j est rare plus $P(t_j|nrel)$ est basse et vice versa. Le deuxième mode d'estimation consiste à établir un échantillonnage des documents selon leur pertinence. Supposons qu'il existe NDP documents pertinents et que NDP_{t_j} soit le nombre de documents pertinents contenant t_j. On a alors :

$$P(t_j|rel) = \frac{NDP_{t_j}}{NDP} \tag{2.31}$$

Par ailleurs, ND_{t_j} est le nombre des documents contenant t_j et ND est le nombre de documents total. $P(t_j|nrel)$ est calculée par :

$$P(t_j|nrel) = \frac{ND_{t_j} - NDP_{t_j}}{ND - NDP} \tag{2.32}$$

Le passage de l'estimation de la pertinence des documents à l'estimation de la pertinence des termes repose sur l'hypothèse que les termes sont indépendants, ce qui est toutefois problématique. En effet, cette hypothèse ne se vérifie pas dans la pratique, par exemple, le terme « tomate » a certainement une probabilité plus grande d'apparaître dans un texte où apparaît le terme « marché » que dans un texte où apparaît « carburateur ». D'une part Williams [63] répond à cette conjoncture en indiquant que l'indépendance des termes est supposée pour des raisons purement mathématiques, sans quoi la plupart des calculs ne pourraient être réalisés [21], et d'autre part l'étude de Ngouya [37] qui applique le modèle probabiliste appliqué à une tâche de classification des textes a montré que l'on obtient toutefois des résultats de classification satisfaisants à partir de l'approche probabiliste fondée sur une classification bayésienne naïve des textes dès lors que le classificateur est bien entraîné (entraînement réalisé par un nombre suffisant de documents). La technique de classification utilisée par Ngouya a également montré ses preuves dans la détection des spams. À titre de référence, nous pouvons citer les travaux de Paul Graham [22], qui, en utilisant l'approche bayésienne naïve arrive à stopper 99,5% de spams avec moins de 0,03% d'erreurs. Il existe par ailleurs de nombreux modèles de recherche fondés sur le principe probabiliste et qui sont utilisés par des moteurs de recherche. Dans [116], Sparck-Jones réalise une étude comparative de ces différents modèles.

2.4.4 Le modèle utilisant les réseaux de neurones

Il est aussi appelé le modèle connexionniste, son principe repose sur la réalisation du système de recherche à partir d'un réseau de neurones artificiels. Le principe de l'approche neuronale est d'utiliser la requête en tant que stimuli[23] qui se propage dans le réseau jusqu'aux neurones de sortie. L'architecture d'un réseau de neurones en recherche d'information est généralement composée de trois couches qui sont elles-mêmes composées de neurones : une couche d'entrée constituée de noeuds dédiés aux termes de la requête, une seconde couche qui représente l'ensemble des termes de la collection et une dernière couche composée de noeuds correspondant aux documents [59]. Il existe cependant des systèmes dont le réseau est constitué de deux couches seulement : une couche « documents » et une couche « termes d'indexation » [94] [12].

La première représentation est la plus utilisée, et son mécanisme d'appariement de la requête aux documents est relativement simple. L'activation des neurones correspondant aux termes d'une requête excite les neurones de la couche des termes de la collection et l'activation se propage vers la couche de sortie qui correspond à la couche des documents. De ce fait, les documents correspondant aux neurones ayant le potentiel le plus fort sont considérés comme les plus pertinents.

Les réseaux de neurones utilisés dans un but de recherche d'information ne nécessitent pas for-

[21] *« Il ne s'agit pas d'atteindre un résultat vrai ou faux, probable ou improbable, mais seulement profitable ou non profitable. »*

[22] www.paulgraham.com/better.html

[23] Élément qui provoque une réaction dans un système.

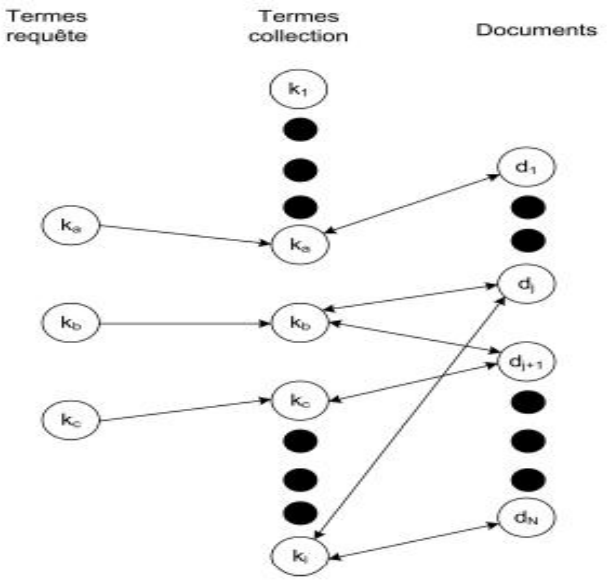


FIG. 2.7 – Architecture du réseau de neurones pour la RI (Schéma établi par Baeza-Yates et Ribeiro-Neto, 1999)

cément d'apprentissage pour être opérationnels. Ils sont initialisés à partir des informations que l'on a sur le corpus. Les neurones de sortie correspondant aux documents sont chargé d'un potentiel qui est calculé en fonction de la distribution des termes modélisés dans le réseau. Certains systèmes utilisent les poids issus de la représentation vectorielle. Par conséquent, ces systèmes donnent des résultats similaires à ceux du modèle vectoriel dans le cas où ils ne subissent aucun apprentissage.

Un traitement itératif du signal par propagation peut affiner les résultats de la première phase d'activation. À l'aide de leurs connexions bidirectionnelles, les noeuds des documents pertinents engendrent un signal en direction des neurones « termes d'indexation ». Ces derniers représentent les termes les plus représentatifs des documents jugés pertinents et envoient à leur tour un signal à la couche « documents ». Cette seconde phase correspond à une forme de retour de pertinence « relevance feedback » ce qui permet le plus souvent de mieux couvrir les attentes de l'usager du système en mettant en évidence des documents dont les termes ne sont pas contenus explicitement dans la requête (dans le cas de liens sémantiques par exemple) [94].

L'apprentissage est utilisé par les réseaux de neurones afin d'améliorer les résultats des appariements. Les poids du réseau peuvent être réajustés selon l'opinion des utilisateurs concernant la pertinence des documents retournés. Par ailleurs, l'intérêt principal de l'utilisation des réseaux de neurones dans le contexte de la recherche d'information est qu'ils permettent la représentation des différents liens entre les différentes entités modélisées. En effet, il est possible à l'aide des structures neuronales, de modéliser les liens entre les termes et les documents (la fréquence d'un terme, son poids, etc.). Elles permettent aussi de représenter les liens entre termes (en fonction des associations sémantiques possibles) [19], et il est également possible de représenter les rela-

tions entre les documents (similarité ou référence) [82].

Cependant, les deux problèmes majeurs des réseaux de neurones sont le fait qu'il est nécessaire de déterminer l'architecture du modèle à estimer [7] (l'ensemble des termes d'indexation et la définition des liens et de leurs poids) et également le fait qu'ils fonctionnent de manière transparente en mode « boîte noire ». À ce propos, Hafedh [7] énonce au sujet du principe des réseaux de neurones : *« c'est comme si on voulait examiner le cerveau de quelqu'un pour savoir ce qu'il pense : on ne peut inspecter et visualiser que les prédictions faites »*. En effet, d'un point de vue critique, les réseaux de neurones de manière générale fournissent des réponses, mais pas des explications. Il est par conséquent complètement impossible à l'utilisateur de l'inspecter dans le cas où ce dernier ne comprend pas pourquoi tel document lui a été retourné.

2.4.5 Le modèle logique

Ce modèle est considéré plus en tant que cadre théorique pour la recherche d'information[24]. Il peut être considéré également comme un méta-modèle de RI. Il est d'un niveau théorique élevé et il est fondé sur le principe de l'incertitude logique. Pour calculer le taux de correspondance d'un document d par rapport à une requête r, le modèle logique défini par Rijsbergen [122] mesure l'incertitude qui existe dans l'implication $d \Rightarrow r$ relative à un ensemble de données K. Le taux de pertinence correspond à l'extension minimale que l'on doit apporter à la connaissance K dont on dispose au moment de l'évaluation de l'implication pour établir la preuve de $d \Rightarrow r$.

Un document est considéré comme étant un ensemble de phrases, il en est de même pour la requête qui est le plus souvent, dans un contexte de RI, formée d'une seule phrase. L'implication à estimer est traduite de la manière suivante : si une ou plusieurs phrases du document impliquent la requête, cette dernière est considérée comme satisfaite. Si l'implication n'est pas satisfaite, il devient nécessaire d'ajouter de nouvelles informations de façon à ce que l'implication devienne vraie dans le nouveau contexte d'informations étendues.

Le modèle logique est certes théorique, néanmoins sa définition permet un certain degré de liberté aux différentes applications de recherche d'information. Son principal avantage est dans le fait qu'il évite de définir un modèle ad-hoc à l'implantation d'une certaine classe d'applications, et par conséquent plus ou moins empirique. Par ailleurs, pour définir de nouveaux modèles, plus spécifiques en tant qu'instances opérationnelles il est nécessaire de donner un sens à « $\Rightarrow$ », à la formalisation de d et r et au support d'informations K. Le formalisme des graphes conceptuels a été proposé en tant que formalisme opérationnel pour le modèle logique. Le modèle logique appliqué au formalisme des graphes conceptuels a été utilisé par Chevallet pour la recherche de composants logiciels [23].

[24] *« A theoretical framework »*.

2.4.6 Les approches issues du TAL

Il est essentiel de ne pas perdre de vue que les modèles que nous venons de présenter ne sont pas les seuls travaux dans le domaine. De nombreuses autres recherches ont été effectuées pour mettre au point des systèmes de recherche d'information performant se rapprochant le plus possible des besoins de l'utilisateur. Parmi ces méthodes celles dont les approches sont issues du TAL. Le principe de ces méthodes est que, au lieu d'avoir des termes en tant que descripteurs, des unités linguistiques plus complexes peuvent être envisagées, telles que par exemple des syntagmes (nominaux ou verbaux) simples ou complexes [112]. La recherche est ainsi orientée vers la détection de patrons spécifiques définissant les syntagmes descriptifs. Ces approches fondées sur le TAL sont coûteuses et complexes de par la nature de la langue elle-même, et qui de surcroît évolue constamment.

2.5 Conclusion

2.5.1 Discussion au sujet de l'indexation

Nous souhaitons dans le cadre de notre travail, utiliser une approche de recherche d'information bien adaptée au contexte et aux particularités de notre application. Par rapport à la nature hétérogène du corpus des plaintes que nous possédons, nous avons implémenté des modèles de recherche aux formalismes différents. Nous nous sommes intéressés à l'application du système vectoriel dans un espace de représentation correspondant à un dictionnaire généraliste de la langue française. S'il est clair que cette procédure d'indexation contrôlée entraîne une perte d'information importante, le résultat est un modèle plus simple à traiter automatiquement qu'un texte brut en langue naturelle. Par ailleurs, nous avons développé d'autres modèles de recherche, entre autres, le modèle fondé sur la proximité floue des termes et dont le procédé d'indexation est libre indépendant de toute ressource lexicale prédéfinie. Nous étudions les capacités de chaque système utilisé, vis à vis notamment de son mode d'indexation, à l'étape d'évaluation au chapitre 6.

Par ailleurs, pour être réaliste, tout traitement automatique de la langue naturelle ne peut être envisagé que pour atteindre des objectifs bien précis de manière à pouvoir développer les outils nécessaires et adéquats à chaque application [128]. Dans ce sens, nous démontrons dans la section 5.8.4 que notre domaine d'application reste vaste. Nous avons alors fait le choix d'utiliser des outils et des ressources existantes (lemmatiseur, dictionnaire, corpus de textes) sans devoir décrire les spécificités morphologiques, grammaticales et syntaxiques du langage, surtout que nous avons pris connaissance à travers la littérature qu'il était impossible de définir une grammaire couvrant la totalité d'une langue [1].

L'indexation n'est pas ici un objectif en soi, nous avons néanmoins orienté notre travail sur les diverses applications de cette technique pour notamment : l'analyse du vocabulaire (section 5.8), la formalisation et la recherche de documents par similarité textuelle (section 5.7) et la

réalisation de synthèses automatiques (section 5.10).

2.5.2 Discussion au sujet des modèles d'appariement

Le choix du système de recherche est par ailleurs fortement lié à la taille du corpus et de ses éléments. Dans le cas des collections de textes de grande taille, la représentation la plus utilisée est la représentation vectorielle, puisque dans ce cas le poids des descripteurs établi notamment à partir du calcul de la force discriminatoire prend tout son sens. Cette représentation est également souhaitable pour les textes longs. D'après Singhal [114], il existe deux propriétés à considérer dans les textes longs par rapport aux textes courts et qui influenceraient le calcul des poids des descripteurs au sein d'une représentation vectorielle :

- Les mots présents ont tendance à avoir des fréquences plus élevées,
- Les textes longs sont plus susceptibles de contenir des mots clés différents.

Par ailleurs, la taille du vecteur des caractéristiques correspond à la dimension du vocabulaire. Ce dernier étant la plupart du temps de taille considérable, il serait fortement souhaitable de trouver une alternative à l'utilisation de ces vecteurs pour la modélisation et l'appariement des textes concis. De plus, la représentation des documents à l'aide du modèle vectoriel exclut toute notion de position et de distance entre les mots. En effet, la position des termes est d'autant plus pertinente dans le cas où il s'agit de textes courts. La représentation vectorielle est d'ailleurs très souvent appelée la représentation en « sac de mots » dans la littérature.

Par rapport aux différents modes d'indexation à étudier, et par rapport à la taille relativement variable des textes en notre possession, il était primordial pour nous d'élargir notre réflexion liée aux modèles de représentation. Dans le chapitre 6 nous argumentons l'utilisation des différents systèmes adoptés et nous les comparons ensuite dans le cadre du chapitre d'évaluation 6.

Chapitre 3

Adaptation structurelle et sémantique des systèmes de recherche

La recherche d'information qui tient compte de la dimension structurelle dans les documents (structurés ou semi-structurés) tels que la galaxie XML (section 3.2) est une application dont l'intérêt est double. Dans un sens il existe une masse de plus en plus importante de données au formalisme XML[1] et pour laquelle il est nécessaire d'adapter les techniques de recherche d'information classiques par rapport aux exigences des mêmes applications que celles qui manipulent des documents non structurés. Dans un autre sens, la prise en considération de la structure des documents permet d'améliorer le niveau de précision des résultats de la recherche. En effet, l'arborescence des documents semi-structurés XML, ou documents *XMLisés*, donne la possibilité d'accéder à des éléments plus « fins » que le document « plat » pris en entier, et permet d'envisager une recherche « focalisée ».

Dans ce chapitre nous allons d'abord aborder de manière générale les formes de structure possibles et connues à ce jour des documents (sections 3.1 et 3.2). Des réglementations concernant la structure des documents et qui existent depuis l'avènement du support informatique sont énoncés également. Après avoir introduit le langage XML, nous discutons de la compagne INEX et de son rôle dans la promotion des travaux autour des bases XML (section 3.3). Dans la section suivante des initiatives d'adaptation de SRI classiques à la recherche dans des corpus XML sont présentées. Des adaptations du modèle vectoriel (section 3.4) et du modèle probabiliste (section 3.5) sont détaillées. Des travaux s'inscrivant dans cette démarche, initiés à partir d'autres systèmes de recherche de base sont également cités (section 3.6). Enfin, des études concernant l'intégration de la sémantique à l'aide de ressources externes dans des SRI classiques sont exposées (section 3.7).

[1] le moteur de recherche Google annonçait le 11 décembre 2001 qu'il permettait l'accès à 2 milliards de pages.

3.1 Prise en compte de la structure

Les travaux en recherche d'information s'intéressant à la problématique des documents structurés peuvent être séparés en deux catégories : les travaux de « recherche de passages » de documents, et ceux liés à la « recherche de structure ».

- L'approche par passage [81] consiste à considérer les passages (et des recouvrements de passage) fixés par un nombre donné de termes comme un document indépendant et se limite très souvent à réutiliser des modèles usuels de la RI (le vectoriel par exemple).

- L'approche par structure, [127, 24] utilise un modèle usuel, le modèle vectoriel standard notamment, avec la différence que les taux de pertinence globale sont évalués en combinant les pertinences au sein des parties atomiques et au sein du document.

L'étude mise au point par James Callan [22] compare les deux approches, en réalisant des tests sur des documents de bases légales. Dans le cadre de notre étude, nous nous intéressons aux approches par structure. Le but étant ici de bénéficier de l'organisation des documents qui sont présentés sous forme structurée initialement.

3.2 La structure des documents

En vue de définir les différents modes de recherche dans un contexte de documents structurés, nous présentons tout d'abord la notion de structure dans un document, notion sur laquelle il est indispensable de s'accorder. Il est communément admis qu'un document quelconque peut être structuré selon deux critères distincts : un document obéit à une structure logique et à une structure physique. Schématiquement, la structure physique reflète des caractéristiques typographiques (ou le point de vue d'un typographe). Elle s'intéresse à organiser un document en fonction de sa mise en page, c'est à dire de la police des caractères, du découpage en blocs de textes (ou en pages) et leur agencement les uns par rapport aux autres. Alors que la structure logique met en exergue une structure plus axée sur le contenu (ou le point de vue d'un auteur). Elle met en valeur le rôle et la nature de chaque élément ainsi que les liens hiérarchiques qui les lient (titre de l'article, le nom de l'auteur, etc). En tout, une structure physique bien conçue aura pour principale qualité de rendre lisible la structure logique du document, et c'est cela qui peut d'ailleurs amener à confondre les deux aspects. En effet, les deux organisations peuvent conduire à un découpage équivalent d'un seul document.[2]

Plus récemment, et en fonction des besoins, la structure sémantique est prise en compte. Elle consiste à mettre en évidence des rubriques de conversation dans la réalisation de la structure.

[2] La preuve qu'il s'agit également d'un document électronique Un site Web peut être décrit comme un document avec une structure logique et physique (isdn.enssib.fr/archives/axe3/axe3_annexes_RFVGAGc.pdf).

Par exemple la représentation de la structure sémantique d'une notice médicale peut se présenter à l'aide des champs suivants : composition, voie d'administration, posologie, etc. D'autres structures deviennent nécessaires à prendre en compte, notamment la structure temporelle des documents multimédia par exemple, et la structure générique-spécifique qui consiste à avoir une structure générique régie par une norme par exemple fusionnée à plusieurs autres structures spécifiques le tout associé au sein d'une même organisation. Par exemple l'organisation des thèses peut être effectuée suivant une même structure globale, dite générique. Cependant chaque thèse a sa propre structure individuelle (nombre de chapitres, etc.), dite spécifique.

Depuis l'avènement des documents électroniques, des normes de représentation et d'échange de documents structurés ont été proposées [83]. La norme ODA (Open Document Architecture)[3], définit les deux structures logiques et physiques d'un document. Elle a été supplantée par la norme SGML (Standard Generalized Markup Language)[4]. SGML s'intéresse à la structure logique des documents, mais il est possible d'utiliser des balises (marqueurs, éléments structurants ou « doxel » pour « document element » en anglais, et par analogie avec le pixel) décrivant des aspects de présentation, comme c'est dans le cas de HTML qui est le langage du Web issu des normes SGML. XML (eXtensible Markup Language) [5] qui est une autre application de SGML est plus souple et plus expressive que HTML. En effet, une des grandes nouveautés par rapport à HTML est la possibilité de créer ses propres balises pour permettre à l'utilisateur de structurer aussi finement qu'il le désire son document. XML ne se préoccupe que de la structuration logique et sémantique du document. Ainsi, aucune mention de présentation physique n'apparaît directement dans un document XML. La présentation physique d'un document est confiée à un outil indépendant, XSL (Extensible stylesheet language)[6], qui fonctionne sur le principe des feuilles de style[7].

3.3 L'initiative INEX

INEX[8] (INitiative for the Evaluation of Xml retrieval) [48], au même titre que TREC (Text REtrieval Conference) [123] ou CLEF (Cross Language Evaluation Forum) [92] pour les documents plats[9], est une campagne internationale ayant pour but de constituer un corpus de documents XML qui permette l'évaluation de systèmes de Recherche d'Information Structurée (RIS). Elle fournit également un ensemble de requêtes et des « jugements de pertinence », c'est-à-dire les estimations humaines concernant les réponses pertinentes se rapportant aux jeux de requêtes proposés. Cette campagne a vu le jour en avril 2002, et lors de sa création elle utilisait une collection de 12107 articles de IEEE, publiés entre 1995 et 2002 et transcrits en XML et dont

[3] Architecture de document ouverte, Norme ISO/CEI 8613.

[4] langage normalisé de balisage généralisé, Norme ISO 8879 :1986.

[5] langage de balisage extensible.

[6] XSL est le langage de description de feuilles de style associé à XML.

[7] C'est un document permettant de mettre en forme un autre document rédigé dans un langage de balisage (HTML, XML, etc.). Les feuilles de style sont rédigées dans un langage spécifique (XSLT, CSS,SL-FO, etc.).

[8] inex.is.informatik.uni-duisburg.de

[9] non-structurés.

chaque document contient en moyenne 1500 doxels et la collection contient au total 8 millions de noeuds et 180 balises distinctes. En 2005, la collection XML a été étendue à environ 17000 articles provenant de 21 magazines ou revues différentes. En 2006, le corpus INEX évolue encore pour atteindre le nombre de 659388 documents en anglais extraits de l'encyclopédie en ligne Wikipedia [33].

Les « objets de recherche », appelés aussi « topics », désignent les requêtes soumises aux évaluations. Ces topics peuvent porter sur le contenu seulement (Content Only topics « CO ») ou bien sur, à la fois, le contenu et la structure (Content And Structure topics « CAS »). Dans le premier cas, une requête envoyée au système de recherche d'un participant INEX est une simple expression de mots clés, assez similaire à une requête pour un moteur de recherche standard. Le résultat retourné au participant n'est pas un classement de documents mais une liste d'éléments XML, dont la granularité est déterminée par le modèle expérimenté. Concernant les topics CAS, ils définissent l'information recherchée en plus de certaines contraintes de structure (la structure sur le document à retourner et/ou sur le type d'élément à retrouver). L'objectif de l'évaluation initiée à partir de topics CAS est d'avoir un classement des résultats précis en fonction du contenu et que la granularité des réponses retournées soit en adéquation avec les besoins de l'utilisateur.

Les systèmes de Recherche d'Information Structurée (RIS) sont dans l'ensemble fondés sur des modèles dérivés des modèles de recherche classiques que nous avons étudiés dans la section 2.4 du chapitre précédent. Dans la section suivante nous présentons, entre autres, le modèle EVSM (Extended Vector Space Model) [44]. Ce dernier est le premier modèle fondé sur le principe du modèle vectoriel ayant pris en considération l'aspect structurel des documents.

3.4 Adaptation structurelle du modèle vectoriel

3.4.1 Le modèle Extended Vector Space Model

L'idée de ce modèle est de représenter les requêtes et les documents par des sous-vecteurs, chacun représente une classe d'information différente (*concept-terms* notée *C-type*). La similarité entre un document D et une requête R est une somme des similarités calculées au niveau de chaque classe d'information (ou sous-vecteur) pondérée par des coefficients différents.

$$Sim(R, D) = \sum_{i=1..n} \alpha_i Sim(C_i^R, C_i^D) \tag{3.1}$$

Où α_i est le coefficient qui correspond à la classe d'information C_i. Pour la prise en compte de la structure et le contenu des documents il est nécessaire de déterminer les poids des termes en fonction de leur contexte d'apparition et en fonction de leur distrinution dans le corpus. La mesure d'appariement doit tenir compte de la nouvelle unité de structure.

3.4.2 Le moteur de recherche JuruXML

Carmel [30] a adapté son système de recherche Juru à la structure des documents XML. Les index ne sont plus des termes mais des couples (terme, contexte d'apparition du terme). La similarité entre une requête R et un document D est modélisée par la formule suivante :

$$S(R,D) = \frac{\sum_{(t_i,C_i)\in R} \sum_{(t_i,C_i)\in D} W_R(t_i,C_i) \times W_D(t_i,C_k) \times cr(C_i,C_k)}{|R| \times |D|} \tag{3.2}$$

Cette écriture permet d'augmenter le taux de similarité entre une requête et un document dans le cas où un même terme t n'apparaît pas dans un même contexte dans la requête et dans le document. Si un terme t apparaît dans deux contextes différents mais jugés proches par la notion de ressemblance de contextes $cr(C_i, C_k)$, le poids de ce terme permet d'améliorer le taux de pertinence du document.

$$W_X(t,C) = TF_X(t,C) \times IDF(t,C) \tag{3.3}$$

$$TF_X(t,C) = \log\left(freq_X(t,C) + 1\right) \tag{3.4}$$

$$IDF(t,C) = \log\left(\frac{|N|}{|N_{(t,C)}|}\right) \tag{3.5}$$

Où X désigne le document ou la requête, $|N|$ désigne la taille du corpus et $|N_{(t,C)}|$ correspond au nombre de documents où une occurrence de t apparaît au moins une fois dans le contexte C. La proximité entre une requête et un document est calculée à partir d'un appariement croisé des contextes du moment qu'ils sont similaires. Cette similarité est évaluée par cr.

Carmel calcule la similarité de contextes $cr(C_i, C_k)$ par la formule suivante :

$$cr(C_1,C_2) = \\ \alpha LCS(C_1,C_2) + \beta POS(C_1,C_2) - \gamma GAPS(C_1,C_2) - \delta LD(C_1,C_2) \tag{3.6}$$

La similarité entre deux contextes C_1 et C_2 est une combinaison linéaire des quatre facteurs suivants :

LCS : c'est le rapport entre la taille de la sous-chaîne commune d'éléments la plus longue entre les contextes C_1 et C_2, $l(C_1, C_2)$ et la taille du chemin de la requête (nous considérons C_1 comme un chemin de la requête).

$$LCS(C_1, C_2) = \frac{l}{|C_1|} \tag{3.7}$$

POS(C_1,C_2) : ce facteur est fonction de l'élément AP (pour Average position) qui calcule la position moyenne des éléments communs de C_1 et C_2. Le but est de favoriser le contexte ayant les éléments communs avec le contexte de la requête situés en tête du chemin puisqu'ils sont dans ce cas des éléments plus discriminants.

GAPS : plus il y a d'éléments différents « gaps » dans le chemin par rapport au chemin de la requête plus la valeur de GAPS est grande. Sa valeur est proportionnelle à la taille de la sous-chaîne commune la plus longue. Dans le cas de deux chemins identiques, gaps et GAP prennent la valeur 0.

$$GAPS = \frac{gaps}{gaps + l(C_1, C_2)} \tag{3.8}$$

LD : afin de favoriser les chemins C_2 des documents D dont la taille se rapproche le plus de la taille du chemin C_1 de la requête R, l'élément LD est calculé de la manière suivante :

$$LD(C_1, C_2) = \frac{|C_2| - l(C_1, C_2)}{|C_2|} \tag{3.9}$$

Le tableau 3.1 montre le principe de calcul de chaque facteur ainsi que ses effets sur le taux de similarité des contextes cr. L'exemple considère un contexte de la requête C_1 et un contexte de document C_2, avec $C_1 = book/chapter/title$.

3.4.3 Le modèle universel pour la recherche d'information structurée en XML

Azevedo a aussi présenté une approche d'extension du modèle vectoriel dans un contexte de recherche de documents XML. Elle considère la balise XML en tant que nouvelle unité de

	Effet de *lcs* sur *cr*				
C_2	**lcs**	AP	gaps	ld	**cr**
media/**book/chapter/title**/number	**3**	3	0	2	**0.84**
media/**chapter/book/title**/number	**2**	3	0	3	**0.53**
media/**title/chapter/book**/number	**1**	2	0	4	**0.29**
magazine/volume/artice/**title**/number	**1**	4	0	4	**0.19**
	Effet de *AP* sur *cr*				
C_2	lcs	**AP**	gaps	ld	**cr**
book/chapter/title/subtitle/number	3	**2**	0	2	**0.92**
media/**book/chapter/title**/number	3	**3**	0	2	**0.84**
media/catalog/**book/chapter/title**	3	**4**	0	2	**0.75**
	Effet de *gaps* sur *cr*				
C_2	lcs	AP	**gaps**	ld	**cr**
media/catalog/**book/chapter/title**/subtitle/number	3	4	**0**	4	**0.78**
catalog/**book**/chapters/**chapter**/section/**title**/number	3	4	**2**	4	**0.68**
	Effet de *ld* sur *cr*				
C_2	lcs	AP	gaps	**ld**	**cr**
book/chapter/title/subtitle/subtitle/number/bullet	3	2	0	**4**	**0.88**
book/chapter/title/subtitle	3	2	0	**1**	**0.95**

TAB. 3.1 – Effets des différents facteurs sur le taux de similarité des contextes *cr*

recherche au même titre que le document dans le modèle vectoriel [76].

$$\rho(Q,D) = \sum_{t_i \in Q \cap D} \frac{W_Q(t_i) \times W_D(t_i,e) \times fxml(t_i,e)}{|Q| \times |D|} \tag{3.10}$$

De manière analogue aux pondérations TF-IDF des termes des documents non structurés, les poids des termes des documents XML $W_X(t_i,e)$ sont calculés comme suit :

$$W_D(t_i,e) = \log(tf(t_i,e)) \times \log(\frac{N}{df(t_i,e)}) \tag{3.11}$$

$$fxml(t_i,e) = fnh(t_i,e) \times fstr(t_i,e) \times focr(t_i,e) \tag{3.12}$$

Le fnh (Nesting Factor ou degré d'implantation) désigne la pertinence d'un terme en fonction de sa position dans l'arbre XML.

$$fnh(t_i,e) = \frac{1}{1+nl} \tag{3.13}$$

Où nl est le nombre de niveaux qui séparent l'élément e de son sous élément où apparaît t. L'intérêt du facteur $fnh(t_i,e)$ est d'éviter que les poids des termes apparaissant directement dans un élément aient moins de poids dans ce dernier que dans les éléments englobant et distants situés plus haut dans l'arbre.

Le fstr (Structure Factor) est calculé par la forme suivante :

$$fstr(t_i,e) = \frac{common\_markups + 1}{nr\_qmarkups + 1} \tag{3.14}$$

$Common\_markups$ dénote le nombre de balises communes entre la structure de la requête et la structure du contexte de l'élément e dans le document où apparaît le terme t. $nr\_qmarkups$ est le nombre total de balises dans la structure de la requête. L'intérêt de fstr est d'affecter un poids plus important aux termes situés dans des contextes dont la structure se rapproche le plus de la structure de la requête, ce qui est évident.

Le dernier facteur focr (Co-occurrence Factor) désigne la relation entre les balises et leur contenu, il est donné par la formule suivante :

$$focr(t_i,e) = cf(t_i,e) \times idf(t_i,e) \times N \times icf(e) \tag{3.15}$$

Où $cf(t_i, e)$ est le nombre de fois où l'étiquette m, délimitant l'élément e contient au moins une fois le terme t dans le corpus. $idf(t_i, e)$ est l'inverse du nombre d'éléments e contenant au moins une fois t. $icf(e)$ est l'inverse du nombre de fois où l'étiquette m apparaît dans la collection. N est la taille du corpus. L'intérêt de focr est de quantifier un certain lien sémantique entre le terme et l'élément où il apparaît en fonction de l'évaluation simultanée de la répartition du couple (terme, élément) dans le corpus et de la pertinence de l'élément.

3.4.4 Adaptation structurelle de Zargayouna : 1ère version

Pour la modélisation et l'appariement des documents *XMLisés* dans [132] et [133], Zargayouna adapte le modèle vectoriel de Salton en fonction de la structure des documents. Elle part du principe que la structure d'un document est un élément sémantique non négligeable. Par cette méthode, un vecteur des poids des termes est lié à chacune des balises. Au final, une matrice des scores des termes est attribuée à chaque document. Par conséquent, le poids des termes est recalculé de façon à respecter ses nouvelles distributions. Plusieurs nouvelles dimensions de distribution sont alors considérées. Le score TF-ITDF « term frequency-inverse tag and document frequency » désigne le poids d'un terme t pour un élément[10] associé à la balise b dans un document d. Dans la suite, nous ne ferons plus la distinction entre la notion de balise et celle d'élément.

$$TF - ITDF(t, b, d) = TF(t, b, d) \times ITF(t, d) \times IDF(t, b) \tag{3.16}$$

TF(t,b,d) « term frequency » est la fréquence d'apparition du terme t dans la balise b du document d. ITF(t,d) « inverse tag frequency » désigne la valeur inverse du nombre de documents qui contiennent le modèle de balise b dans laquelle le terme t apparaît au moins une fois ; DF(t,b) . IDF(t, b) « inverse document frequency » correspond à la valeur inverse du nombre de balises dans le document d dans lesquelles le terme t apparaît au moins une fois (TagF(t,d)).

$$ITF(t, d) = \log\left(\frac{|D_b|}{DF(t, b)}\right) \tag{3.17}$$

$$IDF(t, d) = \log\left(\frac{|B_d|}{TagF(t, d)}\right) \tag{3.18}$$

Les paramètres $|D_b|$ et $|B_d|$ désignent, respectivement, le nombre total des documents où le modèle de balise b apparaît dans la structure et le nombre total des balises dans le document d .

Concernant le modèle vectoriel étendu, il est nécessaire de commencer d'abord par calculer

[10] Zargayouna définit également un élément comme étant du texte délimité par une balise ouvrante et une balise fermante.

des similarités locales correspondant à l'ensemble des balises. Pour l'estimation des similarités locales, les scores TF-ITDF sont utilisés par la formule du Cosinus de la même façon que les poids TF-IDF. Les similarités locales correspondent au degré d'appariement des balises du document cible (requête) r et du document source d deux à deux. Le score de similarité intégrale correspond à l'agrégation des similarités locales.

$$Sim(r,d) = \sum_{i=1..n} Cosinus(b_{ri}, b_{di}) \tag{3.19}$$

3.4.5 Adaptation structurelle de Zargayouna : 2ème version

Dans la deuxième adaptation du modèle vectoriel pour le traitement des documents semi-structurés [131], Zargayouna ne considère pas les éléments comme des unités indépendantes. En effet, les éléments appelés « contextes » peuvent être imbriqués et donc être liés par des relations de spécialisation/généralisation. Elle souhaite introduire cet aspect structurel supplémentaire dans le calcul des poids des termes.

Pondération des termes

Le poids d'un terme est calculé en fonction de sa fréquence, de sa représentativité et en fonction du niveau de son statut discriminant. La fréquence TF d'un terme t au sein d'une instance n du contexte C dans un document d est calculée comme suit :

$$TF(t,d,C,n) = \sum_{C_i \in SP_C \cap C_d} \left(\frac{1}{1 + dist(C, C_i)} \times \sum_{n \in inst(C_i) \cap SP_n} count(t,n) \right) \tag{3.20}$$

count(t,n) désigne le nombre d'occurrences de t dans l'instance n.

Étant donné un contexte C appartenant à un document d, pour estimer le degré de représentativité du terme t par rapport à ce contexte précis au sein du document d, le facteur CF est calculé.

$$CF(t,d,C) = \log \left(\frac{|SP_c^t \cap C_d|}{|C_d^t|} + 1 \right) \tag{3.21}$$

$|SP_c^t \cap C_d|$ est le nombre de spécialisations du contexte C dans le document d contenant au moins une fois le terme t. $|C_d^t|$ est le nombre de contextes dans d qui contiennent au moins une fois t.

Étant donné un contexte existant dans un corpus de documents, afin d'estimer le taux de représentativité du terme t par rapport à ce contexte dans le cadre général du corpus, la force discriminatoire du terme t par rapport au contexte C est calculée par IDF.

$$IDF(t,C) = log\left(\frac{|D|_c}{|D|^t_{SP_c}} + 1\right) \tag{3.22}$$

$|D|^t_{SP_c}$ est le nombre de documents qui comportent des spécialisations de c et où une occurrence de t apparaît au moins une fois. $|D|_c$ est le nombre de documents dans le corpus qui comportent c.

Inspirée de la mesure TF-IDF, Zargayouna définit le poids TF-I**C**DF (par rapport à sa première version TF-I**T**DF) d'un terme t dans un contexte C au sein d'un document d de la manière suivante :

$$TF - ICDF(t,d,C,n) = TF(t,d,C,n) \times CF(t,C,d) \times IDF(t,C) \tag{3.23}$$

Nous remarquons alors que dans sa première version, les poids des termes relatifs aux documents structurés étaient tributaires des éléments (balises ou tags) considérés en tant qu'unités atomiques à un seul niveau, dans sa dernière version Zargayouna prend en considération la structure interne des éléments et leurs éventuelles instances pour définir les poids des termes. Ces poids renseignent les vecteurs descriptifs des contextes constituant les documents *XMLisés* permettant leur appariement.

Plusieurs autres réalisations se sont intéressées à utiliser le principe vectoriel pour la recherche au sein des collections structurées. Parmi ces approches nous pouvons citer [72], [58], [28], [5], en plus d'autres travaux mis au point à IBM par l'équipe de Mass et Carmel dans le cadre du moteur JuruXml [39] et [129].

3.5 Adaptation structurelle du modèle probabiliste

Une version du modèle de recherche Okapi[11] [118] a été adaptée pour la recherche de documents structurés dans le cadre du projet Outilex[12][93]. Okapi est un des modèles de recherche les plus performant, il est fondé sur le principe probabiliste. Dans le cadre du projet Outilex, l'équipe de Piwowarski a adapté Okapi à la recherche dans les corpus structurés en définissant le score attribué à un élément X pour une question q de la manière qui suit :

$$Okapi(q,X) = \sum_{j=1}^{Longueur(q)} w_{j,X} \; \frac{(k_1+1)\; tf_{X,j}}{k_X + tf_{X,j}} \; \frac{(k_3+1)\; tf_{q,j}}{k_3 + tf_{X,j}} \tag{3.24}$$

[11] Okapi est le nom d'un système de recherche documentaire expérimental basé à la City University de Londres.

[12] Ce projet vise à mettre à la disposition de la recherche, du développement et de l'industrie une plate-forme logicielle de traitement des langues naturelles ouverte et compatible avec l'utilisation de XML, d'automates finis et de ressources linguistiques.

Où k_1 et k_3 sont des constantes[13], par rapport au modèle Okapi de base la longueur moyenne d'un document et le nombre de documents dans lesquels apparaît un terme sont adaptés à la recherche structurée.

- $w_{j,X} = \log\left(\frac{N-n_j+0,5}{n_j+0,5}\right)$. N est la taille de la collection et n_j est le nombre de documents contenant le terme j. Il existe différentes manières d'adapter n_j à la recherche structurée. Par exemple il est possible de calculer le nombre d'éléments d'un type donné (par exemple calculer la fréquence au sein des éléments (balises), paragraphes, sections, etc.).

- $k_x = k_1((1-b) + b\ \frac{dl}{avdl})$. Dans Okapi classique, b est une constante [14], dl est la longueur du document et avdl est la longueur moyenne des documents. Dans un corpus structuré, les longueurs peuvent correspondre à la taille des éléments ou des éléments du même type que X ou ceux situés au niveau de la même hiérarchie.

Suite à une évaluation réalisée sur le corpus d'INEX 2004, le modèle Okapi étendu en fonction de la structure a pu être un peu plus précisé. Les résultats ont montré que l'un des meilleurs modèles calculait la valeur de n_j en fonction du nombre d'éléments dans lesquels apparaît le terme j. Concernant le paramètre advdl, la meilleure configuration a été réalisée en lui attribuant la moyenne des éléments de même type. Néanmoins, et comme le suggère Piwowarski [93], il est tout à fait possible d'ajuster ce modèle en fonction des besoins des applications ou de proposer des modèles plus complexes capables d'apprendre leurs paramètres.

3.6 Adaptation structurelle d'autres modèles de recherche

D'autres approches tentent de définir de nouveaux modèles pour l'indexation et la recherche d'éléments de structure. Myaeng [84] utilise un modèle probabiliste, Paradis [91] s'est basé sur un modèle booléen pondéré et Mulhem [83] s'est intéressé à étendre un modèle logique de RI en l'adaptant à la structure des éléments de recherche.

Parmi les meilleures évaluations obtenues lors des campagnes INEX, on trouve le moteur JuruXML 2005 [39] qui est une adaptation du modèle vectoriel (section 3.4.2) et on trouve également le travail de Sigurbjörnsson [8] qui s'inscrit dans la discipline probabiliste. Bien que INEX soit actuellement une référence en matière d'évaluation des systèmes de recherche adaptés à la structure, il faut prendre les résultats avec prudence. En effet, il s'agit d'une collection particulière et de requêtes particulières. De plus, la définition de la nature de la requête ou du topic à utiliser (CO ou CAS) peut être un critère de choix entre l'utilisation d'un système au détriment

[13] Dans le cadre du projet Outilex les valeurs standards ont été utilisées pour les paramètres d'Okapi (k_1=1.2, k_3=7).

[14] Dans Okapi standard b=0,75.

d'un autre.

3.7 Les modèles de recherche sémantique

Nous avons vu dans la section 3.4 qui précède qu'il existait des modèles de recherche qui tentent de modéliser l'information sémantique inhérente des textes en tenant compte des co-occurrences des termes et de leurs contextes d'apparition dans les corpus de référence. Dans cette partie nous allons définir des modèles de similarité qui prennent en compte la sémantique en employant des moyens externes autres que les corpus. Ces ressources sont utilisées pour calculer les distances de « sens » entre les termes (ou les concepts auxquels ils se rattachent) afin de les utiliser directement dans le cadre des modèles d'appariement classiques entre documents.

3.7.1 Adaptation structurelle et sémantique du modèle vectoriel

Le modèle vectoriel sémantique étendu de Zargayouna : 1ère version

Le modèle vectoriel standard et le modèle vectoriel étendu attribuent un score aux termes en fonction de leur distribution directe dans les documents et dans les balises en faisant abstraction de l'aspect sémantique. Dans les travaux de Zargayouna cité dans [132] et [133], la fonction SemW(t,b,d) (formule 3.25)réévalue la pondération des termes en tenant compte des séparations sémantiques (selon les rubriques) et d'une ontologie du domaine (section 4.2.2). Au moyen de cette nouvelle évaluation des scores, le poids d'un terme qui n'apparaît pas directement dans une unité sémantique peut être augmenté en fonction des scores des termes appartenant à ce contexte et qui sont sémantiquement liés au terme considéré.

$$SemW(t,b,d) = TF-ITDF(t,b,d) + \frac{\left(\sum_{i=1..n} Sim(t,t_i) \times TF-ITDF(t_i,b,d)\right)}{n} \quad (3.25)$$

Avec comme condition la valeur de $Sim(t,t_i)$ qui doit être supérieure au degré de similarité entre le concept correspondant au terme t et celui auquel est rattaché la balise. Un avantage de cette dernière restriction est la réduction en partie du problème de l'ambiguïté du langage naturel. D'un point de vue général, un terme ambigu est un terme ayant plusieurs sens possibles, et d'un point de vue conceptuel, un terme ambigu est un terme rattaché à au moins deux concepts dans la représentation ontologique. Le score d'un terme ambigu est augmenté uniquement par le degré de similarité du concept le plus proche des concepts correspondant aux termes co-occurrents au niveau de la même unité sémantique. Le concept à l'origine de l'ambiguïté est éloigné de l'ensemble des concepts liés aux termes du même contexte, et donc il n'intervient pas dans le nouveau score du terme.

Par exemple, supposons le cas d'un document *XMLisé* où le terme « avocat » n'apparaît pas

dans une balise notée « <profession> ». Supposons également que les deux termes « juge » et « fruit » existent au sein de cette balise. Le poids du terme homonymique avocat qui désigne aussi bien le fruit que le défenseur et qui n'existe pas dans la balise sera augmenté du poids des termes contenus dans cette dernière. Pour cela, il est nécessaire que le taux de similarité de ces termes avec le terme auquel est rattachée la balise soit au moins aussi élevé que la valeur de la similarité entre le terme « avocat » et « profession », dans ce cas nous vons le terme « juge ». Il est bien évident qu'aucun lien sémantique ne peut exister entre « profession » et « fruit », le taux de ce lien est donc inférieur à celui qui relie « avocat » à « profession ». Par conséquent, le poids de « fruit » n'intervient pas dans l'augmentation.

Le modèle vectoriel sémantique étendu de Zargayouna : 2ème version

Zargayouna définit également une extension de sa seconde version de l'adaptation structurelle (bi-dimensionnelle) du modèle vectoriel (section 3.4.5) [131], et ce, en tenant compte de la dimension sémantique. Nous rappelons ici que dans sa seconde adaptation structurelle du modèle vectoriel, Zargayouna ne considère plus les éléments structuraux en tant qu'unités atomiques indépendantes mais elle tient compte des « contextes » qui peuvent apparaître dans une structure hiérarchique. Les contextes peuvent donc être liés par des relations de spécialisation/généralisation. Pour calculer le poids sémantique d'un terme conformément à ce modèle trois mesures sont à déterminer :

- La fréquence sémantique du terme
  Dans le cadre du document : $SemTF(t, d, C, n)$[15] désigne la fréquence du terme t dans l'instance n du contexte C au niveau du document d, enrichie par la similarité des termes co-occurrents au sein du même contexte w_c. Le degré de similarité des termes co-occurrents au sein du même contexte que t est pondéré par leurs fréquences.

$$SemTF(t, d, C, n) = TF(t, d, C, n) + \sum^{t_i \in w_c} TF(t_i, d, C, n) \times Sim(\varphi_C(t), \varphi_C(t_i)) \quad (3.26)$$

  La notation $\varphi_C(t)$ caractérise le terme t dans son contexte C (son sens)[16]. L'ensemble des termes sémantiquement proches du terme t doit être déterminé en fonction d'un seuil correspondant à un taux de similarité optimal au-delà duquel un terme est considéré significativement sémantiquement proche du terme t. La valeur minimale de $Sim(\varphi_C(t), \varphi_C(t_i))$ est à déterminer de manière ad-hoc en fonction aux résultats de chaque réalisation.

  Dans le cadre du corpus : $SemTF(t, C)$ désigne le nombre d'occurrences de t dans le contexte C au sein du corpus augmenté des fréquences des termes sémantiquement proches

[15] Semantic Term Frequency.

[16] Il désigne plus exactement le marqueur ou l'élément structurant ainsi que le reste des termes qui constituent le vocabulaire délimité par le marqueur en question.

de t.

$$SemTF(t,C) = TF(t,C) + \sum^{t_i \in w_c} TF(t_i,C) \times Sim(\varphi_C(t), \varphi_C(t_i)) \tag{3.27}$$

– La représentativité sémantique du terme
Dans le cadre du document : $SemCF(t,d,C)$ [17] désigne le degré de rattachement (ou de représentativité) sémantique du terme t vis à vis du contexte C dans le cadre du document d. Autrement dit, cette mesure s'intéresse à la fréquence d'apparition du sens de t ($\varphi_C(t)$) dans le contexte C dans le document considéré.

$$SemCF(t,d,C) = \log\left(\frac{|SP_c^{\varphi_C(t)} \cap C_d|}{|C_d^{\varphi_C^(t)}|} + 1\right) \tag{3.28}$$

$|SP_c^{\varphi_C(t)} \cap C_d|$ est le nombre de spécialisations de c dans d contenant au moins une fois $\varphi_C(t)$ et $|C_d^{\varphi_C(t)}|$ désigne le nombre total des contextes dans d contenant au moins une fois $\varphi_C(t)$.
Dans le cadre du corpus : SemCF(t,C) indique la représentativité d'un sens $\varphi_C(t)$ pour un contexte C par rapport aux différents contextes du corpus.

$$SemCF(t,C) = \log\left(\frac{|SP_C^{\varphi_C(t)}|}{|C^{\varphi_C(t)}|} + 1\right) \tag{3.29}$$

$|SP_C^{\varphi_C(t)}|$ exprime le nombre de spécialisations du contexte C dans le corpus contenant au moins une fois $\varphi_C(t)$ et $|C^{\varphi_C(t)}|$ est le nombre total de contextes comportant au moins une fois $\varphi_C(t)$.

– Le pouvoir discriminatoire de la sémantique du terme
$SemIDF(t,C)$[18] désigne la force discriminatoire du sens $\varphi_C(t)$ et qui est proportionnellement inverse au nombre de documents contenant le contexte C et dans lequel $\varphi_C(t)$ apparaît au moins une fois ; $|D_C^{\varphi_C(t)}|$.

$$SemIDF(t,C) = \frac{|D_c|}{|D_C^{\varphi_C(t)}|} \tag{3.30}$$

Sachant que $|D_c|$ est le nombre de documents de la collection contenant au moins une fois le contexte C.

[17] Semantic Context Force.
[18] Semantic Inverse Document Frequency.

Ainsi, de la même façon que le poids des termes présenté dans la partie 3.4.5 a été défini, la pondération d'un terme t (SemTF-ICDF) dans le cadre d'un document d dans un contexte C augmenté par son voisinage sémantique, est définie en multipliant les trois mesures sus-citées :

$$SemTF - ICDF(t, d, C, n) = SemTF(t, d, C, n) \times SemCF(t, d, C) \times SemIDF(t, C) \quad (3.31)$$

3.7.2 Adaptation sémantique d'autres modèles de recherche

Des modèles de recherche sémantique à base de méta données ont également été développés. Ce ne sont plus les termes auxquels on s'intéresse mais plutôt le sens qu'ils véhiculent. Par conséquent, les documents sont automatiquement indexés suivant les concepts qu'ils renferment, les requêtes sont analysées et leur sens est comparé au sens des documents pour les apparier.

Corese (COnceptual REsource Search Engine) [27] est un moteur de recherche sémantique. Il permet de rechercher des ressources représentées par des URI[19] qui ont été préalablement annotées au moyen d'un vocabulaire conceptuel (ontologie). Corese construit une représentation interne de ces informations sous forme de graphes conceptuels. Pour interroger une base d'informations cela revient à interroger une base de graphes (ou méta données), ceci est réalisé à l'aide de requêtes sous forme de graphes également. La projection du graphe requête peut être totale ou approchée. La projection approchée est réalisée dans le cas où la projection standard (totale) de la requête ne fournit pas de réponses. Dans ce cas, Corese implémente un algorithme de recherche approchée qui retourne les meilleures approximations. L'estimation du taux de pertinence des réponses est réalisée en calculant des distances sémantiques dans l'ontologie ce que nous présentons dans la section 4.3.

D'autres moteurs de recherche implantent la sémantique tout en restant concentrés sur l'expression des besoins en langue naturelle. Parmi ces réalisations on peut citer Verity[20], Intuition et Excalibur[21]. Verity est fondé sur le principe probabiliste et réalise un auto-apprentissage en tenant compte des choix des usagers. Par conséquent, il met à jour sa représentation des concepts qui est initialement implémentée dans son système, ce qui lui permet d'améliorer progressivement la qualité de ses recherches. Par ailleurs, à l'aide du moteur Intuition[22], l'analyse sémantique est réalisée par un algorithme déterministe et non via un réseau de neurones. La liste des réponses produite suite à la soumission d'une requête sera la même que celle issue d'une soumission ul-

[19] Uniform Resource Identifier, est une courte chaîne de caractères identifiant une ressource physique ou abstraite sur un réseau.

[20] Le moteur de l'entreprise Autonomy, principal concurrent de Sinequa réalisateur du moteur Intuition.

[21] Produit de l'entreprise Convera : www.convera.com/products/excalibur/

[22] Produit de l'entreprise Sinequa. C'est une plate-forme de recherche pour entreprises, son approche est fondée sur des analyses statistiques, morphosyntaxiques, sémantiques, etc.

térieure de la requête initiale. Plus récemment (au mois de mai 2008), la start-up Powerset a annoncé publiquement le lancement de son nouveau moteur de recherche, en version bêta[23]. Ses particularités sont non seulement le fait qu'il sera à terme un système fonctionnant en ligne mais aussi le principe de son algorithme qui analyse le sens des phrases en entrée et de celles constituant les pages web qu'il indexe. Pour sa version bêta, les créateurs du moteur se sont limités à intégrer les pages de Wikipedia et de la base de données libres Freebase[24].

3.8 Conclusion

En plus de la variété des modèles de recherche de base implémentés par ces différents systèmes, leurs applications se distinguent par rapport au cadre sémantique qu'elles considèrent. Le cadre sémantique accommode deux notions complémentaires : la ressource sémantique et le modèle de mesure de similarité entre concepts. Dans le chapitre suivant nous définissons le concept « sémantique » en prenant appui sur un certain nombre de paradigmes de ressources sémantiques et leurs procédés de calcul d'appariement conceptuel adaptés.

[23] Consultable en ligne à l'adresse : www.powerset.com/explore/go/
[24] www.freebase.com

Chapitre 4

La sémantique et ses ressources

Ce chapitre a pour but de définir la notion de « sémantique », de présenter la nature de ses principales ressources ainsi que les modèles de calcul permettant d'évaluer le degré de rapprochement entre leurs composants.

Nous commençons ce chapitre par une définition du domaine de la sémantique (section 4.1), ensuite nous exposons une panoplie de ses ressources (section 4.2). Les mesures de similarité entre composants de ces ressources sont présentées (section 4.3). Nous détaillons le principe de l'ensemble des mesures dont on a pris connaissance durant notre étude, et cela en mettant l'accent sur leurs principales différences.

4.1 La sémantique

La sémantique est une branche de la linguistique qui étudie le sens des mots. Le mot « sémantique » existe depuis le XIXème siècle et c'est le linguiste français Michel Bréal qui l'a inventé. L'étude sémantique possède différents domaines d'étude dont :

- La signification des mots (et des mots composés)
- Les rapports de sens entre les mots (relations d'homonymie, de synonymie, d'antonymie, de polysémie, d'hyperonymie, d'hyponymie, etc.)
- La distribution des actants au sein d'un énoncé
- La pragmatique, qui est très souvent considérée comme une branche de la sémantique (section 2.2.4).

On parle de liens sémantiques entre deux termes lorsqu'il existe entre eux un lien de sens alors que leur forme n'a rien en commun (pas de lien morphologique). Ces variations recouvrent donc les termes qui se correspondent dans une relation de synonymie voire d'hypéronymie. L'hyperonymie (Opposé de l'hyponymie) est la relation hiérarchique entre termes suivant leur sens. Le

concept (ou le terme) de plus haut niveau englobe l'extension du second, plus spécifique. Le terme plus générique est dit hyperonyme et le plus spécifique est l'hyponyme (appellation dans le cadre d'une hyponymie).

De manière générale, le principal intérêt de la gestion de la sémantique dans le cadre global des systèmes de recherche est de pouvoir contourner les problèmes résiduels de l'indexation automatique qui sont à l'origine de la baisse des performances des SRI. Ces difficultés sont dues notamment à la synonymie et à la polysémie (et l'homonymie). Lorsqu'un utilisateur exprime une requête en langue naturelle (ou par une série de mots clés) il utilise son propre vocabulaire voire des expressions très imagées. Pour un système de recherche classique établissant un appariement terme-à-terme, il est nécessaire qu'un document pertinent ait un lexique (forme morphologique) similaire (baisse de la performance du rappel). Et inversement, dans le cadre des termes polysémiques. En effet, ce même système de recherche estimera un document utilisant des termes morphologiquement semblables aux termes de la requête et ayant des sens différents comme très pertinents (baisse de la performance de la précision).

En plus de la plysémie et l'homonymie, les mots d'une langue entretiennent un réseau riche de relations sémantiques : hyperonymie/hyponimie, méronymie(lien de composition : « volant » est une pièce de « voiture »), antonymie (malin-benin), ... etc. La construction des ressources sémantiques est fondée sur cette notion de structure des connaissances, ces bases sont utilisées en tant que support des analyses sémantiques des textes ou d'acquisition lexicale (pour la reconnaissance des mots composés par exemple).

4.2 Les ressources sémantiques

Nous allons citer dans ce qui suit les principaux paradigmes existant pour la représentation du sens.

4.2.1 Les thésaurus et les taxonomies

Un thésaurus[1] est un vocabulaire contrôlé hiérarchisé. Il est le résultat d'un long processus d'indexation qui aboutit à la sélection de deux types de termes, des termes descripteurs pour l'indexation et des termes non-descripteurs qui ne ressortent pas du corpus. Les non-descripteurs sont néanmoins employés dans le thésaurus puisqu'ils renvoient à des descripteurs terminologiques issus du langage documentaire.

Les thésaurus les plus répandus concernent généralement un domaine précis où l'on utilise un langage opératif. À titre d'exemple nous pouvons citer la hiérarchie thématique du MeSH[2]. Le

[1] Les deux orthographes thesaurus et thésaurus sont admises par les dictionnaires, mais la forme francisée (avec accent) semble la plus fréquente dans la littérature.

[2] Medical Subject Headings.

MeSH est le thésaurus de la base de données bibliographiques MEDLINE[3], il est le thésaurus de référence dans le domaine biomédical. Le système IRAIA [57] utilise un thésaurus spécialisé pour l'interrogation et la gestion de documents économiques. Par ailleurs, Wordnet (section 4.2.1) est un thésaurus à caractère généraliste pour la langue anglaise dont l'organisation dépend du bon sens humain.

Une taxonomie[4] est un vocabulaire contrôlé organisé sous forme hiérarchique simple. Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation donnant ainsi un sens supplémentaire à un terme. La différence avec le thésaurus, c'est que ce dernier permet de parcourir la hiérarchie de manière connexe permettant ainsi de « restreindre » ou de « spécialiser » le champ de connaissance. Sommairement, un thésaurus est une taxonomie qui fonctionne dans les deux sens.

WordNet

De par sa complétude et sa libre disponibilité WordNet[5] [42] est une ressource (ou base de données) lexicale de référence dans le cadre des travaux du TAL. Sa conception est inspirée par les théories psycholinguistiques de la mémoire lexicale humaine. En effet, dans la compilation WordNet, les mots de la langue anglaise sont hiérarchisés et mis en relation manuellement selon les différents types de liens sémantiques qui peuvent exister entre eux. Schématiquement, WordNet se présente sous forme d'une arborescence de concepts appelés « senset ». Un senset est la composante atomique du réseau, il est constitué d'un jeu de termes dénotant un sens ou un usage commun. WordNet se compose de quatre réseaux qui représentent les quatre super catégories syntaxiques principales : noms, verbes, adjectifs et adverbes.

WordNet en français

WordNet a inspiré des travaux visant à réaliser une ressource similaire pour d'autres langues. La Global WordNet Association[6] répertorie des compilations dans le genre de WordNet pour plus de 50 langues. On peut citer EuroWordNet pour certaines langues d'Europe de l'Ouest [124] ou BalkaNet pour l'Europe de l'Est [121]. Néanmoins, la version anglaise reste la plus complète à ce jour. En effet, EuroWordNet par exemple, ne comporte que des lexèmes verbaux et nominaux, aucun adjectif ni aucun adverbe n'y sont mentionnés. De plus, il n'a pas été largement utilisé, principalement en raison de problèmes d'acquisition de licence. Jacquemin [21] a même mentionné en 2007 qu'aucun projet n'a pris le relais pour poursuivre l'extension et l'amélioration de EuroWordNet français. Ce n'est que très récemment (en 2008) qu'il existe une tentative d'élaborer un WordNet pour le français. Cette initiative se poursuit aujourd'hui grâce à des

[3] medline.cos.com/

[4] Ou taxinomie, taxis pour « ordre » et nomos pour « nom ».

[5] Projet développé depuis 1985 par des linguistes de l'université de Princeton.

[6] www.globalwordnet.org/

travaux de recherche [38]. Cette ressource est nommée WOLF (WordNet Libre du Français), elle exploite des ressources multilingues librement disponibles pour construire automatiquement un WordNet français à large couverture à disponibilité[7] libre également. Cette ressource est réalisée automatiquement, son évaluation est réalisée par rapport à l'EuroWordNet français. Néanmoins, et par rapport à la réalisation automatique de WOLF, une validation et des enrichissements manuels sont nécessaires. Ceci est envisagé dans l'avenir par l'équipe de recherche en charge de la réalisation de WOLF.

4.2.2 Les ontologies

À l'origine, "ontologie"[8] est un terme philosophique qui désigne « science de l'être » [9]. Alors qu'un thésaurus ou même une taxonomie structurent des termes, l'ontologie elle, structure des concepts[9]. De plus, lorsque l'on établit une hiérarchisation de concepts dans une ontologie, on établit des dépendances entre ces concepts. Ces hiérarchisations ont un sens en dehors du vocabulaire lui-même. L'objectif à travers une ontologie est de modéliser explicitement la connaissance à un niveau conceptuel en utilisant un langage « formel » et « commun » offrant une sémantique plus ou moins rigoureuse permettant une utilisation non-ambiguë du domaine. Par « formel » il faut entendre que l'ontologie doit être compréhensible par les outils informatiques et par l'adjectif « commun » il faut comprendre la prise en compte d'un savoir consensuel, c'est à dire qu'une ontologie n'est pas l'objet d'un individu mais qu'elle doit être reconnue par un groupe.

On peut ainsi dire que le lexique sémantique (taxonomie et thésaurus) structure le vocabulaire, tandis que l'ontologie structure le monde[10] par des concepts et fait appel au vocabulaire pour « étiqueter » ceux-ci [100]. Les définitions de l'ontologie dans le domaine de l'intelligence artificielle abondent dans la littérature. Les définitions que nous avons pu consulter, dans leur diversité, offrent des points de vue à la fois différents et complémentaires concernant l'ontologie. Toutefois la définition la plus célèbre et la plus citée est celle de Gruber, elle est formulée par : « une ontologie est une spécification explicite d'une conceptualisation ».

Avec l'intérêt grandissant du public pour le web sémantique et celui des entreprises pour les systèmes de recherche performants, les ontologies possèdent aujourd'hui une forte dynamique. Plusieurs ontologies sont disponibles gratuitement sur le net comme Ontolingua[11] ou partiellement disponibles comme l'ontologie de haut niveau et qui fournit des concepts très génériques du sens commun américain Cyc[12]. Les ontologies connues, dans leur plus grand nombre, ont été mises au point par des compagnies pour leur propre utilisation et ne sont ,par conséquent, pas disponibles au grand public.

[7] WOLF peut être téléchargée à l'adresse suivante : alpage.inria.fr/ sagot/wolf.html

[8] Du grec « ontos » et « logos » qui signifient respectivement « essence » et « vie ».

[9] Il arrive que les définitions des ontologies aient été diluées, en ce sens que les taxonomies sont considérées comme des ontologies complètes.

[10] Ou plus rigoureusement « un certain modèle d'un certain monde ».

[11] www.ksl.stanford.edu/software/ontolingua/

[12] www.cyc.com/

Parmi les ontologies linguistiques générales indépendantes de tout domaine et de tout type de tâche, nous pouvons citer : le Generalized Upper model (GUP)[53]. Comme son nom l'indique, le Generalized Upper model a été prévu pour être transféré dans plusieurs langues, c'est la raison pour laquelle il ne contient que les notions linguistiques principales et leurs organisations possibles dans toutes les langues. Dans le domaine des ontologies d'ingénierie, les ontologies Eng-Math [119] et PhysSys [18] ont été mises au point pour la modélisation des bases conceptuelles des mathématiques et de la physique respectivement.

4.2.3 Les dictionnaires

Au même titre qu'un thésaurus, les dictionnaires électroniques représentent des ressources lexico-sémantiques. Qu'ils soient généralistes ou spécialisés, les dictionnaires sont des ouvrages qui reprennent un maximum du vocabulaire courant ainsi que ses différentes significations. Il s'agit là d'un atout majeur pour un système de recherche d'information implantant la sémantique, car la ressource sémantique assurée par le dictionnaire se rapproche ainsi d'une certaine exhaustivité, tant au niveau des termes à considérer qu'au niveau des liens possibles entre mots.

Dans ce contexte, nos pouvons citer l'initiative de la Vidéothèque de Paris qui propose depuis 1989 une interrogation en langue naturelle grâce à un dictionnaire gérant les associations entre le langage documentaire utilisé par les documentalistes de la Vidéothèque et les interrogations des lecteurs. Le système de recherche de la vidéothèque tente ainsi d'enjamber le fossé séparant termes contrôlés et langue naturelle [77].

Un phénomène lexical important est la synonymie, et naturellement, pour qu'un système informatique puisse établir ce lien, il est indispensable qu'il dispose des ressources nécessaires, comme notamment un dictionnaire des synonymes ou une autre ressource sous une forme équivalente. Le dictionnaire des synonymes français (DICTIONNAIRE) du laboratoire CRISCO du CNRS est une ressource en ligne depuis 1998. Il se présente sous la forme d'un fichier texte de 48881 lignes[13], où chaque entrée correspond à un mot-vedette suivi de la liste de ses synonymes, comme par exemple ci-dessous pour le mot « maison » :

maison : abri, appartement, asile, baraque, bas-lieu, bâtiment, bâtisse, bercail, bicoque, boîte, bouge, branche, building, bungalow, cabane, cahute, campagne, case, cassine, chacunière, chalet, château, chaumière, chez-soi, clapier, clinique, commerce, construction, couronne, couvert, demeure, descendance, domesticité, domestique, domicile, dynastie, édifice, entreprise, établissement, famille, ferme, feu, firme, foyer, galetas, gens, gîte, gourbi, habitacle, habitation, home, hôpital, hôtel, hutte, immeuble, institut, institution, intérieur, lares, lieu, lignée, logement, logis, maisonnée, maisonnette, manoir, masure, ménage, monde, naissance, nid, nom, origine, palais, parents, pavillon, pénates, pigeonnier, place, prison, propriété, race, réduit, résidence, retraite, séjour, serviteur, suite, taudis, temple, toit, trône, villa.

[13] Nombre d'entrées calculé à la fin de l'année 2007.

Ces mots-vedettes et leurs synonymes proviennent de la compilation de sept dictionnaires, dont des dictionnaires de synonymes : René Bailly[14], Henri Bénac[15], Henri Bertaud du Chazaud[16], François Guizot[17], Pierre-Benjamin Lafaye[18], Le grand Larousse de la Langue Française[19] et Le grand Robert[20].

DICTIONNAIRE est accessible en ligne au grand public[21], cependant une convention est nécessaire pour l'obtention des droits d'accès au fichier des vedettes. Cette convention est élaborée sous forme d'un règlement d'ensemble pour la cession de données à des fins de recherche entre universités.

Les ressources sus-citées mettent le vocabulaire à plat et l'organisent selon les liens possibles entre ses termes. Afin de calculer la valeur de la distance sémantique entre termes, des mesures adaptées à ces différentes ressources sont employées.

4.3 Les mesures de similarité sémantique

4.3.1 Méthodes appliquées aux structures hiérarchiques

Nous allons faire un survol des recherches classiques menées dans le cadre du graphe conceptuel WordNet où les nœuds[22] sont sensés représenter des concepts dans les travaux cités. Dans le cadre de ces derniers, les chemins[23] des graphes sont utilisés pour mesurer la distance entre les concepts. Seuls les liens d'héritage dans la hiérarchie des noms de WordNet (liens « is-a) sont pris en compte dans ces mesures.

La mesure de Rada [101]
Pour Rada cette démarche est la plus intuitive. Cette mesure utilise une mesure, $dist(C_1, C_2)$, pour calculer la distance minimum en nombre d'arcs à parcourir pour aller du concept C_1 au concept C_2.

$$Sim_{Rada}(C_1, C_2) = \frac{1}{1 + dist(C_1, C_2)} \tag{4.1}$$

[14] 1964 aux éditions Larousse.
[15] 1956 aux éditions Hachette.
[16] 1971 aux éditions Robert.
[17] 7ème édition apparue en 1864 aux éditions Didier.
[18] 1858 aux éditions Hachette.
[19] 1971 chez les éditions Larousse.
[20] 1985 chez les éditions Robert.
[21] elsap1.unicaen.fr/dicosyn.html
[22] Senset dans WordNet.
[23] Enchaînement d'arcs.

La mesure de similarité conceptuelle de Leacock et Chodorow [65]
Elle est fondée elle aussi sur la distance minimale entre deux concepts dans le graphe conceptuel et sur la profondeur maximale du graphe.

$$Sim_{LC}(C_1, C_2) = -\log(\frac{dist(C_1, C_2)}{2 \times (profmax)}) \tag{4.2}$$

Où $profmax$ est la profondeur maximale de la taxonomie conceptuelle considérée (égale à 16 dans WordNet 1.7).

La démarche de Wu et Palmer [90]
Wu et $Palmer$ utilisent la notion du généralisant commun de C_1 et C_2 de profondeur maximale (que nous noterons dans la suite LCS[24]) et la profondeur des concepts dans la hiérarchie pour évaluer la similarité conceptuelle.

$$Sim_{WP}(C_1, C_2) = \frac{2 \times prof(LCS)}{prof(C_1) + prof(C_2)} \tag{4.3}$$

La mesure de Zargayouna [133]
Dans le cadre de son travail d'indexation des documents semi-structurés, Zargayouna définit une mesure de similarité conceptuelle inspirée des travaux de Wu et Palmer. Elle souhaite privilégier toujours les liens père-fils aux autres liens de voisinage en adaptant la mesure de Wu et Palmer qui pénalise, dans certains cas, les fils d'un concept par rapport à ses frères. Pour ce faire, elle augmente le dénominateur de la mesure de Wu et Palmer de la fonction de pénalisation des concepts qui ne sont pas de la même lignée, spec.

$$Sim_{Zarga}(C_1, C_2) = \frac{2 \times prof(LCS)}{prof(C_1) + prof(C_2) + spec(C_1, C_2)} \tag{4.4}$$

$$spec(C_1, C_2) = dist(C_1, LCS) \times dist(C_2, LCS) \times prof(LCS) \tag{4.5}$$

$spec(C_1, C_2)$ est égale à 0 si C_1 et C_2 se trouvent sur la même lignée.
Cette mesure n'utilise pas le principe du contenu informatif parce qu'elle est utilisée conjointement avec la mesure distributionnelle du modèle vectoriel des termes dans les documents pour les apparier.

[24] Least Common Subsumer : en anglais le généralisant de proche en proche commun à C_1 et C_2 le plus éloigné de la racine du graphe.

La mesure de Corby [27]
Dans le cadre du développement du moteur de recherche Corese (section 3.7.2), Corby détermine et implémente un algorithme de calcul de distance sémantique entre nœuds dans un graphe conceptuel. La similarité entre un concept requête et un concept cible correspond à la plus petite somme des longueurs l entre chacun des types des concepts et leur LCS. La longueur d'un arc entre un fils et un père de profondeur d est $1/2^d$, faisant ainsi en sorte que la distance entre concepts décroît avec la profondeur. Ce qui est vrai, étant donné que dans un réseau conceptuel les différentes zones ne sont pas de densité homogène. En effet, il est connu que les concepts frères situés aux premiers niveaux sont sémantiquement plus éloignés que les concepts frères situés à des niveaux plus profonds.

$$dist(C_1, C_2) = \min\left(l(t_1, LCS), l(t_2, LCS)\right) = \min\left(\sum_{x \in <t_1, LCS>, x \neq t_1} 1/2^{d_x}, \sum_{x \in <t_2, LCS>, x \neq t_2} 1/2^{d_x}\right) \tag{4.6}$$

4.3.2 Méthodes basées sur le contenu informatif

Le contenu informatif d'un concept CI a été introduit en premier dans les travaux de Resnik.

Le CI selon Resnik [97]
Selon Resnik, le CI d'un concept est calculé à partir d'un corpus d'apprentissage de spécialité et traduit le degré de spécificité d'un concept dans ce corpus. Resnik évalue le contenu informatif d'un concept C, CI(C), en appliquant l'entropie de Claude Shannon [111]. Le CI est en relation inverse avec p(C), sachant que p(C) est la probabilité de rencontrer le concept C ou un de ses descendants dans le corpus. Le contenu informatif d'un concept C est calculé par la mesure entropique de la théorie de l'information comme suit :

$$CI(C)_{Resnik} = -\log(p(C)) \tag{4.7}$$

Quand p(C)=0, le contenu informationnel du concept C est indéfini.

Le CI selon Seco [110]
Dans les travaux de Seco, le CI des nœuds est calculé uniquement à partir du réseau conceptuel WordNet. Le principe de cette approche est que plus un concept a de descendants (hyponymes), moins il est informatif, ce qui est évident, puisque il est moins spécifique (pertinent) que ses

concepts fils.

$$CI(C)_{WordNet} = 1 - \frac{\log(hypo(C)+1)}{max_{WordNet}} \tag{4.8}$$

La valeur de hypo(C) correspond au nombre de descendants dont dispose le concept C, et $max_{WordNet}$ est une constante qui indique le nombre total des concepts dans WordNet (la division est une sorte de normalisation de la proportion des concepts fils de c au sein de WordNet).

Ainsi, pour calculer la proximité entre deux concepts C_1 et C_2 à partir de leur contenu informatif, il est nécessaire de trouver l'ensemble des concepts qui les généralisent, soit E cet ensemble. La proximité entre C_1 et C_2 correspond à la valeur maximum du CI des éléments de E.

$$Sim(C_1, C_2) = \max_{c \in E} [CI(C)] \tag{4.9}$$

4.3.3 Méthodes hybrides

Elles consistent en une approche d'appariement conceptuel mixte qui combine la notion des arcs de chemin et celle du CI des nœuds.

La mesure de Jiang et Conrath [54]
Cette approche attribue un poids aux arcs du graphe. Dans une taxonomie le poids d'un lien dépend de plusieurs éléments tels que la densité locale des différentes parties de la taxonomie et la profondeur des concepts. En raison de ces différents critères, le poids d'un arc wt reliant un concept fils f à son concept père p est calculé par Jiang et Conrath de la manière suivante :

$$wt(f,p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}[CI(f) - CI(p)]\,T(f,p) \tag{4.10}$$

Où d(p) désigne la profondeur du concept père p, E(p) est la densité locale de p [25], $\overline{E}$ est la densité moyenne de la taxonomie et T(C,p) le nombre de types de liens reliant le nœud fils au nœud père [26]. Les paramètres α et β gèrent l'influence de la profondeur et la densité locale sur le calcul du poids des arcs.

[25] Il est évident que si un nœud parent a beaucoup de fils, il est nécessaire d'augmenter la distance entre père et fils.

[26] Suivant les types de liens, les distances parcourues sont variables.

Étant donné que la plupart des applications s'intéressent exclusivement aux relations hiérarchiques is-a, la valeur de T(f,p) est égale à 1. Les réseaux sémantiques sont très souvent considérés de structure homogène, par conséquent les paramètres α et β sont mis respectivement à 0 et à 1. La distance globale du nœud C_1 au nœud C_2 est calculée de proche en proche (entre deux concepts quelconques C_1 et C_2 et leur LCS, les CI des concepts intermédiaires s'annulent en agrégeant les poids des arcs).

$$dist_{JC}(C_1, C_2) = CI(C_1) + CI(C_2) - 2\,(LCS(C_1, C_2)) \tag{4.11}$$

$$Sim_{JC}(C_1, C_2) = \frac{1}{dist_{JC}} \tag{4.12}$$

La mesure de Lin [70]
Elle combine elle aussi l'information contenue dans les nœuds et la structure de la hiérarchie.

$$Sim_{Lin}(C_1, C_2) = \frac{2 \times CI(LCS)}{CI(C_1) + CI(C_2)} \tag{4.13}$$

4.3.4 Méthodes appliquées aux dictionnaires des synonymes

La représentation du dictionnaire des synonymes « DICTIONNAIRE » permet de calculer la distance entre les différentes paires de mots en fonction du nombre de synonymes communs [74]. La distance de Jaccard est la métrique utilisée pour calculer le taux de similitude entre ensembles. L'indice de Jaccard est le rapport entre la cardinalité de l'intersection des ensembles et la cardinalité de l'union sans doublons des ensembles. De cette façon, dans DICTIONNAIRE les taux de similarité entre deux mots sont calculés à l'aide du coefficient de Jaccard en considérant comme ensembles la totalité des termes synonymes associés à une vedette en intégrant la vedette en question.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.14}$$

Ainsi, pour la paire de sommets A et B du graphe des synonymes, considéré dans la figure 4.1, l'indice de similitude est défini par le nombre de sommets communs (qui correspond à 4), divisé par le nombre de sommets en relation avec l'un des deux membres de la paire (12). L'indice de similarité selon le coefficient de Jaccard pour cet exemple entre A et B est de 4/12.

4.4 Conclusion

« L'indexation du sens » des textes rédigés en langue naturelle est une tâche ardue qui nécessite des ressources sémantiques adéquates. En effet, ces éléments de connaissance doivent s'adapter au langage usité. Notre corpus des plaintes provient de sources diverses. Les agents en charge de récupérer les plaintes, le plus souvent, reprennent le vocabulaire du plaignant dans la description de la situation et les complètent, dans un vocabulaire plus ou moins dédié, d'informations constatées sur site. Par conséquent, la question qui se pose face à cette réalité, et qui de

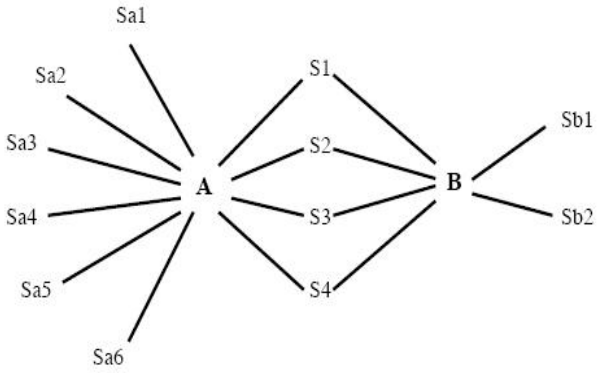


FIG. 4.1 – Exemple de configuration synonymique entre deux sommets A et B

manière générale est source de polémique dans le domaine, quelle ressource parmi celles sus-citées allons nous utiliser au détriment des autres, et pour quelles raisons ? Dans le chapitre suivant du mémoire nous argumentons la stratégie de notre approche pour contourner les difficultés dues notamment à la nature du langage et aux objectifs que nous nous sommes assignés.

Chapitre 5

Notre approche

Nous réalisons dans le cadre de cette thèse un système informatique pour automatiser l'analyse des situations de pollution intérieure exprimées dans des textes de plaintes pour leur apporter des solutions. Le système que nous proposons s'appuie principalement sur une base archive évolutive constituée de plaintes écrites résolues. Cette mémoire, constituée de problèmes expliqués par leurs solutions, est destinée à se développer au fur et à mesure que de nouveaux cas seront saisis, résolus et vérifiés à travers l'applicatif proposé.

La première partie de ce chapitre est consacrée à la présentation du modèle de raisonnement des personnes qui assurent les diagnostics (section 5.1). En second lieu, nous exposons le principe, les avantages et les inconvénients de quelques méthodes à base de connaissances que nous avons examinées dans le cadre de notre recherche (sections 5.2, 5.3 et 5.4). Nous présentons ensuite le schéma de l'approche que nous adoptons pour apporter des solutions automatiques aux plaintes écrites (section 5.5). Nous analysons d'abord avec précision la nature de notre corpus des plaintes air intérieur constituant une base de connaissance essentielle dans notre démarche (section 5.6). Ensuite, nous détaillons la composition du module fonctionnel qui est au coeur de notre système. Nous mettons en évidence dans cette partie du chapitre nos contributions dans le cadre des systèmes de recherche, notamment concernant l'application de la sémantique (section et 5.7.2) ainsi que la définition d'un nouveau modèle fondé sur la théorie du signal (section 5.7.3). Suivra, l'étude du vocabulaire des plaintes que nous avons mis en œuvre pour la caractérisation de la ressource sémantique nécessaire (section 5.8). Dans la section suivante, nous exposons la ressource sémantique sélectionnée et notre apport quant à son adaptation aux besoins d'appariements sémantiques reposant sur la sélection des racines des lemmes (section 5.9). Enfin, nous exposons la méthode de construction de la base des scénarios possibles en pollution intérieure à partir des plaintes d'un échantillon représentatif (section 5.10). Nous présentons plus précisément l'adaptation de la catégorisation des textes aux différents formalismes des textes utilisés.

5.1 Philosophie de l'approche

Nous souhaitons développer un système informatique dont le modèle de raisonnement s'apparente le plus possible à la logique des experts. Les experts enquêtent sur les sites en veillant à respecter les normes en vigueur dans le domaine (section 1.7), toutefois leurs observations et leurs déductions dépendent d'autres facteurs propres à chaque sujet expert influençant la variabilité des jugements. L'élément essentiel déterminant le jugement expertal pour la réalisation des diagnostics est certes l'« intime conviction » du diagnostiqueur. L'intime conviction est une méthode de jugement basée essentiellement sur l'« expérience ». Par expérience nous entendons la pratique qui est obtenue en étudiant des cas variés participant ainsi à enrichir le savoir faire des agents experts. Spontanément, ces derniers capitalisent leurs connaissances en les utilisant et en les mettant en synergie pour l'analyse de nouvelles situations.

La mise en valeur de l'historique expérimental dans l'activité quotidienne est aussi l'approche adoptée par les médecins. La métaphore médicale est significative dans notre contexte. D'une part, le domaine de la pollution de l'air intérieur est un domaine intimement lié à la santé publique, et d'autre part le diagnostiqueur des sites concernés par le phénomène de la pollution atmosphérique est sensé avoir certaines notions médicales et surtout pouvoir associer l'état du logement et de son environnement à l'état sanitaire des occupants. Qu'il s'agisse du raisonnement purement médical ou du raisonnement utile au diagnostic de la qualité de l'air intérieur, nous empruntons le mode de jugement expert sus-décrit pour définir notre approche dédiée à la réalisation du système d'aide à l'analyse des cas de pollution domestique à partir des plaintes écrites.

Les premières applications informatiques fondées sur le raisonnement des sujets experts sont les systèmes basés sur les règles ou encore les systèmes experts.

5.2 Les systèmes experts

Le « Système Expert », ou « SE », est un outil doté des mécanismes cognitifs d'un expert d'un domaine particulier. À partir d'un ensemble de faits donnés, il est sensé permettre l'exploitation d'une base de règles pour réaliser le raisonnement. Le SE permet par conséquent de répondre à un utilisateur sans que ce dernier ait besoin d'assimiler les connaissances du domaine modélisées au moyen de la base de règles. Prenons le syllogisme classique d'Aristote formalisé par la règle suivante : si a est b et b est c, alors a est c. Supposons que l'usager du système énonce les deux faits suivants : « homme est mortel » « Socrate est un homme ». La déduction établie par le système régi par la règle sus-citée est comme suit :

Socrate est un homme → homme est mortel → Socrate est mortel.

Les systèmes experts réalisent de bonnes performances dans de nombreux domaines d'applications, cependant ils présentent un certain nombre de limitations.

5.2.1 Limites des systèmes experts classiques

À l'instar de la plupart des outils informatiques et malgré la richesse des bases de connaissances, le SE manque du «bon sens» humain [25]. En effet, il est régi par une base de règles établie pour la résolution de problèmes spécifiques, le SE n'a donc pas la capacité de remettre en question sa stratégie [71]. Un SE n'est généralement pas capable de fournir autre chose que ce que la séquence des règles lui permet. Par ailleurs, un système basé sur des règles ne peut pas apprendre, son évolution ne peut s'effectuer que par la modification des règles qu'il utilise ce qui nécessite l'intervention d'un spécialiste [25, 71].

Enfin, écrire une base de règles est un processus relativement coûteux. Le but des systèmes experts de seconde génération est de contourner cette difficulté en substituant la base de connaissances des systèmes experts primitifs, subissant la rigidité des règles, par un ensemble de cas. Ainsi, la base de connaissances des nouveaux systèmes experts est constituée de problèmes expliqués par des solutions et validés par les spécialistes du domaine.

5.2.2 Les systèmes experts de seconde génération

Sous l'appellation « Systèmes Experts de Seconde Génération », ou « SESG », se trouvent les « Systèmes à Bases de Connaissances », ou « SBC », actuels. Le but commun de l'ensemble de ces travaux est de dépasser les difficultés inhérentes essentiellement de la tâche de construction des bases de connaissances fondées sur le formalisme de règles. Le cas est un formalisme de connaissance récent. Le premier avantage des modèles fondés sur les cas est dans la facilité de constitution d'une base de cas par rapport à une base de règles classiques. Par ailleurs, la configuration fondée sur le cas se rapproche encore plus du raisonnement humain. Le fondement de cette famille de systèmes est que, plus un cas passé est pertinent par rapport à un cas courant à traiter, plus le cas ancien apparaît comme un « cas d'école » exprimant une vraie expertise, permettant ainsi d'atteindre la solution.

5.3 Le RàPC

Le « Raisonnement À Partir de Cas », ou « RàPC », est une approche d'implémentation des SBC. Il se situe dans le cadre des travaux sur les systèmes experts de seconde génération. Son principe consiste à résoudre de nouveaux problèmes en utilisant des expériences passées. L'ensemble des expériences forme une base de cas. Un cas de la base (cas source) est constitué d'au moins deux parties : une description de la situation représentant le « problème », et la « solution » qui fut utilisée pour remédier à la situation en question. Pour résoudre un problème nouveau (cas cible) le RàPC consiste à produire une nouvelle solution en adaptant celle de la situation similaire, supposée transposable, au problème à résoudre.

5.3.1 Le cycle du RàPC

Les étapes successives assurant la réponse aux cas au moyen du paradigme du RàPC sont :

1. Élaboration (ou formalisme) des cas
   Il existe trois modèles établis pour formaliser l'ensemble des cas [13].
   - Dans *le modèle structurel*, les cas sont des ensembles de paires (attribut, valeur). Les valeurs des attributs sont le plus souvent des symboles, des entiers, des réels ou des booléens.
   - *Le modèle conversationnel* est semi-structuré. Il est utilisé lorsqu'il s'avère difficile de caractériser un cas à l'aide de valeurs numériques ou symboliques. Un cas conversationnel est une série de rubriques textuelles obéissant à une structure sémantique 3.2 donnée.
   - Dans *le modèle textuel*, les travaux en RàPC portent sur des expériences dont la description est purement textuelle.
2. La phase d'appariement
   Ce module consiste à extraire de la mémoire archive le cas source le plus similaire au cas cible. Pour cela, des mesures de similarité sont établies dépendamment du formalisme de représentation des cas.
3. La phase d'adaptation
   Comme son nom l'indique, ce module adapte la solution du cas jugé le plus similaire au contexte du cas courant.
4. La phase d'évaluation et de correction
   Comme nous l'avons cité précédemment (chapitre 2), le traitement automatique de la langue (TAL) est un domaine où il est difficile de fournir des fonctions d'évaluation automatique (lorsqu'il s'agit de représentation semi-structurée ou textuelle). Une correction et une validation humaine des cas, remémorés et adaptés automatiquement, est nécessaire.
5. La phase d'apprentissage et maintenance de la base de cas
   Elle consiste à capitaliser en mémoire l'expérience d'un nouveau cas jugé valide par l'expert.

5.3.2 Limites du RàPC

Le RàPC repose ainsi sur des connaissances de transposition qu'il faut acquérir et représenter. Bien que l'étape d'appariement (correspondant aux systèmes de recherche définis dans le chapitre 2) ait été abondamment étudiée et formalisée, la phase d'adaptation a été que très peu abordée. Cette situation est due en partie à la nature et aux objectifs des applications réalisées dans le cadre du RàPC. La plupart de ces réalisations sont inscrites dans de domaine des systèmes supports de formations à distance [55], des forums interactifs à base de cas et autres foires aux questions FAQ. Néanmoins, l'étude de Lamontagne [62] s'est orientée plus particulièrement vers la mise en œuvre de nouvelles solutions textuelles par la modification du contenu des anciennes

solutions.

Pour son application fondée sur le principe du RàPC textuel et qui est dédiée à la réponse automatique aux courriers électroniques, Lamontagne [62] définit un processus pour réutiliser la solution ancienne dans la construction d'une nouvelle. Une des étapes de ce processus est la phase de sélection des extraits pertinents. Ces passages pertinents sont sujets à modification, et pour les distinguer, Lamontagne propose de réaliser une extraction des entités nommées, qui est une technique du TAL nécessitant des ressources externes appropriées 2.2.3. Alors que en RàPC les cas sont supposés transposables [113, 87], réaliser l'adaptation de cas textuels ou semi-structurés exprimés de façon libre, nous semble à ce jour irréalisable à l'aide de cette technique modélisant l'adaptation.

Par ailleurs, les problèmes résolus sont supposés être des cas **typiques**. Par conséquent, les domaines notamment les plus récents, où il est encore difficile d'attester de manière absolue l'existence d'une régularité concernant la nature des cas, ne peuvent être que partiellement modélisables au moyen du paradigme RàPC. La condition portant sur la régularité des cas est très importante, puisqu'elle traduit l'exhaustivité de la base des cas. Ce critère absolu s'avère un frein, toutefois, afin d'y apporter une solution plusieurs propositions ont été avancées : la création de plusieurs bases de cas, l'utilisation de « cas virtuels » ou l'utilisation d'ontologies pour maîtriser les différences sémantiques entre plusieurs bases de cas [125, 98]. Ces contributions n'améliorent la difficulté que sur un plan théorique, puisque concrètement les ressources citées et qui sont nécessaires à un bon raisonnement sont très souvent incomplètes, complexes et difficiles à la mise en œuvre.

5.4 Les systèmes experts de seconde génération et la réponse aux plaintes

Il a été démontré [ref bulletin sécurité] qu'il n'est pas toujours nécessaire d'effectuer des mesures ou des prélèvements des contaminants pour résoudre les problèmes relatifs à la qualité de l'air intérieur. Une évaluation initiale de l'état du logement, dans la plupart des cas, permet de délimiter le problème. Cette information nous a offert au début de notre thèse une piste de réflexion importante pour l'automatisation du processus de résolution des cas de pollution domestique à partir des textes des plaintes des particuliers.

Cependant, et comme nous l'avons vu dans le chapitre 2, l'extraction de connaissance à partir des textes pour l'analyse du contenu de la plainte est une tâche complexe. L'approche n'est pas aussi déterministe que le traitement de questionnaires renseignés, comme ceux établis lors de l'enquête logement de l'Observatoire de la Qualité de l'Air Intérieur OQAI. Dans la section suivante, nous exposons le niveau d'adaptation des questionnaires pour la modélisation des conditions domestiques en lien avec la qualité de l'air intérieur. Ensuite, nous positionnons les propriétés du RàPC et des SESG par rapport aux particularités des plaintes liées au domaine de

la pollution de l'air dans les logements.

5.4.1 Réflexion sur les questionnaires

Les questionnaires de la campagne logements de l'OQAI portent sur la description du bâtiment et de son environnement, sur les occupants et leurs activités et sur l'état de santé au niveau respiratoire et allergique. Ces formulaires permettent de saisir des contextes de vie dans des bases de données fortement structurées. Les enquêtes de l'OQAI ont été réalisées au sein de logements constituant un échantillon représentatif du parc de logements français[1]. Ces habitations ne connaissent donc pas toutes forcément des situations de malaises dues à la qualité de l'air intérieur. Les champs des formulaires sont renseignés par des valeurs numériques ou symboliques.

Pour des cas de pollution exprimés entièrement de manière formalisée à l'aide de formulaires, un système expert doté d'une base de règles permettra de déduire la conclusion. Cependant, est-il possible et évident d'établir une base de règles capitalisant la connaissance du domaine de la pollution domestique ? L'expertise de terrain du domaine de la pollution intérieure est encore très peu formalisée. À ce jour, il est encore impossible de dresser une liste exhaustive des sources de pollution, dans une base de règles notamment. Cela est lié d'une part à la nouveauté du domaine, et d'autre part à la particularité de chaque foyer, de ses équipements possibles et de chaque style de vie.

5.4.2 La mise en correspondance des SESG et la réponse aux plaintes

Une approche telle que les systèmes experts de seconde génération, dépourvue de formalisme de connaissances rigide, semble mieux adaptée. En effet, étant donné qu'aujourd'hui il n'existe ni théorie précise ni modèle formalisé permettant d'appréhender les circonstances de pollution de l'air au sein des lieux de vie, l'expérience acquise à ce jour est prédominante.

Dans le contexte de notre thèse, cette expérience se présente sous forme de plaintes écrites traduisant des cas concrets de pollution domestique. Il est nécessaire que cette base archive contienne le plus de documents possibles témoignant de l'expérience pratique des professionnels. Plus la base est complète, mieux le système répond aux nouvelles situations. Il est également important que l'ensemble des plaintes recensées ait donné lieu à des enquêtes, à des explications de la nature de la gêne et à un ensemble d'options de gestion à entreprendre pour améliorer les conditions de vie. Cette dernière condition assure la validité de la base archive dont les éléments correspondent aux cas (présentés dans la section 5.4) permettant la constitution de nouvelles solutions aux nouveaux problèmes. Un autre avantage des SESG est leur évolution en fonction du développement du niveau de connaissances issues du domaine d'application. Effectivement,

[1] Chambre d'étudiant, foyer Sonacotra (Société nationale de construction de logements pour les travailleurs), loge de gardien d'immeuble, etc.

la liste des situations de pollution intérieure possibles est illimitée, et toute nouvelle situation introduite en mémoire peut déboucher sur la mise au point de la résolution d'un futur nouveau cas.

5.4.3 La mise en correspondance du RàPC et la réponse aux plaintes

Au début de notre thèse nous avons commencé à modéliser les étapes du raisonnement à partir de cas pour l'élaboration de notre applicatif dédié à la résolution des plaintes AI. Mais comme nous venons de le citer dans la section précédente, la liste des situations de pollution intérieure possible est illimitée. En effet, ce constat nous a été confirmé par les professionnels que nous avons rencontrés dans le cadre de cette étude, alors que nous avons mentionné dans la section 5.3.2 dédiée aux limites du RàPC, que les problèmes sources sont supposés être des cas **typiques**. Par ailleurs, une des exigences de cette thèse est d'automatiser le plus possible le processus de réponse aux plaintes. À ce jour, l'adaptation est réalisée manuellement lorsqu'il s'agit de système implémentant l'approche du RàPC textuel. Nous définissons par conséquent notre propre méthodologie. Dans la section suivante nous décrivons la démarche que nous adoptons pour réaliser notre application, en prenant en compte la nature du domaine étudié et les spécificités du corpus en notre procession.

5.5 Schéma synoptique de l'approche proposée

À l'instar des SESG, la méthodologie que nous proposons est guidée par l'expérience. Étant donné que l'absence de cas typiques en pollution de l'AI fut un obstacle important à l'élaboration d'une approche fondée sur le RàPC, nous avons souhaité analyser cet aspect à partir d'un échantillon représentatif de documents du corpus des plaintes.

Par ailleurs, dans les SESG de manière générale, et en RàPC particulièrement, il est tout à fait habituel de considérer qu'il y a autant de classes de solutions que de solutions différentes dans la base de cas[2]. Par conséquent, nous avons souhaité connaître le nombre et la nature des classes de plaintes possibles reflétant le domaine de la pollution domestique. Cette réalisation est établie à partir de l'échantillon représentatif de la base de plaintes la plus exhaustive que nous possédons. Ce travail revient à situer le concept « cas » des SESG dans le cadre de notre approche.

Le principe de cette réalisation est exposé dans la figure 5.1 et dans la section suivante. Nous présentons ensuite les modules que nous avons définis et qui sont nécessaires au traitement des plaintes.

5.5.1 Étude de la régularité thématique

La réalisation de l'analyse des textes nous a été inspirée par un constat noté suite à la lecture du corpus des plaintes dans sa totalité. En effet, bien qu'à l'origine, les textes des plaintes

[2]Cours de Alain MILLE : liris.cnrs.fr/ãmille/enseignements/

sont non structurés et sont d'un niveau de précision imprévisible propre à chaque auteur, une certaine régularité thématique implicite existe néanmoins. Les parties problèmes du texte des dossiers sont décrites avec des mots différents aussi bien lexicalement que sémantiquement, au moment où les rapports établis par les experts en guise de réponse aux problèmes se présentent sous forme d'un ensemble de modèles de lettres répétitifs.

Ainsi, nous faisons l'hypothèse de l'existence d'une régularité thématique des plaintes, et cela en nous appuyant sur la régularité constatée à partir des réponses apportées aux plaintes. Afin de vérifier notre hypothèse, nous réalisons d'abord une segmentation (ou classification) automatique d'un échantillon représentatif de plaintes. Ensuite l'ensemble des classes automatiques sera interprété par des sujets experts. Ces derniers, à chaque classe, affecteront une étiquette correspondant au titre de la thématique abordée communément par les éléments de chaque synthèse automatique. En effet, les experts n'ont pas conscience a priori de ces classes mais peuvent interpréter les regroupements.

La mise en place de l'ensemble des synthèses des diverses situations possibles en matière de pollution domestique correspond à la conception d'une base d'« exemples ». Ces derniers correspondent à des plaintes regroupées dans une base de scénarios, où chaque scénario correspond à une classe thématique de plaintes à laquelle est associé un rapport de solution type. Pour situer notre travail par rapport aux « cas », ces derniers correspondent aux scénarios définis dans notre approche.

Dans la suite, nous désignerons par « plainte-exemple » toute plainte résolue, faisant partie d'un scénario, et située en mémoire. Pour la réalisation des solutions, il était nécessaire de faire appel à la contribution de différents professionnels du domaine de l'air intérieur de manière générale ainsi que ceux spécialisés dans d'autres domaines connexes (ventilation, médecine générale, etc). Ces solutions se présentent sous forme de solutions génériques regroupant les points de vue des différents experts contactés. Les rapports indiqueront, de la manière la plus exhaustive possible, les actions à entreprendre dans le contexte du scénario considéré. Ainsi, au final, à chaque scénario correspond une solution unique. Dans la section 5.10 nous exposons les méthodes que nous avons implémentées pour construire la base de scénarios.

5.5.2 Le module fonctionnel

Nous pouvons considérer la réalisation des solutions génériques comme une alternative à l'adaptation du RàPC. Rappelons que le but de notre travail est de répondre à une plainte à partir de son texte. Dans notre approche, les textes sont écrits en langue naturelle. Pour réaliser l'assignation d'une solution appropriée à une plainte courante à traiter, un système de recherche implémenté selon les principes des SRI classiques (chapitre 2) doit être utilisé. Cela revient à attribuer une plainte écrite à un scénario.

Nous appelons « module fonctionnel » le module chargé d'apparier le texte de la plainte nouvelle

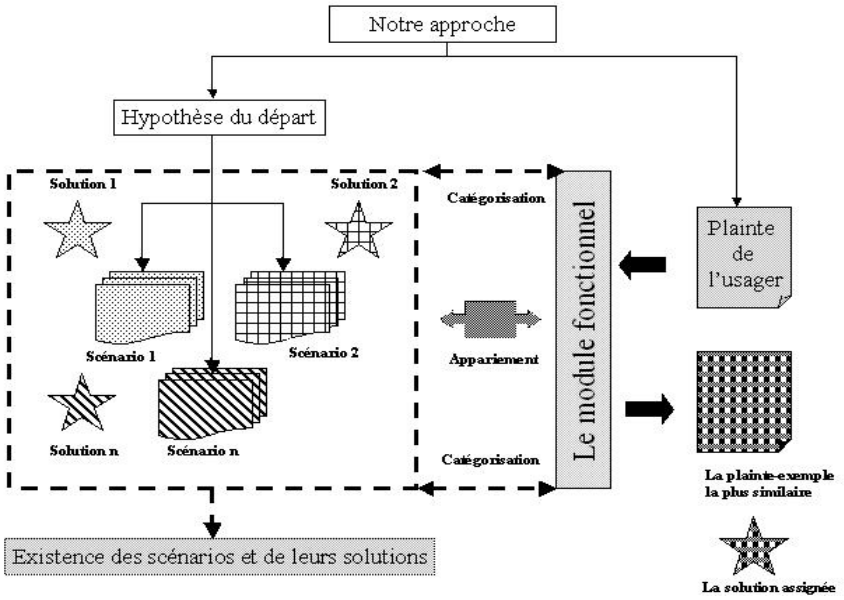


FIG. 5.1 – Architecture synoptique de l'approche proposée

avec les textes des plaintes résolues regroupées dans des scénarios et se trouvant en mémoire. Le rôle du module fonctionnel consiste à générer d'abord un modèle de représentation adapté au système de recherche utilisé. Dans le cadre de cette thèse nous avons étudié différents modèles de recherche. Nous avons élargi notre rayon de recherche en raison de la nature hétérogène des éléments de notre corpus. En effet, l'hétérogénéité apparaît selon deux aspects. Le premier aspect concerne la taille des documents, et le second concerne la quantité d'information non utile au raisonnement expert existant dans la plainte. Par conséquent, nous maintenons au coeur du module fonctionnel les différents modèles de recherche étudiés. Nous décrivons dans la section 5.7 les différents modèles retenus pour l'implémentation du module fonctionnel. On insistera sur leurs spécificités adaptées à la nature des documents que nous possédons et à la nature de ceux que l'on doit traiter.

5.5.3 Assignation automatique de solution

Après que le module fonctionnel ait apparié la plainte de l'usager à une plainte-exemple, le système affecte la plainte courante au scénario de la plainte-exemple. Ainsi, la solution attribuée au scénario d'affectation est assignée à la nouvelle plainte.

L'approche que nous proposons repose beaucoup sur le corpus des plaintes. Dans la section suivante 5.6, nous allons décrire la composition de cette base ainsi que ses caractéristiques.

5.6 Mise en place d'une base d'exemples

Notre corpus est constitué de 655 plaintes. Il est issu de divers organismes en charge de recueillir les plaintes et de les résoudre : 4% des documents sont issus du département de microbiologie et du département de pollution chimique du laboratoire d'hygiène de la ville de Paris, *LHVP*, 88% des dossiers proviennent du Service des Analyses en Milieux Intérieurs de Liège en Belgique, *Sami*, et 8% des plaintes ont été reçues au Centre Scientifique et Technique du Bâtiment, *CSTB*[3]. Ces documents se présentent sous formats électroniques divers (Word, PDF, TXT, etc) et aussi sur des supports papier. Le contenu de ces textes est initialement sous un apect brut, il est écrit en langue naturelle et apparaît sous forme non-structurée. La taille des plaintes varie ; les dossiers issus du Sami de Liège sont très brefs, alors que les dossiers provenant du LHVP sont beaucoup plus détaillés et donc beaucoup plus longs.

5.6.1 Les conditions nécessaires à la prise en compte de la « plainte air intérieur »

Pour tenir compte des demandes des particuliers, les intervenants du Sami de liège ou ceux des deux départements (microbiologie, chimie) du LHVP exigent des occupants de faire objectiver leurs doléances par un médecin (généraliste ou spécialiste). Ce dernier adresse une demande d'investigation des lieux de vie du patient à l'organisme, attestant de l'état de santé de son patient dont il suppose la dégradation liée à une viciation de l'air intérieur.

Le témoignage du médecin est, pour le moment, le seul moyen à permettre de statuer sur l'aspect de la plainte en amont de l'enquête. En effet, les établissements sus-cités n'ont pas compétence pour faire enquêter sur des situations de malaise dues à des rapports conflictuels. En effet, l'origine d'une plainte peut être simplement un désir de relogement, un litige de voisinage ou bien simplement un désaccord entre l'occupant et le propriétaire des lieux. Néanmoins, cette démarche n'est pas toujours claire dans la demande de l'occupant.

Par conséquent, nous devons définir avec beaucoup de rigueur le cadre d'utilisation de notre système pour avoir une certaine assurance vis à vis de l'« authenticité » des plaintes air intérieur à traiter. Cela consiste à définir avec une haute précision trois facteurs essentiels : le profil des « utilisateurs » de l'applicatif (grand public, personnel de santé, opérateurs d'organismes spécialisés, etc.), le « mode d'utilisation » du système (logiciel en ligne ou application centralisée), et les motifs de conception des éléments de la « ressource » à considérer.

[3]Nous avons consulté les données du Laboratoire Central de la Préfecture de Police (LCPP) également. Cette ressource se présentait sous forme d'une base de données fortement structurée. Dans cette base, un agent de saisie enregistrait les interventions suites à des cas d'intoxication oxycarbonnée notamment. Etant donné que nous définissons notre approche de résolution dans le cadre des plaintes écrites en langue naturelle, nous n'avons pas sélectionné ces enregistements pour cette étude.

Les témoignages constituant la plainte	
Certificat du généraliste	Je vois madame XXX âgée de 58 ans qui présente des symptômes de toux et d'irritation trachéale. Son état s'aggrave lors du séjour dans une pièce contenant une paroi isolée en laine de verre. La possibilité d'une irritation par cette laine de verre n'est pas exclue et un test local sera préférable.
Certificat du spécialiste	Je certifie que madame XXX souffre de toux et d'irritation trachéale.
Informations de l'occupant	– Présence de poussière et de moisissure – Odeur d'égout dans la salle de bain – Présence d'animal de compagnie – Présence de plantes – Usage d'aspirateur – Présence de trous et de fissures dans les murs – Laine de verre dans le grenier

Tab. 5.1 – Matérialisation de la plainte

Les 3 facteurs sus-cités (utilisateur, mode d'utilisation et ressource) sont, dans notre cas, en inter-dépendance. Concernant les motifs de la « ressource », nous nous intéressons aux plaintes liées aux ambiances intérieures polluées affranchie de tout aspect juridique ou conflictuel, attestées par un médecin à l'instar du mode de fonctionnement classique. Le facteur « ressource » est donc lié à la formation de l'« utilisateur ». Par conséquent, cet opérateur peut être le médecin ou bien un interlocuteur des services spécialisés ayant en main les preuves (documents écrits ou appels téléphoniques) médicales nécessaires à la prise en compte de la demande d'intervention. Parallèlement, la précision du profil « utilisateur » implique les « conditions d'utilisation ». Ce dernier facteur ne peut, par conséquent, pas correspondre à une configuration en ligne, en accès libre.

5.6.2 Composition d'une plainte

Comme nous venons de le citer, qu'il s'agisse du LHVP ou du Sami, l'expert en charge de l'enquête établit une demande d'intervention en prenant en compte l'avis des différents médecins consultés ainsi que celui de l'occupant. Dans le tableau 5.1, nous exposons un cas de pollution concret recensé par les services du Sami de Liège. Dans ce dossier nous constatons qu'il y a le rapport du médecin généraliste, le certificat du spécialiste ORL et des informations recueillies auprès de l'occupant par l'expert suite à une communication téléphonique.

Présentation finale de la plainte	
Partie problème	Suite à votre demande et dans le cadre de l'activité du SAMI (Service d'Analyse des Milieux Intérieurs) de la province de Liège, nous avons rendu visite ce 30/07/04, à votre patiente qui présente des problèmes respiratoires et ORL de type irritation trachéale. Lors de notre visite, nous avons pu constater que les parois étaient isolées au moyen de laine de verre. De plus, après une observation de l'habitation et des conditions de vie, nous avons constaté les éléments suivants : – Présence de poussière et de moisissure – Odeur d'égout dans la salle de bain – Présence d'animal de compagnie – Présence de plantes – Usage d'aspirateur – Présence de trous et de fissures dans les murs

TAB. 5.2 – Forme finale de la partie problème du dossier de la plainte

La réunion de ces éléments constitue une demande d'intervention, ce que nous appelons dans le cadre de cette thèse « la plainte ». Le diagnostiqueur (du Sami dans le cas de cet exemple) réalise un « assemblage » de ces différentes informations dans un seul document auquel est joint le rapport établi suite à l'intervention du diagnostiqueur au niveau du logement. Dans le tableau 5.2 nous indiquons la forme finale de la partie problème du dossier, reconstituée à partir des différentes déclarations prises en compte et exposées dans le tableau 5.1. La plainte indiquée, les certificats médicaux mentionnés ainsi que l'enregistrement des informations établi suite à la télécommunication avec l'occupant font partie des dossiers qui nous ont été transmis par le Sami de Liège. Les dossiers issus du LHVP obéissent au même principe. Les services du LHVP reprennent également les résultats d'examens des spécialistes de la santé ainsi que les renseignements recueillis sur place auprès des occupants. Par ailleurs, il est très important de noter que les plaintes résolues dont nous disposons ont fait l'objet d'enquêtes expertes. Ces dernières, comprennent une visite approfondie du logement ainsi qu'un entretien avec l'occupant et/ou le gestionnaire de l'habitat (une demi-journée à une journée par site). Ces affaires ont été soldées par des diagnostics scientifiques appropriés. Un suivi est assuré par les services spécialisés. Le Sami de Liège, par exemple, se charge de contacter par téléphone les occupants après une durée moyenne de 2 mois suivant la date du diagnostic. Le Sami rapporte qu'une moyenne de 88% des particuliers se disent soulagés, et que la situation à l'origine du malaise soit améliorée, après que l'occupant se soit conformé aux conseils et aux aménagements recommandés par le diagnostiqueur.

5.6.3 Structure d'une plainte

Concernant la structure discursive des plaintes, malgré le fait qu'aucune régularité conversationnelle n'est explicite sur l'ensemble des textes au départ, une certaine structure « rhétorique »[4] apparaît de manière fréquente à travers le corpus. En effet, la structure est une donnée qui est fournie ici implicitement par l'auteur de la plainte et nous pouvons en tenir compte. Nous avons constaté à travers le fond documentaire provenant de structures différentes, que le plaignant parle le plus souvent de ses symptômes, de son habitat, ses matériaux, ses équipements et de ses activités principales au sein de son logement. Il décrit également l'environnement extérieur de son domicile qu'il incrimine d'ailleurs très souvent.

Pour vérifier notre hypothèse portant sur la structure logique et sémantique de la plainte, nous nous sommes entretenus avec un groupe de 5 experts du CSTB s'intéressant à la recherche liée à la problématique air intérieur et à ses répercussions. Le groupe d'experts est constitué plus précisément d'un expert en équipements intérieurs de ventilation, de deux chercheurs/ingénieurs chimistes ainsi que de deux chercheurs/ingénieurs microbiologistes. Les experts se sont unanimement prononcés en faveur de la collecte des items suivants en précisant une série d'exemples en lien avec les items à renseigner :

- Description des symptômes : perception et description de la pathologie par le plaignant,
- Description de l'environnement extérieur : évoquer par exemple l'existence d'une rue à fort trafic, d'usines, de travaux, de sources de pollens, etc,
- Description de l'habitat : décrire son équipement, le mobilier, les systèmes de chauffage, l'usage de produits chimiques, etc.

Cette catégorisation des informations rejoint notre constat établi à partir des revues du corpus sur le plan structurel. Ainsi, pour établir un formalisme de sauvegarde standard des plaintes des usagers et des éléments de la base d'exemples nous avons retenu les trois rubriques conversationnelles, en l'occurrence les rubriques : symptômes, habitat et environnement extérieur. Ces 3 champs correspondent plus formellement à des éléments structurants sémantiquement pertinents permettant de conserver les plaintes sous forme XML. Les balises XML délimitent le contenu de chaque partie de la plainte. Les balises portent le nom du champ de conversation en relation avec la rubrique délimitée. Nous appelons cette partie la "structure de contrôle" (Fig. 5.2).

Les plaintes constituant notre corpus ont été saisies manuellement pour être insérées en mémoire. La structure citée est unique, elle est respectée par la totalité des plaintes en mémoire et également par les nouvelles plaintes à traiter. Une plainte est ainsi reprise dans un format XML à granularité[5] fixe. Les balises sont connues, leur nombre est fixe pour la totalité des plaintes

[4] La Théorie de la Structure Rhétorique (en anglais RST) concerne la structure des usages langagiers, et plus spécifiquement la structure discursive des textes écrits.

[5] La granularité correspond généralement à l'élément ou à l'attribut

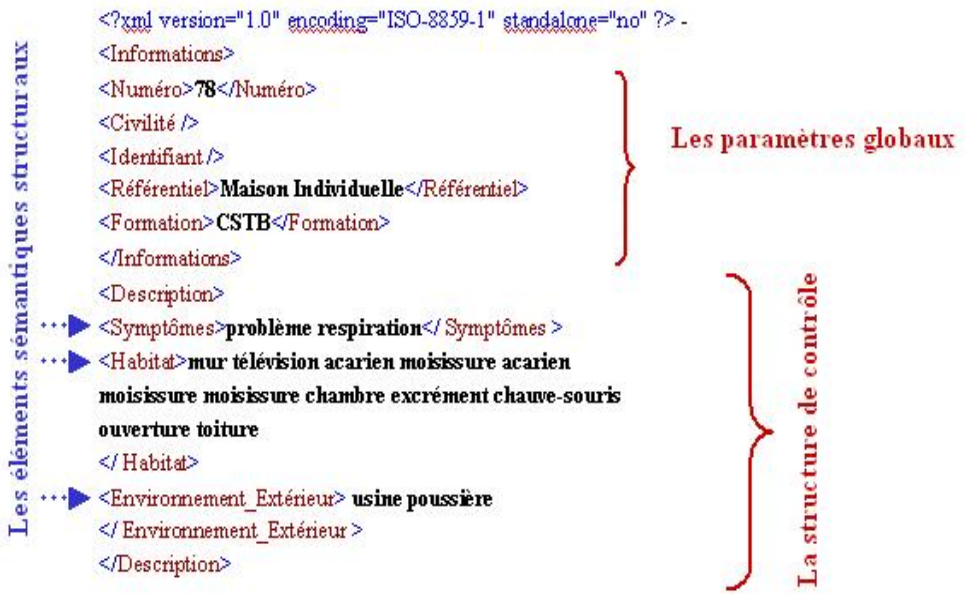


FIG. 5.2 – Exemple d'une plainte structurée au format XML

saisies (les balises concernant les rubriques non renseignées dans un document sont vides mais existent néanmoins dans la structure).

Dans le domaine de la représentation des données et des connaissances, pour structurer des documents issus d'une enquête, les variables correspondent à des questions [66]. L'interface usager de l'applicatif que nous avons développée et qui est dédiée à l'enregistrement des plaintes se présente sous forme de trois champs de saisie où l'usager (expert d'un domaine connexe au domaine de la pollution intérieur et à terme le particulier) répond à trois questions principales : parlez moi le plus précisément possible de vos problèmes de santé, de votre logement et de vos habitudes de vie, de votre environnement extérieur. Les autres informations périphériques comme le numéro de la plainte ou bien le nom de l'organisme en charge du dossier renseignent les "paramètres globaux". Ce raisonnement n'est bien entendu valable que si l'auteur considère, lors de la saisie de son cas, avec sérieux la structure et le contenu pour l'expression de sa situation.

Pour saisir les plaintes se présentant initialement sous format brut et que nous avons réuni dans le cadre de cette thèse pour établir notre base d'exemples, la structure conversationnelle a été repérée manuellement. Nous avons utilisé l'interface usager dédiée à la saisie des plaintes, et nous avons réuni dans chaque champ de saisie les différentes parties du texte en lien avec le titre de la rubrique concernée. Cette réalisation a eu lieu sur la totalité du corpus initial composé des 655 dossiers provenant des différentes sources sus-citées.

5.6.4 Difficultés inhérentes à la structure des plaintes

Dans un contexte structuré ou brut, plus un recueil des observables est complet, plus la conclusion du raisonnement (humain ou automatique) est valable. Dans le contexte dans lequel notre système évolue, une situation décrite à l'aide d'informations issues uniquement du champ de conversation symptôme ne peut être considérée en tant que « plainte air intérieur ». En effet, cette dernière par définition met en évidence un éventuel lien entre la situation d'un logement et un état sanitaire. Les experts que nous avons contactés et réunis statuent unanimement sur l'impossibilité de distinguer l'origine de la pathologie sur la base d'un cas fondé uniquement sur des symptômes non spécifiques et donc ambigus.

Néanmoins, il existe des cas de figure où le lien entre le symptôme et la nature de la pollution est suffisamment explicite. Nous pouvons citer par exemple le cas d'une personne qui dit être atteinte d'un cancer de la plèvre (section 1.4.2). Dans ce cas précis, le symptôme est très exclusif, le lien avec une exposition à l'amiante est évident. Toutefois, ce cas de figure est quasiment unique, et force est de constater que dans la majeure partie des plaintes, les symptômes avancés ne sont pas spécifiques. Les manifestations les plus souvent citées touchent la sphère ORL (bronchite, asthme, irritation, maux de tête, etc), les poumons et les affections cutanées. Les origines de ces syndromes pouvant être très diverses, un appariement de faits fondé uniquement sur ces informations est impraticable. Par conséquent, il devient nécessaire de prendre en compte d'autres constatations concernant l'habitat, son équipement, les activités domestiques, les éventuels chantiers extérieurs, etc. Par conséquent, d'un point de vue application, l'interface utilisateur ne doit pas permettre la prise en compte d'une description fondée uniquement sur l'aspect sanitaire.

Après avoir présenté les spécificités de la base documentaire permettant d'élaborer la base d'exemples, nous allons justifier nos choix portant sur l'application des systèmes de recherche pour la réalisation du module fonctionnel. Malgré le fait que la régularité des plaintes est notre hypothèse initial pour la définition de la méthode, ce n'est que par la suite que l'étude de la régularité thématique du domaine est exposé. Nous procédons dans cet ordre, étant donné que les systèmes de recherche, en plus de leur fonction dans l'approche d'assignation des solutions aux nouvelles plaintes à traiter, ils sont également à l'origine de la catégorisation des textes.

5.7 Réalisation du module fonctionnel

Dans le chapitre 2 nous avons étudié les systèmes de recherche d'information. Nous les avons étudiés pour les utiliser précisément dans le cadre du module fonctionnel défini dans notre approche. Compte tenu de la structure conversationnelle apparaissant dans les textes des plaintes et la possibilité de réaliser une recherche plus fine et mieux focalisée à l'aide des systèmes de recherche adaptés à la structure (chapitre 3), nous avons développé quelques systèmes de recherche modélisés pour la prise en compte de cet aspect.

	symptômes	habitat	environnement extérieur
mot_1	x_{11}	x_{12}	x_{13}
mot_2	x_{21}	x_{22}	x_{23}
...	...	...	...
mot_n	x_{n1}	x_{n2}	x_{n3}

TAB. 5.3 – Représentation matricielle du texte

5.7.1 Les systèmes de recherche mis en œuvre

Les textes des plaintes sont de taille variable. Ils sont parfois courts, comme c'est le cas des demandes d'intervention enregistrées au Sami, et peuvent être longs, comme c'est le cas des dossiers enregistrés au LHVP. Plus le texte « plat » d'une plainte est long, plus au moins une rubrique parmi celles renseignées contient un texte long dans le format XML. Par conséquent, le critère de taille est à prendre en considération dans le contexte des textes structurés également. Nous exposons dans le chapitre 6 les résultats de notre analyse portant sur la dépendance des résultats des systèmes de recherche implémentés par rapport à la taille des documents traités.

Comme nous l'avons mentionné dans la section 2.5.2, le modèle vectoriel est plus adapté au traitement des textes longs. Par conséquent, nous avons choisi de le mettre en œuvre en utilisant une de ses adaptations au formalisme XML proposées dans la section 3.4. Nous avons utilisé la première version de Zargayouna [132], définie dans la section 3.4.4, et qui s'intéresse à l'adaptation de la pondération TF-IDF à la distribution des termes dans un contexte structuré.

Application du modèle de Zargayouna

Rappelons que la vectorisation à l'aide de ce modèle consiste à représenter un texte par plusieurs vecteurs. Dans notre étude, chacune des 3 balises retenues (symptômes, habitat et environnement extérieur) est représentée par un vecteur. L'ensemble du texte correspond par conséquent à une matrice (tableau 5.3).

Dans le cadre de notre étude, nous choisissons en tant qu'espace de représentation vectorielle approprié, un dictionnaire généraliste de la langue française. Nous utilisons les vedettes du dictionnaire des synonymes, DICTIONNAIRE (section 4.2.3), en tant que primitives vectorielles caractérisant les textes des différentes rubriques. À l'exemple de la plupart des SRI vectoriels, l'indexation est en mode « contrôlé » (section 2.3) puisqu'elle utilise les entités répertoriées dans le vocabulaire autorisé, en l'occurrence ici le dictionnaire des synonymes.

	poids direct	poids sémantique
...	...	...
champignon	$p_{champignon}$	$p_{champignon}$
cloison	$p_{cloison}$	$p_{cloison}$
toiture	$p_{toiture}$	$p_{toiture}$
moisissure	0	$0 + Sim_{moisissure,champignon} \times p_{champignon}$
mur	0	$0 + Sim_{mur,cloison} \times p_{cloison}$
toit	0	$0 + Sim_{toit,toiture} \times p_{toiture}$
...	...	...

TAB. 5.4 – Comparaison entre les poids calculés selon le modèle vectoriel direct et le modèle vectoriel sémantique

Application du modèle sémantique de Zargayouna

Nous avons employé la première version du modèle sémantique défini par Zargayouna et qui, pour calculer le poids sémantique d'un terme, prend en compte l'aspect structure des documents et la sémantique des autres termes situés dans les balises du terme en question. Cette mesure est détaillée dans la section 3.7.1.

Le tableau 5.4 expose des poids calculés selon les deux modèles vectoriels adaptés à la structure. Les poids des termes « moisissure », « mur » et « toit »,qui n'existent pas directement dans les textes modélisés, augmentent à l'aide de la mesure sémantique Sim permettant d'évaluer la similarité entre les termes. Ainsi, les termes présents directement dans les textes, en l'occurrence « champignon », « cloison » et « toiture », contribuent à l'augmentation des poids des termes absents mais qui leur sont sémantiquement proches.

Pour les modèles directs et sémantiques proposés par Zargayouna (sections 3.4.4, 3.4.5 et 3.7.1), nous avons pris en compte les premières versions au détriment des secondes. Nous avons effectué ce choix puisqu'il s'accorde plus avec la structure XML de nos plaintes. De plus, les premières versions sont plus faciles à la mise en en œuvre. En effet, la structure des plaintes que nous proposons repose sur des balises indépendantes, et de profondeur fixe. Au delà des trois balises prises en compte dans le formalisme des plaintes, l'applicatif ne permet pas d'insérer d'autres structures au sein des éléments prédéfinis. Alors que pour les secondes versions des modèles (le modèle direct et le modèle sémantique), Zargayouna considère que les balises sont dépendantes, liées par des relations de spécialisation/généralisation. Par conséquent, le calcul des poids se complique, puisque ce dernier doit tenir compte des différents contextes d'apparition des termes.

L'avantage de ces deux derniers modèles réside principalement dans leur formalisme de base (vectoriel) qui est relativement simple. Il suffit de munir l'espace vectoriel d'une distance entre vecteurs pour calculer la similarité entre textes. Cette mesure peut correspondre au cosinus de l'angle formé par les deux vecteurs (formule 5.1) ou bien à la norme Euclidienne qui calcule la

différence entre deux vecteurs par application du théorème de Pythagore (5.2). Ainsi, des similarités locales sont calculées entre les deux documents à comparer. Ces mesures locales sont calculées entre les textes délimités par des balises identiques. Ensuite une agrégation est réalisée pour mesurer la similarité entre les documents en entier. Nous utilisons dans notre application la valeur du cosinus des vecteurs pour calculer la distance entre vecteurs parce que cette grandeur nous semble plus intuitive. En effet, la distance Euclidienne est utilisée lorsque les deux vecteurs sont de tailles très différentes. La taille des vecteurs dans notre cas varie entre 0 et 1, puisque les poids des termes sont normalisés. Concernant l'agrégation, pour mesurer la similarité totale entre deux documents, nous effectuons une moyenne des similarités locales.

$$cosinus\ \alpha = \frac{v \times u}{||v|| \times ||u||} \tag{5.1}$$

$$d^2 = \left| ||v||^2 - ||u||^2 \right| \tag{5.2}$$

Les principes théoriques de ces modèles sont formalisés dans les sections qui leurs sont consacrées (sections 3.4.4 et 3.7.1). Les résultats sont exposés dans le chapitre 6. Ces résultats dépendent évidemment des composants (ou primitives) de la représentation vectorielle choisie. Les traits pertinents correspondent aux termes du dictionnaire considéré. La notion de terme, qui désigne un mot abstrait, prend en compte les mots simples ainsi que les mots composés. Par ailleurs, il est préférable que les traits ne soient pas trop nombreux, car ils fixent la dimension de l'espace vectoriel en jeu. Cependant, le problème des méthodes lexicales est qu'on aboutit souvent à un très grand nombre de traits : plusieurs centaines ou plusieurs milliers. Bien sûr, cet inconvénient est encore plus important lorsqu'il s'agit du modèle sémantique, puisque ce dernier, pour chaque entrée lexicale prend en compte l'ensemble de ses termes sémantiquement proches. Dans le cas de DICTIONNAIRE, 48881 entrées sont prises en compte. Ceci pose bien entendu de sérieux problèmes de calcul. Ce qui est d'ailleurs appelé dans le domaine des SRI vectoriels la « malédiction dimensionnelle » [78]. On a alors affaire à des vecteurs très creux avec une grande majorité de composantes nulles.

Il devient par conséquent important d'essayer de réduire la dimension de l'espace en diminuant le plus possible le nombre de traits pertinents, ou en trouvant une transformation vectorielle appropriée [78]. De plus, les modèles vectoriels implémentés reposent sur une indexation contrôlée guidée par un vocabulaire généraliste. Les textes des plaintes comportent des termes généraux de la langue française ainsi que des termes spécifiques aux domaines annexes de la pollution intérieur. Nous avons par conséquent ressenti le besoin d'utiliser des SRI fondés sur une indexation libre (section 2.3). Par rapport à ces réalités des systèmes vectoriels d'une part, et par rapport à la non prise en compte de la position des termes les uns par rapport aux autres dans le modèle vectoriel d'autre part, nous avons utilisé un autre système de recherche dont la représentation des textes ne prend en compte que les termes présents au sein des textes. Ce système est le modèle de proximité floue présenté dans la section 2.4.2. Dans la section suivante nous exposons

une de nos contributions dans le cadre de cette thèse. Cette contribution a consisté à augmenter le modèle de proximité floue à l'aide d'une mesure sémantique et à l'ajuster dans un contexte structuré.

5.7.2 Adaptation sémantique et structurelle du modèle de proximité floue

La mesure de Mercier (section 2.4.2) est très avantageuse. En effet, par rapport au modèle vectoriel qui ne tient pas compte de la position des termes les uns par rapport aux autres, le modèle de Mercier évalue la densité des termes de la requête au sein du document en tenant compte de leurs positions. De plus, ce modèle est facile à mettre en œuvre et comme nous l'avons cité dans la section précédente dédiée au modèle vectoriel, l'application du modèle de proximité floue nécessite un temps de calcul raisonnablement court (nous en discutons dans le chapitre 6).

Toutefois, cette mesure ne tient pas compte de la sémantique des termes. En effet, son modèle est limité par la relation de co-occurrence directe des termes et ne prend pas en compte les éventuels liens sémantiques qui peuvent exister entre les termes de la requête et ceux du document. L'intégration d'une mesure sémantique entre termes dans ce modèle est nécessaire. Pour cela, nous définissons notre contribution qui consiste à formaliser l'adaptation du modèle fondé sur le principe de la proximité floue aux besoins des appariements sémantiques.

Modélisation de la pertinence sémantique locale

Le principe de notre modèle augmenté est de considérer dans le document non seulement les termes de la requête mais également les termes situés à une certaine distance sémantique de ces derniers. Ainsi, au lieu de considérer uniquement les positions prises par un terme t de la requête dans l'ensemble $d^{-1}(t)$, les positions prises par des termes sémantiquement proches de t dans l'ensemble $d^{-1}(Sem(t))$ sont prises en compte, où Sem(t) désigne l'ensemble des termes proches sémantiquement de t. Nous utilisons ici les notations de base du modèle de Mercier sans les définir, puisque nous les avons déjà présentées dans la section 2.4.2 consacrée au modèle de proximité floue.

$$\mu_t^d(x) = \max_{i \in d^{-1}(t)} \left(\max(\frac{k - |x - i|}{k}, 0) \right) \tag{5.3}$$

Rappelons, néanmoins, que le modèle de Mercier calcule la densité des termes de la requête en évaluant pour chaque terme t de la requête, aux différentes positions x, du document d, un taux de pertinence locale $\mu_t^d(x)$ (formule 5.3). Par ψs_t^d nous nommons la nouvelle formule permettant de calculer le taux de pertinence sémantique locale d'un terme t de la requête au sein d'un document d. La formule se présente ainsi :

$$\psi s_t^d(x) = \max_{i \in d^{-1}(Sem(t))} \left(\max(\frac{k - |x - i| \times Sim(t_i, t)}{k}, 0) \right) \tag{5.4}$$

À l'aide de ce modèle, les positions i des termes sémantiquement proches du terme t (appartenant à Sem(t)) sont prises en compte, mais leur influence doit dépendre du degré de similarité qu'ils partagent avec le terme t de la requête. Ceci est la raison pour laquelle nous multiplions la pertinence locale par $Sim(t_i, t)$. Cette dernière mesure n'excédant jamais la valeur 1, la pertinence locale d'un terme de la requête à une position prise par un terme qui lui est sémantiquement proche atteint au plus la valeur de pertinence dans le cas où la position considérée est occupée par le terme lui-même.

Notre modèle augmenté attribue un score de pertinence locale au moins aussi élevé que le taux de pertinence locale calculé par le modèle direct, puisque $d^{-1}(t) \subset d^{-1}(Sem(t))$ [6]. Par exemple, dans le cas où il n'existerait pas de termes sémantiquement proches des termes de la requête au sein du document (à part les termes eux mêmes), notre modèle donnerait des résultats semblables à ceux du modèle de base. Par conséquent, l'application de notre modèle ne risque pas de provoquer une perte d'information par rapport au modèle de Mercier.

Un seuil de similarité est nécessaire pour caractériser l'ensemble Sem(t). Ceci revient à déterminer la valeur de similarité $Sim(t_i, t)$ à partir de laquelle le terme t_i est considéré comme étant sémantiquement proche de t. Ce seuil est déterminé empiriquement lorsqu'il s'agit de documents (et requêtes) plats. Lorsqu'il s'agit de documents structurés, ce seuil peut être défini empiriquement ou il peut correspondre au degré de similarité entre le terme t et le terme auquel est rattaché le marqueur (ou la balise) de l'élément structurant au sein de la requête. Cela dépend du degré de granularité du document XML. Nous désignons par document structuré à granularité fine un document fortement structuré, dans ce cas le contenu de la balise a de fortes chances de correspondre sémantiquement au titre de la balise. Inversement, un document XML (ou balisé de manière générale) est de granularité faible dans le cas où les textes délimités par les balises sont très détaillés et donc les termes présents correspondent à des niveaux différents du champ sémantique du titre de la balise. Par conséquent, dans le cas où il s'agit d'une structure à granularité fine la similarité du terme en question avec le terme marqueur peut être utilisée comme seuil, et dans le cas d'une structure à faible granularité une évaluation empirique du seuil est nécessaire.

À l'origine, la mesure de proximité floue est appliquée pour évaluer la pertinence d'un document quelconque (non structuré) par rapport à une requête. La dernière spécification concernant l'évaluation du seuil de similarité sémantique dans le cadre des documents structurés à granularité plutôt fine met en jeu la structure dans la définition du modèle. En plus de l'intégration de l'aspect sémantique, la condition que nous venons de citer, représente notre contribution dans l'adaptation du modèle de proximité floue selon le formalisme structuré des documents et des requêtes, en l'occurrence ici XML. Par ailleurs, pour évaluer la pertinence totale d'un élément

[6] Étant donné qu'un terme est aussi sémantiquement proche de lui même.

(contenu d'une balise) du document par rapport à un élément du même nom de la requête (puisque la requête et le document obéissent à la même structure), nous respectons le principe du modèle de Mercier. L'adaptation du modèle que nous proposons s'applique aux éléments de requêtes obéissant au schéma booléen classique, c'est à dire que l'élément d'une requête est une combinaison de conjonctions et/ou de disjonctions de mots clés.

Illustration de l'intérêt de l'intégration de la sémantique

À travers le tableau Tab. 5.5 nous illustrons une requête contenant les 3 mot-clés suivants : moisissure, mur et toit. Le document est formé, entre autres, par les mots clés : champignon, cloison et toiture. Les termes de la requête, selon le dictionnaire des synonymes DICTIONNAIRE, sont des synonymes des termes cités du document. Les taux de pertinence locale, μ, implémentant le modèle de Mercier concernant les termes de la requête considérés séparément ou bien pris en compte dans des compositions à l'aide d'opérateurs logiques ont des scores égaux à zéro aux différentes positions de l'ensemble d^{-1}. Par contre, les taux de pertinence sémantique locale ψ indiquent des valeurs plus élevées selon les différents cas de figure cités pour les taux μ. Ces valeurs relativement significatives correspondent aux taux de similarité calculés au moyen de l'indice de Jaccard (formule 4.14) selon la configuration synonymique dans DICTIONNAIRE ; Sim(champignon, moisissure)=0.18, Sim(mur, cloison)=0.28, Sim(toit, toiture)=0.12.

Ainsi, le taux de similarité selon le critère de densité élaboré par Mercier entre la requête disjonctive (moisissure OU mur) et le document donnés en exemple est de 0%, alors que le taux de similarité selon notre modèle augmenté à l'aide de la sémantique indique un score de 22,2% pour la même requête. Les scores indiqués dans le tableau de l'exemple correspondent aux résultats d'appariements locaux. La généralisation de ces mesures au niveau « document » est établie en effectuant une agrégation des similarités locales, puisque les apariements établis dans cet exemple sont réalisés entre élements structuraux (balises). La notion d'agrégation est à définir selon les besoins de l'application comme nous en avons discuté dans la section (3.4).

Nous étudions le modèle de proximité, entre autres, afin de l'utiliser pour apparier le texte d'une plainte nouvelle à analyser avec les éléments de la base d'exemples. Rappelons également qu'une plainte est saisie en langue naturelle par les soins de l'usager du système. L'utilisation de requêtes booléennes pour supplanter l'expression naturelle des besoins de manière générale, même enrichies de connecteurs variés, nécessite une mise en forme manuelle soigneuse et coûteuse des requêtes.

En effet, nous n'avons pas connaissance d'outils permettant de traduire automatiquement une expression en langue naturelle vers son interprétation sous forme booléenne. Le raisonnement induit par les connecteurs logiques est différent selon que ces connecteurs logiques sont utilisés en tant qu'opérateurs booléens ou en tant qu'opérateurs dans le langage naturel.

x	1	2	3	4	5	6	7	8
d	champignon		cloison			toiture		champignon
$\psi^d_{moisissure}(x)$	0.18	0.16	0.15	0.13	0.13	0.15	0.16	0.18
$\psi^d_{mur}(x)$	0.22	0.25	0.28	0.25	0.22	0.19	0.16	0.14
$\psi^d_{toit}(x)$	0.06	0.07	0.08	0.1	0.1	0.12	0.10	0.10
$\mu^d_{moisi\ OU\ mur}(x)$	0	0	0	0	0	0	0	0
$\mu^d_{(moisi\ OU\ mur)ETtoit}(x)$	0	0	0	0	0	0	0	0
$\psi^d_{moisi\ OU\ mur}(x)$	0.22	0.25	0.28	0.25	0.22	0.19	0.16	0.18
$\psi^d_{(moisi\ OU\ mur)ET\ toit}(x)$	0.06	0.07	0.08	0.1	0.1	0.12	0.10	0.10

TAB. 5.5 – Comparaison entre les valeurs de pertinence locale directe et sémantique

Dans son étude expérimentale [35] sur l'usabilité des requêtes booléennes en vue de recherche d'information, Dinet positionne la logique impliquée par les opérateurs booléens par rapport à la logique utilisée **habituellement** par un individu. Par exemple, l'opérateur ET implique une inclusion dans le langage naturel alors qu'il implique une exclusion (des résultats) dans le langage documentaire. Par exemple, lorsqu'une personne demande un croissant ET un café, elle espère avoir les deux. Par contre, dans le cadre d'une « recherche documentaire » demander « croissant ET café » correspond à une restriction du champ de réponses. En effet, pour une recherche initiée à partir de la requête « croissant ET café » seuls les documents contenant les deux termes « croissant » et « café » à la fois sont attendus. En d'autres termes, on aura pas les documents s'agissant du terme « croissant » sans le terme « café », et ceux concernant le terme « café » et ne contenant pas le terme « croissant ».

De même, l'opérateur OU implique une inclusion dans le langage documentaire alors qu'il implique une exclusion (une restriction) dans le langage naturel. Parfois, dans la vie courante, il faut choisir : « boire » OU « conduire ». Dans le « langage documentaire », l'utilisation du OU entre deux termes signifie que l'on effectue une recherche simultanée sur les deux termes. C'est à dire, pour la requête « croissant OU café » un SRI fournit l'ensemble des documents comprenant le terme « croissant » ainsi que la totalité des documents contenant le terme « café ».

Compte tenu du fait que nos requêtes (nouvelles plaintes à traiter) sont écrites en langue naturelle, et vis à vis de cette incohérence entre les langages, nous ne pouvons utiliser parfaitement les opérateurs logiques en les relevant directement des expressions naturelles des usagers de notre application. Et par rapport aux avantages qu'offre le modèle de proximité floue, nous adaptons le principe de densité de la requête dans le document au modèle filtré lemmatisé exempt d'opérateurs logiques. Pour cela, nous nous sommes inspirés du principe générique de la théorie du signal. Ainsi, nous présentons un nouveau modèle de recherche en émettant une hypothèse. Cette

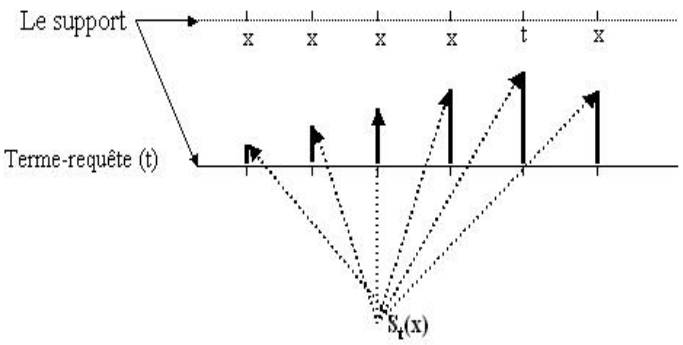


FIG. 5.3 – Intensité d'un signal émis par un terme t

dernière consiste à supposer que les termes de la requête émettent des signaux au sein des positions possibles dans les documents. À l'aide de ce modèle, la superposition des signaux émis par les termes de la requête au sein des documents détermine la densité de la requête dans ces documents.

5.7.3 Le modèle de recherche fondé sur la superposition des signaux

Dans le contexte de notre adaptation, les signaux à traiter proviennent de sources connues et sont engendrés dans les documents dont on souhaite évaluer la pertinence. Ces sources correspondent aux termes de la requête. Le support du signal correspond à l'intervalle discret borné $[1, |d^{-1}|]$, où d^{-1} désigne l'ensemble des positions pouvant être prises dans un document d.

Le principe du modèle fondé sur la superposition des signaux

Le principe de ce modèle se résume comme suit : d'abord, les signaux émis par les sources d'intérêt, en l'occurrence ici les termes de la requête, sont évalués aux différentes positions dans le document. Ensuite, les intensités du signal global sont estimées aux différentes régions d'intérêt situées sur le support considéré, en l'occurrence ici les positions où se trouvent les termes de la requête dans le document.

L'intensité du signal émis par un terme t de la requête à une position x du document correspond au degré d'influence ζ du terme t au niveau x. L'intensité maximale d'un terme t à une position x est de 1 lorsqu'il existe une occurrence de t en x. Le degré d'intensité diminue au prorata de la distance (figure 5.5).

$$\zeta_t^d(x) = \sum_{i \in d^{-1}(t)} \left(\max(1 - \frac{|x-i|}{|d^{-1}|}, 0) \right) \tag{5.5}$$

x	1	2	3	4	5
d	champignon	champignon			
$\zeta^d_{champignon}(x)$	1.8	1.8	1.4	1.0	0.6

TAB. 5.6 – Mise en évidence des valeurs d'intensité de signaux émis par plusieurs occurrences d'un seul terme

x	1	2	3	4	5
d	champignon			toiture	
$\zeta^d_{champignon}(x)$	1	0.8	0.6	0.4	0.2
$\zeta^d_{toiture}(x)$	0.4	0.6	0.8	1	0.8
$\zeta^d_{champignon\ toiture}(x)$	1.4	1.4	1.4	1.4	1

TAB. 5.7 – Mise en évidence des valeurs d'intensité de signaux émis par des termes de requête moyennement proches

En plus du positionnement des formules booléennes par rapport à l'expression naturelle des documents-requêtes, que nous avons exposé à la fin de la section précèdent 5.7.2, nous mettons en évidence ici les particularités de notre modèle par rapport aux paramètres de la formule de Mercier. Notre formule 5.5 ne tient pas compte du paramètre k permettant d'évaluer la pertinence locale. Ce paramètre définit la taille de la zone d'influence d'un terme. Dans notre modèle, la taille du support $|d^{-1}|$ est prise en compte. Dans le contexte précis de notre étude, les plaintes sont enregistrées à l'aide de balises délimitant les rubriques maintenues. Étant donné que leur taille varie, il était plus intéressant pour nous d'adapter l'étendue d'influence des occurrences des termes de la requête en fonction de la taille des rubriques, plutôt que d'utiliser un paramètre fixé a priori. Par ailleurs, pour un terme de la requête situé à une position donnée x du support nous réalisons une somme des signaux (ce qui respecte la métaphore du signal et de ses propriétés) émis à partir des différentes occurrences du terme en question (par rapport à l'application du maximum dans le cadre de la pertinence). Autrement dit, ce cas de figure est vérifié dans le cas où un terme de la requête apparaît plus d'une fois au sein d'un document (exemple dans le tableau 5.6).

Pour évaluer la pertinence totale d'un document, le protocole d'appariement prend en considération le principe de superposition des signaux aux différentes positions actives, $d^{-1}(r(t))$. Rappelons que ces dernières correspondent aux positions x du document où apparaît une occur-

x	1	2	3	4	5
d		champignon	toiture		
$\zeta^d_{champignon}(x)$	0.8	1	0.8	0.6	0.4
$\zeta^d_{toiture}(x)$	0.6	0.8	1	0.8	0.6
$\zeta^d_{champignontoiture}(x)$	1.4	1.8	1.8	1.4	1

TAB. 5.8 – Mise en évidence des valeurs d'intensité de signaux émis par des termes de requête très proches

rence d'un terme de la requête (r(t) désigne l'ensemble des termes de la requête). La somme des intensités des interférences aux positions actives 5.6, correspond au niveau de pertinence de la requête par rapport au document ; Signal(r,d). Comme l'indique la formule 5.7, il est également possible de prendre en compte la moyenne des intensités des superpositions des signaux aux positions actives. Dans l'exemple du tableau 5.7, la moyenne globale des intensités locales est de 1,4. Dans l'exemple exposé dans le tableau 5.8, dans lequel les termes de la requêtes apparaissent plus proches que dans le premier exemple, la moyenne des intensités est de 1,8.

Ne prendre en considération que les positions actives est important. En effet, nous souhaitons faire en sorte que la variation des densités des termes de la requête ne soit pas fondée uniquement sur l'effet de bord imposé par la taille du document. Comme nous l'avons cité précédemment, il était essentiel pour nous d'adapter l'influence des occurrences des termes à la nature hétérogène, d'un point de vue taille, des plaintes que nous devons traiter en associant la zone d'influence à la taille des documents. Pour cela, nous prenons en compte uniquement l'intensité des superpositions des signaux aux positions prises par les occurrences des termes de la requête. Ainsi, indépendamment de l'effet de bord, plus ces occurrences sont proches plus les intensités locales sont importantes (tableau 5.8), et plus elles sont éloignées plus les intensités locales du signal sont faibles (tableau 5.7).

$$Signal_{Somme}(r,d) = \sum_{x \in d^{-1}(r(t))} \left(\sum_{t \in r(t)} \zeta^d_t(x) \right) \tag{5.6}$$

$$Signal_{Moyenne}(r,d) = \frac{\sum_{x \in d^{-1}(r(t))} \left(\sum_{t \in r(t)} \zeta^d_t(x) \right)}{|d^{-1}(r(t))|} \tag{5.7}$$

Précisons que la mesure de similarité (ou de distance) que nous introduisons ici est bien évidemment asymétrique puisque la requête constitue la référence[7]. En effet, nous calculons la densité de la requête dans le document et pas l'inverse. Le même constat est de mise concernant le modèle de Mercier. Ce critère est en inadéquation avec les caractéristiques des autres modèles de similarité symétriques employés, dont le modèle vectoriel. Néanmoins, cette configuration asymétrique est maintenue en tant que mesure de similarité implémentée par le module fonctionnel, mais nous employons également la formule résultant d'une « symétrisation » des modèles asymétriques par moyenne des valeurs réciproques. Nous jugerons à travers les résultats des expérimentations de l'intérêt de cette harmonisation des mesures par rapport à la version asymétrique des formules orientées requêtes et cela à travers les performances du module fonctionnel dans le chapitre 6.

5.7.4 Adaptation sémantique du modèle fondé sur la superposition des signaux

Comme nous l'avons réalisé et appliqué pour le modèle de proximité floue et comme nous l'avons utilisé pour le modèle vectoriel étendu selon la structure XML en vue bidimensionnelle (ou matricielle), notre objectif est aussi de permettre à notre modèle inspiré du principe du signal de prendre en compte la sémantique des termes.

Étant donné que le modèle que nous proposons est de formalisme proche de celui du modèle de Mercier, nous procédons ici comme nous l'avons fait précédemment pour adapter le modèle de proximité floue au principe de la sémantique. Nous entendons par la notion de « signal sémantique » d'un terme de la requête tout signal émis au sein du document et initié par l'ensemble des termes sémantiquement proches du terme de la requête considéré et qui sont présents au sein du document. La formule 5.8 permet de calculer l'intensité du signal sémantique ζs émis par les termes sémantiquement proches des termes de la requête en appliquant une mesure sémantique entre termes, Sim.

$$\zeta s\,{}^{d}_{t}(x) = \sum_{i \in d^{-1}(Sem(t))} \left(\max \left(Sim(t_i, t) \times \left(1 - \frac{|x - i|}{|d^{-1}|} \right), 0 \right) \right) \tag{5.8}$$

À l'instar du modèle direct inspiré du principe du signal, seules les intensités aux positions actives sont prises en compte pour le calcul de la pertinence d'un document par rapport à la requête (en réalisant une somme ou une moyenne). Les modèles fondés sur le principe du signal peuvent être appliqués dans des contextes quelconques (structurés ou non-structurés). Dans le cas des textes structurés une agrégation est nécessaire. De la même manière que pour la pertinence locale sémantique, l'ensemble des termes sémantiquement proches des termes de la requête

[7] D'un point de vue mathématique une similarité est par définition symétrique [16]. Lorsque la fonction d vérifie les propriétés suivantes : d(a,a)=0 et d(a,b)=d(b,a) elle est appelée « indice de dissimilarité » (notion inverse de la similarité). Nous employons ici le mot similarité dans le sens plus général de « ressemblance ». Par ailleurs, on peut parler de « pseudo-dissimilarité » ou « pseudo-distance » dans la mesure où la propriété de symétrie n'est pas assurée.

doit être déterminé en fonction de la nature des documents et de la requête (granularité) (cf. 5.7.2).

Dans le cadre de cette étude, nous avons implémenté l'ensemble des modèles de recherche sémantique que nous avons cités dans ce chapitre (les modèles directs ont été implémentés également). Ces systèmes utilisent des ressources externes pour mesurer la similarité sémantique entre les termes. Dans la section suivante, nous allons parler des ressources que nous avons utilisées afin de réaliser une analyse automatique des plaintes tenant compte de la sémantique inhérente.

5.8 Choix de la ressource sémantique

Le choix de la ressource sémantique est tributaire des possibilités offertes par le vocabulaire du corpus. Par conséquent, nous réalisons une analyse du vocabulaire des textes du corpus qui permettra de mettre en évidence les tendances terminologiques des textes des plaintes. Ces dernières, ne sont pas formulées uniquement par des experts dans un langage technique et précis mais elles contiennent des passages des discours des particuliers en langage naturel (section 5.6). En effet, le vocabulaire de notre corpus est très vivant, de nombreux noms de marques sont utilisés, des noms d'espèces fongiques, d'acariens et d'autres diminutifs apparaissent de manière fréquente. Par conséquent, nous nous sommes interrogés quant à la possibilité de construire une ontologie terminologique[8]. Nous avons alors souhaité étudier d'abord le vocabulaire des textes des plaintes dont nous disposons.

5.8.1 Lemmatisation et filtrage des textes du corpus

Pour lemmatiser les textes des plaintes nous avons utilisé le lemmatiseur TreeTagger[9] adapté au français (section 2.3.4). Suite à la lemmatisation de chacune des plaintes nous récupérons en sortie uniquement les lemmes correspondant aux différentes flexions originelles de chaque document. Un dictionnaire d'arrêt est utilisé afin d'éliminer automatiquement les mots vides de sens à partir de la forme lemmatisée des textes.

5.8.2 Traitement des mots composés

La reconnaissance et la prise en compte des mots composés sont essentielles (exemple gêne-respiratoire). Nous utilisons le dictionnaire des synonymes DICTIONNAIRE pour examiner la présence des mots composés dans les textes traités. Cette analyse est réalisée à partir des textes lemmatisés à l'aide de TreeTagger et filtrés des éléments du dictionnaire d'arrêt.

[8] Ontologie terminologique (exemple : les lexiques) spécifie les termes utilisés pour représenter les connaissances dans un domaine. Elle est construite notamment à partir des corpus de spécialité.

[9] Nous conservons ici l'adjectif lemmatiseur (et non étiqueteur) pour TreeTagger puisque c'est les lemmes qui nous intéressent dans le cadre de cette analyse et non pas les étiquettes. Aucun traitement autre que la simple reconnaissance des lemmes n'est nécessaire ici.

Une liste des mots composés de DICTIONNAIRE est établie à partir des mots vedettes de ce dernier. Une copie de cette liste est conservée pour servir de référence. À partir de la première liste, de chaque mot composé vedette nous retirons les signes de ponctuation (traits d'union ou apostrophes). Ensuite nous appliquons le filtre à l'aide du dictionnaire d'arrêt pour enlever les mots vides de sens de l'ensemble des mots composés considérés. Suite à cela, nous réalisons un appariement exact entre la forme filtrée exempte de signes de ponctuation avec les textes filtrés lemmatisés du document. Pour valider la présence d'un mot composé dans un texte, ce dernier doit contenir la chaîne de lemmes conservés dans la forme finale du mot composé de DICTIONNAIRE. Au final, nous maintenons le mot composé de référence ainsi que sa forme décomposée.

5.8.3 Traitement des abréviations et des mots clés scientifiques

D'autres catégories de mots ne peuvent être reconnues directement par TreeTagger, par exemple : les sigles, les abréviations et autres acronymes en relation avec le domaine de la pollution domestique. Pour chaque mot en entrée, inconnu de TreeTagger, ce dernier propose la sortie <unknown>. L'analyseur TreeTagger est un analyseur général et donc non adapté au vocabulaire dédié concernant toutes les spécialités qui se greffent au domaine de l'air intérieur[10]. À titre d'exemple, nous pouvons citer le cas de l'abréviation BPCO qui désigne « Broncho-Pneumopathie Chronique Obstructive » et qui signifie un groupe de maladies chroniques systémiques d'origine respiratoire atteignant les bronches. Cette abréviation est très souvent citée dans les plaintes, notamment dans les passages extraits des résultats d'examens des spécialistes de la santé. Cette abréviation ou ses variantes[11] ne sont évidemment pas reconnues automatiquement par TreeTagger. Le même problème se pose pour des termes « atomiques » mais concernant un domaine scientifique spécialisé, comme « Aspergillus » ou bien « Cladosporum » ou encore « Alternaria ». Ces termes indiquent les types de moisissures les plus couramment relevées dans le bâtiment, connus par les experts, les médecins et parfois même par les occupants, en faire abstraction provoquerait une importante perte d'information.

Pour compenser cela, nous avons récapitulé l'ensemble des mots du corpus (constitué des 655 documents) inconnus par TreeTagger. L'objectif de cette étape est de substituer ces termes par des quasi-synonymes ou des termes plus génériques, compréhensibles par le lemmatiseur dans le cas où il s'agit de marques de produits ménager, ou une sorte très spécifique de champignons des milieux intérieurs par exemple. Cette liste correspond à un fichier texte où chaque entrée de ligne est un mot spécifique à substituer suivi du terme générique auquel il correspond et qui est compréhensible par l'outil TreeTagger. Nous avons réalisé cette fiche de substitution conjointement avec un professionnel du Sami (Dr Alain NICOLAS) qui est non seulement expert du domaine mais aussi médecin. Par rapport à la l'allure de la courbe des l'évolution du nombre des mots inconnus par le lemmatiseur (figure 5.4), et par rapport aussi à la taille réduite de notre corpus, la liste des mots inconnus accompagnés de leur substituants est évidemment incomplète. Pour

[10] La médecine, la chimie, la microbiologie, l'architecture du bâtiment, etc.

[11] Par exemple BPCO's.

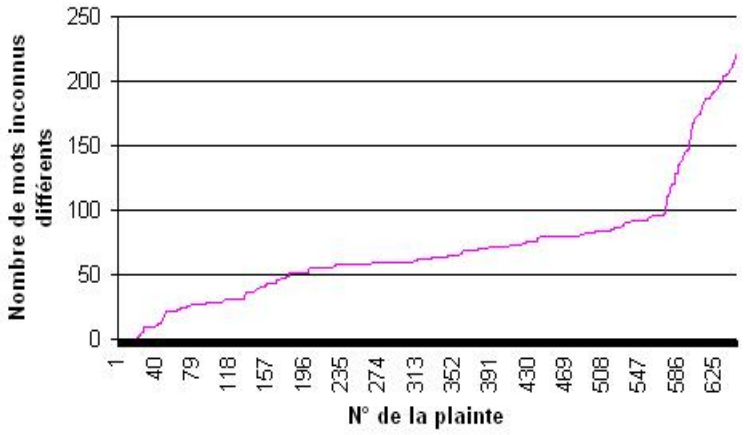


FIG. 5.4 – Évolution du nombre de mots inconnus différents en fonction du nombre des plaintes analysées

cela, nous avons dédié un fichier pour cumuler les mots inconnus de TreeTagger lors des futures utilisations de l'application. Cette fiche doit être mise à jour régulièrement par les spécialistes du domaine.

Rappelons que le but de cette analyse des textes est de connaître les propriétés du vocabulaire et de savoir s'il est possible de construire une ontologie (ou réseau sémantique) couvrant le domaine à partir du corpus. Par conséquent, suite à la mise en évidence des mots composés dans les textes des plaintes, et après que l'on ait substitué les termes inconnus du lemmatiseur, il devient possible et important d'analyser le vocabulaire pour se positionner par rapport à la réalisation d'un réseau sémantique du langage utilisé dans l'expression des plaintes.

5.8.4 Allure du vocabulaire

La compréhension de la langue naturelle est spécifiée formellement par la notion d'ontologie (4.2.2). Pour étudier la pertinence du recours à une ontologie dans le cadre de notre étude, nous analysons l'évolution du vocabulaire conformément aux méthodes classiques utilisées dans ce sens en RI (5.8.1). Le constat schématisé par l'allure de la courbe de la figure 5.5 est une preuve de l'insuffisance d'une éventuelle ontologie gérant de façon contrôlée et intégrée la sémantique et le vocabulaire terminologique issu du corpus des plaintes. En effet, la courbe ne cesse de croître. On peut observer également que la richesse du vocabulaire utilisé dépend de l'origine du laboratoire recevant la plainte. Les points encerclés sur la courbe désignent des passages d'un corpus appartenant à un laboratoire d'analyse des milieux intérieurs à un autre.

Afin de permettre l'usage de la langue naturelle dans la description des plaintes, il devient nécessaire de construire un réseau conceptuel généralisé de façon à comprendre le langage naturel. À ce jour, en sémantique, aucune classification universelle n'existe. La constitution d'une clas-

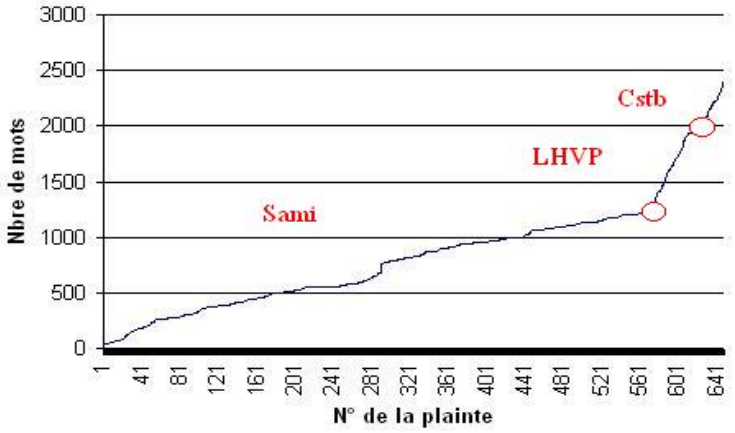


FIG. 5.5 – Évolution du nombre de mots différents (connus par TreeTagger) en fonction du nombre des plaintes analysées

sification universelle risque même d'être théoriquement impossible [40]. En effet, les contraintes de la modélisation cognitive sont simplement incompatibles avec le postulat des structures et des procédures qu'un système conceptuel à représentations permanentes exigerait. Néanmoins, la base de données lexicale WordNet (4.2.1) aurait pu servir de base pour représenter les termes et les relations entre termes, mais le point essentiel sur lequel nous insistons est que notre corpus est rédigé intégralement en français et que le système sera dédié au traitement des plaintes écrites en langue française.

La traduction manuelle de WordNet, ou de Cyc (section 4.2.2), en français serait une tâche énorme très coûteuse en temps et en ressources. De plus, le passage d'une langue à l'autre pose de gros problèmes en général. Une des difficultés inhérentes de la traduction est l'adaptation de la polysémie. Par exemple, dans le cas où un mot est polysémique d'une façon dans une langue et d'une façon différente dans une autre langue réaliser la traduction consisterait à revoir la structure de données. Les difficultés sont certes moins importantes dans le cas d'un domaine précis et technique, mais dans le cadre de notre étude et à la vue de la courbe de la figure 5.5, nous devons nécessairement tenir compte de la langue naturelle française générale tenant compte également des concepts (termes) issus du domaine de l'air intérieur et ses spécialités connexes (la médecine, le bâtiment, etc).

Rappelons qu'un WordNet pour le français a été développé dans le cadre du projet EuroWordNet (section 4.2.1). Cependant, deux raisons nous ont empêchés de nous en servir. La première, est que EuroWordNet ne comporte que des lexèmes verbaux et nominaux, mais ni adjectif ni adverbe. Le deuxième point pour lequel nous devions abandonner cette solution est dû aux problèmes de licence. Par ailleurs, WOLF 4.2.1 est une initiative très récente, elle nécessite néanmoins encore une validation manuelle indispensable.

Renoncer à toute idée de représentation symbolique pour expliquer l'interprétation sémantique

des énoncés langagiers des plaintes n'est pas une démarche satisfaisante dans le cadre de cette thèse. Nous avons alors été amenés à réfléchir aux différents intérêts que nous pouvons avoir en utilisant des dictionnaires électroniques des synonymes pour le contrôle de la sémantique tout en assurant une couverture la plus exhaustive possible du lexique des plaintes que nous devons traiter. Face à ce manque de standardisation d'un point de vu « réseau conceptuel », les SRI actuels implémentant la sémantique établissent leurs ressources sémantiques en partant depuis un thésaurus vide et exploitent les processus d'analyse automatique du lexique « métier » pour crée un dictionnaire adapté à leurs besoins[12].

On peut conclure que l'utilisation d'un dictionnaire généraliste en tant que ressource sémantique couplé d'une fiche de substitution mise à jour par le personnel expert nous semble la solution la mieux appropriée. De manière générale, ce constat est de mise pour les nombreux domaines qui ne disposent pas de thésaurus préétablis. Ceci est d'autant plus vrai concernant les domaines ayant trait à la science ou à des techniques récentes. En effet, en utilisant la version que nous possédons du dictionnaire des synonymes on s'est aperçu qu'il existait des termes connus du lemmatiseur mais qui sont absents dans DICTIONNAIRE (exemple : acarien). Comme nous l'avons réalisé dans le cadre de TreeTagger, et en collaboration avec le docteur Alain NICOLAS, nous avons créé une fiche de substitution des termes du corpus inconnus de DICTIONNAIRE (exemple : le substituant de acarien est araignée). Nous avons choisi l'option des fiches de substitution en évitant d'apporter des modifications dans DICTIONNAIRE. En effet, cette opération aurait eu inévitablement des conséquences sur la structure de données du dictionnaire. Ainsi, la totalité du corpus a été automatiquement revu pour remplacer les termes inconnus du dictionnaire des synonymes. Cette vérification est directement appliquée par notre système pour chaque nouvel enregistrement de plaintes à traiter.

De plus, DICTIONNAIRE ne tient pas compte du lien du « sens » entre les mots issus de la « même famille » (pas de lien sémantique entre « polluer » et « pollution »). Ce qui est normal, puisque des termes de catégories grammaticales différentes ne peuvent être ni synonymes directs, ni même pseudo-synonymes par transitivité (synonymes d'autres synonymes). Pour évaluer la sémantique commune entre les mots d'une même famille, nous avons utilisé DICTIONNAIRE.

5.9 Racinisation, DICTIONNAIRE et sémantique

Au coeur des problèmes du TAL, se pose le choix des concepts (pour la définition de nouveaux thésaurus notamment). Salton avait préconisé en 1966 l'automatisation de ces tâches car leur réalisation manuelle est coûteuse et non déterministe (à l'origine de solutions hétérogènes)[103]. Le système ANA (Apprentissage Naturel Automatisé) développé par Enguehard [36] et qui est directement inspiré par l'apprentissage humain de la langue maternelle, est un bon exemple d'implémentation réussie de cette approche.

[12] www.clever-age.com/veille/clever-link/organiser-sa-gestion-documentaire-deuxieme-partie.html

5.9.1 L'heuristique d'Enguehard

Enguehard a choisi de travailler avec un minimum de connaissances, sans analyseur syntaxique, sans dictionnaire mais uniquement par observation des textes pour la sélection des concepts primitifs. Enguehard détermine le concept à partir des différentes flexions des termes de la langue française en employant le postulat suivant : « la forme canonique correspondant à un terme est la sous-chaîne de caractères rassemblant les premières lettres qui le composent jusqu'à l'obtention de deux voyelles non consécutives ».

Ce besoin de réunir des termes de graphies différentes et issus d'une même famille (ou racine) est considéré tantôt comme une entrave à l'efficacité de la reconnaissance, tantôt comme une richesse permettant d'acquérir davantage de termes. En effet, la traduction des termes par des codes ne se fait pas sans heurt. Les concept-racines (le concept qui englobe les termes issus de la même famille) définis par ce postulat sont des formes appauvries de la langue, et surtout pour une langue aussi riche par ses variations morphologiques que le français[13], ce qui est là en effet une problématique que nous avons évoquée au début de ce mémoire à la section 2.2.2. Néanmoins, nous avons choisi d'implanter le principe de racinisation d'Enguehard appliqué au français, d'une part par rapport à l'aisance de sa mise en œuvre et d'autre part par rapport à la réussite de son implémentation dans d'autres applications [60].

5.9.2 Application de l'heuristique d'Enguehard dans DICTIONNAIRE

Dans le cadre de notre travail, la prise en compte des variations des racines est traitée postérieurement à l'extraction des candidats-termes (section 5.8.1). Pour déterminer le taux de similarité entre les candidats termes issus d'une même racine nous avons réalisé une extrapolation du rapport des indices de communauté à l'instar des appariements sémantiques entre les vedettes de DICTIONNAIRE. Pour chaque paire de mot-vedette A et B de DICTIONNAIRE issus d'une même racine nous avons effectué un échange de synonymes en introduisant un degré d'influence de $\frac{1}{2}$. Cet échange des synonymes n'est pas toujours possible dans les dictionnaires. En effet, la synonymie est une relation pseudo transitive, la polysémie est un frein à la transitivité dans la synonymie. Si X est synonyme (ou traduction) de Y, et que Y est synonyme (ou traduction) de Z, alors ou bien X et Z sont synonymes, ou bien Y (l'élément transitoire) est polysémique.

Prenons un exemple similaire dans DICTIONNAIRE : « fenêtre » est synonyme de « baie », « baie » est synonyme de « golf », « fenêtre » n'est pas synonyme de « golf », et pour cause ; baie est polysémique. Cependant, nous appliquons la transitivité pour réaliser un échange mutuel de synonymes entre les termes de la même racine. Cet échange consiste à considérer les synonymes d'un terme A de la même racine qu'un terme B en tant que synonymes de ce dernier. Nous

[13] Pour l'allemand, les travaux de Marko [56] se portent en faveur de la racinisation et pour l'arabe, une étude réalisée par Al-Kharashi [52] montre également que l'utilisation de racines pour les indexes est une meilleure méthode que l'utilisation de lemmes ou de mots.

jugeons la transitivité possible dans ce cas de figure, puisque le problème de polysémie est levé en considérant les deux termes de la même famille en tant qu'une seule racine. C'est à dire, les synonymes du terme A sont synonymes du terme B puisque A et B (issus de la même racine) partagent un même sens (sans polysémie) comme le stipule l'hypothèse initiale d'Enguehard.

L'extrapolation de l'indice de communauté concernant l'intersection des ensembles nous permet de considérer le résultat de l'intersection entre deux ensembles dont les éléments sont ponctués de degrés d'influence différents. Le résultat d'une telle intersection correspond au résultat d'une intersection classique mais dont les éléments communs aux deux ensembles sont ponctués de la valeur d'influence minimale initiale. Nous étendons cette extrapolation également sur la cardinalité des ensembles. La cardinalité d'un ensemble constitué d'un élément ponctué d'un certain degré d'influence correspond à ce dernier.

Dans l'exemple suivant on suppose que les deux entrées du dictionnaire des synonymes A et B sont de la même famille. Nous supposons également que A' et A" (B', B") sont les deux synonymes de A (respectivement de B). La similarité entre A et B est calculée comme suit :

$$Sim(A,B) = \frac{|\{\frac{1}{2}A, \frac{1}{2}A', \frac{1}{2}A'', \frac{1}{2}B, \frac{1}{2}B', \frac{1}{2}B''\}|}{|\{A, A', A'', B, B', B''\}|} = \frac{1/2\ |\{A, A', A'', B, B', B''\}|}{|\{A, A', A'', B, B', B''\}|} = 1/2$$

Par conséquent, nous pouvons généraliser et dire que le taux de similarité entre les termes de toute paire issue d'un même concept-racine est de $\frac{1}{2}$.

Cette généralisation de représentation fondée sur un système de racinisation défini par un postulat ad-hoc (adapté à la langue) est particulièrement efficace lorsque les textes se réfèrent à un domaine technique, où les textes sont généralement écrits dans un langage opératif comportant peu d'homographes ou de synonymes [41].

Rappelons que nous utilisons les mesures et les ressources sémantiques pour enrichir les systèmes de recherche que nous employons dans le cadre du module fonctionnel (section 5.5.2). Les modèles de recherche implémentés sont à l'origine des systèmes de classification réalisés[14] (section 5.5.1). Par conséquent, tenir compte de la sémantique permettrai d'une part de mettre en exergue des scénarios pertinents et plus faciles à interpréter et d'autre part de réaliser une assignation de solution mieux adaptée à la plainte à traiter .

[14] La distinction entre ces deux domaines n'est pas toujours très facile à établir [117].

5.10 Construction de la base de scénarios

La forme et le contenu répétitifs que nous avons pu apercevoir dans l'ensemble des rapports, établis manuellement par les diagnostiqueurs pour répondre à des plaintes, témoignent d'une régularité des situations de pollution domestique (section 5.5.1). Prendre en compte la régularité permettrait de dépasser en quelque sorte les limites «individuées» pré-supposées des plaintes compromettant la faisabilité d'une approche automatique pour leur traitement (section 5.4.2). En effet, expliquer la nature des classes résultant d'une catégorisation automatique des textes des plaintes permettrait d'organiser ces dernières sous forme d'une base de scénarios où à chaque scénario est attribué un rapport de solution type. Ainsi, une approche automatique de réponse aux plaintes devient envisageable. La catégorisation automatique des textes utilise des mesures de similarité inter-textuelle. Une question se pose cependant : quelle classification des textes convient-il d'adopter pour l'ensemble des modèles implémentés ?

5.10.1 Construction des scénarios par application de l'algorithme des K-moyennes

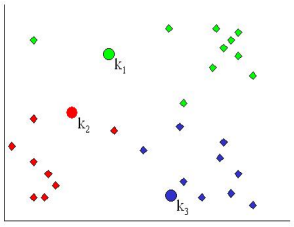
Suite au filtrage et à la lemmatisation des textes exprimés en langue naturelle, chaque modèle de représentation (et de recherche) utilise des attributs caractéristiques spécifiques à sa propre définition pour modéliser les textes. Les modèles algébriques implémentés, en l'occurrence le modèle vectoriel étendu en vue bidimensionnelle direct et le modèle vectoriel étendu sémantique, utilisent DICTIONNAIRE en tant qu'espace de représentation, et ce, pour chaque rubrique renseignée. L'algorithme de quantification vectorielle des k-moyennes [73] classe les objets selon leurs descriptions vectorielles en k classes. Nous adaptons l'algorithme des k-moyennes à la description matricielle des textes des plaintes. Le nombre de classes est fixé a priori par l'utilisateur et représente le seul paramètre de l'algorithme. Pour notre catégorisation de l'échantillon représentatif des plaintes, nous avons souhaité réaliser des catégorisations allant de 3 à 8 scénarios. L'algorithme implémenté commence par choisir aléatoirement[15] k plaintes d'observation **différentes** dans l'échantillon représentatif des plaintes. L'algorithme est ensuite itératif, chaque itération est composée de deux étapes :

Pour chaque point d'observation une classe est constituée. Cette classe regroupe les plaintes qui sont jugées plus proches de la plainte d'observation considérée que les k-1 autres plaintes référentes (autres points d'observation). Les taux de similarité entre plaintes sont calculés en appliquant une agrégation des similarités locales évaluées à partir des vecteurs caractéristiques.

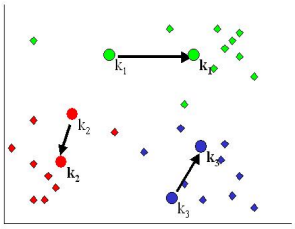
Une plainte-centroïde (ou centroïde)[16] est ensuite agrégée pour chaque classe à partir de ses constituants pour remplacer le centroïde précédent. Plus précisément, un nouveau centroïde est pris en compte, les coordonnées de ses vecteurs caractéristiques sont calculées en réalisant une

[15] Il se peut que le choix des k centres initiaux soit orienté par une heuristique.

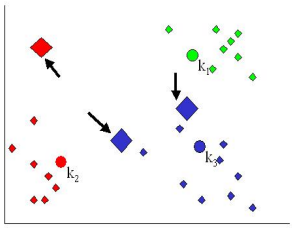
[16] La moyenne des vecteurs ne correspond plus (à part exception) à un document du corpus.



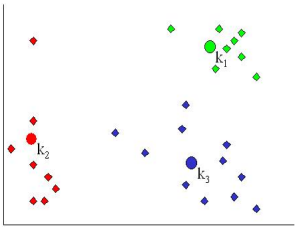
1- Centroïdes aléatoires et 1ère classification



2- Réévaluation des centroïdes



3- Les points qui changent de classe



4- Recalcul des centroïdes

TAB. 5.9 – Classification des objets dans un espace vectoriel

moyenne des valeurs scalaires des vecteurs représentant les rubriques de l'ensemble des plaintes de la classe considérée.

S'ensuit une alternance entre association des points au centroïde le plus proche et agrégation des centroïdes des classes inhérentes jusqu'à convergence (Tableau 5.9). L'algorithme converge lorsque l'on obtient pour deux itérations consécutives les mêmes centroïdes ce qui correspond à dire que c'est aussi lorsque plus aucune plainte ne change de classe. Nous avons adapté l'algorithme des k-moyennes pour la catégorisation des plaintes dont la représentation est vectorielle. Les modèles de recherche dont le formalisme tient compte de la forme filtrée lemmatisée des textes ne peuvent pas être à l'origine d'une classification employant les centroïdes issus d'agrégations vectorielles. Dans la section suivante nous présentons la méthode appliquée pour la classification fondée sur le modèle de proximité floue (direct (section 2.4.2) et sémantique (section 5.7.2)) et le modèle des superpositions des signaux (direct (section 5.7.3) et sémantique (section 5.7.4)).

5.10.2 Construction des scénarios par sélection des référents optimaux

Cette méthode est fondée sur un principe proche de l'algorithme des k-moyennes. Seulement, démunis de mesures vectorielles, les référents (centroïdes dans la configuration géométrique) correspondent à une plainte existante. Pour qu'une plainte soit retenue en tant que barycentre de son groupe, il est nécessaire que la somme des distances (inter-textuelles) entre cette plainte d'observation et les autres plaintes de la même classe soit la plus petite en comparaison avec les autres éléments de la même classe. De la même manière que la méthode des k-moyennes, la convergence de cet algorithme est atteinte dès que pour une nouvelle mise à jour des référents les groupes restent identiques.

Les classes issues des différentes catégorisations des plaintes, réalisées à l'aide des modèles de recherche implémentés, correspondent aux scénarios que nous confrontons aux classifications des experts résultant du même échantillon des plaintes. Les résultats de ces comparaisons sont présentées dans la section 6.4 du chapitre d'évaluation.

5.11 Conclusion

Dans les pages qui précèdent, nous avons cherché à exposer d'une part, notre approche méthodologique pour apporter des solutions automatiques à des plaintes écrites en langue naturelle, et d'autre part, d'évoquer nos réflexions et nos apports personnels dans le domaine de la recherche d'information et la sémantique.

Par rapport aux connaissances en matière de pollution domestique (normes, données, niveau de connaissances, etc.) d'une part, et par rapport à la nature des plaintes d'autre part, nous avons été contraints d'abandonner des méthodes classiques des systèmes à base de connaissance que nous avons implémentées au début de cette thèse, et de définir une approche plus adaptée. Cependant, nous sommes conscients que les formalismes définis par les modèles utilisés ainsi que les moyens sémantiques employés dans le cadre de notre approche adaptée ne fournissent que des solutions partielles aux différents problématiques du TAL et de la sémantique. Malgré nos efforts, le cadre de modélisation des textes des plaintes, de recherche et de la sémantique est assez restreint. Cette imperfection résulte du fait que les terminologies, en l'occurrence DICTIONNAIRE et les fiches de substitutions réalisées, n'offrent qu'une possibilité de description (indexation) limitée aux concepts réels recouverts par les mots clés utilisés. En effet, c'est selon la terminologie utilisée que la description est plus ou moins fine.

Un autre aspect ineffable des systèmes de recherche actuels de manière générale est la non-prise en compte de la négation. Si « j'ai une moquette dans le salon » a une représentation vectorielle à l'aide de DICTIONNAIRE et une traduction dans les modèles fondés sur la proximité des termes, nous ne savons pas interpréter « je n'ai pas de moquette dans le salon ». Toutefois la négation dans certaines expressions sommaires lexicalisées peuvent être traitées en introduisant l'opérateur logique « NOT » dans la définition des modèles booléens adaptés à ce besoin. Cependant, connaître les dépendances des adverbes de négations par rapports aux mots clés retenus est un problème qui subsiste inévitablement aujourd'hui. À l'instar de cette insuffisance, nous avons pris connaissance du manque de modélisation des localisations temporelles, exemple : hier, aujourd'hui, l'année dernière. De même, les informations numériques ne sont pas prises en compte, exemple : « je passe l'aspirateur 4 fois par jour ».

Au vu de ces différentes constatations concernant les insuffisances inhérentes de l'utilisation de thésaurus et de dictionnaire d'arrêt simplifiant la formalisation de la langue naturelle, il est nécessaire d'évaluer notre approche par confrontations aux jugements des experts. Les évalua-

tions que nous présentons dans le chapitre qui suit concernent les comparaisons des composants (modèles) du module fonctionnel aux jugements des experts. Ce qui consiste à évaluer la capacité de notre système à réaliser des assignations exactes de solutions. Étant donné que les scénarios dépendent considérablement de ces modèles, une confrontation des scénarios automatiques aux classes des experts issues du même échantillon est aussi nécessaire.

Chapitre 6

Expérimentations et évaluations

Nous avons présenté dans le chapitre précédent notre approche pour l'attribution de solutions automatiques à des plaintes écrites. Dans ce chapitre, nous présentons les résultats des expérimentations menées à partir d'un corpus représentatif de plaintes pour connaître les effets concrets de notre applicatif. Nous présentons d'abord l'environnement technologique qui nous a permis d'élaborer notre prototype (section 6.1). Ensuite, dans la section 6.2, nous exposons les caractéristiques du corpus expérimental des plaintes que nous utilisons afin de mettre en évidence, en fonction de la pratique, le potentiel de notre système.

Dans la partie 6.3, nous présentons les résultats des analyses comparatives des modèles de recherche implémentés dans le cadre du module fonctionnel. Dans cette section, nous évaluons les expérimentations des modèles en fonction de leur principe théorique. Ensuite, pour chaque modèle expérimenté, nous indiquons une comparaison entre les approches par recherche directe et par recherche sémantique.

Nous exposons dans ce chapitre les résultats de nos expérimentations ayant porté sur l'étude de la régularité thématique des plaintes à travers une comparaison des scénarios automatiques par rapport aux scénarios des spécialistes du domaine (section 6.4). Pour l'évaluation générale de notre applicatif, nous exposons dans la section 6.5 le taux de réussite des assignations relativement aux propositions des experts, et cela pour l'ensemble des systèmes de recherche implémentés . Enfin, afin de juger de l'adaptabilité des modèles de recherche implémentés en fonction de la taille des documents, ce que nous avons énoncé théoriquement dans les chapitres précédents (2 et 5), nous indiquons les résultats d'une analyse statistique témoignant du degré de dépendance existant entre la réussite des assignations et la taille moyenne des rubriques des documents *XMLisés* (section 6.6).

6.1 L'environnement applicatif

6.1.1 Le langage de programmation

Pour développer notre applicatif, nous avons utilisé le langage de programmation Java (1.6.0). De plus, notre choix s'est porté sur ce langage par rapport notamment à sa disponibilité gratuite sur le Web.

6.1.2 Scilab

Pour élaborer les courbes rappel-précision permettant de discriminer les modèles de recherche entre eux, nous avons utilisé le logiciel Scilab. À l'aide de sa fonction *plot*, nous avons pu réaliser de manière très souple les courbes. Notre choix s'est porté sur Scilab car il permet de dessiner dans une même fenêtre graphique des courbes différentes suivant des abscisses et des ordonnées entièrement hétérogènes.

6.1.3 Utilisation et gestion de TreeTagger

Pour la lemmatisation des textes des rubriques renseignées à partir de l'interface usager, nous utilisons le lemmatiseur TreeTagger adapté au Français. Comme nous l'avons déjà indiqué dans la présentation de TreeTagger, nous avons été obligés de reprendre la forme de base des enregistrements des rubriques des plaintes. Le fichier en entrée de TreeTagger doit contenir un mot par ligne pour que la totalité du texte soit prise en compte, autrement, c'est uniquement le premier mot qui est considéré. En sortie, nous récupérons uniquement les lemmes des formes initiales des mots. Par ailleurs, nous avons constaté que TreeTagger ne tient compte que des 200 premiers mots du fichier en entrée. Par conséquent, nous avons fragmenté automatiquement les textes des rubriques renseignées et dont la taille est supérieure à 200 mots en un ensemble de fichiers de taille inférieure à 200 chaînes de caractères. Ces fragments des textes initiaux ont été validés manuellement, assurant ainsi le maintien de la composition syntaxique initiale des phrases. La non-prise en compte de cet aspect aurait pu être à l'origine d'incohérences d'étiquetage et par conséquent d'erreurs de lemmatisation.

6.2 Spécificités du corpus

6.2.1 La taille du corpus expérimental

Comme nous l'avons annoncé dans la section 5.6, notre corpus des plaintes dans sa totalité est composé de 655 éléments. Nous avons choisi de réaliser les expérimentations de ce chapitre sur un échantillon de 100 plaintes. Cela nous permet de conserver un nombre conséquent de plaintes pour des vérifications ultérieures. Ce corpus de travail est certes réduit mais représentatif. En effet, en accord avec les experts, en respectant les proportions du corpus de base selon l'organisme de provenance, nous avons constitué manuellement le corpus expérimental en s'assurant de maintenir

l'exhaustivité initiale du corpus de base. De plus, la majorité des évaluations présentées dans ce chapitre sont réalisées à partir de comparaisons entre les résultats des applications automatiques et les résultats issus des jugements des experts. Par conséquent, afin de réduire la difficulté des réalisations manuelles nous avons été contraints d'alléger la taille du corpus.

6.2.2 Génération de la structure du corpus

Le corpus expérimental des 100 plaintes *XMLisées* est un corpus XML obéissant à une structure unique. Cette structure est générée au moyen de l'interface usager (figure 6.6). Pour constituer le corpus structuré en mémoire archive, nous avons saisi manuellement les énoncés des plaintes que nous avons extraits directement à partir des textes bruts et nous les avons intégrés dans l'interface utilisateur. Les rubriques XML des plaintes séparées par les balises « symptôme », « habitat » et « environnement_extérieur » se présentent sous forme d'une succession de mots issus des traitements du lemmatiseur TreeTagger. Il est important de noter ici que l'ordre dans l'usage de la stop-liste et du lemmatiseur est important. En effet, il est primordial d'utiliser la stop-liste en aval de l'outil TreeTagger. Les mots outils sont essentiels pour reconnaître la catégorie grammaticale des termes. Pour l'expression suivante : « j'ai mal à la gorge », le terme « gorge » est reconnu directement par TreeTagger et cela en fonction de son rôle dans la phrase. Cependant, dans le cas où le filtre des mots vides de sens est appliqué en amont de TreeTagger, l'expression en entrée serait « mal gorge ». Dans ce cas, TreeTagger présente en sortie « mal gorger ».

6.3 Évaluation du module fonctionnel

Dans le cadre de notre étude, l'aspect expérimentation a occupé une place particulière. On voulait tester plusieurs méthodes de recherche applicables à notre problématique afin de vérifier leurs effets concrets et de pouvoir appliquer le système le mieux adapté pour assurer un maximum d'assignations réussies. La méthode d'évaluation des systèmes de recherche d'information la plus usuelle actuellement tient compte des deux critères d'évaluations : « le rappel » et « la précision »[1].

6.3.1 Mesure du taux de rappel et du taux de précision

Dans le cadre de notre approche, nous calculons les deux critères (rappel et précision) à partir d'un classement, établi par un système de recherche, des documents par rapport à une requête. Aux différentes rangs du classement nous calculons les deux taux rappel et précision. À une position P d'un classement d'un système considéré, nous calculons le rapport entre le nombre des documents pertinents classés en tête de liste jusqu'à l'élément considéré et le nombre des documents pertinents existant en mémoire archive (formule 6.1). Nous mesurons le taux de précision aux différentes positions du classement en question également. À une position donnée

[1] C'est sur cette expérience que s'est basée National Institute of Science and Technolog (NIST) pour créer la campagne d'évaluation Text REtrieval Conference (TREC) en 1992.

le taux de précision correspond au rapport entre le nombre des documents pertinents trouvés jusqu'à la position concernée, P, et le nombre total des positions considérées jusqu'au rang en question (formule 6.2). Un cours remarquable de Catherine Berrut [14] nous a permis de bien comprendre le principe de cette méthode d'évaluation des SRI.

$$Rappel_P = \frac{Le\ nombre\ des\ documents\ pertinents\ jusqu'au\ rang\ P}{Le\ nombre\ total\ des\ documents\ pertinents} \tag{6.1}$$

$$Prcision_P = \frac{Le\ nombre\ des\ documents\ pertinents\ jusqu'au\ rang\ P}{P} \tag{6.2}$$

Le rappel et la précision prennent des valeurs entre 0 et 1. En plus de ces deux facteurs, d'autres mesures peuvent être calculées pour évaluer les systèmes de recherche. Nous pouvons citer le silence (complément du rappel, « 1-rappel ») (formule 6.3), le bruit (complément de la précision, « 1-précision ») (formule 6.4), l'hallucination (formule 6.5) et l'élimination (complément de l'hallucination) (formule 6.6). Par ailleurs, la généralité (ou bien la densité du thème) est une mesure non qualitative par rapport au système étudié, elle évalue néanmoins le degré de difficulté de trouver les documents pertinents à partir de la proportion de ces derniers dans le corpus (ou bien la précision moyenne que l'on obtiendrait en sélectionnant les documents aléatoirement) (formule 6.7).

$$Silence = \frac{Le\ nombre\ des\ documents\ pertinents\ non\ extraits}{Le\ nombre\ total\ des\ documents\ pertinents} \tag{6.3}$$

$$Bruit = \frac{Le\ nombre\ des\ documents\ non\ pertinents\ extraits}{Le\ nombre\ des\ documents\ extraits} \tag{6.4}$$

$$Hallucination = \frac{Le\ nombre\ des\ documents\ non\ pertinents\ extraits}{Le\ nombre\ total\ des\ documents\ non\ pertinents} \tag{6.5}$$

$$\acute{E}limination = \frac{Le\ nombre\ des\ documents\ non\ pertinents\ non\ extraits}{Le\ nombre\ total\ des\ documents\ non\ pertinents} \tag{6.6}$$

$$G\acute{e}n\acute{e}ralit\acute{e} = \frac{Le\ nombre\ total\ des\ documents\ pertinents}{Le\ nombre\ total\ des\ documents\ du\ corpus} \tag{6.7}$$

Le rappel et la précision sont les deux notions les plus souvent utilisées. Un système de recherche parfait doit avoir une précision et un rappel d'une valeur égale à 1. Ce cas de figure est souvent contradictoire, puisque dans la réalité une très forte précision ne peut être obtenue qu'au prix d'un rappel faible et vice-versa [117]. La mesure de performance la plus communément utilisée est la courbe rappel-précision. Pour discriminer, entre eux, les systèmes que nous avons développés, nous comparons leurs courbes rappel-précision. Pour élaborer ces courbes,

nous avons choisi un échantillon de 15 plaintes *XMlisées* (de taille variée et d'organismes différents) dans notre corpus expérimental de 100 documents. Pour l'ensemble des systèmes de recherche implémentés, nous avons réalisé un classement des éléments du corpus en fonction de la chaque requête de l'échantillon des 15 requêtes sélectionnées. Ainsi, pour chaque modèle nous avons calculé les taux de rappel et de précision (tenant compte des jugements de pertinence des experts) pour chaque requête aux différents rangs des classements. Ensuite, pour chaque modèle, nous avons établi, aux différentes positions du classement, une moyenne des précisions et des rappels calculés pour l'ensemble des requêtes. Cette méthode d'agrégation des taux du rappel et de la précision initiés d'un jeu de requêtes est appelée « la méthode de la moyenne » ou encore en anglais « user oriented recall average »[2] [14]. Dans la suite de cette section, nous utilisons uniquement les courbes rappel-précision, car elles nous apportent une information suffisante sur le comportement des systèmes développés.

6.3.2 Évaluation des modèles de recherche par approche comparative

Dans cette section nous présentons les courbes rappel-précision mettant en évidence des confrontations entre modèles de recherche ayant opéré à l'aide d'un même jeu de requêtes sur un même corpus. Dans le cadre de notre application, la précision est privilégiée par rapport au taux de rappel. En effet, ce constat est de mise par rapport à la réalisation de l'assignation de solution à la plainte à traiter qui est effectuée en fonction de l'élément positionné en tête de liste dans le classement du modèle de recherche employé. Par conséquent, pour juger de la performance des systèmes à partir des courbes rappel-précision nous analysons les positions des courbes les unes par rapport aux autres aux premiers taux de rappel.

Nous réalisons des comparaisons entre les modèles en fonction de leur principe théorique. D'abord, nous montrons les résultats des systèmes basés sur la proximité des termes en tenant compte de la symétrisation par moyenne des valeurs réciproques dont nous avons discuté dans le chapitre 5 de ce mémoire.

- **Mise en évidence de l'effet de la symétrisation des modèles « orientés requête »** En effet, le modèle de Mercier (section 2.4.2) et le modèle fondé sur la superposition des signaux que nous avons présentés dans la section 5.7.3, sont des modèles orientés requêtes, donc asymétriques. Nous souhaitons montrer dans cette section l'intérêt concret de la symétrisation de ces approches. La figure 6.1 montre les courbes rappel-précision pour les méthodes de proximité floue implémentant l'opérateur logique « OU » mettant en évidence l'avantage de la symétrisation du modèle de Mercier dans le cadre de notre application.

[2] Une deuxième méthode d'agrégation existe. Elle est connue sous le nom de « la méthode de fusion des requêtes » ou bien en anglais « system oriented recall average ».

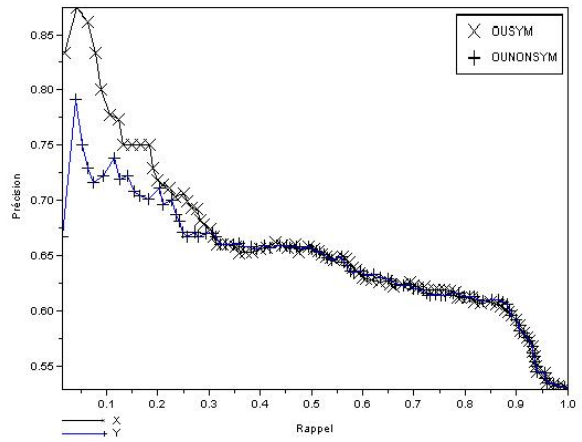


FIG. 6.1 – Évaluation de l'intérêt de la symétrisation du modèle de proximité floue

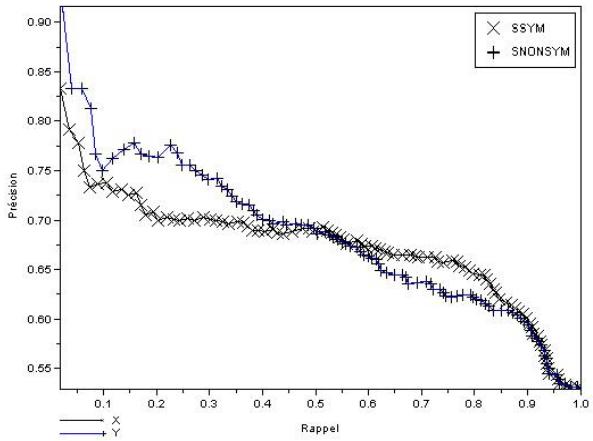


FIG. 6.2 – Évaluation de l'intérêt de la symétrisation du modèle de la superposition des signaux

À l'opposé de cette constatation, les courbes de la figure 6.2 sont la preuve que le modèle fondé sur le principe de la superposition des signaux est plus efficace dans le cadre asymétrique.

– **Analyse de l'intérêt de la mesure TF-ITDF des vedettes dans le cadre précis de notre application**
Par rapport à la taille de notre corpus nous étions conscients que l'évaluation du niveau de la force discriminatoire du modèle vectoriel de Salton (et de Zargayouna en particulier) n'était pas une information très pertinente. Par conséquent, nous avons développé en parallèle le modèle vectoriel binaire dans l'espace de DICTIONNAIRE. Nous utilisons le terme binaire par rapport aux valeurs scalaires correspondant aux primitives vectorielles issues du dictionnaire des synonymes. En effet, la valeur scalaire correspondant à une vedette du dictionnaire est égale à 1 si une occurrence de la vedette en question est dans le

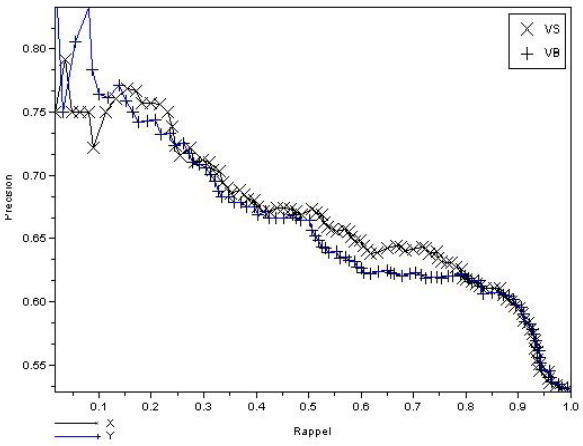


FIG. 6.3 – Évaluation des modèles vectoriels directs

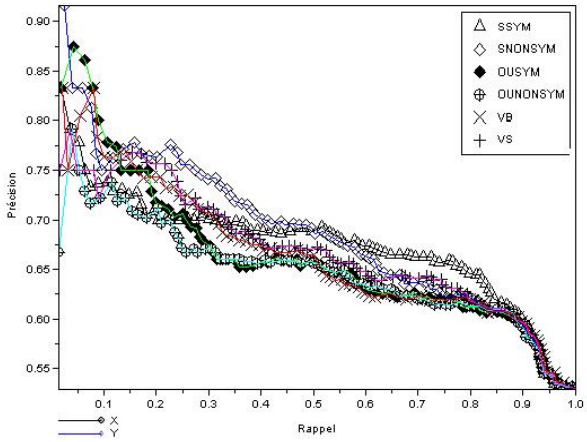


FIG. 6.4 – Évaluation de l'ensemble des modèles directs

texte modélisé, et elle prend la valeur 0 sinon. La figure 6.3 montre en effet, une légère amélioration du classement des documents au moyen du modèle binaire par rapport au modèle de Zargayouna.

– **Évaluation générale des modèles de recherche directs**
La figure 6.4 indique le classement des modèles développés en fonction de leur niveau de pertinence indiqué par les courbes rappel-précision. Selon le besoin de notre application, le modèle fondé sur la théorie du signal non symétrisé est le meilleur. Derrière, nous remarquons le modèle de proximité floue implémenté sous sa forme symétrique. Cette expérience qui prouve l'avantage pratique du modèle de proximité floue symétrisé par rapport à sa version asymétrique témoigne de l'intérêt de notre contribution par rapport à l'application de la moyenne des valeurs réciproques du modèle de Mercier dans le cadre de notre étude.

	Modèle du signal		Proximité floue		Vect-binaire	Vect-Zarga
Mode	NonSym	Sym	NonSym	Sym		
Moyenne	0,875	0,813	0,730	0,833	0,792	0,771
Classement	1^{er}	$3^{ème}$	$6^{ème}$	$2^{ème}$	$4^{ème}$	$5^{ème}$

TAB. 6.1 – Classement des modèles directs conformément aux moyennes de précision

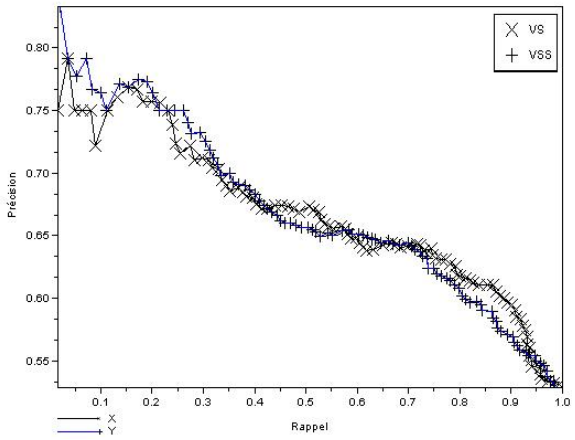


FIG. 6.5 – Évaluation de l'intérêt de la sémantique dans le modèle vectoriel bidimensionnel de Zargayouna

Concernant la suite du classement, et par rapport à des soucis de visibilité concernant la position des courbes les unes par rapport aux autres, nous avons réalisé une moyenne des précisions jusqu'au taux de rappel de 5%. Le tableau 6.1 indique le classement des modèles conformément à la moyenne des précisions au taux de rappel retenu.

Ainsi, nous constatons à partir des courbes de la figure 6.4 et du tableau 6.1, que notre modèle asymétrique initié du principe des interférences des signaux émis par les occurrences des termes de la requête dans les documents, est le plus performant, du moins dans le contexte précis de notre corpus et de notre application.

6.3.3 Évaluation de l'intégration de la sémantique

À l'exemple des analyses discriminantes pour l'évaluation des modèles de recherche, nous nous basons sur les courbes rappel-précision pour mettre en exergue l'effet de la prise en compte de la sémantique dans le comportement des modèles de recherche. La figure 6.5 témoigne de l'intérêt de l'utilisation de la sémantique, déterminée à partir de DICTIONNAIRE et de l'heuristique des codes, dans le cadre du modèle vectoriel de Zargayouna étendu en vue bidimensionnelle.

Dans le tableau 6.2 nous exposons le classement général des modèles directs et sémantiques. En première position se trouve notre modèle inspiré du principe du signal asymétrique non-sémantique. Avec une moyenne de pertinence imperceptiblement plus faible, notre modèle du

		Signal		Flou		Vect-bin	Vect-Zarga
	Mode	NonSym	Sym	NonSym	Sym		
2*Direct	Moyenne	0.875	0.813	0.73	0.833	0.792	0.771
	Rang	1^{er}	$5^{ème}$	$12^{ème}$	$3^{ème}$	$7^{ème}$	$9^{ème}$
2*Sémantique	Moyenne	0.854	0.75	0.771	0.833	0.771	0.813
	Rang	$2^{ème}$	$8^{ème}$	$9^{ème}$	$3^{ème}$	$9^{ème}$	$5^{ème}$

TAB. 6.2 – Classement général des modèles conformément aux moyennes de précision

signal asymétrique sémantique se positionne en seconde position. À un même niveau de performance, le modèle de proximité floue symétrisé direct et le modèle de proximité floue symétrisé sémantique se situent en troisième position. Avec un degré de pertinence de 81,3%, au taux de rappel de 5%, nous trouvons le modèle vectoriel de Zargayouna sémantique et le modèle flou asymétrique direct en $5^{ème}$ position. En fin de classement nous observons d'abord le modèle vectoriel binaire, dans la suite du classement nous trouvons sa version sémantique. À la même position que ce dernier modèle, et avec une pertinence de 77,1%, le modèle flou asymétrique sémantique présente une performance meilleure que sa version non-sémantique positionnée en fin de classement juste derrière le résultat de l'application du modèle du signal symétrisé tenant compte de la sémantique.

En résumé, nous retenons du tableau 6.2 que les modèles implémentés dans le cadre de notre corpus et de notre application et pour qui l'avantage de l'intégration de la sémantique est probante sont : le modèle vectoriel de Zargayouna et le modèle flou asymétrique. Par rapport au contexte non-sémantique (tableau 6.1) et le contexte général (tableau 6.2) ces approches ne sont pas les plus appropriées aux besoins de notre application (par rapport notamment aux jugements des experts). Pour les modèles les plus en accord avec le classement des experts, en l'occurrence le modèle des superpositions des signaux et le modèle flou, l'intégration de la sémantique n'apporte pas une amélioration probante par rapport à leurs applications non-sémantiques. En effet, leurs adaptations sémantiques les maintient aux mêmes rangs dans le classement sans les améliorer notablement.

Ces résultats appellent plusieurs commentaires :

1. Concernant les modèles directs, nous avons remarqué que les jugements de pertinence des documents portent essentiellement sur la présence, dans ces derniers, des mots clés du document-requête. Le problème majeur est que ces mots clés ne sont pas tous de pertinence égale. En effet, dans notre domaine, il existe des termes plus subtils que d'autres, plus décisifs à la discrimination des documents par rapport au sens de la requête. Par exemple, les mots clés : « maison », « enfant », « problème » sont des termes très souvent rencontrés, desquels on doit néanmoins tenir compte. Par contre, les mots clés « moisissure », « fuite », « fibre » indiquent beaucoup plus aux experts la nature du scénario à l'origine de la plainte. Et bien évidemment, d'autres mots clés sont encore plus déterminants que les autres. En

effet, nous avons constaté pendant nos réunions de lecture et d'analyse des plaintes avec les experts du CSTB que lorsqu'un spécialiste aperçoit le mot clé « cancer-de-la-plèvre » dans le texte d'une plainte, sa déduction du phénomène est automatique. Dans ce cas précis, l'expert est catégorique par rapport au fait qu'il s'agisse d'un scénario d'exposition aux fibres, et plus particulièrement à l'amiante. Pour le modèle vectoriel classique, cette information est prise en compte par la mesure IDF, qui évalue le degré de représentativité d'un terme par rapport à sa fréquence dans le corpus. Cette donnée a plus de sens dans les corpus de très grande taille. Par rapport aux modèles fondés sur le principe de la densité des termes de la requête dans les documents (modèle du signal et le modèle flou), leur définition doivent tenir compte des poids des termes. Dans le cadre de notre application, les poids d'intérêt des mots clés peuvent être définis par les experts du domaine. Nous avons également pensé à permettre aux usagers de pondérer leurs mots au moyen d'une interface adaptée. Ce cas de figure ne nous a pas paru pratique par rapport à l'utilisateur, et notamment lorsque son texte est long. Par conséquent, la taille réduite de notre corpus était pénalisante par rapport aux modèles vectoriels implémentés, et la prise en compte des jugements des experts par rapport aux poids des termes aurait pu améliorer fortement les résultats de notre modèle asymétrique fondé sur la théorie du signal. En effet, ceci aurait certainement aidé à mettre plus en valeur l'intérêt de notre contribution dans le cadre de la symétrisation du modèle flou.

2. La non prise en compte de la négation, comme nous en avons discuté à la fin du chapitre 5, a altéré nos résultats. En effet, dans le cadre de cette étude, nous n'avons pas pris en compte l'analyse de la négation par rapport à ses formes d'expression multiples. Nous avons été obligés d'intégrer l'ensemble des adverbes de négation dans la stop-liste. Par conséquent, nous retenons le concept autour duquel l'auteur de la plainte a exprimé la non-existence à l'aide de la forme de négation automatiquement inhibée à la suite de la phase de lemmatisation. Par exemple, l'expression « je n'ai pas de moisissures » est retenu de la même manière que l'affirmation « j'ai des moisissures », sous la forme filtrée lemmatisée « avoir moisissure ». Par conséquent, la réalité des plaintes modélisées ainsi par les systèmes automatiques s'éloigne de la lecture faite par les experts, ce qui fait baisser les performances de nos systèmes de recherche.

3. Ces deux aspects des systèmes sont encore plus pénalisants dans les modèles sémantiques que dans les modèles directs. En effet, l'intégration de la sémantique amplifie le sens des concepts visés par les textes. Dans le cas où un terme exprimé, suite à une négation, est retenu sous sa forme positive, l'utilisation de la sémantique va ramener l'ensemble des termes proches de la forme inverse exprimée à la base. Par conséquent, cette augmentation rapproche les documents jugés par les experts à l'opposé du sens initial. Ainsi, dans le cas de la négation, l'utilisation de la sémantique est défavorable à la performance des systèmes. Concernant la non prise en compte des poids des termes, l'utilisation de la sémantique

peut aussi être dommageable. En effet, l'application de la sémantique autour des mots clés retenus et non très pertinents pour le processus d'appariement, augmente le bruit. L'utilisation de la sémantique sur les termes dont on souhaite baisser l'influence par procédés de pondération, provoque l'effet inverse. En effet, la sémantique ramène l'ensemble des termes proches des termes-bruit suscitant ainsi des rapprochements sur la base de termes impertinents.

4. Par ailleurs, nous avons fixé un seuil ad-hoc pour l'enrichissement sémantique. Une valeur de seuil de 10% a été retenue pour intégrer les termes sémantiquement proches de ceux existant directement dans les textes formalisés. En effet, la valeur de ce seuil est un élément déterminant qui reste à évaluer. En consultant les mesures d'appariement entre termes à partir de DICTIONNAIRE, certaines valeurs nous ont paru faibles alors que dans la pratique (dans notre domaine notamment) les deux termes en question sont très souvent confondus, cependant pour certaines autres vedettes nous avons établi le constat inverse. En effet, le rapprochement sémantique fondé sur le nombre des synonymes communs entre les vedettes dépend de la richesse du vocabulaire et des variations du sens existant autour des vedettes en question. Par conséquent, le seuil fixe que nous appliquons peut être soit trop bas soit trop haut. Dans le cas où le seuil est trop bas, donc trop permissif, l'augmentation sémantique peut amener à retourner des documents qui comportent des termes assez éloignés de ceux recherchés. Et dans le cas où le seuil est trop haut, l'exploration sémantique peut être incomplète. De même, pour l'appariement sémantique entre termes générés du même code par l'heuristique expliquée dans la section 5.9, et dont nous fixons la valeur à 50%, il est nécessaire de réaliser des tests au moyen de valeurs différentes.

5. À cause de l'ambiguïté de la langue véhiculée par DICTIONNAIRE, dans certains cas, l'augmentation sémantique peut occasionner inévitablement un sens erroné pour les textes modélisés. Par exemple, les rapprochements entre le terme « porte » à la vedette « introduction », ou le terme « problème » de l'expression « problème de santé » au mot « histoire » provoquent une modification du contexte de conversation. En effet, à partir de la série de tests que nous avons réalisés, et en compagnie d'un expert du Sami de Liège, nous avons constaté ces deux exemples comme étant essentiellement à l'origine de rapprochements sémantiques non-pertinents par rapport à l'application de la version directe pour les modèles du signal et les modèles flous[3]. Dans le cas où la sémantique participe sous cette forme désavantageuse, le classement des appariements automatiques s'éloigne plus des avis des spécialistes.

[3]Pour l'ensemble des versions symétriques et asymétriques.

6.4 Évaluation de la segmentation thématique automatique

Comme nous l'avons cité dans la présentation de notre approche, nous souhaitons étudier la possibilité de construire une base de scénarios pour rendre effective les assignations de solutions aux nouvelles plaintes (section 5.10). Pour ce faire, nous réalisons une classification des textes du corpus expérimental et représentatif.

La réalisation de l'étiquetage thématique des classes des textes par les experts, directement à partir des différentes classifications réalisées, nous a paru périlleuse. En effet, nous avons établi une multitude de catégorisations. D'une part, pour l'ensemble des systèmes de recherche implémentés nous avons réalisé des catégorisations. D'autre part, les algorithmes des nuées dynamiques demandent de choisir, a priori, un nombre de classes. Et pour élargir le champ de notre évaluation, nous avons étudié les partitions produites pour un nombre allant de 3 à 8 classes. Par conséquent, nous avons obtenus un grand nombre de groupes de numéros de plaintes correspondant à des catégorisations différentes mettant en évidence l'existence d'éléments hétérogènes, d'un point de vue des experts, au sein de certaines classes à interpréter.

Ainsi, à défaut de mettre des étiquettes à des classes difficilement compréhensibles par les expert, et pour évaluer la qualité des scénarios composés automatiquement, nous les comparons aux catégorisations des experts. Trois experts du CSTB ont regroupé, dans un nombre de classes de leur choix, les éléments du corpus de tests selon les thématiques qu'ils constatent. Pour comparer les résultats des différentes catégorisations, nous utilisons l'indice de Rand-corrigé [47, 64].

6.4.1 L'indice de Rand pour la comparaison des partitions

Pour comparer les partitions automatiques aux partitions des experts, nous avons utilisé l'indice de Rand. Pour comparer deux partitions, $Part_1$ et $Part_2$ d'un même ensemble de données, nous avons besoin de déterminer les facteurs suivants :

- a= le nombre de paires dans une même classe dans $Part_1$ et dans $Part_2$
- b= le nombre de paires séparées dans $Part_1$ et séparées dans $Part_2$
- c= le nombre de paires séparées dans $Part_1$ et ensemble dans $Part_2$
- d= le nombre de paires ensemble dans $Part_1$ et séparées dans $Part_2$

Dans le cadre de deux partitions, avec un même nombre de classes, l'indice symétrique brut de Rand est calculé en fonction de l'ensemble des accords, a+b, comme l'indique la formule 6.8.

$$Rand = \frac{a+b}{n^2} \tag{6.8}$$

Sachant que n désigne le nombre d'individus à classer. Lorsque la partition $Part_1$ a plus de classes que la partition $Part_2$, l'indice asymétrique brut de Rand est calculé par la formule 6.9. En effet, dans ce cas, on considère que deux éléments qui ne sont pas classés dans une même

classe dans $Part_1$ peuvent l'être dans $Part_2$. Par conséquent, le nombre des accords devient a+b+c. Nous utilisons la notion d'« asymétrie » par rapport aux nombres de classes qui diffèrent entres les deux partitions à comparer.

$$Rand = \frac{a+b+c}{n^2} \tag{6.9}$$

6.4.2 L'indice de Rand-corrigé

En appliquant les formules de Rand, présentées dans la section précédente, entre une partition experte et une partition aléatoire, la valeur de l'indice défini dans les formules précédentes, n'est pas nulle comme on pourrait s'y attendre. L'indice de Rand-corrigé défini par Hubert [64] réduit la mesure brute de Rand. Il évalue d'abord le Randespéré. Ce paramètre correspond à l'indice brut de Rand entre la partition experte de référence et une partition aléatoire respectant les critères de la partition à évaluer (le nombre de ses classes notamment)(formule 6.10).

$$Rand - corrigé = \frac{Rand - Randespéré}{1 - Randespéré} \tag{6.10}$$

Cet indice prend la valeur 0 lorsque la valeur du Rand-brut (Rand simple présenté précédemment) entre la partition experte et la partition évaluée est équivalente au Rand-brut entre la partition de référence et la partition aléatoire. Et inversement, l'indice du Rand-corrigé d'une partition par rapport à la classification experte reste égale à 1 lorsque la valeur brute de l'indice de Rand est de valeur 1 par rapport à la classification de référence.

6.4.3 Évaluation des scénarios automatiques par comparaison aux scénarios des experts

À partir d'un même corpus de plaintes, nous avons laissé la liberté aux experts de choisir le nombre de leurs classes. Le premier expert, spécialiste en chimie, a extrait six classes, le deuxième expert, spécialiste en micro-organismes, a mis en évidence l'existence de 7 classes, le dernier expert, ingénieur en ventilation, a fait ressortir 8 classes de plaintes à partir du même échantillon des 100 plaintes. Dans le tableau 6.3, nous montrons les étiquettes attribuées par chaque expert à chaque classe qu'il a extrait. Dans le tableau 6.4, nous exposons les résultats des comparaisons entre les partitions automatiques établies à l'aide des modèles de recherche utilisés et les partitions des 3 experts. Nous avons utilisé l'indice de Rand-corrigé (symétrique et asymétrique selon le nombre des classes) pour évaluer les différences entre les catégorisations des plaintes. Dans le tableau 6.4, nous présentons pour chaque expert et pour chaque modèle l'indice de Rand le plus élevé par rapport aux différentes classifications réalisées (différentes en fonction du nombre des classes pris en compte). Nous avons constaté que, pour un modèle donné, l'indice de Rand atteint son maximum par rapport aux différentes catégorisations des experts pour une même partition automatique (avec un même nombre de classes). Le tableau 6.4 indique des pourcentages d'accord entre les partitions automatiques d'un côté et les partitions des experts d'un autre coté. Pour évaluer concrètement ces valeurs, nous avons besoin des résultats d'une comparaison de référence. Pour cela, nous avons appliqué l'indice de Rand-corrigé pour confronter les jugements

	Expert N°1	Expert N°2	Expert N°3
Nbr de classes	6	7	8
8*Étiquettes	Moisissures	Moisissures	Moisissures
	Fibres	Fibres	Fibres
	Contamination chimique	COV	COV
	Moisissures et/ou acariens	Moisissures et acariens	Moisissures et/ou acariens
	Moisissures et contamination chimique	COV et autres facteurs[5]	Multipolluants[4]
	Sans cause apparente	Cause non-identifiée	Cause indéterminée
		Moisissures autres facteurs[6]	Polluants chimiques
			COV et fibres

TAB. 6.3 – Partitions établies par les experts à partir du corpus expérimental de 100 plaintes

	Vectoriel	Vectoriel-Sém	Vect-Binaire	Vect-Binaire-Sém
	5	4	3	4
Expert N°1	0,1925	0,2962	0,3426	0.0631
Expert N°2	0,2412	0,2938	0,3953	0.1382
Expert N°3	0,2457	0,2948	**0,4539**	0.1563
	Flou-Sym	Flou-NonSym	Flou-Sém-Sym	Flou-Sém-NonSym
	3	4	3	3
Expert N°1	0,1676	0,3289	0,0997	0.3831
Expert N°2	0,2887	0,3705	0,1970	**0.4325**
Expert N°3	0,2924	0,3634	0,1976	**0.4366**
	Signal-Sym	Signal-NonSym	Signal-Sém-Sym	Signal-Sém-NonSym
	5	5	4	5
Expert N°1	**0,4414**	0,3701	0,3215	**0.9140**
Expert N°2	**0,4642**	0,3949	**0,4007**	**0.9264**
Expert N°3	**0,4925**	**0,4206**	**0,4396**	**0.9328**

TAB. 6.4 – Évaluation des partitions automatiques par rapport aux partitions des experts par application de l'indice de Rand-corrigé

	Expert N°1	Expert N°2	Expert N°3
Expert N°1	1	0.5927	0.6591
Expert N°2		1	0.7717

TAB. 6.5 – Comparaison entre les partitions des experts

des experts entre eux (tableau 6.5). Les résultats des confrontations des classements des experts sont asymétriques étant donné que la valeur du Rand-corrigé dépend du taux d'accord entre la partition de référence et une partition aléatoire respectant les critères de la deuxième partition à évaluer. Par conséquent, dans le tableau 6.5, nous affichons les taux d'évaluation les plus élevés uniquement. En comparant les deux tableaux (6.4 et 6.5), les comparaisons entre l'ensemble des partitions de référence et les partitions automatiques construites à l'aide du modèle du signal sémantique dans sa version asymétrique donnent des indices de correspondance très élevés. En effet, en comparant ces derniers taux aux niveaux des accords entre les partitions expertes entre elles, nos résultats sont très encourageants. On constate une nette domination de notre modèle du signal sémantique non symétrisé dans le cadre de notre application avec une catégorisation à 5 classes à un taux dépassant les 90% par rapport à l'ensemble des jugements des experts. Au niveau global, les modèles directs et sémantiques inspirés du principe des interférences des signaux présentent des valeurs de correspondance relativement intéressantes autour des catégorisations à 4 et 5 classes. Ces résultats témoignent de l'existence d'une bonne approximation entre les partitions automatiques et 5 scénarios observés par les experts.

6.4.4 Mise en correspondance entre les scénarios automatiques et les partitions des experts

L'expérience réalisée par les experts n'est pas issue des mêmes conditions que les partitions automatiques. En analysant les partitions de l'expert N°1 ayant communiqué le nombre de scénarios le plus proche de la taille des partitions automatiques les plus pertinentes (tableau 6.3), nous constatons qu'il est question à la base de 5 classes et que la sixième catégorie regroupe les plaintes pour lesquelles la décision sur la nature du problème n'a notifié aucun motif apparent. Les plaintes jugées « sans cause apparente », dans leur majorité, sont décrites à l'exemple de toute plainte convenable au traitement et à la prise en compte, néanmoins nos experts les ont situées dans cette classe « de rejet » en s'inspirant des rapports qui accompagnaient les plaintes et qui révélaient la non identification du malaise à partir des séries des prélèvements effectués sur site. Par conséquent, nos systèmes de recherche fondés sur les mots clés des parties problèmes (des plaintes) ne peuvent pas relever cette différence.

Toujours à partir du tableau 6.3, l'expert N°2 a mis en évidence l'existence de la classe « COV ». Cette dernière est dérivée de la classe « contamination chimique » de l'expert N°1. En effet, comme nous l'avons spécifié dans le cadre du chapitre 1, les COV sont des facteurs de risque d'origine chimique. Les deux classes « COV et autres facteurs » et « moisissures et autres facteurs » sont indiquées par l'expert N°1 sous l'appellation sommaire « moisissures et contamination

chimique ». En effet, nous avons constaté à partir de l'ensemble des plaintes classées dans ces 2 catégories, qu'il était beaucoup plus question de contamination due aux moisissures et aux composés organiques volatils (formaldéhydes notamment) que de fibres. En effet, les passages succincts de ces plaintes concernent les particules de poussières plutôt que les fibres minérales provenant des équipements d'isolation. Par conséquent, et grâce à la contribution discriminante de l'expert N°1, nous pouvons dire qu'en dehors de la classe « de rejet » il est également question de 5 scénarios dans la partition de l'expert N°2.

De même, les plaintes de la catégorie « COV et fibres » désignée par l'expert N°3 comprennent des extraits très brefs exprimant l'existence de poussières sans préciser leurs éventuelles sources de provenance. La classe dénotée « multipolluants » par le troisième expert, regroupe l'ensemble des plaintes traitant de polluants d'origine chimique et de problèmes dus aux moisissures et aux poussières. De ce fait, nous pouvons affirmer l'existence au sein de notre corpus représentatif des 5 scénarios indiqués par l'expert N°1, en l'occurrence, les scénarios suivants : « Moisissures », « Fibres », « Contamination chimique », « Moisissures et acariens », « Moisissures et contamination chimique ».

6.4.5 Analyse des thématiques abordées dans le corpus des tests

Le scénario « Moisissures et contamination chimique » correspond à des situations complexes. Il regroupe l'ensemble des plaintes, où il est question de problèmes d'origines diverses, en l'occurrence chimique et micro-biologique. Cependant, le scénario « Moisissures et acariens » est un cas de figure issu d'un certain nombre de conditions communes, dont notamment l'humidité. En effet, l'humidité élevée favorise la croissance des moisissures et des acariens. Par ailleurs, les plaintes de notre corpus faisant état de situations de malaises liés à la présence des fibres ne concernent pas parallèlement d'autres typologies de pollution domestique. Dans notre corpus, que nous avons souhaité représentatif du corpus initial des 655 documents, le scénario « Moisissures » est le plus répandu et celui traitant de la « Contamination chimique » concerne essentiellement les problèmes de santé liés aux COV et plus particulièrement aux formaldéhydes. Nous ne distinguons pas de scénario d'origine physique (en lien avec les facteurs de risques physiques cités dans la section 1.3.3) pour la simple raison que les organismes que nous avons contactés dans le cadre de cette étude, ne traitent pas (ou pas encore pour certains) de cette catégorie de pollution domestique.

Par conséquent, nous insistons sur le fait que les scénarios mis en évidence dans le cadre de cette étude sont issus du domaine de la pollution de l'air au sein des lieux de vie et témoignent d'une exhaustivité relative. En effet, la base de plaintes initiale provenant des organismes contactés et dont est issu le corpus représentatif ne peut prétendre réunir tous les cas possibles issus des différentes conditions du bâtiment pouvant être à l'origine de problèmes sanitaires. Le domaine est récent, les organismes en charge de traiter les plaintes des particuliers s'intéresseront probablement, dans peu, à de nouveaux facteurs de risque.

	Direct		Sémantique	
Le vectoriel de Zargayouna	81,93%		79,52%	
Classement	5		7	
Le vectoriel binaire	83,13%		78,31%	
Classement	4		8	
Le flou	NonSym	Sym	NonSym	Sym
	81,48%	87,95%	83,13%	89,16%
Classement	6	2	4	1
Le signal	NonSym	Sym	NonSym	Sym
	86,75%	75,90%	87,95%	74,70%
Classement	3	9	2	10

TAB. 6.6 – Taux de réussite des assignations de solutions au moyen des modèles automatiques

6.5 Évaluation de l'assignation automatique

Évaluer notre système revient à évaluer le procédé des assignations automatiques de solutions à des plaintes écrites. Pour cela, nous avons sélectionné 96 nouvelles plaintes du corpus initial des 655 documents. Nous avons fait en sorte que ces nouveaux documents ne fassent pas partie de l'échantillon de référence des 100 éléments étiquetés par les experts. Pour l'ensemble des modèles de recherche développés, nous évaluons le niveau de réussite des assignations effectuées. L'assignation de la solution correspond à l'affectation de la plainte courante à un des scénarios déterminés à partir de l'échantillon expérimental des 100 plaintes. Cet exercice d'assignation des solutions a été réalisé parallèlement par un groupe de trois experts du CSTB. Pour évaluer le niveau d'accord entre les attributions automatiques réalisées par notre applicatif et les assignations des spécialistes, nous calculons le pourcentage des attributions des plaintes aux bons scénarios. Nous estimons qu'une assignation automatique est valide si au moins un des trois experts a mentionné l'appartenance de la plainte en question au scénario indiqué. Le tableau 6.6 expose les pourcentages de réussite des assignations établies par l'ensemble des modèles développés (directs et sémantiques) par rapport aux avis de référence. Il est bien sûr nécessaire d'avoir des taux d'assignation de référence pour évaluer concrètement les affectations automatiques de solution. Pour ce faire, nous nous basons sur les désaccords entre les avis des experts qui ont participé à l'exercice des assignations des 96 nouvelles plaintes aux scénarios déterminés au sein de l'échantillon des 100 plaintes.

Le tableau 6.7 indique les pourcentages des accords concernant les affectations des 96 plaintes aux scénarios entre les différents avis-experts. Ce tableau exprime le fait qu'il n'est pas question d'unanimité des avis sur l'ensemble des 96 plaintes testées, mais que les taux des accords entre les experts considérés deux à deux vont de 88,54% à 88,75%. En comparant les taux de réussite des assignations automatiques (tableau 6.6) aux résultats du tableau 6.7, et à partir de nos données, nos résultats semblent globalement (toutes les méthodes) favorables à l'automatisation

	Expert N°1	Expert N°2	Expert N°3
Expert N°1	0	88,54%	88,75%
Expert N°2	88,54%	0	88,54%
Expert N°3	88,75%	88,54%	0

TAB. 6.7 – Pourcentage des accords entre les experts concernant les affectations des plaintes aux scénarios

des réalisations des solutions aux plaintes écrites.

Le classement des systèmes de recherche présenté dans le tableau 6.6 est différent de celui du tableau 6.1. Cette différence est due essentiellement aux données. Le classement exposé par le tableau 6.6 repose sur des données différentes de celles utilisées pour obtenir les résultats du tableau 6.1. En effet, le classement fondé sur les premiers niveaux de pertinence utilise des requêtes extraites du corpus des 100 plaintes. Par contre, le classement des modèles du tableau 6.6 est fondé sur les résultats des assignations réalisées pour des plaintes nouvelles différentes du contenu étiqueté par les experts. En effet, utiliser une donnée existante dans l'ensemble labellisé donnera lieu immanquablement à une assignation réussie étant donné qu'il est question dans ce cas d'un appariement automatique parfait entre les éléments identiques.

À partir du classement des modèles selon les résultats de leurs assignations nous constatons que les modèles fondés sur le principe de la densité des termes de la requête au sein des documents se situent en tête du classement avec le modèle flou symétrisé en première position. Une fois de plus, nous constatons l'intérêt de notre contribution dans le cadre de la symétrisation du modèle flou. Dans la suite du classement, nous trouvons d'abord la version asymétrique de notre modèle du signal augmenté sémantiquement suivie de sa version non-sémantique. À l'exemple de notre analyse des systèmes de recherche au moyen des courbes rappel-précision, et après nos observations dans le cadre des partitions automatiques évaluées au moyen de l'indice de Rand-corrigé, une fois de plus, les résultats du tableau 6.6 témoignent de la convenance de l'approche asymétrique de notre modèle inspiré du principe des interférences des signaux pour l'évaluation de la pertinence.

6.6 Analyse statistique de la dépendance des assignations de la taille des documents

Comme nous l'avons déjà exprimé dans ce mémoire, nous avons pris connaissance à partir de la littérature [114], que les résultats des systèmes de recherche peuvent être liés à la taille des documents traités. Pour analyser l'effet de cet aspect, nous tenons compte des résultats des assignations réalisées au moyen des modèles développés. Nous examinons la dépendance entre la taille moyenne des rubriques de l'ensemble des 96 requêtes *XMLisées* employées pour les assignations et le succès ou l'échec de l'affectation au bon scénario.

L'analyse de la variance ou ANOVA (pour ANalysis Of VAriance)[7], est une méthode statistique permettant de comparer l'espérance mathématique de deux échantillons ou plus. Elle est appliquée pour savoir si une variable[8] appelée variable à expliquer, est en relation avec une variable[9] explicative. Dans le cadre de notre analyse, la variable à expliquer est « le résultat de l'assignation », qui peut être en l'occurrence un succès ou un échec, et la variable explicative désigne la taille moyenne des rubriques renseignées des 96 requêtes XML.

Ainsi, pour déterminer à quel point la taille est une sorte de cause qui pilote la performance des réponses de notre applicatif implémentant un système de recherche donné, nous réalisons pour le modèle en question un fichier sous forme d'un tableau de correspondance entre la taille moyenne de la requête appliquée (des entiers) et le niveau de réussite de l'affectation au scénario approprié (1 pour succès et 2 pour échec). Nous importons ce fichier de données dans l'outil Tanagra[10] qui est utilisé dans le cadre de l'analyse des variances.

Le nom ANOVA de la méthode s'explique par sa façon de procéder. En effet, l'approche consiste à décomposer la variance totale de l'échantillon en deux variances partielles, la variance inter-classes et la variance résiduelle, ensuite on compare ces deux variances. Pour définir le principe de l'ANOVA, nous désignons par p le nombre de groupes d'observations. Pour chaque groupe, k, des observations ($X_{k,1}$,...,$X_{k,nk}$) nous notons l'espérance mathématique μ_k. Le nombre total des valeurs observées est de N. ANOVA teste l'hypothèse H_0 suivante : « les espérances des p groupes considérés sont égales ». Le test de cette hypothèse se déroule comme suit :

1. Pour chaque groupe la moyenne empirique est calculée :
$$m_k = \frac{x_{k,1} + x_{k,2} + ... + x_{k,n_k}}{n_k}$$
2. On calcule la moyenne empirique totale de l'échantillon :
$$M = \frac{n_1 m_1 + n_2 m_2 + ... + n_p m_p}{N}$$
3. On calcule la variance empirique de chaque groupe :
$$V_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{k,i} - m_k)^2$$
4. On calcule la moyenne des variances (variance intra-classes) :
$$V_{intra} = \sum_{k=1}^{p} \frac{n_k}{N} V_k$$

[7] En français, on parle parfois d'Anavar pour analyse de la variance. Ce terme n'est pas courant.

[8] ou plusieurs variables dépendantes.

[9] ou plusieurs autres variables.

[10] TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

	BSS	WSS	Fisher's F	P-value	Classement
VB[15]	2,4791	6428,6567	0,030465	0,861886	$3^{ème}$
VZ[16]	1,4843	6429,6515	0,018237	0,89292	$2^{ème}$
OUSym[17]	0,3809	6430,7549	0,004679	0,945638	1^{er}
OUNONSym[18]	8,6116	6422,5242	0,105926	0,745691	$4^{ème}$
SSym[19]	47,1694	6383,9664	0,58371	0,447138	$6^{ème}$
SNONSym[20]	24,3358	6406,8	0,300076	0,585379	$5^{ème}$

TAB. 6.8 – Tableau d'analyse de la variance

5. On calcule la variance des moyennes (variance inter-classes) :

$$V_{inter} = \sum_{k=1}^{p} \frac{n_k}{N} \ (m_k - M)^2$$

6. Enfin, la variable de test est calculée comme suit :

$$F_{p-1,N-p} = \frac{V_{inter}/p-1}{V_{intra}/N-p}$$

La valeur de F, appelée Fisher's F sous Tanagra, est une variable de test qui est comparée à une valeur critique. Dans le cas où la F-mesure est supérieure à la valeur seuil, l'hypothèse d'égalité initiale est rejetée. En effet, plus le rapport entre la variance inter-classes et la variance intra-classes est élevé plus les groupes sont différents, et par conséquent l'hypothèse portant sur l'égalité des espérances n'est pas vérifiée. Le tableau 6.8 présente les valeurs des inerties [11] inter-groupes[12], les valeurs des inerties intra-groupes [13] et la mesure F de Fisher[14]. Ces évaluations statistiques sont calculées à l'aide du logiciel Tanagra à partir des résultats des assignations liées à la taille des requêtes et réalisées au moyen de l'ensemble des modèles de recherche utilisés. De même, mais dans un processus de raisonnement opposé à celui de la F-mesure, la p-value (ou probabilité critique) renvoie la probabilité de vérification de l'hypothèse d'égalité. Il convient donc de comparer cette probabilité avec le seuil qu'on se fixe. Traditionnellement, les applications d'ANOVA fixent un seuil allant de 1% à 5%. Dans le cas où la p-value est inférieure à ce seuil alors logiquement on rejette l'hypothèse d'égalité. Et dans le cas contraire, où la p-value est supérieur au seuil fixé, on ne peut rejeter H_0. Les p-value du tableau témoignent de probabilités fortes d'égalité des espérances des deux groupes d'assignation établis par l'ensemble des modèles. Nous pouvons modifier le seuil selon notre degré d'exigence. Cependant, les valeurs des probabilités critiques restent malgré tout significativement élevées. Ainsi, et selon l'échantillon à partir duquel nous avons réalisé notre analyse, les résultats des assignations sont indépendants de la taille des requêtes.

[11] En anglais et sous Tanagra l'inertie correspond à « Sum of square » ou SS.
[12] Between-group Sum of Squares ou BSS.
[13] Within-group Sum of Squares ou (WSS).
[14] Fisher's F

Pour schématiser ce constat et visualiser le niveau d'influence de la taille des documents sur les résultats des assignations pour chaque modèle, nous avons utilisé les diagrammes de Tukey (ou boîtes à moustaches). Ces diagrammes sont un moyen d'observation des populations clair et puissant. En résumé, pour expliquer ce mode de représentation, la boîte à moustaches utilise cinq valeurs qui résument les données, en l'occurrence : l'extrême minimum (le cercle inférieur), les trois quartiles Q_1 (trait inférieur de la boite), Q_2 (médiane de la boîte représentée par le trait horizontal au centre des boîtes), Q_3 (trait supérieur de la boîte), l'extrême maximum (le cercle supérieur) et les deux « moustaches ». Les moustaches, inférieure et supérieure, sont représentées ici par les petits rectangles verticaux de part et d'autre des boîtes. Elles délimitent les valeurs dites adjacentes[21] qui sont déterminées à partir de l'écart inter-quartile $(Q_3 - Q_1)$ [49].

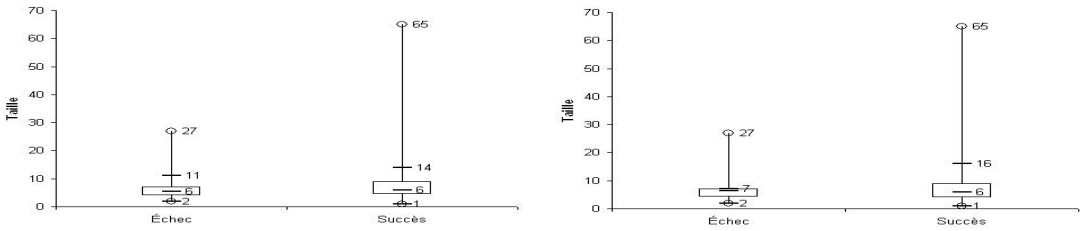
Les diagrammes de la figure 6.9 illustrent la répartition des tailles des requêtes autour de leurs médianes par rapport aux 2 groupes des résultats des assignations, en l'occurrence échecs et réussites, réalisées par les modèles appliqués. Nous constatons en premier, que la totalité des diagrammes montrent une ressemblance significative des médianes des deux groupes opposés[22]. Les valeurs adjacentes entre les deux groupes pour les différents modèles sont similaires, faisant ainsi en sorte que les boîtes paraissent de tailles quasiment identiques dans les différents graphiques. Par ailleurs, pour l'ensemble des systèmes de recherche, leur succès a concerné des plaintes de tailles extrêmes (la plus petite taille est égale 1 et la plus grande est de 65 termes). Ces valeurs sont extrêmes et non représentatives, il est donc important de se baser sur les valeurs médianes (entre les deux moustaches) pour comparer les deux échantillons.

Ainsi, et à l'exemple des résultats que nous avons obtenus au moyen du logiciel Tanagra, l'efficacité des modèles de recherche que nous avons définis et/ou utilisés ne semble pas dépendre de la taille des textes. Nous réalisons cette constatation suite à nos expériences menées dans le cadre d'un corpus spécifique. En effet, la taille moyenne la plus élevée des rubriques des plaintes est de 65 termes. Il est tout à fait possible qu'une application des modèles développés pour l'appariement de textes beaucoup plus longs et dans le cadre de corpus plus consistants puisse trouver un lien de causalité entre la taille des documents et l'efficacité de ces mêmes modèles.

Nous n'avons pas exploité les modèles sémantiques dans cette étude, étant donné qu'on souhaitait vérifier le niveau de dépendance entre la taille des documents et l'efficacité des systèmes par rapport à leur principe de base et non par rapport à l'intégration ou la non prise en compte de la sémantique. En effet, nous avons souhaité vérifier les résultats de nos modèles vis à vis de

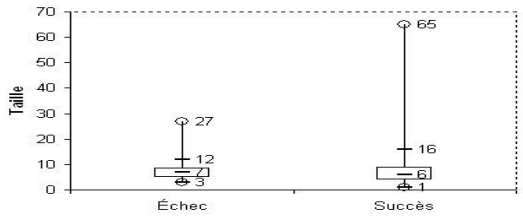
[21] La valeur de la frontière basse est de : $Q_1 - 1,5 \times (Q_3 - Q1)$ et la valeur de la frontière hausse est de : $Q_3 + 1,5 \times (Q_3 - Q_1)$.

[22] La médiane des distributions n'est pas toujours centrée dans la boîte. En effet, plus les grandeurs entre les deux moustaches sont asymétriques plus la médiane s'écarte de la moyenne. Lorsque la distribution est plus allongée vers les grandes valeurs, la médiane est inférieure à la moyenne. Lorsque la distribution est plus allongée vers les petites valeurs, la médiane est supérieure à la moyenne.

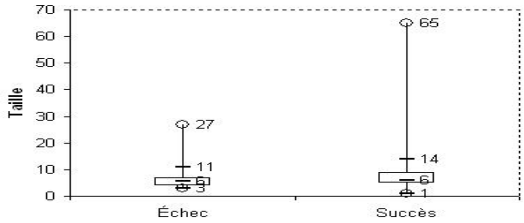


1- Modèle vectoriel booléen

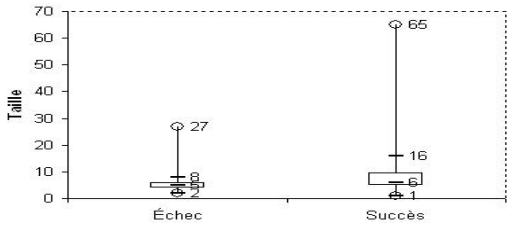
2-Modèle vectoriel de Zargayouna



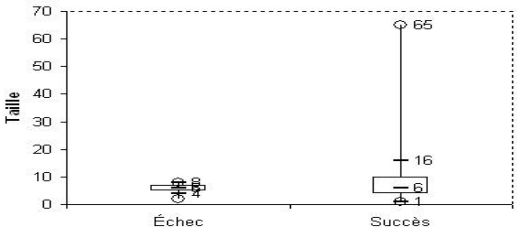
3- Modèle flou asymétrique



4- Modèle flou symétrique



3- Modèle signal asymétrique



4- Modèle signal symétrique

TAB. 6.9 – Diagrammes de Tuckey témoignant du lien de causalité entre la taille moyenne des rubriques des requêtes et les résultats des assignations établies par l'ensemble des modèles appliqués

la taille des requêtes par rapport aux déclarations de la littérature (qui ne tient pas compte de la sémantique). Rappelons que nous avons analysé la dépendance entre le résultat de l'assignation et la taille moyenne des rubriques de la plainte-requête. En effet, nous ne tenons pas compte de la taille des éléments du corpus étant donné que le but de notre travail est d'avoir un maximum de plaintes libres (d'un point de vue taille également) résolues en mémoire afin de se rapprocher encore plus de l'exhaustivité, et se restreindre à certaines tailles serait défavorable à notre applicatif. Ainsi, en respectant cette condition, nous avons tenu compte de la taille des requêtes seulement, pour la simple raison que lorsqu'une nouvelle plainte se présente au traitement, sa taille est connue (et unique). Par conséquent, nous pouvons choisir le système le plus approprié à l'assignation dans le cas où un lien de causalité entre la taille et l'adaptabilité des modèles aurait pu être prouvé. Par ailleurs, pour les modèles orientés requêtes, en l'occurrence le modèle flou asymétrique et le modèle du signal asymétrique, il n'y a que la taille de la requête qui importe. En effet, le principe de densité est traditionnellement utilisé dans le cadre des requêtes courtes sans se soucier de la taille des documents en mémoire archive. Il a été néanmoins démontré à travers notre corpus, que l'efficacité des modèles de proximité asymétrique, à l'exemple des modèles vectoriels, ne dépendaient pas de la taille des requêtes.

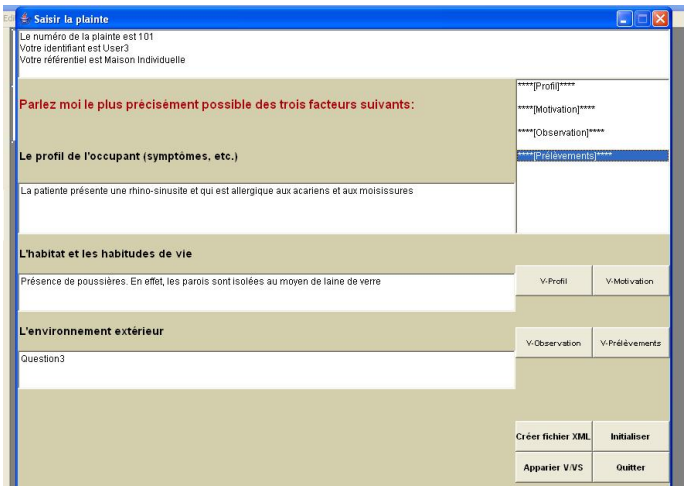


FIG. 6.6 – Interface graphique de saisie des plaintes à traiter

6.7 Démonstration de l'applicatif

À partir de l'interface usager de la figure 6.6, la plainte courante est appariée aux plaintes *Xmlisées* en mémoire au moyen de l'ensemble des modèles appliqués. Nous sauvegardons les résultats des appariements dans des fichiers au format Comma-separated values (CSV) en mémoire. Nous avons désactivé les différentes interfaces dédiées à l'affichage des résultats des différents modèles que nous avons développées au début de ce travail. Pour des problèmes de visibilité, notamment par rapport aux dimensions des interfaces, à la taille du corpus, et pour assurer une pérennité des données (notamment pour calculer les taux de rappel et de précision ainsi que pour la création des scénarios), nous avons dédié un répertoire spécifique à chaque modèle. Pour chaque modèle et à chacune de ses versions nous avons créé un répertoire pour la sauvegarde des résultats des processus de recherche initiés par l'ensemble des plaintes-requêtes de nos corpus des tests.

6.8 Conclusion

Dans ce chapitre nous avons présenté les résultats des expérimentations menées par notre applicatif à partir d'un corpus de plaintes issues de trois organismes différents. Nos résultats ont mis en évidence un intérêt important dans l'intégration du cadre sémantique que nous avons défini pour certains modèles de recherche implémentés. En plus de l'analyse que nous avons apportée dans la section 6.3, nous ajoutons que la taille réduite de notre corpus, ainsi que le peu de diversification des sources de provenance des plaintes, n'agit pas grandement sur l'hétérogénéité du vocabulaire nécessitant l'intégration de ressources sémantiques. Ainsi, l'apport de certains modèles de recherche augmentés sémantiquement n'a pu être plus mis en valeur. Nous faisons la même réflexion sur la qualité des scénarios obtenus par segmentations automatiques guidées par les systèmes de recherche sémantique.

Par ailleurs, beaucoup d'éléments restent à évaluer. Les résultats de nos modèles combinés au cadre sémantique que nous avons défini peuvent être plus probants si nous déterminons, au

moyen d'une série d'expérimentations, la valeur optimale du seuil de similarité sémantique. De plus, utiliser des coefficients adaptés aux classes d'information (les rubriques) permettrait un appariement plus en rapport avec le raisonnement expert.

Nous avons développé des systèmes qui en théorie sont plus appropriés à la modélisation des textes en fonction de leur taille. Considérant la taille des requêtes comme axe de projection des classes (réussite ou échec de l'assignation au bon scénario) nous avons constaté que les deux classes n'étaient pas bien séparées. Les tests de l'ANOVA affirment cette constatation, au moyen d'une probabilité critique relativement élevée témoignant d'une indépendance des résultats d'assignation obtenus par rapport à la taille moyenne des rubriques renseignées. Réaliser l'ensemble des expérimentations sus-citées dans un cadre non-structuré donnerait certainement des résultats différents. De plus, tenir compte de la taille globale de la requête peut être plus discriminant à la convenance des modèles.

Dans le cadre de cette thèse, nous avons étudié les niveaux de performance de l'ensemble des modèles présentés, mais au final un seul système devrait être appliqué pour réaliser l'affectation des solutions. Dans le cas où la taille de la requête aurait été un facteur déterminant à contrôler, un principe commutateur au sein du module fonctionnel devrait être établi pour déclencher le système adéquat à l'appariement pour l'assignation. Et au delà du choix dépendant de la taille des requêtes, les résultats des évaluations devraient être concluantes par rapport à l'utilisabilité de chaque modèle. Dans notre contexte précis, nous avons apporté la preuve, à partir d'un corpus spécifique des plaintes, que dans l'absolu le modèle que nous avons défini, en l'occurrence le modèle inspiré de la théorie du signal asymétrique direct, est le modèle le plus pertinent. Dans le cadre de la recherche d'information sémantique, ce même modèle couplé au cadre sémantique défini et respecté dans cette étude est également sensiblement le plus performant.

Conclusion

Dans ce travail, nous avons cherché à étudier le degré de faisabilité de l'approche automatique de résolution de plaintes spécifiques écrites en français et en langue naturelle. Ces plaintes révèlent des problèmes de santé dus à la qualité de l'air au sein des ouvrages d'habitation. Plus spécifiquement, nous avons tenté d'explorer les conséquences d'une hypothèse portant sur l'existence de scénarios de pollution témoignant de la régularité des motifs des plaintes du domaine de l'air dans les bâtiments. Une fois cette hypothèse vérifiée, au moyen d'un système de recherche d'information hébergeant les représentations associées aux expressions des plaintes résolues, une assignation automatique de solutions aux plaintes futures à traiter peut être envisagée.

Les réalisations de recherche automatique présentées dans ce mémoire se situent dans le contexte général de la recherche d'information et de l'ingénierie des connaissances. Les systèmes que nous avons appliqués, se positionnent plus particulièrement dans le cadre de la recherche d'information structurée (*Xmlisée*). À travers les applications de recherche employées, nous avons essayé de démontrer l'intérêt de l'intégration d'un cadre sémantique général pour assurer une représentation et une recherche plausible des plaintes. Le cadre sémantique que nous avons utilisé est caractérisé par un dictionnaire des synonymes de la langue française et d'une heuristique de racinisation des termes. En plus de la définition du cadre sémantique adapté à notre ressource documentaire, nous avons proposé une formule d'extension sémantique pour le modèle de recherche fondé sur le principe de la proximité floue des termes des requêtes au sein des documents. Nous avons également proposé un système de recherche tenant compte des signaux de présence générés par les termes des requêtes. Établi aussi sur le principe de la densité des textes, par rapport au modèle de proximité floue, notre système présente l'avantage (entre autres) essentiel de pouvoir s'appliquer à partir des formes filtrées lemmatisées des textes envisagées dans cette approche.

6.8.1 Limitations de notre travail

Le travail d'évaluation exposé dans le dernier chapitre de ce mémoire comporte une importante partie critique, dont le but était de souligner les limitations de nos ressources et des modèles implémentés sur le plan cognitif. En effet, nous avons confronté les résultats des différents modèles implémentés (directs et sémantiques), la qualité des scénarios résumés automatiquement, ainsi que les résultats des exécutions de notre applicatif (assignation des solutions) aux avis d'un groupe de spécialistes du domaine. Vis à vis de cet impératif, nous avons été contraints d'utiliser un échantillon de plaintes expérimentales relativement réduit. En plus de la taille du

corpus des tests, la variété négligeable des sources de provenance des plaintes a été en défaveur de la diversité du vocabulaire employé. En effet, cette particularité à mis un frein à la mise en évidence de l'intérêt de l'application de la sémantique dans le cadre de certains modèles appliqués à travers les différentes réalisations effectuées (appariement, classification et assignation). Par ailleurs, dans notre thèse et de manière plus générale, la non-représentation de la négation, des paramètres numériques, des données temporelles et spatiales est une entrave inévitable à la performance des modèles fondés sur des formalisations sommaires (filtrées) de la langue naturelle.

6.8.2 Perspectives

Ces critiques nous motivent à dresser, à ce niveau, un certain nombre de perspectives. En plus des perspectives réalisables à court terme, spécifiques aux évaluations, et citées dans la conclusion du chapitre 6.8, nous proposons ici des alternatives générales à notre approche et à ses ressources. Dans le chapitre 5 de ce mémoire, nous avons mentionné qu'il était impossible d'enfermer la totalité des termes ainsi que leurs relations sémantiques dans un vocabulaire (ou conceptualisation) fixe et permanent. Cependant, nous avons appris que très récemment, l'existence de la ressource WOLF[23]. En effet, l'application de cette initiative en tant que base sémantique hiérarchisée de la langue française permettrai probablement de mettre en évidence des liens sémantiques manquant par rapport aux niveaux de correspondance entre les termes évalués à partir de leurs configurations synonymiques

Notre système doit être reçu comme un outil d'aide à la décision pour mieux servir l'approche traditionnelle de la réponse aux plaintes. En effet, les résultats des assignations automatiques de solutions à des plaintes réelles écrites en langue naturelle sont très favorables à l'automatisation que nous proposons. Les résultats obtenus ont été validés et témoignent avec une relative justesse de l'adaptabilité de notre approche par rapport au processus classique de la réponse aux plaintes. Ainsi, à travers cette démarche, et par rapport aux nombres des plaintes qui restent sans réponse dans la pratique, les organismes en charge d'accueillir ces demandes peuvent pousser plus loin leur gestion. Par conséquent, ces organismes pourront plus cerner et mieux répondre aux problématiques des populations en lien avec la qualité de l'air intérieur.

L'application qui découle de notre thèse, doit être perçue comme une réalisation destinée à un raffinement continu. Cette précision est un avantage accordé par notre méthode permettant à notre système de tenir compte de la caractérisation de nouvelles pistes d'action (scénarios) dans le domaine de la qualité de l'air dans les logements. En effet, notre analyse a porté sur des situations de pollution domestique identifiées et ayant fait l'objet de plaintes et d'enquêtes. La prise en compte des futures interventions issues de conditions nouvelles non traitées par notre base de scénarios est essentielles à la pérennité de l'outil. Cette propriété prometteuse, consiste à intégrer au sein du corpus l'ensemble des plaintes de condition nouvelle organisées sous forme de scénarios (mini-bases) accompagnés de leur solutions génériques adaptées à chaque classe d'infor-

[23] Publication datée de Juin 2008.

mation hébergée. Ainsi, une intégration circonspecte des nouvelles conditions par un spécialiste du fonctionnement du système permettra de réaliser de nouvelles décisions pour de nouvelles conditions pour le mieux-être de la population.

Liste de nos publications

1. Zoulikha Bellia Heddadji, Nicole Vincent, Georges Stamon et Séverine Kirchner. Système d'aide à la décision pour la surveillance de la qualité de l'air intérieur. *In actes de Extraction et gestion des connaissances, EGC.* Lille, EGC 2006.

2. Zoulikha Bellia Heddadji, Nicole Vincent, Georges Stamon et Séverine Kirchner. Extension sémantique du modèle de similarité basé sur la proximité floue des termes. *In actes de Extraction et gestion des connaissances, EGC.* Namur, EGC 2007.

3. Zoulikha Bellia Heddadji, Nicole Vincent, Georges Stamon et Séverine Kirchner. Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. *In actes de Extraction et gestion des connaissances, EGC.* Sophia Antipolis, EGC 2008.

4. Zoulikha Bellia Heddadji, Nicole Vincent, Georges Stamon et Séverine Kirchner. Mise en évidence de la sémantique dans un système d'aide au diagnostic des plaintes écrites associées à des situations de pollution de l'air intérieur. *In actes Terminologie et Ontologie : Théories et applications, Toth.* Annecy, Toth 2008.

Bibliographie

[1] P. Blache A. Abeillé. *Ingénierie des langues.* Hermès.

[2] M. Ott A. Buchmann, F.De Blay. *La pollution intérieure des bâtiments : la connaître pour la prévenir.* WEKA, France, 2002.

[3] AFNOR. Lignes directrices pour le traitement des réclamations dans les organismes. Norme iso 10002, St Denis la plaine, France, Décembre 2004.

[4] JD Anderson and J Perez-Carballo. The nature of indexing : how humans and machines analyze messages and texts for retrieval. part i : Research, and the nature of human indexing. *Information Processing and Management*, 2 :231–254, 2001.

[5] V.N. Anh and A. Moffat. Compression and an ir approach to xml retrieval. In *Proceedings of INEX 2002 Workshop, Dagstuhl, Germany,2002*, 2002.

[6] Présidence Luxembouregoise au conseil de l'UE. Indoor air quality : Elaboration d'un standard européen. Technical report, Conférence de Luxembourg sur le plan d'action "Environnement et Santé" le 13-15 juin 2005, 2005.

[7] Hafedh El Ayech. Les réseaux de neurones artificiels pour la prévision du trafic aérien de passagers. Technical report, 2003.

[8] J. Kamps et M. de Rijke B. Sigurbjörnsson. Mixture models, overlap, and structural hints in xml element retrieval. pages 196–210, 2005.

[9] Djida Bahloul. *Une approche hybride de gestion des connaissances basée sur les ontologies : application aux incidents informatiques.* PhD thesis, Institut national des sciences appliquées de Lyon, Lyon, France, 2007.

[10] Michel Beigbeder and Annabelle Mercier. Application de la logique floue à un modèle de recherche d'information basé sur la proximité. In *Proceedings de la 12es rencontres francophones sur la Logique Floue et ses applications*, pages 231–237. CEPADUES, 2004.

[11] Michel Beigbeder and Annabelle Mercier. An information retrieval model using the fuzzy proximity degree of term occurences. In *SAC '05 : Proceedings of the 2005 ACM symposium on Applied computing*, pages 1018–1022, New York, NY, USA, 2005. ACM Press.

[12] Rik K. Belew. Adaptive information retrieval : Using a connectionist representation to retrieve and learn about documents. In *Proc. of SIGIR '89*, pages 11–20, Cambridge MA, 1989.

[13] Zoulikha Bellia. Modélisation d'un système informatique pour la gestion des demandes d'intervention dans le domaine des ambiances intérieures : Une approche basée sur le raisonnement à partir de cas, 2004.

[14] Catherine Berrut. Recherche d'informations. Technical report, Cours pour Master 2 Recherche en Système d'Informations. Université Joseph Fourier. Grenoble., 2006.

[15] Romaric Besançon and Martin Rajman. Validation de la notion de similarité textuelle dans un cadre multilingue. In *Proceedings of 6èmes Journées internationales d'Analyse statistique des Données Textuelles, JADT 2002*, 2002.

[16] Gilles Bisson. Une approche symbolique/numérique de la notion de similarité. In *Actes de la 4èmes journées sur l' induction symbolique/numérique*, pages 93–96, Orsay, France, Mars 1994.

[17] F.De Blay, F.Lieutier-colas, P.Krieger, S.Casel, and G.Pauli. Asthme, allergie et polluants de l'habitat (à l'exception du tabac). *Revue Française d'Allergologie et d'Immunologie Clinique 2000*, pages 193–215, 2000.

[18] Willem NIco Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Tweenty. Enschede, The Netherlands, 1997.

[19] Mohand Boughanem. *Système de recherche d'informations : d'un modèle classique à un modèle connexionniste*. PhD thesis, Université Paul Sabatier, Toulouse, France, 1992.

[20] Brigitte Buege. Des médecins accusent notre environnement. *Viva*, pages 12–13, 2005.

[21] L. Monceaux C. Jacquin, E. Desmontils. French eurowordnet lexical database improvements. In *Proc. of CICLing'07*, Mexico, 2007.

[22] James Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pages 301–310, 1994.

[23] Jean Pierre Chevalet. *Un Modèle Logique de Recherche d'Informations applique au formalisme des Graphes Conceptuels*. PhD thesis, Univerité Joseph Fourier, Grenoble, France, 1992.

[24] Yves Chiaramella, Philippe Mulhem, and Franck Fourel. A model for multimedia information retrieval. Technical Report 4-96, University of Glasgow, 1996.

[25] William J. Clancey. The epistemology of a rule-based expert system : a framework for explanation. *Artificial Intelligence*, 20 :215–251, 1983.

[26] C.Olivo, A.Stoebner, N.Chautard, and N.Terral. Comment les relations entre habitat et santé sont-elles prises en compte par les architectes, médecins et travailleurs sociaux ? *Santé publique 8ème année*, 1 :17–26, 1996.

[27] Olivier Corby, Rose Dieng.and Catherine Faron-Zucker, Fabien Gandon, and Alain Giboin. Querying the semantic web with the corese search engine. In *Proc. 15th ECAI/PAIS*, Valence ESpagne, Août 2004. IOS Press.

[28] C. J. Crouch, S. Apte, and H. Bapat. An approach to structured retrieval based on the extended vector model. In *Proceedings of INEX 2003*, pages 89–93, 2004.

[29] FFB CSTB, ADEME. Bien-être et santé dans les constructions. Technical report, CSTB, ADEME, FFB, 2000.

[30] G. Landau Y. Maarek et Y. Mass D. Carmel, D. Efraty. An extension of the vector space model for querying xml documents via xml fragments in xml and information. In *Actes du "Retrieval workshop" de SIGIR*, 2002.

[31] Direction de la santé publique de Montréal. Environnement : garder notre monde en santé. Technical report, Direction de la santé publique de Montréal, 2003.

[32] Scott Deerwester, S Dumais, G Furnas, T Landauer, and R Harshman. Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 6(41) :391–407, 1990.

[33] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[34] Inspection Générale des Affaire Sociales-IGAS. Gestion administrative des plaintes- guide de bonne pratique. Technical report, 2006.

[35] Jérôme Dinet. La pertinence des outils d'experts au service des non-experts en recherche d'informations : un exemple avec les opérateurs booléens. *Revue de l'EPI, n° 99*, 2000.

[36] Chatla Enguehard. *ANA, Apprentissage Naturel Automatique d'un réseau sémantique.* PhD thesis, L'Université de Compiègne, Compiègne, France, 1992.

[37] Serge Mayaya Eric Ngouana. Classification bayesienne naîve de textes. Technical report, 2005.

[38] Sagot Benoît et Fiser Darja. Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de Traitement Automatique du Langage Naturel (TALN)*, Juin 2008.

[39] Y. Mass et M. Mandelbrod. Component ranking and automatic query refinement for xml retrieval. pages 73–84, 2005.

[40] C. Jacquemin et P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. In C. Garbay (Eds.) Jacques Le Maitre, J. Charlet, editor, *Le document multimédia en sciences du traitement de l'information*, pages 71–110. Cépaduès, 2000.

[41] Pierre Falzon. *Ergonomic cognitive du dialogue.* Presses Universitaires de Grenoble, Sciences et Technologies de la connaissance, Lausanne, Zwitserland, 1989.

[42] Christiane Fellbaum. *Wordnet. An Electronic Lexical Database.* Massachusetts Institute of Technology, US, 1998.

[43] Martin Hervé Florence Sédes. Recherche d'information multimédia. décembre 2002.

[44] Edouard Fox. *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types.* PhD thesis, Cornell University, New-York, 1983.

[45] Séverine Frère, Isabelle Roussel, and Aymeric Blanchet. Les pollutions atmosphériques urbaines de proximité à l'heure du développement durable. *Revue Développement Durable et Territoires*, pages 12–13, 2005.

[46] G.Boussin, A.M.Rajon, J.M.Gat, and J.Pous. Application de l'indicateur de santé perceptuelle de nottingham (ispn) à l'analyse des plaintes pour nuisances de l'habitat dans l'agglomération toulousaine. *Santé publique 5ème année*, 2 :34–42, 1993.

[47] Genane Youness Gilbert Saporta. Une méthodologie pour la comparaison de partitions. *Revue de Statistique Appliquée*, 1 :97–120, 2004.

[48] N. Govert and G. Kazai. Overview of the initiative for the evaluation of xml retrieval (inex) 2002. In *Proceedings of the 1st Workshop of the (INitiative for the Evaluation of XML Retrieval (INEX)), Dagstuhl, Germany, December 8-11, 2002*, pages 1–17. ERCIM, 2003.

[49] Monique Le Guen. La boîte à moustaches de tukey, un outil pour initier à la statistique. *Revue de Statistiquement votre*, 4, 2001.

[50] M Halleb and A Lelu. Hypertextualisation automatique multilingue à partir des fréquences des n-grammes. *Hypertextes et hypermédias*, 1 :275–287, 1997.

[51] R. Hooper. Indexer consistency tests : origin, measurement, results and utilization. Technical report, IBM Corporation, 1965.

[52] M.W Evens I.A Al-Kharashi. Comparing words, stems and roots as index terms in an arabic information retrieval system. *JASIS*, 45(8) :548–560, 1994.

[53] G. Fabris J Bateman, B. Magnini. The generalized upper model knowledge bare : Organization and use. In *Towards Very Large Knowledge Bases*, pages 60–72. IOS press, 1995.

[54] Jay Jiang and David Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of international conference on computational linguistics*, pages 133–138, Tapei, Taiwan, 1997. ROCLING X.

[55] Céline Joiron. *Une contribution aux systèmes supports de la Formation Médicale Continue à distance et d'apprentissage entre pairs : Conception et expérimentation du forum DIACOM (Discussions Interactives à bAse de Cas pour la fOrmation Médicale)*. PhD thesis, l'Université de Picardie Jules Verne, France, 2003.

[56] P Daumke S. Schultz U. K, Marko and Hahn. Cross-language mesh indexing using morpho-semantic normalization. In *Proc. AMIA Symp. 2003*, pages 425–429, 2003.

[57] J. Mothe K. Englmeier. Iraia : A portal technology with a semantic layer coordinating multimedia retrieval and cross-owner content building. In *In Proceedings of the International Conference on Cross Media Service Delivery*, page 181.

[58] V. Kakade and P. Raghavan. Encoding xml in vector spaces. In *Proceedings of ECIR 2005, Saint Jacques de Compostelle, Spain, 2005*, 2005.

[59] K. Kwok. A neural network for probabilistic information retrieval. In *Proc.of SIGIR '89*, pages 21–30. Cambridge MA, 1989.

[60] N. Vincent L. Serradura, M. Slimane. Classification semi-automatique de documents web à l'aide des chaînes de markov cachées. In Florence Sèdes, editor, *actes de INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision.*

[61] Denise Lacasse. *Introduction à la microbiologie alimentaire*. Editions Saint-Martin, Québec, 1995.

[62] Luc Lamontagne. *Une approche CBR textuel de réponse au courrier électronique*. PhD thesis, Faculté des arts et des sciences, Montréal, Canada, 2004.

[63] G. N. Lance and W. T. Willams. A general theory of classification sorting strategies. 1. hierarchical systems. *Comp. J.*, (9) :373–380, 1967.

[64] Phipps Arabie Lawrence Hubert. Comparing partitions. *Journal of Classification*, 2 :193–218, 1985.

[65] Claudia Leacok and Martin Chodorow. Combining local context and wordnet similarity for word sense identification). *WordNet : An Electronic Lexical Database*, pages 265–283, 1998.

[66] Yves Lechevalier. le tableau de données, une structure unique, des réalités multiple. In *Recueil des présentations de la journée RDC'05*, pages 21–50, Paris, France, Mars 2005.

[67] Kurt Leininger. Interindexer consistency in psycinfo. *Journal of Librarianship and Information Science*, 1, 2000.

[68] L. E Leonard. Inter-indexer consistency studies, 1954-1975 : a review of the literature and summary of study results. *University of Illinois Graduate School of Library Science Occasional Papers*, 1977.

[69] David Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217, 1992.

[70] Dekang Lin. Automatic retrieval and clustering of similar words. In *ACL98*, volume 2, pages 768–774, 1998.

[71] George F. Luger and William A. Stubblefield. *Artificial Intelligence : Structures and Strategies for Complex Problem Solving*. 1997.

[72] M. Rijke M. Marx, J. Kamps and B. Sigurbjornsson. The importance of morphological normalization for xml retrieval. In *Proceedings of the First Workshop of the INiative for the Evaluation of XML REtrieval(INEX). Dagsthul, Germany, December 2002*, pages 41–48, 2002.

[73] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–296, Berkeley, University of California Press, 1967.

[74] Jean-Luc Manguin. La dictionnairique internet : l'exemple du dictionnaire des synonymes du crisco. *CORELA Cognition, Représentation, Langage, Numéro spécial*, 2005.

[75] Olivier Lemaitre Marc Bruant. Diagnostic qualité d'air intérieur des logements. méthode, principes d'analyse, présentation de l'outil. Technical report, CETE, 2007.

[76] Nívio Ziviani Maria IzabelAzevedo, Lucas Pantuza Amorim. A universal model for xml information retrieval. In *Proceedings 3rd Iniciative for the Evaluation of XML Retrieval (INEX 2004)*, Lecture Notes in Computer Science. Springer-Verlag, 2004.

[77] Jean Pascal Martin. *Description sémiotique de contenus audiovisuels.* PhD thesis, L'universtité de Paris XI, Orsay, France, 2006.

[78] Daniel MEMMI. Le modèle vectoriel pour le traitement de documents. Technical report, Les Cahiers du Laboratoire Leibniz n° 14, 2000.

[79] Mathilde Merlo. Qualité de l'air intérieur et habitat : Analyse des plaintes et leur suivi, 2000.

[80] du développement et de l'aménagement durables Ministère de l'écologie. recenser, prévenir et limiter les risques sanitaires environnementaux dans les bâtiments accueillant des enfants. Technical report, République Francaise, 2007.

[81] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models,. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pages 318–327, 1994.

[82] Fabienne MOREAU. *Revisiter le couplage traitement automatique des langues et recherche d'information.* PhD thesis, Université de Rennes 1, Rennes, France, 2007.

[83] Philippe Mulhem. *HDR - Vers l'Indexation et la Recherche Interprétative de Documents Multimédia.* PhD thesis, l'Université Joseph Fourier, Grenoble, France, 2002.

[84] Sung Hyon Myaeng, Dong-Hyun Jang, Kim Mun-Seok, and Zong-Cheol Zhoo. A flexible model for retrieval of sgml documents. pages 138–145, 1998.

[85] Fiammetta Namer. Flemm : Un analyseur flexionnel du français à base de règles. *Revue Traitement Automatique des Langues*, 41, 2000.

[86] Aurélie Neveol. *Automatisation des tâches documentaires dans un catalogue de santé en ligne.* PhD thesis, INSA Rouen, Rouen, France, 2005.

[87] Raymond HOUE NGOUNA. *Modélisation des connaissances normatives en vue l'évaluation de la recyclabilité d'un produit en conception : des normes aux contraintes.* PhD thesis, L'institut national polytechnique de Toulouse, Toulouse, France, 2006.

[88] Hector Nunez, Miquel Sanchez, Ulises Cortes, Joaquim Comas, Montse martinez, Igansi Rodriguez, and Manel Poch. A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. 26(1) :33–44, 2003.

[89] Province de Liège OMS Europe, République et Canton de Genêve. Habitat et santé. Technical report, République et Canton de Genêve, OMS, Province de Liège, 2000.

[90] Martha Palmer and Zhibiao Wu. Verbs semantics and lexical selection. In *Proceedings of the 32nd Association for Computational Linguistics*, pages 133–138, Las Cruces, 1994. New Mexico State University.

[91] François Paradis. *Un modéle d'indexation pour les documents textuels structurés.* PhD thesis, l'Université Joseph Fourier, Grenoble, France, 1996.

[92] C. Peters. What happened in clef 2005. In *In Working Notes for the CLEF 2005 Workshop, Vienna, Austria, Sept. 2005*, 2005.

[93] B. Piwowarski. Rapport outilex. Technical report, University of Paris 6, 2006.

[94] T. Osborn R. Wilkinson, P. Hingston. Incorporating the vector space model in a neural network used for document retrieval. *Library Hi Tech*, (10) :69, 1992.

[95] M. Rajman and A. Bonnet. Corpora-base linguistics : new tools for natural language processing. In *Proceedings of the 1st Annual Conference of the Association for Global Strategic Information*, 1992.

[96] AM Rajon and J. Belin. Des nuisances liées à l'habitat ou une étude de terrain à toulouse. *Les cahiers médico-sociaux*, pages 151–158, 1988.

[97] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In C. S. Mellish, editor, *IJCAI-95*, pages 448–453, 1995.

[98] Guillermo CORTES ROBLES. *Management de l'innovation technologique et des connaissances : synergie entre la théorie TRIZ et le Raisonnement à Partir de Cas.* PhD thesis, L'institut national polytechnique de Toulouse, Toulouse, France, 2006.

[99] L. Rolling. Indexing consistency, quality and efficiency", year = 1981, journal = Information Processing and Management, volume = 2, pages = 69-76,.

[100] Mathias Rossignol. *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle.* PhD thesis, l'Université de Rennes 1, Rennes, France, 2005.

[101] R.Rada, H.Mili, E.Bicknell, and M.Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1) :17–30, 1989.

[102] Arnon Rungsawang. *Recherche Documentaire à base de sémantique distributionnelle.* PhD thesis, Ecole nationale supérieure des télécommunications, ENST, Paris, France, 1997.

[103] G. Salton. Automatic phrase matching. In *Proceedings of I.C.C.L. International conference on computational linguistics*, 1965.

[104] Gerald Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5) :513–523, 1988.

[105] Gerard Salton. *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, Reading, MA, 1989.

[106] Gerard Salton. The smart project in automatic document retrieval. In *Proc. SIGIR*, pages 356–358. ACM Press, 1991.

[107] Gerard Salton, Edward A. Fox, and Harry Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11) :1022–1036, 1983.

[108] Gerard Salton and M J Mac Gill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[109] Helmut Schmidt. Probablistic part-of-speech tagging using decision trees. In *Actes de the First International Conference on New Methods in Natural Language Processing (NemLap-94)*, pages 44–49, Manchester, England, 1994.

[110] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the European Conference on Artificial Intelligence*, pages 1089–1090. IOS Press, 2004.

[111] Claude Shannon. A mathematical theory of communication (part I). 27 :379–423, 1948.

[112] Sahbi Sidhom. *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances.* PhD thesis, L'Université Claude Bernard Lyon 1, Lyon, France, 2002.

[113] G. Simon. Knowledge acquisition and modeling for corporate memory : from experience. In *Proc. of KAW'96*, pages 411–418, 1996.

[114] A. Singhal. Pivoted length normalization. In *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 21–29, 1996.

[115] Karen Sparck-Jones. Reflections on trec. *Information Processing and Management*, 31 :291–314, 1995.

[116] Karen Sparck Jones, S. Walker, and Stephen E.Robertson. A probabilistic model of information retrieval : Development and comparative experiments. *IPM*, 36(6) :779–808, 809–840, 2000.

[117] Mathieu Stricker. *Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filtres d'informations.* PhD thesis, L'universtité de Paris VI, Paris, France, 2000.

[118] S.Walker and Stephen Robertson. Okapi keenbow at trec-8. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of NIST Special Publication 500-246 : The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, USA, 1999.

[119] R. Olsen T. Gruber. An ontology for engineering mathematics. Technical report, Knowledge Systems Laboratory, Stanford University, 1994.

[120] Nathalie Tchilian. Questionnaire d'enquête sur les plaintes concernant la qualité de l'air intérieur : Principaux résulats. Technical Report DGS/SD7C/2004/354, DGS, Juillet 2004.

[121] D. Tufis. Balkanet design and development of a multilingual balkan wordnet. *Romanian Journal of Information Science and Technology*, 2000.

[122] C.J van Rijsbergen. A new theoretical framework for information retrieval. In *Proceedings of ACM Conference on Research Development in Informational Retrieval*, pages 194–200, 1986.

[123] E. Voorhees. Overview of trec 2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC), Nov.2004*, 2004.

[124] Piek Vossen. *EuroWordNet : A multilingual database with lexical semantic networks.* Kluwer Academic Publishers, 1998.

[125] I. Watson. Knowledge management and case-based reasoning : A perfect match. In *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society conference*, pages 118–122, 2001.

[126] R.M Bauer M.R WARE W.B Lyles, K.W Greve. Sick builging syndrome. *Southern medical journal*, 84 :65–71, 1991.

[127] Ross. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval, Dublin, Ireland*, pages 311–317. ACM Press/Springer, 1994.

[128] J. Ambroziak William Woods. Natural language technology in precision content retrieval. In *Proceedings NLP+IA'98*, Moncton, New Brunswick, CANADA, 1998.

[129] E. Amitay Y. Maarek Y. Mass, M. Mandelbrod and A. Soffer. Juruxml-an xml retrieval system at inex 2002. In *Proceedings of the First Workshop of the INiative for the Evaluation of XML REtrieval(INEX), Dagsthul, Germany, Decemder 2002*, pages 73–80, 2002.

[130] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[131] Haïfa Zargayouna. *Indexation sémantique de documents XML*. PhD thesis, l'Université Paris XI Orsay, Orsay, France, 2005.

[132] Haïfa Zargayouna and Sylvie Salotti. Contexte et sémantique pour une indexation de documents semi-structrés. In *Proceedings de la COnference en Recherche d'Information et Applications (CORIA'O4)*, 2004.

[133] Haïfa Zargayouna and Sylvie Salotti. Mesure de similarité sémantique pour l'indexation de documents semi-structurés. In *Proceedings du 12ème Atelier de Raisonnement à Partir de Cas*, 2004.