CHAPITRE I

Introduction

Table des matières du Chapitre I

A. MOTIVATIONS

La genèse de cette étude puise à trois sources. La première naît d'un engouement naturel pour l'étude du langage. Tel un minéral qui, en se cristallisant, a adopté une géométrie remarquable, parfaitement ajustée à sa composition, mais aussi avec sa régularité et sa fantaisie propre, chaque langue apparaît comme une formation « optimale », une solution possible répondant à son essence profonde de langage humain. La nature a mystérieusement forgé et poli cette matière, qui accueille la représentation, la communication, l'évocation et la mémoire des connaissances, des idées, des sentiments.

La deuxième source à l'origine de ce travail est le cadre applicatif dans lequel il s'inscrit. Dans le monde professionnel actuel, l'information à considérer augmente et se renouvelle dans des proportions quasi inhumaines. La complexité des projets et des réalisations conduit à nouer des collaborations franchissant les repères qui cernaient les domaines de compétence et d'activité. L'organisation et la circulation de l'information prennent un caractère stratégique pour l'entreprise : il en va de sa vitalité, de sa cohésion, et de son déploiement possible. L'idée prend corps, de systèmes capables de donner des vues sélectives d'un ensemble de documents, et de guider la recherche d'interlocuteurs compétents sur un sujet. A la Direction des Etudes et Recherches d'Electricité de France, une équipe a commencé à concevoir un outil relayant les contacts entre les chercheurs à l'échelle du Centre de recherche. Les premières applications sont le « *Qui Fait Quoi ?* » (signalant à chacun les collègues ayant des projets d'activité voisins) et la diffusion ciblée (aidant à repérer les destinataires potentiellement les plus concernés par une information). La technique est un calcul d'associations entre documents (ADOC), dont les premiers résultats sont prometteurs, mais qui demande à être affinés.

La troisième source est issue de l'observation du domaine dit du Traitement Automatique du Langage Naturel (TALN). Beaucoup d'études se focalisent sur la signification du mot, l'articulation de la phrase. Mais le rôle du texte comme contexte global et unificateur, la compréhension synthétique que garde le lecteur, apparaissent encore méconnus des techniques classiques. Pourtant, une branche de la linguistique confirme cette intuition de l'importance du niveau textuel, et montre que l'étude des textes n'est pas réductible à celle des mots ou des phrases qui le composent. L'impact de ce changement de point de vue est majeur : car le texte est la forme effective des documents « réels ». C'est bien du document de travail le plus banal dont on veut pouvoir rendre compte, non de phrases calibrées.

B. POINTS FONDAMENTAUX DU SUJET : UNE LECTURE DU TITRE

1. La diffusion ciblée

a) Principes

L'outil de diffusion ciblée, tel qu'il a été mis au point à la Direction des Etudes et Recherches d'EDF, est un logiciel qui, pour un document que l'utilisateur lui soumet, présente en réponse une sélection des personnes potentiellement les plus intéressées ou concernées parmi l'ensemble des personnes d'un organisme.

Son nom de *diffusion* fait écho à l'application envisagée dès l'origine. L'outil a d'abord été conçu pour favoriser la circulation de l'information : il aide à répertorier les *destinataires* à qui signaler, adresser ou faire suivre un document. Mais le document soumis au système peut également être représentatif d'une problématique sur laquelle on recherche l'avis ou l'aide d'un expert. L'outil de diffusion ciblée est alors utilisé non en vue de l'envoi d'un document, mais pour l'identification d'*interlocuteurs* compétents sur un sujet donné.

Quant au terme *ciblée*, il est essentiel pour rappeler qu'il ne s'agit pas d'une diffusion à tout va, redoutée par tous, mais bien d'un envoi sélectif, qui ne doit apporter que de l'information utile et nécessaire. L'adjectif *sélective* aurait pu également convenir, s'il n'était déjà utilisé pour désigner une application voisine (la diffusion sélective de l'information, ou DSI), avec laquelle il ne faut pas entretenir de confusions. Dans la lignée de ce terme, les documents envoyés ont été appelés *projectiles*, et les profils *cibles*. Cette terminologie balistique, offensive à l'égard des malheureuses cibles, a été abandonnée.

Techniquement, le principe est le suivant : chaque document à diffuser est confronté à une base de « profils » caractérisant les membres de l'organisation, destinataires potentiels. Une mesure détermine alors les meilleurs rapprochements document-profil. Dans son principe, la diffusion ciblée suppose donc : la caractérisation de l'ensemble des destinataires (profils), la caractérisation des documents, et la définition d'une mesure évaluant la pertinence des rapprochements document-profil.

b) Premier pilier : l'automatisme

La diffusion ciblée se caractérise tout d'abord par l'automatisation complète des procédures de caractérisation des documents et de construction des profils, ainsi que de l'opération de recherche des profils correspondant à un document présenté. L'automatisme de l'ensemble de la chaîne de traitement repose sur le fait que toutes les entrées, à savoir aussi bien la soumission des documents que la création et l'actualisation des profils, se font sous forme de textes (cf. deuxième pilier). C'est une des différences majeures avec la plupart des applications faisant usage de profils (notamment les DSI), qui demandent l'élaboration manuelle des profils des personnes à partir de mots-clés que l'on rassemble, et de la même façon la description des documents à l'aide de mots-clés bien choisis.

L'interactivité ne semble concevable qu'à une seule étape, celle où un utilisateur recherche les personnes concernées par un document, et où il souhaite ajuster ou approfondir sa recherche. C'est une étape semi-automatique ponctuelle, à laquelle ne se consacre qu'une personne, pour une recherche. Il n'en serait pas de même si par exemple la définition des profils était conçue par un mécanisme d'apprentissage. En effet, l'apprentissage suppose des retours, qui valident ou corrigent la représentation que construit le système ; et ces retours, pour être exploitables, devraient d'une part se répartir sur l'ensemble des profils et être en nombre suffisant pour chacun d'entre eux. Le volume général des contributions serait considérable, et vraisemblablement inenvisageable, pour des raisons au moins économique (mobilisation générale conséquente de tout le personnel) et psychologiques et sociales (certains sont motivés par l'élaboration de leur profil, mais pas tous).

L'automatisme conditionne en fait la faisabilité même de l'application de diffusion ciblée, qui, pour avoir un sens, suppose un déploiement important : la diffusion ciblée serait ridicule à l'échelle d'une petite entreprise de dix personnes, elle devient un relais irremplaçable à l'échelle d'un

organisme de plusieurs milliers de personnes. La solution économiquement et techniquement réaliste, pour trouver un équilibre entre traitement de masse et coût, passe par la prise en charge de tous les traitements par la machine, sans nécessiter de préparations particulières ou de retouches. Ainsi, des modifications manuelles (sur un profil par exemple) ne devraient concerner que des phénomènes locaux (à savoir à l'échelle de quelques destinataires) ; inversement, l'automatisation permet à une amélioration d'avoir une incidence générale sur tous les profils. Une éventuelle montée en volume (développement des besoins en matière de circulation de l'information, extension de la base de profils à d'autres branches de l'entreprise) est envisageable sereinement : l'essentiel de l'investissement est initial (achat d'un serveur), le coût marginal est négligeable.

Une réalisation informatique efficace permet également d'assurer une rapidité de traitement, bénéfique à de multiples points de vue. Chacun s'accorde à penser qu'il est nécessaire de communiquer et de contribuer à une bonne circulation de l'information dans l'entreprise, qu'une veille technologique à large spectre est essentielle et complexe, mais en même temps qu'il ne faut pas s'y consacrer (c'est une tâche généralement présentée comme secondaire ou auxiliaire). Une application automatisée et rapide procure un gain de temps dans chaque recherche d'interlocuteurs, et facilite la communication sans grever l'activité principale. La circulation de l'information est alors moins mise en concurrence avec les autres tâches, et peut être mieux assurée (information plus complète et plus sélective).

Autre apport essentiel de la rapidité de traitement : la fraîcheur de l'information, point d'autant plus sensible que se multiplient les documents à durée de vie brève, et que s'accélère l'avancée des techniques et savoirs liés aux projets. La fraîcheur de l'information a une double incidence : sur les documents et sur les profils. Pour un document soumis au système, une proposition de destinataires est faite immédiatement : le document peut être transmis ou les personnes contactées dans les plus brefs délais, directement, sans que l'affaire ne s'attarde dans les méandres des circuits standards et généraux. Quant aux profils, l'automatisation permet de mettre à jour l'ensemble des profils au rythme de l'évolution des activités, en minimisant voire annulant le décalage entre les activités réellement en cours et leur prise en compte par le système.

L'analyse automatique des textes assure une homogénéité du traitement sur l'ensemble des textes, et une qualité régulière des descriptions[1]. Les représentations calculées sont fidèles, non pas qu'elles soient objectives (elles reflètent des choix faits lors de la conception du système), mais elles rendent compte uniformément de l'information présente.

Les utilisateurs doivent être sensibilisés à tirer le meilleur parti des propositions calculées, car un tel système n'est jamais infaillible. Il ne peut égaler un professionnel de l'information sur la qualité des informations fournies. Mais en l'occurrence, les services que rend la diffusion ciblée ne sauraient être assurés par une équipe de documentalistes. L'automatisation prend en charge le fastidieux (décliner la combinatoire des requêtes sous-jacentes aux différents aspects d'un document) et surtout le démesuré (avoir la vue d'ensemble, précise et actuelle, de la base des profils des destinataires). Les propositions calculées ont également l'avantage d'apporter une vision neuve, non biaisées par ce qui est déjà connu ou ce qui est le plus évident.

Enfin, le service rendu par un automate tolère tous les excès : il n'y a pas de complexe à répéter (refaire une recherche dont on a égaré les résultats), à soumettre de grosses charges de traitement (un grand nombre de documents le même jour). Du point de vue de l'usager, le service est disponible à tout moment, et il peut l'utiliser sans craindre de monopoliser une ressource dont d'autres auraient besoin (le traitement est suffisamment rapide pour ne pas générer d'attente sensible).

c) Deuxième pilier : l'utilisation directe des textes

La seconde caractéristique majeure et spécifique de la diffusion ciblée est que *toutes les données que reçoit le système sont des textes*. Ceci est doublement novateur, car concerne aussi bien les documents, avec lesquels on lance directement une recherche de destinataires (requête textuelle

[1] Ceci est à nuancer lorsque l'on fait appel à des ressources terminologiques et lexicales : le dictionnaire ou le thesaurus utilisé ne couvre pas nécessairement tous les domaines des textes concernés avec un même niveau de détail, ce qui engendre une qualité inégale des descriptions obtenues.

plutôt que des mots-clés), que la construction des profils, opérée automatiquement à partir de documents électroniques existants.

L'interrogation de l'application de diffusion ciblée se fait au moyen d'un texte libre, sans contrainte de forme (une page Internet), de vocabulaire (termes techniques, néologismes, noms de marques ou de produits), de syntaxe (une liste de mots qui ne forme pas une phrase, des bribes de textes) ou de longueur (texte de plusieurs dizaines de pages). L'application répond à un impératif de robustesse et de simplicité : elle doit pouvoir être consultée avec le moins de contraintes possibles, qui la rendraient en définitive inopérante. Concrètement, à EDF-DER, l'utilisateur, à son poste de travail, active son navigateur Web et se positionne sur le site Intranet DECID. Sur la page d'entrée de la requête, une fenêtre est affichée pour donner le document à comparer aux profils. Plusieurs manière de faire sont possibles : entrer quelques lignes de synthèse au clavier ; mieux, procéder par copier / coller (par exemple depuis son traitement de texte, ou depuis une autre page Web) ; autre solution encore, indiquer un fichier, accessible sur le poste de travail. La requête est donc un texte pris tel quel, que le système de diffusion ciblée analyse directement et compare aux profils de l'ensemble des destinataires définis dans la base.

Les profils eux aussi sont définis à partir de textes, avec la même souplesse. Ils sont générés automatiquement à partir d'un corpus, réunissant des textes représentatifs des centres d'intérêts et des compétences acquises de toutes les personnes considérées. En partant de textes existants, aucune surcharge de travail n'est demandée aux personnes pour la définition de leur profil : il n'y a ni à mobiliser la personne pour qu'elle s'efforce de caractériser son activité à l'aide de mots-clés, ni à mobiliser un professionnel de l'information pour aider à exprimer et mettre en forme le profil dans un format adéquat et efficace. Une personne (destinataire potentiel) peut être décrite par un ou plusieurs textes, sans limitation de sujet, de degré de rédaction (notes qui ne forment pas des phrases) ou de longueur. A EDF-DER, les profils des 1 200 responsables de projet sont calculés deux fois par an, à partir des textes établissant les programmes d'activité rédigés à l'intention de la Direction.

L'utilisation directe des textes présente de nombreux avantages. D'une part –côté document– elle fait de la diffusion ciblée un outil souple et ergonomique (pas de langage d'interrogation ésotérique et réducteur), d'autre part –côté profils– elle assure une description systématique et riche de l'ensemble des destinataires. L'équilibre et l'homogénéité des représentations des documents (requête) et des profils (base de recherche) optimise les mécanismes de confrontation et le calcul de similarité. Enfin et surtout, le texte présente toute une richesse sémantique qui profite à la recherche : gain en précision (bonne contextualisation, qui fait généralement défaut aux requêtes par mots-clés), réduction du silence (le texte déploie lui-même le sujet et esquisse éventuellement certaines perspectives attenantes).

d) Troisième pilier : des profils formant une base

La dernière caractéristique définitoire de la diffusion ciblée, et qui mérite autant d'attention que les deux premières, est la cohérence d'ensemble des profils. Les destinataires forment une base, complète et équilibrée.

La représentation des destinataire est donc systématique : l'ensemble des activités de l'organisme (ici la Direction des Etudes et Recherches d'EDF) est décrit, sans exception. La base ne comporte pas uniquement quelques profils de personnes demandeuses d'information, mais elle rend accessible à chacun la connaissance des compétences de tous. La base reflète une organisation réelle, pas une partie mal définie de cette organisation. Autrement dit, la diffusion ciblée renseigne de façon assez sûre sur l'intégralité des thèmes d'activité, présents ou passés, effectivement suivis dans le cadre de l'organisme. Et quand on sollicite la diffusion ciblée pour suggérer des destinataires concernés par un document donné, son rôle n'est pas d'être capable de fournir quelques noms (il est rare que l'on n'ait pas déjà en tête au moins un ou deux destinataires possibles), mais de faire une exploration de toute la base pour être à même d'indiquer des personnes auxquelles il aurait été plus difficile de penser (nouveaux arrivants, activité inattendue par rapport au Département de rattachement, etc.).

Penser l'ensemble des profils comme formant une base donne une perspective différente de celle d'un système qui gère une série de profils indépendants les uns des autres. Par exemple, dans un

système documentaire de veille par DSI (Diffusion Sélective de l'Information), chaque profil est décrit pour lui-même, et peu importe de savoir ce que reçoit le voisin ou ce qui caractérise le profil d'à côté. Ici au contraire, les profils ont leur contenu propre, mais sont également définis les uns relativement aux autres[2]. Un destinataire est caractérisé en tenant compte des aspects de son activité qu'il a en commun avec d'autres (domaine de son Département) et de ce qu'il apporte de spécifique. Clairement, une compétence originale ou exceptionnelle doit être valorisée pour permettre son repérage en cas de besoin. La diffusion ciblée est également sensible au regroupement ou à la dispersion d'une activité par rapport à l'organisation structurelle : par exemple, si l'on diffuse quatre exemplaires d'un document, il est plus utile qu'ils touchent quatre secteurs concernés différents, que quatre personnes d'une même équipe qui se voient quotidiennement.

Nous avons insisté sur le fait que, et les documents, et les profils, sont définis à partir de textes. Le traitement des profils et des documents n'est cependant pas symétrique. La base des profils regroupe un nombre de destinataires potentiels important : c'est la raison d'être et le point d'appui du système. Les documents sont eux considérés un à un. Les profils se modifient au fil du temps : les destinataires évoluent et changent et les activités se renouvellent, essentiellement au rythme des années. Un document est considéré ponctuellement, au moment où il est présenté au système. La préparation de la base n'est pas visible de l'utilisateur, elle s'opère *en batch*, alors que le traitement d'un document s'effectue *en direct*, interactivement. C'est donc du côté des profils que l'on peut soigner davantage l'analyse du texte, de sorte que chaque profil soit en mesure de sélectionner avec précision les éléments qui lui correspondent, dans une description même un peu moins bien ajustée d'un document. Les documents reçoivent un traitement rapide, moins fouillé. Leur prise en compte dans le cadre d'une session interactive oblige une attention aux délais, aux temps de calcul, aux files d'attente.

2. Voie explorée : la linguistique textuelle

a) Le texte, point de départ et unité de travail de la diffusion ciblée

Les textes sont *la* source essentielle d'informations pour la construction des profils des destinataires, comme pour la caractérisation des documents. L'enjeu est donc bien de savoir en tirer le meilleur parti, et donc d'exploiter autant que possible la richesse des textes.

Irréductibilité aux mots et aux phrases

Comme il s'agit de définir, à partir des textes, une représentation permettant le calcul des rapprochements, l'objet d'étude est d'abord le texte avant d'être les mots ou les phrases qui le constituent. De même, ensuite, le « grain » souhaité pour les rapprochements est celui du document (ou de l'extrait de document), non celui des mots ou de la phrase.

Ce n'est pas que le travail sur les mots ou les phrases soit rejeté : au contraire, il prend sens dans l'étude du texte, qui l'intègre. L'analyse textuelle unifie les informations et résultats linguistiques locaux. Le risque serait par exemple de partir d'emblée sur un traitement lexical, en omettant la perspective plus large qui situe le lexique comme une dimension du texte parmi d'autres. Cela aurait une cascade de conséquences regrettables. Tout d'abord, une altération de la qualité de la représentation est à prévoir, puisque les mots ne sont pas indépendants de leur contexte. A cela

[2] Le passage d'une conception individuelle et isolée des profils, à une société de profils, a une portée plus grande qu'il n'y paraît, en tant que modèle opératoire :
« le terminal relié à un réseau et via un autocommutateur à d'autres terminaux est finalement une métaphore toute prête pour interpréter notre statut de sujet [le statut d'un usager], en traitant l'individu comme un terminal, comme un point et la communication comme un système point à point.
Or, quelle révélation, l'homme n'est pas un terminal, il n'est pas un point isolé ! Il n'est pas non plus à l'opposé un termite (!), c'est-à-dire qu'il ne vit pas sous une forme agrégative complexe où le statut même de l'individu peut être questionné. Ce serait plutôt un intercom ou mieux encore un autocommutateur, si l'on veut rester dans la métaphore téléphonique, dans la mesure où il n'existe que de la mise en relation de segments divers de la société. Ce qui nous est définitoire, c'est plutôt l'interaction, l'interlocution, l'interdiction ou l'intersection. » (Boullier 1997, p. 38)

s'ajoute l'appauvrissement prématuré de l'information de caractérisation –deux textes peuvent avoir reçu les mêmes mots-clés, sans pour autant être équivalents aux yeux de l'utilisateur. La singularité du document ainsi émoussée entraîne une chute de la précision des diffusions proposées.

Rester dans le champ de la linguistique

L'approche la plus réaliste consiste aussi à ne pas quitter ce domaine initial, la nature linguistique et sémiotique des données. En faisant appel à la linguistique, et plus particulièrement à la sémantique, on choisit de construire des représentations et de travailler au niveau du sens plutôt que des connaissances (Rastier, Cavazza, Abeillé 1994, p. 14). Le sens se situe du côté de la langue et des possibilités d'interprétation d'un texte. Les connaissances concernées varient avec la situation, les lecteurs, l'utilisation du document, etc. Une représentation sémantique (c'est-à-dire du sens) semble donc plus adéquate pour ne pas réduire *a priori*, par une analyse précoce et prédéterminée, l'éventail des connaissances possibles, selon les destinataires et les circonstances. Techniquement aussi, le traitement a tout intérêt à se passer d'une base de connaissances de type description du monde encyclopédique, dont tout à la fois la réalisation, le coût et l'utilisabilité seraient problématiques. L'option est donc celle d'une linguistique autonome, telle que la sémantique différentielle.

> [La sémantique différentielle est] issue de la linguistique structurale européenne [...]. Elle définit la signification comme un rapport linguistique entre signes, plus précisément entre signifiés. Les signifiés ont à leur tour des corrélats psychologiques, voire physiques, mais ces corrélats ne les définissent pas en tant que tels. [...]
>
> [La sémantique différentielle] nous paraît apte à traiter des textes, car elle ne définit pas la signification par des relations entre le texte et d'autres réalités non linguistiques. Partant, elle n'exige pas le recours à une psychologie (comme la sémantique cognitive) ou à une ontologie (comme la sémantique formelle).
>
> (Rastier, Cavaza, Abeillé 1994, §II.1)

b) Une discipline difficile à cerner

Il n'y a pas de discipline clairement constituée répondant au nom de *Linguistique Textuelle*. Des courants très différents quant à leur inspiration et à leur objet peuvent s'en réclamer, et inversement les désignations voisines abondent : *linguistique des textes*, *sémantique des textes*, *linguistique du discours*.

Sous l'étiquette « Linguistique Textuelle », il faut entendre ici, de façon simple, les travaux qui ont le double appui, de l'étude de la langue et de celle des propriétés des textes. Parmi les branches de la linguistique, la sémantique a ici une importance primordiale, puisqu'elle donne accès au sens, au « contenu », y compris à travers des marques morpho-syntaxiques. Pour cerner plus avant ce qui est visé ici, nous allons donner quelques exemples de voies explorées par la Linguistique Textuelle, qui retiennent notre attention dans la perspective de la diffusion ciblée.

La Linguistique Textuelle considère de façon unifiée les paliers du mot, de la phrase et du texte. Ces trois paliers dessinent les effets de contexte et de dominance. Ils entrent en coopération pour la désambiguïsation des mots-clés extraits du texte (un mot pris isolément se prête à de multiples interprétations divergentes). Le palier textuel n'est manifestement pas une simple extrapolation des autres paliers : un paquet de mots ne fait pas un texte, une collection quelconque de phrases n'est pas reconnue comme texte. Le texte forme un tout, une unité cohérente et organisée.

Qui dit texte dit aussi lecture(s) : la représentation que l'on construit du texte est toujours l'élaboration d'une interprétation, avec ses attentes et ses partis-pris. Les stratégies et tactiques de lecture, alliées aux contraintes linguistiques qui guident l'interprétation, distinguent les statuts de diverses zones (découpage interne du texte), classent et pondèrent les éléments à retenir. La linguistique textuelle considérée ici est donc une linguistique interprétative.

Enfin, le texte prend place dans un ensemble, qu'il s'agisse de l'étagère que l'on a sous les yeux, de ses lectures antérieures, des autres textes avec lesquels il entre en débat, du genre dans lequel il s'inscrit. De même qu'il est artificiel d'isoler le mot hors du texte, c'est ignorer la réalité intertextuelle que de considérer un texte indépendamment de tout corpus. L'ensemble des textes qui définissent les profils de destinataires est notamment à envisager en termes de corpus.

En somme, l'approche textuelle permet d'envisager l'appréhension des phénomènes globaux et le rééquilibrage des effets locaux, complémentairement et en interaction avec les analyses lexicales et phrastiques (sur lesquelles sont basés les outils TALN).

Le rôle de la Linguistique Textuelle est ici de définir ce qui doit être retenu comme représentation des textes pour l'application de diffusion ciblée. L'analyse statistique des données textuelles et les outils TALN sont alors des moyens (respectivement quantitatifs et qualitatifs) de construire, automatiquement, cette représentation, dans un format approprié au calcul (pondération, étiquetage, distance...).

c) Domaine : des documents à dominante scientifique et technique

Le domaine des textes concernés par la diffusion ciblée dans un centre de recherche industriel ne s'étend pas aux textes dans leur plus grande diversité. Les propriétés plus spécifiques des textes littéraires ou religieux par exemple ne nous intéressent pas ici. Les documents juridiques, qui peuvent circuler en entreprise, ne sont pas envisagés dans notre contexte. Les documents administratifs sont très présents, mais ne semblent pas en soi constituer l'objet prioritaire d'une application d'aide à la diffusion et de ciblage.

Sont concernés au premier chef les documents à dominante scientifique et technique. Le « à dominante » permet d'élargir le corpus des documents consultés, échangés et rédigés dans le travail quotidien des chercheurs de la Direction des Etudes et Recherches d'EDF, aux documents relatant l'activité de recherche, dans les circuits administratifs ou la communication d'entreprise par exemple.

Concrètement, les corpus qui rentrent dans ce champ s'étendent des très diversifiées notes techniques (résultats expérimentaux, document de travail, rapport de stage, compte-rendu de congrès ou de réunion, thèse...) aux descriptifs de programme de travail (actions annuelles, projets pluriannuels et transversaux, contrats de Groupes), en passant par les fiches issue de rapports d'activité de laboratoires de recherche, les contrats de collaboration externe (pour leur partie technique), les publications scientifiques.

d) Langue française, langue anglaise, multilinguisme

Le travail a été réalisé pour des documents en langue française. Les besoins pressent aussi pour le traitement de documents en langue anglaise.

En effet, l'anglais est à présent la langue couramment employée pour les échanges internationaux. C'est elle qui est utilisée pour s'adresser à la communauté internationale la plus large. Son poids est renforcé par le fait qu'elle domine généralement les autres langues, en servant de langue-pivôt. La situation est au point que la question du multilinguisme est court-circuitée dans la pratique par l'usage généralisé de l'anglais (un sondage sur les pages Web le montre sans difficulté), alors qu'elle devient pressante et gagne en acuité pour les défenseurs des autres communautés linguistiques.

Dans un centre de recherche comme celui d'EDF, le volume brassé d'informations en langue anglaise est significatif : publications dans les revues, communications dans les congrès, données enregistrées dans les bases documentaires, informations glanées sur Internet. Pour autant, la langue de communication interne reste unanimement le français. Ceci a pour conséquence que les documents relatifs à l'activité de l'entreprise sont rares en anglais.

Le travail présenté ici porte donc d'abord sur le français. Le cas de l'anglais est envisagé pour le moyen terme : s'il n'est pas traité dans un premier temps, les choix posés antérieurement ne doivent pas compromettre l'extension de l'application à l'anglais.

3. Application concrète : évolution des versions du système DECID

a) Contexte : une visée opérationnelle

La thèse s'est déroulée dans le contexte d'une application opérationnelle, le système DECID (*Diffusion Electronique Ciblée d'Informations et de Documents*), accessible à tout agent EDF d'abord par messagerie (en 1995), puis par une interface Web, sur l'Intranet de l'entreprise (à partir de 1996).

C'est un service en exploitation, qui compte environ 500 utilisateurs et reçoit une dizaine de requêtes par jour.

Ces chiffres ne sont que des moyennes, qui donnent une image grossière de la réalité. En particulier, les comportements des usagers sont contrastés et inégaux. Certaines personnes, dont le cœur même de l'activité consiste à rechercher des experts ou à aiguiller des documents, sont de gros utilisateurs de DECID. D'autres connaissent DECID et s'en servent à l'occasion, ponctuellement : par exemple pour faire relire et valider un projet de coopération technique externe, ou au moment de la préparation de l'ordonnancement, pour s'informer de projets d'activité analogues en cours ou en perspective. C'est une utilisation relativement saisonnière, reflétant le rythme annuel, voire hebdomadaire[3], des activités professionnelles. La troisième catégorie d'utilisateurs sont des personnes qui ont essayé DECID, en tant que nouveauté à découvrir, mais n'ont pas intégré la diffusion ciblée dans leur pratique de travail ; ils pourraient se trouver en situation de rechercher un interlocuteur par rapport à un document, sans avoir le réflexe de penser à DECID.

b) Une intégration graduelle des apports de la thèse

La thèse représente en grande partie la composante recherche du projet *diffusion ciblée*, dans le prolongement d'un stage de DEA. A ce titre, elle contribue à l'application dans ses différentes versions.

Certaines propositions et réalisations font partie de la version en exploitation. Il s'agit des premiers modules mis au point (construction de profils synthétiques pour les structures Groupes, Départements, Services ; identification automatique du vocabulaire général des textes descriptifs des programmes d'activité), et qui visent une amélioration de la qualité des propositions du système dans le cadre du modèle initial adopté (*Vector Space Model* avec pondérations, inspiré directement des travaux de Salton en *Information Retrieval*). Il n'y a pas de remise en cause de la technique utilisée, mais des adjonctions compatibles avec ce modèle (le mécanisme des *catégories*, pour articuler informations quantitatives et informations qualitatives, ces dernières n'ayant pas leur place dans les pondérations numériques). L'optique est d'introduire une meilleure prise en compte de la textualité des données.

Un travail était également à réaliser au niveau de l'interface, le passage à une interface Web ouvrant des possibilités nouvelles. La thèse a essentiellement contribué à la définition de modes d'affichage des textes[4], appropriés à leurs caractéristiques (conciliation d'une vision globale de l'ensemble d'un texte, et d'une vision locale, au point de lecture ; positionnement des textes entre eux), et adaptés aux lectures que fait l'utilisateur à travers DECID (mise en valeur contextuelle ou personnelle de certains éléments, confrontation de textes et caractérisation d'un texte par rapport à un autre). Ces nouvelles fonctionnalités d'affichage et d'interaction avec les textes sont pour l'essentiel testées et intégrées dans une version prototype de DECID, qui pourrait soit succéder à la version actuelle, soit être une version plus évoluée et plus puissante, parallèlement à une version de base simple et rapide. Ce « DECID+ » serait surtout utile aux gros utilisateurs de l'application.

La réflexion théorique et linguistique développée au cours de cette recherche a mis en évidence les limites intrinsèques des techniques initiales, davantage centrées sur les mots que sur les textes. L'enjeu est alors de proposer un nouveau modèle, faisant place à la problématique textuelle, sans perdre de vue la nécessité de parvenir à des implémentations logicielles réalistes. Ce travail de fond est également au cœur de cette thèse, et a abouti à la conception de modules informatiques. Certains de ces modules sont achevés et testés (traitements sur la structure des textes, telle qu'elle est représentée dans les fichiers), d'autres sont réalisés mais pas encore intégrés et expérimentés (classification multiclasse non exhaustive, maquettes d'interfaces particulières). Le développement du module qui réalise l'analyse et la description des textes est arrêté au stade d'une première version, qui montre sa faisabilité mais ne couvre pas encore les aspects les plus intéressants et les plus novateurs. La redéfinition du moteur de calcul des rapprochements est encore en phase de conception ; elle peut

[3] Il est connu par exemple que les messageries sont surchargées le lundi matin.

[4] Ce travail n'aurait pu se concrétiser sans la collaboration active et créative de Laurent LUCIANI (société DECILOG), maître d'œuvre de l'interface Web de DECID.

être mise en œuvre ultérieurement, l'ancien moteur pouvant déjà être utilisé avec profit avec les nouvelles unités.

Les implémentations informatiques s'échelonnent donc, de la réalisation opérationnelle à la conception des versions ultérieures. L'ancrage disciplinaire linguistique de la thèse (plutôt qu'informatique) s'est traduit par le fait que l'objectif premier n'a pas été la livraison d'une version aboutie complète, mais l'investigation approfondie de nouveaux fondements linguistiques pour le système, des modules informatiques en donnant de premières illustrations. Toutes ces implémentations cependant répondent à un même mot d'ordre : réalisme et adéquation au contexte industriel de DECID.

C. CONTEXTE INITIAL

1. Un pôle de recherche autour des nouvelles technologies de l'information

a) Le traitement automatique des langues naturelles (TALN)

Le projet ASTREE (*Atelier Standardisé de TRaitement Electronique de l'Ecrit*) (Monteil 1994) a coordonné[5] les travaux pour doter EDF d'une boîte à outils linguistique aussi complète que possible. On dispose de *ressources* (dictionnaires, thesaurus EDF) et d'*outils* (étiquetage morphologique et lemmatisation, extraction de groupes nominaux, repérage de l'expression des actions et de la causalité). Les applications sont diversifiées : indexation automatique, aide au repérage de la terminologie et des concepts d'un domaine, assistance au résumé par la sélection de phrases importantes.

2. Historique du projet

a) L'outil ADOC

Le contexte initial du projet de diffusion ciblée à EDF est marqué par une bonne connaissance des travaux de Salton. Au plan théorique, l'équipe poursuit les recherches dans ce domaine, et a mis en place de nouvelles techniques combinant statistiques et linguistiques, notamment pour les classifications de documents. Au plan expérimental, des applications d'analyses factorielles sur des textes de l'entreprise démontrent l'apport de l'analyse des données pour fournir des lectures visuelles, synthétiques et révélatrices, de corpus volumineux. Les calculs de rapprochement document-requête, utilisées outre-atlantique en *information retrieval*, sont bien connus. L'idée[6] vient alors de les étendre aux confrontations texte-texte[7], avec la perspective d'exploiter les textes descriptifs de l'activité des chercheurs comme premier ensemble de textes. Tout document pourrait ainsi être rapproché de l'activité de certains chercheurs, on aurait un moyen de repérer automatiquement les personnes les plus « proches » d'un document.

---

[5] Trois branches du Centre de recherche s'intéressent au TALN pour leurs applications (Monteil 1994) :
- au Service *Informatique et Mathématiques Appliquées*, le Département concerné (TIEM) s'occupe de gestion des connaissances et d'aide à la décision, avec des applications d'extraction et de modélisation des connaissances à des fins de conduite et de maintenance des installations techniques EDF ;
- au Département *Systèmes d'Information et de Documentation*, il s'agit d'exploiter au mieux le fonds documentaire de l'entreprise, en s'appuyant sur de multiples applications d'analyse de l'information textuelle comme l'analyse documentaire (indexation, résumé), la diffusion, la synthèse, la veille technologique et stratégique ;
- au Département *Groupe de recherche Environnement, Technologie et Société*, on effectue le dépouillement des enquêtes d'opinion et l'analyse de l'image de l'entreprise à travers la presse, soit donc la conception et l'utilisation d'applications d'analyse de l'information textuelles à des fins sociologiques.
Ceci reflète bien la diversité des secteurs dans lesquels le TALN peut intervenir dans les entreprises.

[6] Cette idée, à l'origine de la diffusion ciblée à EDF, est celle de Georges HÉBRAIL, alors chef du Groupe *Ingénierie des Systèmes d'Information* (ISI), au Département *Systèmes d'Information et de Documentation* (SID) de la *Direction des Etudes et Recherches* (DER) d'EDF

[7] En effet, contrairement au modèle de la recherche documentaire par équation booléenne, le *Vector Space Model* de l'*Information Retrieval* propose un formalisme de représentation unifié pour les requêtes et les documents, c'est-à-dire concrètement équivalent pour les interrogations en quelques mots et pour les textes des documents (Salton 1988).

L'idée prend consistance. Un logiciel est réalisé, qui calcule les similarités de documents deux à deux[8]. Ce logiciel est baptisé ADOC : *Associations entre DOCuments*.

Le traitement considère en fait trois ensembles :

- le *corpus de référence*, qui sert au calcul des pondérations intrinsèques des mots ; il se veut représentatif des familles de textes qui seront comparées, plus précisément du vocabulaire employé et de son utilisation (rareté, spécificité ou au contraire emploi courant et répandu, au moins pour le domaine considéré) ;
- le *corpus de requête*, à savoir l'ensemble des documents dont on veut trouver des documents proches ;
- le *corpus de recherche*, qui est l'ensemble des documents confrontés à chacun des documents du corpus de requête.

Les rôles des corpus de recherche et de requête sont formellement semblables dans le principe. Mais la pratique peut introduire une dissymétrie d'usage, qui fait que ces corpus sont interprétés différemment. D'abord, les résultats sont ordonnés suivant le corpus de requête. Ensuite, le corpus de recherche est en quelque sorte la base stable à laquelle on peut vouloir confronter plusieurs vagues de corpus de requêtes. Cela transparaît dans la réalisation informatique : les corpus de référence et de recherche sont compilés, pas le corpus de requête.

La distinction de ces trois statuts ne préjuge pas du non recouvrement des corpus eux-mêmes. En effet, il est courant de se donner comme référence le corpus de recherche (*corpus de référence = corpus de recherche*) ; et une application particulière identifie de surcroît le corpus de requête au corpus de recherche (*corpus de référence = corpus de recherche = corpus de requête*). C'est le cas de la première principale application d'ADOC : le « Qui-Fait-Quoi ? ».

b) Le « Qui-Fait-Quoi ? »

Dès que l'on dispose des documents descriptifs de l'activité des chercheurs pour l'année suivante, ADOC est utilisé pour calculer les similarités d'un projet à l'autre. Ainsi, selon l'optique de la Direction, nul ne peut ignorer qu'il se prépare une recherche voisine ou très liée dans un autre Département, le cas échéant. Il est alors encore temps de se mettre en relation, d'orienter les travaux en complémentarité, de mettre en commun des moyens, de réévaluer les priorités de chacun en évitant les redondances au niveau de l'ensemble du centre de recherche. Le « Qui-Fait-Quoi ? » (orthographe peu à peu préférée à *Kiféquoi*) se présente donc comme une aide à la définition de l'ordonnancement, à la connaissance partagée de l'évolution prévue de l'activité, et à la synergie entre les différentes équipes.

Chaque responsable d'entité reçoit à l'automne un fascicule qui regroupe les projets d'activités extérieurs calculés comme les plus proches de chacun des projets qu'il supervise. Il a ainsi des éléments pour mieux positionner l'activité qu'il encadre. Il peut donner suite à tel ou tel rapprochement signalé. Il peut aussi faire circuler le document pour que chacun ait la connaissance précise d'autres initiatives en rapport avec son projet d'activité.

Plusieurs fois, une enquête auprès des destinataires a permis de recueillir les avis pour faire évoluer le document et valider son utilité (Sta 1993) (Vavasseur & Lemesle 1994). Les éditions annuelles successives (de 1992 à 1995) ont été améliorées sur plusieurs points : recueils pouvant être plus détaillés car plus ciblés (envoi personnalisé non plus aux 35 Départements mais aux 180 Groupes à partir de 1993), présentation soignée et nombreuses aides à la consultation et à l'analyse (table des matières, tableaux récapitulatifs, rappel des liens déjà concrétisés sous la forme de projets transversaux, etc.). Ceci est resté possible avec les exigences de qualité, de délais et de coût, grâce à l'automatisation de la chaîne de production (génération du contenu, mise en page, impression, reliure).

[8] Dès 1992, c'est Jean-David STA, ingénieur à SID/ISI, qui défriche les principes de la mise en œuvre de la diffusion ciblée et réalise les premières expérimentations. Le logiciel ADOC est implémenté en 1993 par Arnaud JOURDAN, sous sa direction. En 1995, les interventions répétées pour la maintenance de ce programme en C++, nécessitée par l'évolution constante des compilateurs, conduit à réécrire ADOC en Ada. C'est ADOC *version 2*, conçu et réalisé par Pascal OBRY, ingénieur à SID/ISI.

c) L'application de gestion des profils

Partant de l'idée qu'il faut pouvoir dégrossir et compléter la caractérisation automatique en retouchant son profil, une interface est développée pour éditer un profil courant et assister son remodelage[9].

Le profil se présente sous la forme d'une liste de descripteurs avec leur fréquence (obtenue par l'indexation automatique des descriptifs d'activité selon le thesaurus EDF). L'utilisateur est invité à tester son profil sur un ensemble de documents, pour observer si les documents que le profil sélectionne correspondent bien à son activité. Ce test est utile préalablement au travail de retouche et pour évaluer étape par étape les modifications apportées.

La fonctionnalité de correction la plus simple est la suppression d'un descripteur qui génère des rapprochements non pertinents. Pour l'ajout d'un descripteur, l'utilisateur peut naviguer dans le thesaurus, alphabétiquement (une recherche incrémentale permet d'accélérer le repérage), par filtrage sur une sous-chaîne de caractères, ou par thème et champ sémantique. Il sélectionne avec la souris les entrées qui lui conviennent. Il peut aussi proposer un nouveau document qui complète son profil : le système en extrait immédiatement les descripteurs et leur fréquence. L'ajout peut alors porter soit sur quelques descripteurs, soit globalement sur tous les descripteurs caractérisants le nouveau document.

Intervenir directement sur les fréquences n'est pas souhaitable. C'est en effet une grandeur interne, dont l'utilisateur a une perception faussée : elle est gouvernée par la distribution lexicale (répartition des mots dans les textes) et est exploitée par des calculs complexes. Autrement dit, les tenants et les aboutissants de la fréquence échappent à l'utilisateur. Jouer sur les fréquences perturbe la cohérence de l'ensemble et fait de l'utilisateur un apprenti sorcier. Cependant, le désir est légitime de sortir d'une alternative binaire du tout ou rien, de l'ajout et de la suppression. Pour cela ont été introduites les catégories : elles permettent d'associer une information qualitative à tout descripteur. Le système de catégories proposé distingue les termes issus de l'indexation d'un texte entièrement intégré au profil (leur fréquence est significative), les termes transverses (à dévaluer), des termes complétant le contexte obtenus en resituant le profil dans son Département de rattachement par exemple, des termes ajoutés manuellement (démunis de fréquence, et *a priori* importants car spécifiés explicitement et validés par l'utilisateur).

La mise au point de l'application de gestion des profils a montré la nécessité d'une réflexion sur l'ergonomie, pour médiatiser la perception qu'a un utilisateur des informations à manipuler et les codages et modèles internes sur lesquels s'appuient les calculs. Cette étape a aussi ouvert la réflexion sur la possibilité d'une gestion décentralisée des profils. La question de la faisabilité est double : technique et économique. Technique, car il faut savoir comment gérer ces modifications locales et désynchronisées : on ne peut plus baser le modèle sur le fait que la caractérisation est issue d'un corpus homogène et clos. Economique, car il reste souhaitable de former voire d'accompagner les utilisateurs : la mobilisation d'experts ou de conseillers se chiffrerait à au moins une personne à plein temps, si le service est offert pour les 1 200 profils.

d) Démonstration d'une chaîne intégrée aux Journées Techniques de la DER (novembre 1994)

L'*intégré*, mis au point à l'occasion d'une démonstration, a marqué un tournant sans doute décisif dans la destinée du projet de diffusion ciblée, car il a fortement marqué les esprits et illustré les potentialités du système.

L'*intégré* a consisté à enchaîner trois outils réalisés par le Groupe : la reconnaissance de documents, l'indexation automatique, et la diffusion ciblée. Un document quelconque, par exemple une plaquette d'un stand voisin, est posé sur le scanner. La reconnaissance de document fait davantage qu'une reconnaissance optique de caractères : elle fait la part entre le texte, les schémas et les illustrations, ne se laisse pas perturber par les fonds colorés des mises en page sophistiquées, etc.

[9] Cette application est réalisée par Marc LAMOUREUX, (société ORIGIN), supervisé par Pascal OBRY et Xavier LEMESLE, ingénieurs à SID/ISI. On est alors en 1994, année où la diffusion ciblée est lancée comme un projet à part entière, en faisant l'objet d'une ARD (*Action de Recherche et Développement*), N4607R. Xavier LEMESLE devient le responsable du projet ; c'est lui qui le supervisera jusqu'à son terme.

L'intégré fournit, au bout de quelques instants, une liste de personnes. Les résultats sont parlants : dans le cas de la plaquette, il retrouve les noms de ceux qui tiennent le stand et sont responsables du projet présenté, ainsi que leurs interlocuteurs.

e) Un automate accessible via la messagerie électronique

L'application de diffusion ciblée devient pleinement réalité sous forme d'un automate accessible depuis la messagerie électronique. Les calculs se font sur la base de l'ensemble des profils des chercheurs pour l'année courante. L'utilisation de l'outil de diffusion ciblée est simple : le texte pour lequel on recherche des destinataires est mis dans le corps d'un message électronique ; le message est envoyé à l'adresse de l'automate. Quelques minutes plus tard, l'émetteur reçoit dans son courrier électronique un message de l'automate, contenant l'ensemble des destinataires suggérés. Ces résultats sont ordonnés par proximité décroissante. La diffusion ciblée commence une phase d'exploitation.

Côté recherche, quelques tests sont menés sur des corpus variés, pour observer la qualité des résultats : résumés de Notes internes, CERD (*Contrat Externe de Recherche et Développement*). C'est l'occasion d'étudier systématiquement le comportement du système, sa robustesse et ses points faibles. Le travail de la thèse s'ouvre d'ailleurs par un recensement et une classification par type des erreurs de l'application.

f) DECID - La version Web

La diffusion ciblée a séduit et fait ses premières preuves ; elle est appelée à se développer rapidement. Cela exige de passer d'une réalisation expérimentale à une application industrielle, associée à une offre de service. L'application se dote d'un nom, DECID, acronyme pour *Diffusion Electronique Ciblée d'Informations et de Documents*.

Le choix est fait de mettre en place un serveur Web. Le service par messagerie est plus largement accessible : la messagerie est déployée à l'échelle de l'entreprise et fait partie des canaux de communication courants, alors que la navigation sur le Web n'est pas une pratique aussi répandue et n'a pas le soutien inconditionnel de la hiérarchie (on craint une perte de temps ou une utilisation détournée).

Mais le serveur Web remplace finalement le service de messagerie pour plusieurs raisons :

- la messagerie interne, très chargée, connaît des interruptions de services, dont pâtit alors la diffusion ciblée.
- l'entretien et l'évolution de deux versions parallèles excède les moyens dont dispose le projet.

Ce cap, franchi fin 1995, anticipe déjà sur la période couverte par la thèse. La version initiale reprend le moteur ADOC de calcul de similarités entre textes et la caractérisation des textes par l'indexation automatique ou le simple découpage des chaînes de caractères. En revanche, l'interface Web permet une interaction plus riche ; l'ergonomie du système devient un pôle d'attention mobilisant et dynamique.

3. Diagnostic en 1995 (début de la thèse)

a) Un système robuste et des résultats prometteurs

L'outil réalisé fonctionne sans défaillances et traite efficacement chaque requête. Il arrive à prendre en compte les textes les plus variables, même tronqués, peu ou pas rédigés, et à y répondre sans sourciller.

Les propositions de destinataires sont inégales, mais il est rare de ne pas trouver des rapprochements pertinents[10]. S'il s'agit de destinataires dont on connaissait à l'avance l'intérêt pour le sujet soumis au système, le comportement du système rassure. Mais le plus convaincant quant à

[10] Sauf si l'on soumet un document hors du champ de préoccupation du centre de recherche, on peut même affirmer que l'on trouve *toujours* des rapprochements pertinents... pour peu que l'on ait la patience d'examiner la liste des résultats aussi loin que nécessaire.

l'intérêt du système, c'est de repérer ainsi des personnes concernées et auxquelles on n'aurait pas pensé (faute de les connaître, ou de savoir qu'elles travaillent actuellement sur le sujet).

Les résultats du calcul de proximités texte à texte sont suffisamment satisfaisants pour être utilisés comme données dans d'autres expériences.

Le mode d'interrogation –par un texte– se montre une alternative intéressante à l'interrogation booléenne (mots-clés liés par des ET et des OU). Elle dispense d'essayer de prévoir quels mots peuvent apparaître dans les textes et avec quelles combinaisons, chose qui oblige à procéder à de multiples essais et ajustements. De plus, le texte apporte naturellement un contexte plus riche que un ou deux mots-clés.

Le constat est donc le suivant : l'outil de diffusion ciblée, avec son algorithme de traitement très fruste, fournit déjà des résultats intéressants. Pourquoi ne pas poursuivre sur cette lancée ? Il y a tout à gagner à consacrer un peu plus d'attention au traitement effectué. D'ailleurs, le formalisme vectoriel est souple et se prête à des réaménagements (Salton 1988).

La thèse reste ancrée dans cette réalité modeste mais opérationnelle. L'objectif est de partir de l'existant et d'affiner peu à peu l'outil, de le faire évoluer en restant dans un contexte d'utilisation courante. Il ne s'agit pas de viser à bâtir de toutes pièces un système irréprochable tant au plan de la théorie qu'à celui des résultats.

b) Les limites connues des techniques de Salton

Gestion des échelles

L'utilisation du Modèle de l'Espace Vectoriel (VSM) et des formules de pondérations associées est une extrapolation dans le cadre de DECID v.2. Ces techniques ont été mises au point pour la recherche d'information dans de grandes bases documentaires : les requêtes des utilisateurs et le format des documents enregistrés se sont d'abord traduits dans les faits par des textes relativement courts et de forme identique, de type résumé.

L'entrée de documents de plusieurs pages, en texte intégral, a fait apparaître des difficultés : représentation volumineuse, sous-estimation des similarités. L'éclatement des documents en « passages » a été proposé pour contourner la difficulté et se ramener au cadre initial. Sauf quelques aménagements pour articuler les calculs sur les documents entiers et ceux sur les passages, l'unité du texte est sacrifiée aux besoins du modèle. Or pour la diffusion ciblée, le calcul de similarité de texte à texte s'écarte effrontément et de l'expression condensée de la requête, et de la brièveté habituelle des informations entrées dans la base.

Le contexte de la diffusion ciblée oblige aussi à considérer des documents tout-venant, de genres divers (descriptif d'activité, résumé de note technique, etc.) : elle rencontre là un deuxième point faible des techniques saltonniennes, démunies devant les variations dues aux genres, notamment le rôle du vocabulaire conventionnel et du style propre à chaque genre.

La construction des profils demande souvent d'associer à une seule personne plusieurs textes : là encore, c'est un cas qui ne trouve pas sa place dans le modèle vectoriel. Le modèle ne peut proposer que d'« additionner » les (vecteurs représentatifs des) documents. Une telle opération a une signification discutable, et s'avère peu performante. La représentation obtenue est une moyenne surchargée : les spécificités de chaque texte sont noyées, et le profil résultant se comporte aussi mal que celui d'un texte long. L'autre option simple est de considérer indépendamment chacun des textes concourant à la description d'une même entité[11] : cela semble résoudre les questions liées au nombre de textes et à leur regroupement, mais ce faisant on perd la vue intégrée et la synergie possible entre les différentes activités représentées par les différents textes.

Le modèle de l'espace vectoriel s'appuie fortement sur les fréquences d'apparition des mots dans les documents. Mais les formules ne s'avèrent pas pleinement satisfaisantes. Les hapax

[11] C'est ce qui est fait pour le « *Qui-Fait-Quoi ?* » de 1993, calculé pour les Départements : « la mesure de proximité entre le texte à diffuser et le Département est le plus grand cosinus trouvé (le maximum des cosinus). Ce choix (plutôt qu'un indicateur de tendance centrale) résulte de la constatation qu'un texte peut intéresser un Département s'il est très lié à un seul de ses textes sans être lié aux autres. » (Sta 1993, p. 7)

(fréquence égale à un) sont survalorisés ; inversement, les contrastes dans la gamme de fréquences sont très (trop) vite « écrasés », si bien que le relief que l'on cherchait à rendre est en fait inopérant.

Pour récapituler tout cela, le modèle de l'espace vectoriel est perturbé par les facteurs d'échelle suivants :

- *grandeur* : textes longs et regroupements de textes multiples, fréquences extrêmes (fréquence de 1 et fréquences de l'ordre de 10 et au delà) ;
- *diversité* : variations sur la longueur des textes, sur les genres de textes, sur le nombre de textes associés à un profil (déséquilibre entre profils associés à un texte et profils associés à plusieurs textes), et écarts sur la gamme des fréquences d'apparition des mots.

Or, concrètement, ce n'est pas parce qu'une personne apporterait, pour la construction de son profil, deux fois plus de documents, ou des documents deux fois plus longs, que son profil devrait être deux fois plus volumineux ou deux fois plus influent ! Rien n'interdit en revanche que la représentation y gagne en finesse, en précision. Il faut donc se donner les moyens de sortir de cette logique cumulative, qui se manifeste par une trop grande sensibilité au volume des documents, par une incapacité à synthétiser l'information, par une ignorance des différents niveaux de lecture possibles (survol, lecture sélective...).

Indépendance des termes dans la représentation des textes

Un document est représenté par un vecteur, dans un espace dont chaque dimension représente un terme. Les axes, associés chacun à un terme, sont orthogonaux : autrement dit, une variation de la description par rapport à un terme n'a aucune incidence sur la description par les autres termes.

Le jeu des pondérations permet de faire peser davantage certains termes. Elle peut rendre compte de rapports de force. En revanche, la co-présence d'un petit groupe de termes en interrelation n'est jamais renforcée, ni la conjonction massive de termes peu informatifs dévalorisée. D'où deux phénomènes qui faussent souvent les calculs de similarités.

Premier cas de figure : un seul terme, grâce à sa forte pondération, permet de franchir le seuil de sélection. Il occulte les autres aspects abordés dans le texte, comme le point de vue spécifique du texte par rapport à la notion qu'il représente. Pour peu que ce soit un terme polysémique ou qu'il représente un concept entrant dans des pratiques très différentes, il sélectionne sur le même plan et mêle des textes (pour DECID, des profils) correspondants à des thèmes divergents.

Ce genre de résultat occasionne le même désagrément qu'un interlocuteur sûr de lui et peu écoutant, qui vous coupe la parole alors que vous entamez une explication, et, ayant saisi un mot au vol, part dans un commentaire ininterrompu ou dans un flot de propositions sortant complètement du champ de ce que vous deviez lui exposer.

Le deuxième cas de figure est celui d'une accumulation de termes dépareillés ou peu représentatifs qui, par leur quantité, permettent également de franchir le seuil de sélection[12]. Le critère de sélection est alors dépourvu de signification, la sélection est arbitraire, ou s'attache à des aspects tout à fait mineurs.

La pertinence, assujettie à la détermination d'un ordre pour des documents ou des systèmes en compétition

L'adéquation d'un document comme réponse à une requête est mesurée par un score, une valeur numérique. Les résultats se présentent donc comme une liste de documents ordonnés par similarité décroissante. Première conséquence : l'image de la pertinence est celle d'une succession stricte de documents, réduisant toute comparaison à une supériorité ou une infériorité, sans laisser place aux critères plus riches qui interviennent dans la réalité.

Cette modélisation pose également le problème de la détermination d'un seuil. Si en effet les documents, au fur et à mesure que la similarité décroît, sont de moins en moins pertinents, alors il arrive un stade à partir duquel les documents ne vaudraient plus la peine d'être présentés à l'utilisateur. Or rien ne permet de fixer ce seuil avec assurance, ne serait-ce parce que l'appréciation

[12] Ghislaine CHARTRON (Chartron 1988, §VIII.4.2) mentionne aussi ce phénomène : « il faut tout de même se méfier de ce que les termes ambigus et souvent inadéquats ne cumulent pas des occurrences trop fortes ».

de la pertinence varie d'un utilisateur à l'autre, et que la valeur de la similarité en elle-même n'est pas un indicateur stable d'un degré de pertinence.

La modélisation de la pertinence s'est développée sur ces bases. Pour un seuil donné, le système indique un nombre fixé de documents. Ces documents se répartissent entre ceux qui sont pertinents et ceux qui ne le sont pas (encore une modélisation contestable). On calcule alors la *précision* de la réponse, comme la proportion de documents pertinents parmi ces documents indiqués. Si de plus l'on connaît le nombre total de documents effectivement pertinents présents dans la base, et que le système devrait donc trouver et signaler à l'utilisateur, on détermine également le *rappel*, à savoir la proportion de documents retrouvés par le système parmi tous les documents pertinents. Evidemment, même en admettant que l'on puisse toujours déterminer la pertinence ou non d'un document, établir le nombre total de documents pertinents supposerait de passer en revue tous les documents de la base pour chaque requête, ce qui est infaisable sauf pour toute base de taille normale (une base qui ne comporte que quelques dizaines de documents n'aurait pas besoin d'un système de recherche) ; le rappel est donc toujours entâché d'une approximation mal maîtrisée. Enfin donc, en faisant varier le seuil, on obtient une série de couples de valeur (*précision*, *rappel*), qui permet de tracer un graphique. Et ce graphique est utilisé pour comparer des systèmes entre eux : le meilleur système est celui qui a sa courbe au dessus des courbes des autres systèmes (voir l'annexe sur TREC).

Tout l'effort s'est donc porté pour avoir un moyen de hiérarchiser des systèmes, sans faire grand cas des formes que prend la pertinence dans une recherche réelle.

Concrètement, pour un application comme la diffusion ciblée, le seuil a du être fixé très bas pour éviter de manquer un destinataire concerné. La conséquence est que l'utilisateur commence à examiner les propositions les unes après les autres, mais il ne sait pas vraiment quand arrêter ce parcours, car il est toujours possible qu'il reste une proposition pertinente parmi les similarités les plus faibles (par erreur, et parce que la pertinence n'a pas la figure d'un ordre séquentiel). De plus, si un mot suscite une série de mauvais rapprochements, alors l'utilisateur est obligé de faire défiler toutes ces suggestions erronées, avant de trouver un autre type de rapprochement, mieux motivé cette fois. Le fait que les mauvais rapprochement soient apparus groupés s'explique bien : c'est le cas d'un mot à forte pondération, donc suffisant à lui seul à obtenir une similarité élevée, et pris à contresens faute de contexte.

c) *Des erreurs*

Expérimentalement, on constate que la méthode génère du bruit (mauvaises propositions de destinataires) et du silence (oublis de destinataires) (Sta 1993). Les difficultés se traduisent sur plusieurs plans :

- *fiabilité* : la qualité des résultats est inégale, pour une même valeur de similarité, d'une paire de documents rapprochés à l'autre. La similarité telle quelle n'est pas un indicateur satisfaisant.
- *justesse* : entre les documents ciblés sur un même profil, la discrimination est souvent indécise, les « bons » rapprochements n'étant pas toujours les mieux notés. Autrement dit, l'ordre n'est pas pleinement significatif ; ou plus généralement, les documents *a priori* les plus intéressants doivent être mieux mis en valeur.
- *interprétabilité* : la simple lecture d'une valeur de similarité ne renseigne pas suffisamment sur la nature du rapprochement ; le sens à accorder au coefficient lui-même n'est pas clair (la présence de termes en commun entre deux documents ne détermine pas l'intérêt de leur rapprochement).

La définition des profils se heurte aux écueils suivants :

- *le profil « cumulatif »*, qui additionne les documents le définissant, sans organisation et sans synthèse, si bien que les points saillants sont noyés,
- *le profil « monopolarisé »*, pour lequel les rapprochements ne peuvent se faire que sur une seule facette, prépondérante ; (en l'occurrence, un ou deux termes, et pas toujours les plus pertinents).

Une analyse des facteurs d'erreur (Sta 1993) (Bommier & Lemesle 1996) montre que les erreurs ne sont pas indépendantes les unes des autres. Elle révèle aussi que chaque étape du processus est impliquée : la manière de définir les documents que l'on soumet au système ; le traitement, linguistique et statistique, qui leur est appliqué ; la mise en forme des résultats pour leur utilisation.

Parmi les points les plus délicats et les plus flagrants, on note[13] :

- la qualité informative et représentative du texte : une forme trop brève, liminaire, formelle ou / et évasive, ne fournit pas les données suffisantes pour un traitement satisfaisant ;
- l'importance cruciale de la terminologie utilisée pour caractériser le texte (notamment sa complétude et sa représentativité) ;
- la prise en compte de l'appartenance du texte à un genre, voire de l'existence d'une forme conventionnelle : les constructions et les formules propres à ce genre ne doivent pas être prises « au pied de la lettre » par le système et devenir source de contre-sens dans les rapprochements. Par exemple, dans le « je reste assez ouverte aux différents sujets... » d'une lettre de candidature, il n'est pas question d'orifices... D'autres cas du même acabit sont mentionnés dans (Bommier & Lemesle, 1996).
- la nécessité de (re)contextualiser les termes en les faisant intervenir de façon groupée[14], et sans dissocier les mots composés et les locutions ;
- la clarté des informations fournies pour l'interprétation des résultats : la difficulté à faire la part entre les différents points de vue sur un même thème (phase d'avancement) ;
- des spécificités de l'application de diffusion ciblée, liées au positionnement des destinataires dans l'entreprise.

[13] (Bommier & Lemesle 1996) rassemble toute une gamme d'exemples illustrant les anomalies résumées ici.

[14] Nous défendrons ici que c'est à la fois un moyen de lutter contre le bruit, en évitant des ambiguïtés, et contre le silence, en limitant l'incidence des différences de formulation grâce à des recouvrements partiels.

D. ENJEUX

1. Gains pour l'application DECID

a) Un contexte de mise en œuvre d'outils de Traitement Automatique du Langage Naturel

La diffusion ciblée est une application privilégiée des outils de traitements du langage naturel, utilisés pour analyser les textes. La première version a fait, à bon droit, un usage opportuniste immédiat des outils existants, avec un paramétrage standard. Cependant, les besoins en matière de TALN (*Traitement Automatique du Langage Naturel*) doivent être étudiés spécifiquement pour l'application[15], et intégrés dans une perspective textuelle.

b) Un gain dans la qualité des rapprochements

Les premières expérimentations ont donné des résultats inégaux ; les traitements linguistiques et statistiques conduisent souvent efficacement à des rapprochements intéressants, mais on observe aussi des erreurs de nature et d'impact variés, et qui pour certaines compromettent toute utilisation des résultats, voire découragent l'utilisateur. Le traitement de ces erreurs s'impose donc, pour la viabilité de la diffusion ciblée.

c) Une exploitation des résultats plus efficace

Les travaux à mener sur l'ergonomie et l'interface, notamment en ce qui concerne la présentation des résultats, ont un impact direct sur deux points fondamentaux. D'une part, les résultats sont mieux interprétés : le but est d'identifier rapidement et sans contresens les destinataires les plus concernés. D'autre part, une bonne perception du comportement du système permet une interaction souple et précise : l'utilisateur sait employer le système au mieux de ses performances.

2. Elargissement à d'autres contextes

a) Une définition plus précise du concept de diffusion ciblée

La diffusion ciblée a été lancée au départ comme une exploitation astucieuse de techniques existantes. En 1995, la diffusion ciblée n'est encore considérée que d'un point de vue très technique et immédiat, elle manque de repères théoriques ; or il lui faut se positionner, et bien percevoir ses atouts propres pour les valoriser.

Le développement d'une application (DECID) et la mise en place d'un service demandent de donner à la diffusion ciblée un fondement et une armature. Cela consiste concrètement à expliciter les besoins réels visés, considérer les questions que cela soulève dans le contexte d'une entreprise, préciser les contours d'usages de la diffusion ciblée, dresser les spécifications d'ensemble auquel le système doit répondre. En cernant les propriétés fondamentales de la diffusion ciblée, on repère les degrés de liberté, les potentialités, pour orienter la recherche sur l'application. Et à un niveau d'ensemble, on considère l'exportabilité du concept : faire la part de ce qui est lié à un contexte d'emploi particulier.

[15] Cette nécessité est pleinement reconnue : une méthodologie a été élaborée, dans le cadre du consortium européen GRAAL, pour établir les spécifications de toute application faisant appel à un Traitement Automatique des Langues (Herviou, Coch, Leblond 1995).

b) Des directions méthodologiques

Pour que l'application de diffusion ciblée puisse se poursuivre et être gérée en cohérence avec les évolutions apportées par la thèse, les principes suivis et leur mode de mise en œuvre sont à expliciter. Cela pourrait aussi contribuer à la reprise de certains acquis dans d'autres applications : tout laisse à penser que le secteur des traitements sur les textes en version électronique n'est au début de son développement.

3. Contribution à la Linguistique Textuelle

a) Rassembler les travaux sur les propriétés des textes

Non fédérées par une discipline, les recherches autour du texte sont difficiles à appréhender. On n'en saisit souvent dans un premier temps qu'un aspect. Il s'agit donc de faire un point, en s'appuyant sur de précédents travaux de synthèse, et en prêtant attention à tout ce qui se donne comme objet d'étude le texte et ses propriétés.[16]

Par la force des choses, ce tour d'horizon est limité, et l'inventaire peut tourner au bric-à-brac. L'objectif est donc d'en tirer un modèle, intégrateur et synthétique, qui guide la construction et l'exploitation d'une représentation des textes, en vue de calculs de similarités.

b) Dédramatiser le choix des formules mathématiques

Revenir au texte et à ses propriétés fournit les éléments pour situer et interpréter les formules existantes. Les modélisations mathématiques utilisées dans les systèmes documentaires font l'effet d'une jungle, pour un œil non averti : une faune d'équations terrifiantes se tapit entre les lignes des publications du domaine. Essayons de l'apprivoiser en s'en rendant familier : dans telle propriété de la formule (présence d'un paramètre, variations de croissance ou de décroissance, etc.), on reconnaît telle conception du texte. La description de la faune devient moins impénétrable, des espèces se dessinent, et même des nouveaux venus peuvent être introduits en harmonie, héritant des qualités éprouvées de leur congénères, cotoyant sans jalousie inutile leurs homologues.

c) La linguistique textuelle à l'épreuve de l'informatique

Le caractère exploratoire de cette thèse, outre la nouveauté d'un système de diffusion ciblée et de son principe, réside dans le fait que les concepts de Linguistique Textuelle ont été développés dans un cadre théorique récent et utilisés sur quelques autres applications, mais dans un petit nombre de contextes[17]. Or la sémantique des textes peut être rationnelle, mais en soi elle n'est pas formelle[18].

---

[16] L'analyse des textes appelle un traitement linguistique : mais quelle linguistique ? interroge Dominique LE ROUX ; elle adopte une perspective semblable à la nôtre, en concluant à la pertinence d'une linguistique *(i)* descriptive (plutôt que normative), *(ii)* non gouvernée par l'informatique, et *(iii)* textuelle. En ce qui concerne la linguistique textuelle, « un état de l'art sur les travaux déjà effectués s'avère indispensable » (Le Roux 1992, note 6, p. 7)

[17] En ce qui concerne la mise en œuvre informatique de la *Sémantique Interprétative* (ou *Sémantique Différentielle Unifiée*), dont les concepts théoriques ont été définis par François RASTIER (Rastier 1987) (Rastier, Cavazza, Abeillé 1994) :
- quelques travaux se fondent directement sur le cadre théorique proposé, dans certaines conditions particulières (expérimentation de la théorie) : (Rastier, Cavazza, Abeillé 1994) (Tanguy 1997) (Thlivitis 1998)
- d'autres en reprennent des éléments, mais de façon assouplie ou partielle (adaptation de la théorie) : (Beust 1998) (Cavazza 1991) (Nazarenko 1996) (Vaillant 1997)
- d'autres encore en tirent des principes directeurs pour guider la définition d'un traitement (compatibilité avec la théorie) : (Antoine 1994) (Assadi 1998) (Bonhomme & *al*. 1996) (Malrieu 1997)

Les questions que soulèvent l'implémentation de la *Sémantique Différentielle Unifiée*, en même temps que sa valeur pressentie, restent d'actualité (Habert 1995) (Nazarenko 1995) (Prié 1995), et sa mise en œuvre informatique dans un contexte nouveau constitue un thème de recherche à part entière.

[18] « si la théorie descriptive n'est pas calculatoire, elle peut cependant faire l'objet d'implantations informatiques –qui comprennent naturellement des calculs. Ce n'est pas à la linguistique de se formaliser par principe, mais à

L'enjeu est alors de retranscrire les concepts pertinents dans le cadre d'une application telle que la diffusion ciblée, et d'évaluer ainsi leur caractère opérationnel et implémentable.

l'informatique de proposer des formalismes de transcription expressifs et utilisables pour les applications. » (Rastier, Cavazza, Abeillé 1994, §II.5, p. 37)

E. ORGANISATION DES CHAPITRES DE LA THÈSE

Le rôle du présent chapitre, l'*Introduction* (chapitre I), est d'abord de définir les concepts essentiels sur lesquels se fonde cette recherche. En ce sens, sa lecture donne le cadre et prépare la compréhension de tous les aspects développés dans les chapitres suivants. C'est également en introduction que la thèse est replacée dans la perspective du projet industriel qui l'accueille : la dynamique induite par l'historique du projet, l'existant (réalisations, besoins) au moment du lancement de la thèse en 1995, les enjeux pressentis qui motivent les directions de recherche choisies.

La première étape consiste à *Définir la diffusion ciblée pour l'entreprise* (chapitre II). En effet, une part non négligeable de l'effort de recherche a été consacré à préciser des repères pour la mise en œuvre de la diffusion ciblée. La démarche n'est pas d'abord de partir bille en tête et réaliser une application élaborée, puis de voir comment elle convient. De fait, le tout premier prototype de DECID a déjà montré que l'application était faisable et prometteuse, mais aussi a été l'occasion de comprendre qu'il s'agissait de préciser un nouveau type d'application, irréductible à la diffusion sélective de l'information par exemple. Il faut donc commencer par examiner en quoi la diffusion ciblée répond à un besoin de l'entreprise, spécifier les caractéristiques attendues et évaluer leur importance. Cette étude profite de la « réalité du terrain » : grande attention a été apportée aux réactions des usagers potentiels, qui constituent déjà une première confrontation du concept de diffusion ciblée à la réalité du travail en entreprise. Les fruits du premier chapitre sont un ensemble construit d'informations, qui donnent les contours de l'application, et fournit une pré-évaluation à laquelle confronter les choix techniques.

La diffusion ciblée, avons-nous dit, est une application originale et innovante. Pourtant on entend beaucoup parler d'applications documentaires qui semblent s'appuyer sur des concepts très proches : filtrage, routage, agents intelligents qui sélectionnent de l'information, modèle de l'utilisateur, partage d'informations. *Un panorama des applications de documentation et d'information mettant en œuvre des profils* (chapitre III) est nécessaire pour clarifier ce que recouvrent ces différents termes, et retirer les enseignements d'expériences en partie comparables.

Le pari de la présente recherche est dans l'apport de la linguistique textuelle à l'application de diffusion ciblée, compte-tenu du rôle central des textes dans cette application. Au plan technique, la qualité des propositions d'un système de diffusion ciblée repose sur la prise en compte du caractère non seulement linguistique mais aussi textuel de toutes les données, profils et documents. Le chapitre IV réunit des *Eléments pour une définition de la textualité*. Il consigne d'abord méthodiquement tout un faisceau de propriétés textuelles mises en évidence dans des courants de recherche diversifiés. C'est à partir de cette matière, et en se plaçant dans le contexte de documents écrits à dominante scientifique et technique, que sont définies quatre facettes textuelles. Leur rôle est de synthétiser les dimensions à prendre en compte dans la conception des traitements automatiques et de l'interface. Cette étude sur les propriétés constitutives du texte se prolonge par une réflexion sur ses rapports avec la machine, en particulier en termes de compréhension (la machine peut-elle / doit-elle comprendre les textes ?), de représentation (la formalisation apporte-t-elle, retire-t-elle quelque chose au texte ?), d'interprétation (le rapport au texte n'est-il pas à penser comme une lecture ? et comment concevoir cette lecture ?). La question de la pertinence appelle également des jalons : la diversité des facteurs qui entrent en jeu est telle qu'il faut, pour s'y repérer, en dégager des lignes de force, et identifier les quelques modèles conceptuels sous-jacents.

Ces premiers chapitres ont préparé le terrain des propositions et réalisations dans le cadre de l'application DECID. Si l'on considère les traitements à effectuer dans leur déroulement chronologique, vient d'abord l'étape de *Constitution et codage du corpus* (chapitre V). Le choix du corpus servant à la définition des profils est évidemment décisif : sur quels critères le baser ? Les textes d'Action (descriptifs des programmes de recherche, écrits à l'intention de la Direction) sont à première vue de bons candidats, mais cela ne dispense pas d'une exploration méthodique des autres corpus envisageables. Côté documents soumis au système, la nature des textes à traiter n'est pas choisie, mais il serait bon d'en prévoir certaines caractéristiques. Quoique l'ensemble des textes que

doit traiter le système soit très ouvert, leur dénominateur commun n'est pas seulement d'être une suite de mots : la recherche d'un modèle de structuration générique part de cette conviction pour identifier, par delà le détail de certains formats (des DTD SGML) et l'indigence d'autres (du texte ASCII), des rapports structuraux fondamentaux et significatifs, à reconnaître, retenir, et utiliser, dans une lecture non plus graphique (caractères) mais textuelle du texte. Car si l'on commence le traitement en démembrant le texte en un paquet de mots - chaînes de caractères, bien des facettes textuelles sont manifestement bafouées, et leurs significations perdues.

Un chantier important s'ouvre ensuite : la *Détermination d'unités de traitement* (chapitre VI). Le concours des outils de Traitement Automatique du Langage Naturel est ici précieux, à condition d'être guidé par une perspective textuelle : auxquels de ces outils faire appel ? et comment exploiter leurs résultats d'analyse ? Une relecture de l'expérience de l'utilisation d'un outil évolué d'indexation automatique est instructive. Dès lors que les unités ne sont plus données, fournies par un mécanisme d'indexation que l'on a sous la main, mais à construire, le besoin de repères généraux se fait sentir, pour s'orienter : quelles sont au juste, et pour notre cas, les fonctions de l'indexation ? quelles sont les stratégies qui se présentent ? L'objectif est aussi de prendre acte, concrètement, des avertissements de la linguistique textuelle : notamment, que « le global détermine le local », et encore que les unités sont construite dynamiquement, en contexte.

L'analyse du texte en unités donne une représentation pour pouvoir le confronter à d'autres textes en présence : c'est la *Caractérisation d'un texte dans un corpus*, et l'étude de la manière dont on peut aller *du quantitatif*, des décomptes d'unités, *vers le qualitatif*, des indications susceptibles de faire sens (chapitre VII). Le corpus joue ici un rôle constitutif, il matérialise un univers intertextuel : tout un éventail de critères (représentativité, homogénéité, etc.) explicitent certes des principes pour sa constitution (dans la mesure des données dont on dispose ou qui sont accessibles), mais aussi et surtout des conditions de signification, sans lesquelles les analyses effectuées sont ininterprétables. Pondérations, similarités ou distances : les formules abondent, sans pour autant que soient clairement perceptibles les conceptions du texte qu'elles reflètent. Pour pouvoir choisir ou élaborer une formule porteuse de signification, et qui soit en accord avec les facettes textuelles, il faut se donner les moyens de comprendre la composition des formules existantes, et les comportements intéressants ou regrettables observés. Ce chapitre se clôt sur des pistes et de la matière de travail pour mettre au point le nouveau moteur de calcul des rapprochements document - profil, dont la réalisation doit suivre celle du moteur de construction et d'attribution des unités : il présente donc une phase plus exploratoire que ne le font les autres chapitres.

Une dernière étape cruciale dans une application de diffusion ciblée est la présentation des résultats, à savoir l'indication de personnes potentiellement concernées par un document. L'affichage sous forme de liste (ordonnée par un score de pertinence), linéaire et figée, montre ses limites d'un point de vue pratique, et est d'ailleurs discutable du point de vue d'une sémantique textuelle. Il s'agit de passer *De la pertinence au parcours interprétatif*, où l'utilisateur, par le biais de *l'interface*, navigue efficacement parmi les propositions du système, et se les approprie pour en tirer le meilleur parti, dans sa situation personnelle du moment (chapitre VIII). C'est l'occasion de faire le point sur l'interface dans son ensemble, et les fonctionnalités « textuelles » innovantes dont la munir. Le choix du texte comme mode de communication homme-machine (puisque c'est le mode d'interrogation, et le mode d'alimentation de la base des profils) est ici plus amplement discuté.

La *Conclusion* (chapitre IX) souligne les apports de la thèse, dans les différents domaines finalement rencontrés : les enseignements tirés de l'élaboration d'une nouvelle application et de sa mise en œuvre en entreprise, les avancées proposées dans le champs des moteurs de recherche d'information sur le texte intégral, la place des outils de Traitement Automatique des Langues dans une perspective textuelle, le positionnement et l'apport de cette recherche dans le contexte des multiples courants de Linguistique textuelle. Une attention particulière est portée à la Sémantique Interprétative de François Rastier. Puis vient le temps d'évoquer les pistes qui sont apparues, et dont on pressent l'intérêt et l'importance.