CHAPITRE IX

Conclusion

Table des matières du Chapitre IX

A. UNE RELECTURE DU CHEMIN PARCOURU

1. La diffusion ciblée : réalisme et potentiel

a) Choix techniques et réalités humaines

Malgré les annonces enthousiastes et massives sur la *société de l'information*, le doute peut se faire jour quant à la réalité des apports de nouveaux outils dans les pratiques quotidiennes. Une étiquette ronflante a intronisé les *Nouvelles Technologies de l'Information et de la Communication*. Mais qu'en est-il au juste de cette information personnalisée, sélective, intelligente, collaborative ? Très vite, deux facteurs de déception apparaissent. Soit on retrouve, masquées, rhabillées au goût du jour, de bonnes vieilles techniques (ex. mots-clés). Ces techniques, conçues, éprouvées et affinées dans un certain cadre, perdent en clarté et en efficacité, dans cette transposition abrupte à un contexte étendu. Autre illusion pas toujours entièrement dissipée, celle de concepts novateurs, comme le *filtrage*, dont l'évidente pertinence se fissure dans la confrontation aux réalités non seulement techniques, mais aussi sociales (on n'élimine pas de l'information sans que cela interroge la place de l'expéditeur, l'activité critique, la préservation d'une ouverture d'esprit, etc.).

L'application de diffusion ciblée, à travers sa mise en œuvre à la Direction des Etudes et Recherches d'EDF, est soumise à l'épreuve des faits. Elle est déployée sur une échelle significative (plus d'un millier de profils), dans la durée (plusieurs années), et s'intègre aux pratiques professionnelles courantes. Qu'est-ce qui permet cette exploitation opérationnelle ? La réponse se trouve en grande partie dans les trois piliers identifiés pour la diffusion ciblée : l'*automatisation* des traitements, l'utilisation directe des *textes* (pour la définition des profils et pour la caractérisation des documents), la constitution d'une *base* de profils représentative et contextualisante.

Ces trois atouts sont fondamentaux, mais ne résolvent pas tout. L'expérience de DECID montre que la mise en place d'une application de diffusion ciblée suppose une attention particulière aux équilibres organisationnels, aux relations interpersonnelles, au quotidien des pratiques professionnelles. La circulation de l'information ne se décrète pas ; et la forme des requêtes (un texte plutôt que des mots clés) va à l'encontre d'habitudes documentaires profondément ancrées et qui ont également leur pertinence. L'application doit donc rassurer en se tenant à une déontologie sans faille, et en cultivant la communication et le dialogue avec ses utilisateurs. Il faut dissiper le mystère qui pourrait régner autour de la constitution des profils, expliquer la force des requêtes textuelles, suggérer de premières occasions d'utilisation. Il faut également souligner en quoi elle n'est pas un substitut (aux services documentaires classiques, aux relations informelles et conviviales) mais un outil permettant de maîtriser et bénéficier au mieux d'une circulation de l'information plus vaste et plus intense, et de valoriser la richesse des compétences présentes dans l'entreprise.

b) Noyau : le calcul de liens texte - textes

La diffusion ciblée a été envisageable à partir du moment où l'on s'est doté d'un outil de calcul de similarités entre textes. A partir de cette technique fondatrice s'ouvre un large éventail d'applications, de l'hypertexte dynamique à l'aide à l'interprétation d'un passage obscur par la recherche de passages analogues (méthode herméneutique connue, dite des passages parallèles).

La circulation de l'information dans l'entreprise, par l'aide à la recherche d'interlocuteurs et de destinataires, n'est donc qu'un cas particulier d'utilisation de l'outil développé pour la diffusion ciblée. Son potentiel est à la mesure des corpus et flux de textes électroniques, dans lesquels il fraie des parcours d'exploration, de proche en proche.

2. Adopter une perspective textuelle sans perdre de vue la valeur des techniques existantes

*a) L'*Information retrieval *et les systèmes documentaires*

Les moteurs de recherche (sur Internet) illustrent toute la classe des systèmes de recherche sur le texte intégral. DECID, dans sa version initiale, est proche de ce paradigme : lui aussi indexe les chaînes de caractères, et procède à la sélection de résultats de recherche par le biais d'une modélisation vectorielle. Cette approche se révèle efficace mais limitée ; les propositions de la thèse pour améliorer la qualité des résultats et la puissance de DECID vont donc dans le sens d'une meilleure intelligence du texte.

La première opération est habituellement de réduire le texte en miettes : il n'en reste que des 'mots', désarticulés. Toute trace est perdue des affinités syntagmatiques, des voisinages, des rapports entre parties du texte. La mise au point d'une représentation structurée des « entrées » du système est donc une première étape indispensable pour envisager un traitement un tant soit peu textuel. Pour être utile, le modèle de document proposé (une DTD SGML, appelée *Corpus*) intègre les formes de structuration fondamentales qui apparaissent comme noyau des descriptions des structures textuelles. On y retrouve par exemple l'alinéa (à la base des paragraphes), le rapport entre un titre et le développement qu'il introduit, l'organisation de plusieurs mentions ou passages sous forme de liste. Il s'agit de structures élémentaires à la fois génériques (à l'œuvre dans la plupart des textes) et significatives, jouant un rôle dans la lecture et l'interprétation. Et pour que le modèle défini soit utilisable, des modules de conversion ont été développés et testés. Ils réalisent une interprétation de formats standards, y compris les moins structurés (texte simple, HTML), dans les termes du modèle de document proposé. La voie est préparée pour l'interprétation automatique de structures textuelles dans toutes sortes d'autres formats.

Un des apports majeurs de ce travail de thèse, est l'introduction des deux étapes –*construction* et *élection*– entre l'analyse des textes, et la détermination de sa représentation interne pour le calcul. Les moteurs de recherche suivent les tactiques classiques d'indexation automatique : le texte est analysé linéairement (mot à mot, phrase à phrase, ligne à ligne), et dès qu'une unité est reconnue (un mot qui ne fait pas partie de la liste exclue des mots grammaticaux, un groupe nominal, une manifestation linguistique d'un descripteur de thesaurus), l'unité est directement affectée à la représentation du texte. Des indicateurs numériques, comme la fréquence ou les pondérations, modulent ensuite la représentation, mais celle-ci reste liée aux unités extraites initialement. Ce qui est proposé dans ce mémoire, consiste toujours à partir de l'analyse locale de chaque texte –il serait difficile de faire autrement–, mais d'attendre d'avoir une vue globale de l'ensemble des résultats d'analyse pour *construire* de véritables unités descriptives, attestées et adéquates pour les textes du corpus. Ce faisant, on évite à la fois les hésitations et les erreurs issues d'une analyse locale, par définition amnésique et à courte vue, (cf. le principe de l'*apprentissage endogène*), et aussi on se donne le moyen de définir des unités thématiques et contextuelles dont la manifestation n'est pas de l'ordre du mot, mais concerne des zones de localité comme le paragraphe ou même le texte dans son entier. La seconde étape, l'*élection*, est la composition d'une représentation globale de chaque texte, à partir d'unités descriptives sélectionnées et qualifiées. Là encore, un point de vue global gère l'équilibre et la représentativité de la description. En somme, les termes d'index des moteurs de recherche superposent et confondent encore *unités élémentaires* (issues d'une première analyse locale), *unités descriptives* (utilisables pour la description), et *unités caractérisantes* (composant la représentation des textes) : c'est en les distinguant que peut être gagné un 'recul' textuel.

Les mots, pour les moteurs de recherche, ne comptent qu'à la mesure de leur poids. La seule information sur un mot qui caractérise un texte, ou la seule comparaison possible entre deux mots, tiennent à une valeur numérique : ce mot-ci a un poids fort / faible dans ce texte ; ce mot-ci a un poids beaucoup plus fort que ce mot-là. Ce mode de raisonnement est si fortement présent, que l'on voit quelquefois proposé de pondérer les termes de sa requête par des chiffres (mais sans que la signification opératoire de tel ou tel choix de chiffre soit clairement établie), voire que l'on suggère de répéter un mot, sans autre raison que d'indiquer son importance en gonflant son nombre

d'occurrences. Que la machine se base sur des chiffres pour opérer ses calculs, soit ; mais qu'elle réduise d'emblée toute information en la fondant dans un seul indicateur numérique, et qu'elle impose les chiffres comme moyen d'interaction sans permettre à l'utilisateur de maîtriser la signification de ses interventions possibles, cela n'est guère acceptable. Toute l'interface de DECID est conçue pour permettre à l'utilisateur d'avoir une vue intelligible du traitement, de percevoir l'impact de tel ou tel choix, et de concevoir les actions possibles sur ce qu'il maîtrise : les données (texte soumis) et les paramètres de traitement (clairement identifiés et documentés). Côté utilisateur, donc, pas de chiffres obscurs ou illusoires. Côté machine, le codage des informations associées à une unité (un mot) fait place aussi à des informations qualitatives : les *catégories* sont prévues pour garder une information dans sa nature initiale, quantitative ou qualitative, et d'en fournir la traduction la plus adaptée pour chaque traitement.

Depuis quelques années, une kyrielle de travaux dénoncent la conception réductrice de la pertinence dans les moteurs de recherche : non, la pertinence ne se réduit pas à un score qui mesure la similarité des sujets traités dans la requête et le document ; non, la pertinence ne se laisse pas penser comme une propriété (fixée par une requête), selon laquelle les documents se rangent l'un derrière l'autre, du plus pertinent au moins pertinent. La pertinence fait intervenir de multiples facteurs, et ne se fixe pas par rapport à une requête, indépendamment d'un utilisateur, d'un lecteur, d'une situation. Nous proposons donc un nouveau modèle de pertinence, la *pertinence différentielle*, en remplacement du modèle de la *pertinence linéaire*. Sa mise en œuvre, au niveau de l'interface, consiste non pas à donner la liste des résultats, mais à guider l'utilisateur de façon optimale pour qu'il repère les propositions qui l'intéressent, et construise son propre cheminement à travers la sélection et les indications du système.

b) Les Traitements Automatiques des Langues

Les modules qui nous intéressent ici sont des analyseurs morphologiques et/ou syntaxiques. En effet, la forme et la disposition des unités linguistiques jouent un rôle dans la construction du sens. La morphologie rend compte d'interrelations paradigmatiques. La syntaxe entre en jeu dans la construction des unités descriptives, elle est même constitutive des unités descriptives à ancrage syntagmatique. En revanche, les traitements sémantiques existants sont en contradiction avec l'approche différentielle et textuelle choisie ici (recours à des lexiques généraux hors contexte, ou de spécialité et inadaptés à la diffusion ciblée ; calcul du sens, et compositionnalité).

Toutefois, l'architecture proposée dans ce travail fait place non seulement à des unités extraites par une analyse automatique robuste du corpus (*indexation par érosion*), mais permet également l'apport d'unités de référence définies extérieurement (*indexation par dotation*), par exemple pour la description fine de genres textuels. Pourraient donc être exploitées certaines ressources de type lexique sémantique. L'indexation par érosion est cependant privilégiée : elle fournit l'essentiel des unités, et est toujours effectuée, alors que l'indexation par dotation est optionnelle et dépend de l'existence d'unités prévues pour le contexte considéré. Dans tous les cas, c'est le corpus qui a le dernier mot, en valorisant les propositions d'unités descriptives opportunes, et en les utilisant comme unités caractérisantes ; les unités descriptives correspondant mal aux textes en présence sont négligées et n'interviennent pas dans la caractérisation.

D'une façon générale, l'étude de l'apport des différents outils de Traitement Automatique des Langues nous conduit à adopter une voie médiane, nous détachant des querelles sur la pertinence ou non de telle ou telle opération. Ainsi, il y a les partisans de la lemmatisation (« que le mot soit au singulier ou au pluriel, c'est toujours la même unité lexicale, et potentiellement le même concept »), et les partisans de la non lemmatisation (« il faut se garder de confondre les différentes formes d'un mot : dans une proportion significative de cas, elles correspondent à des usages et des sens différents »). Et de fait, les arguments de chaque bord sont des arguments de valeur, ceux des lemmatisants comme ceux des non lemmatisants. Ce que prévoit la nouvelle architecture proposée ici, c'est d'utiliser les propositions d'un outil de lemmatisation au moment de la construction des unités descriptives : et c'est pour chaque unité, sur un examen global de ses usages dans le corpus, que l'opération de lemmatisation est validée ou suspendue. Autrement dit, sont rapprochées les formes que le corpus convie effectivement à réunir (elles sont employées dans des contextes similaires, et les

lier permet des descriptions plus économiques tout en restant satisfaisantes), et sont maintenues distinctes les formes qu'il fausserait la description de confondre.

Un malentendu pourrait venir du fait que, pour les besoins de l'application de diffusion ciblée, le traitement doit pouvoir être entièrement fait sans l'apport d'aucun module extérieur : il s'agit d'un mode de fonctionnement autonome de l'outil. Cela ne signifie pourtant pas le rejet des outils de Traitement Automatique des Langues existants : leur intervention est conçue de façon modulaire, telle procédure du traitement de base pouvant être facilement remplacée par l'appel à tel outil, plus travaillé et plus performant. Si dans un premier temps, donc, le développement informatique de la nouvelle version du moteur de DECID s'est fait sans le recours au moindre outil externe, la volonté est réaffirmée d'une réalisation ouverte, permettant l'insertion de modules spécialisés.

3. Définir une sémantique textuelle pour la mise en œuvre informatique : une interprétation respectueuse et libre des théories existantes

a) Le défi de la généralité

La vision très large des travaux sur les textes et la textualité, et que nous avons voulue comme point de départ de cette étude, peut surprendre et inquiéter. Elle nous mène de linguistique en philosophie, des sciences de la documentation à l'informatique. Elle n'élude pas la considération d'approches qui ont donné lieu à des implémentations manifestement inadaptées à la diffusion ciblée (parce que centrées sur la phrase, trop lourdes à mettre en œuvre, interactives et non transposables à un traitement automatique). Elle rencontre des théories qui se sont finalement spécialisées dans la description d'une forme de texte (par exemple le récit), apparemment absente des corpus concernés par la diffusion ciblée. Mais c'est en allant au-delà de cette diversité que l'on peut espérer récolter une bonne part des propriétés associées au concept de *texte*, chaque approche mettant en valeur des aspects complémentaires.

Cette recherche nous place pourtant dans un équilibre délicat. Poursuivant des propriétés suffisamment générales pour guider le traitement de textes très diversifiés –nos conditions sont larges : textes écrits, en français, à dominante scientifique et technique–, serions-nous à la quête d'universaux textuels ? Un tel objectif serait évidemment démesuré... et sans doute illusoire. D'où la proposition de quatre facettes textuelles, définies comme repères méthodologiques pour la conception de traitements automatiques, mais ne revendiquant pas une portée théorique fondamentale. Ces quatre facettes sont : *(i)* la langue, comme matériau du texte ; *(ii)* son organisation interne, close et orientée ; *(iii)* l'intertextualité sur laquelle il se profile ; *(iv)* la lecture, constitutive du texte et de sa sémantique.

b) Une implémentation dans l'esprit de la Sémantique Interprétative

La théorie qui guide ce travail comme un fil d'Ariane, est la *Sémantique Interprétative*, synthétisée par François Rastier (Rastier 1987) (Rastier, Cavazza, Abeillé 1994), à la suite de Coseriu, Greimas, Pottier. Pour autant, les modules informatiques conçus et réalisés dans le cadre de la thèse ne sont pas l'implémentation de la Sémantique Interprétative.

Prenons plutôt pour repère l'étude de Yannick Prié (Prié 1995), qui indique la voie d'un passage indirect de la théorie à l'implémentation. Le rattachement à la théorie est une fidélité à sa « trame conceptuelle résistante », à savoir les points généraux fondamentaux de la théorie. En ce qui concerne la Sémantique Interprétative, (Prié 1995) suggère : la détermination du local par le global, qui se manifeste en particulier dans la non compositionnalité du sens ; la dynamique de la construction de l'interprétation, en contexte (par exemple par des opérations d'assimilation ou de dissimilation) ; le non-déterminisme de l'interprétation, qui admet pour un texte une pluralité de sens (multiplicité qui n'est ni unicité, ni infinité) ; le rôle central joué par le concept d'isotopie, à la base de toute construction d'unités sémantiques d'ordre supérieur, et présidant à la plupart des opérations interprétatives. Une fois une telle trame conceptuelle admise, l'implémentation sélectionne et met en

œuvre ce qui est pertinent pour l'application visée, dans le respect de la trame conceptuelle, et en fonction de la faisabilité informatique.

Le présent travail s'inscrit dans une logique similaire, et choisit ses appuis fondamentaux dans une telle trame conceptuelle résistante de la Sémantique Interprétative. Il est en accord avec les principaux points retenus dans (Prié 1995). Il rencontre aussi d'autres dimensions importantes de la Sémantique Interprétative, et éclaire d'un jour nouveau les modélisations possibles de concepts de la théorie.

La *détermination du local par le global*, de prime abord, ne semble pas compatible avec l'architecture compositionnelle des calculs. Ici, la distinction faite entre unités élémentaires, unités descriptives et unités caractérisantes rend possible le comportement attendu. La construction au niveau global des unités descriptives certes dépend des résultats de l'analyse locale en unités élémentaires ; mais elle est en mesure de réajuster toutes les décisions locales. *In fine*, les unités de description sont uniquement des unités construites avec le recul d'une vue globale sur les textes entiers et le corpus. Sur la question de la préséance du global, on peut aussi mentionner l'organisation des résultats selon une pertinence différentielle, qui donne d'emblée à l'utilisateur la vue globale lui permettant ensuite d'organiser son parcours dans les résultats, et de leur donner sens par rapport à l'ensemble.

Il y a plusieurs aspects de DECID qui rendent compte d'une *dynamique de l'interprétation*. Tout d'abord, le fait que les unités descriptives ne soient pas considérées comme *données* mais soit *construites*. Le mode principal de construction des unités descriptives s'opère à partir d'un corpus : les unités sont relatives au contexte textuel, le système s'adapte dynamiquement à l'évolution ou aux variations d'un corpus à l'autre. Il y a un bon opportunisme de la description sémantique, contraire à toute prétention d'universalité et d'absolu : l'unité sémantique est construite parce qu'elle est efficace pour la description, qu'elle se montre pertinente et économique, productive. La dynamique de l'interprétation se retrouve aussi dans le principe du calcul de rapprochements texte - texte. Car il ne s'agit pas de calculer le sens du texte, de le comprendre (la machine ne peut comprendre un texte, elle ne fait que le transformer rapidement et massivement pour le présenter de façon suggestive). Le calcul de rapprochements texte - texte propose des cheminements interprétatifs ; les projections d'un texte sur un autre (mise en œuvre dans l'interface) sont une invitation à découvrir de multiples lectures d'un même texte, au gré des points de vue suscités par les autres textes. On est donc loin d'une saisie (capture) du sens (unique).

La pertinence différentielle, préférée à la pertinence linéaire, pour la présentation de la sélection opérée par le calcul, rend à l'utilisateur son plein rôle d'interprète des propositions du système. Le système ne peut dire à la place de l'utilisateur ce qui est intéressant pour lui, il y a bien un *non déterminisme* de la valeur que peut prendre chaque suggestion ; en revanche, le système peut organiser, donner des indications, développer interactivement telle piste et non telle autre, et accompagner ainsi l'utilisateur dans son travail de dépouillement et d'appropriation des résultats.

La famille d'unités descriptives que sont les Communautés est certainement une formalisation à la fois proche et très innovante des structures descriptives de la Sémantique Interprétative. Les Communautés prennent appui sur le concept fondamental d'*isotopie* sémantique : elles sont définies à partir des unités contribuant à la manifestation d'une isotopie, par réitération d'un sème dans une zone de localité du texte. Une Communauté induit un sème (ou plusieurs, peu importe car les sèmes sont des unités relatives), comme ce qui est en commun entre ses membres. Le sème n'est bien ici, comme ce qu'indique la théorie, qu'une reconstruction *a posteriori*, et en contexte. Il peut être désigné par un nom, mais cette lexicalisation est facultative et n'est jamais ce qui définit le sème. Les sèmes de la théorie ne sont pas non plus, sous couvert des nécessités de l'implémentation, affectés à des unités hors contexte, indépendamment les unes des autres (atomisation du sens), enregistrées dans un lexique en soi figé et limité à ce qui a été déclaré, à ce qui est déjà connu. La description n'est donc plus artificiellement centrée sur l'unité lexicale, mais ancrée dans les manifestations textuelles du sens.

La voie choisie ne traduit donc pas directement les structures de description qui se prêtent plus naturellement à l'organisation d'un lexique : *taxèmes*, *domaines*, *dimensions* –même si les

taxèmes ont des affinités avec les Associations, et les domaines avec les Communautés[1]. Les classes de la Sémantique Interprétative (taxèmes, domaines, dimensions) sont dessinées par la reconnaissance de sèmes spécifiques (distinguant les éléments) et de sèmes génériques (regroupant en classe). Notre modélisation perd l'opposition *générique / spécifique*, mais cette opposition n'est que le statut que prend un sème relativement à un élément (un sémème), et ne concourt pas à la valeur sémantique propre du sème. En revanche, l'opposition *inhérence / afférence* trouve un écho dans la structure interne des Communautés, dans la distinction entre un *noyau* et un ensemble de *satellites*. Les unités descriptives dans le noyau de la Communauté ont un lien fort et stable au sème de la Communauté, c'est une relation de type inhérence ; les satellites ont une relation au sème beaucoup plus sensible au contexte, elle est plus afférente. La Communauté ne reflète en fait que les principales afférences, celles les plus régulières[2]. Cette description nous paraît plus proche de l'intuition d'inhérence ou d'afférence d'un sème[3], que celle qu'oblige une retranscription littérale de la structure des classes sémantiques et des contraintes sur l'attribution des sèmes. (Tanguy 1997), pour avoir expérimenté cette retranscription en intelligence parfaite avec la théorie, constate lui-même que, avec le passage à l'implémentation (logiciel PASTEL), la relation d'afférence devient surtout un moyen de contourner des contraintes formelles et d'exprimer des relations transverses (*ibid*., §II.3.5, p. 63 sq.) ; les sèmes afférents se comportent comme des alternatives à un sème inhérent[4], et tout le jeu interprétatif permis par PASTEL repose d'ailleurs en grande part sur le retournement possible de la structure (*ibid*., §IV.6, p. 137 sq.), le passage d'un point d'équilibre à un autre point d'équilibre, en promouvant par exemple un sème afférent en sème inhérent, ce qui peut avoir pour effet de chasser le précédent sème inhérent vers les sèmes afférents.

Les Communautés sont déclinées en trois types d'unités en fonction de trois paliers de localisation (*période*, *paragraphe*, *texte*) : c'est là une liberté prise par rapport à la théorie, qui s'en tient à une conception unifiée de l'isotopie, quelle que soit sa portée. Cette différence de vue est sensible mais ne nous met pas en porte-à-faux. D'une part, l'attaque de (Rastier 1987, §V.1.2, p. 110) contre une typologie des isotopies par paliers d'analyse s'était portée contre (Eco 1979), pour une distinction essentiellement entre isotopies phrastiques et transphrastiques : or le premier palier que nous proposons, la période, a une définition moins rigide et grammaticale que la phrase. D'autre part, le concept de Communauté rend compte de la généralité du concept d'isotopie, son caractère fondamentalement unifié n'est pas ignoré ; c'est maintenant de l'expérimentation qu'il faut attendre des échos sur le bien-fondé ou non, heuristiquement, de détailler ainsi trois variétés d'isotopies.

Plus largement, la prise en compte, dès le codage des textes, puis dans la construction des unités, de plusieurs paliers de description (l'unité élémentaire, la période, le paragraphe, le texte, voire l'intertexte d'une 'boîte' dans un 'rangement') est conforme à l'esprit de la Sémantique

[1] Sans doute aussi que l'étude de thèmes ne se confond pas avec celle de langue et de ses structures :
« En analysant le champ lexical des sentiments, nous ne postulons pas que ce champ soit uniforme, ni qu'il soit une unité de langue. Il contient sans doute plusieurs taxèmes. Il ne constitue pas un domaine délimité par l'incidence d'une pratique sociale. Il s'agit donc d'un regroupement *ad hoc*, convoqué par la pratique descriptive en cours. » (Rastier 1995a, §II.2, p. 236)

[2] Ceci est à rapprocher de l'observation de (Dupuy 1993), qui projette une grille sémantique sur un texte (les sèmes sont des 'codes'), dans le but de repérer des zones de récurrences sémantiques :
« Seule certitude : tous les vocables pourvus d'un même code sémantique appartiennent à une même isotopie, précédemment définie dans la grille sémantique ; cependant, certains vocables, qui ne sont pas affectés de ce code, peuvent aussi participer, de façon dénotative ou connotative, à cette même isotopie ; autrement dit, l'énumération des vocables supportant un même code ne suffit pas à tracer de façon exhaustive le développement isotopique. » (Dupuy 1993, pp. 341-342).

[3] L'opposition inhérence / afférence ne reflète pas une différence de nature entre les sèmes, mais a des échos dans l'établissement des parcours interprétatifs :
« Les traits réputés inhérents ne sont aucunement donnés, ils sont simplement hérités par défaut du type lexical. La différence entre traits inhérents et afférents n'est donc pas une différence de nature, mais de complexité des parcours interprétatifs qui permettent de les actualiser. » (Rastier 1994, §1.2.b, p. 329)

[4] D'ailleurs, toutes les occurrences d'un même sémème ont les mêmes sèmes afférents... Le rôle de l'afférence est ici davantage d'assouplir la structure que de refléter une propagation locale des sèmes. Ludovic TANGUY a bien conscience de ce biais, il n'hésite pas à y reconnaître un « détournement » du phénomène d'afférence (Tanguy 1997, §IV.4.3.3, p. 129).

Interprétative, qui proclame l'*irréductibilité de chaque palier*, et en particulier dénonce la réduction qui consisterait à rabattre le texte sur un palier 'inférieurs'. C'est justement une sémantique textuelle, en ce que le texte est reconnu dans son intégrité et sa nature propre, en relation avec les ordres lexicaux et syntaxiques, mais non défini ni déterminé par eux. La sémantique est unifiée notamment grâce au phénomène d'isotopie, qui traverse les différents paliers.

Enfin, le parti pris d'une sémantique autonome convient parfaitement à la mise en œuvre informatique de DECID, qui n'exploite pour toutes données que des textes. La Sémantique Interprétative est centrée sur les textes, tout en étant capable de prendre en considération les incidences externes qui participent aussi à la constitution de la textualité : l'auteur, le lecteur ou le lectorat, l'environnement du domaine concerné et des pratiques sociales dans lequel le texte prend place. Elle considère en effet ces trois instances à travers leur trace dans le texte, les trois pôles intrinsèques du texte (Rastier 1996b) : impression référentielle, foyers énonciatifs et interprétatifs, à l'identification desquels participe le genre du texte. Autrement dit, *la sémantique est autonome, mais pas autarcique*. Dans les corpus considérés par DECID, l'organisation du travail et les pratiques professionnelles dans le centre de recherche façonne le texte et ses lectures. Les *rangements* et les *piles*, définies dans le modèle *Corpus* de codage des textes, sont un moyen d'associer au texte électronique des données de son entour pertinentes pour son interprétation. Et dans la typologie proposée d'unités descriptives, l'élargissement des zones, allant de pair avec la relâche de contraintes syntagmatiques, accentue le fait que les Communautés ne soient pas des formations purement linguistiques. Les Communautés brouillent les axes syntagmatique et paradigmatique : la manifestation d'une isotopie a un déploiement syntagmatique, mais le parallèle de ses occurrences dans d'autres localisations et dans d'autres textes est d'ordre paradigmatique. Par leur formation et leurs effets sémantiques, les Communautés reflètent, dans la langue, des réalités socio-culturelles et des connaissances encyclopédiques.

B. ET POUR POURSUIVRE : UN CHOIX DE PISTES ET DE QUESTIONS

1. Dans la lignée des développements amorcés pour DECID : caractérisation de textes et calcul de liens texte - texte

a) Expérimentation de la nouvelle architecture

Une toute première version d'un moteur d'analyse et de caractérisation de textes a été réalisée. Ce moteur est conforme à l'architecture en *unités élémentaires*, *unités descriptives* et *unités caractérisantes*. L'objectif de cette première version est de concevoir l'architecture informatique du programme (notamment délimitation et constitution d'objets, au sens de la programmation objet) et de tester l'instanciation minimale, qui opère le traitement de bout en bout mais fournit une caractérisation très simple.

En l'occurrence, cette première version construit des unités Simples et quelques Solidarités (sur des critères morphologiques (majuscule initiale) et statistiques). Elle sait en outre gérer et reconnaître des Assimilations. Autrement dit, le niveau effectif d'analyse est comparable à celui de l'outil de découpage fruste, avec liste de mots vides.

Les tests révèlent que ce moteur est lent et très consommateur de mémoire pour la construction d'unités à partir d'un corpus (de l'ordre de 48 heures pour un corpus d'environ 7 mégas, avec 64 méga de RAM). En revanche, la caractérisation d'un texte, une fois l'univers d'unités descriptives construit, est quasiment instantanée. Ceci est encourageant pour la suite : c'est en effet sur la phase de caractérisation que se situent les exigences de performance pour DECID. Quant à la phase de construction d'unités, des gains en performance sont envisageables, dans la mesure où la première version a visé la clarté et la fiabilité (notamment avec un mécanisme parfaitement sûr mais lourd pour la gestion des pointeurs), mais ne s'est pas préoccupée outre mesure d'optimisation.

La question peut être posée, de la *complexité* intrinsèque de la modélisation. La constitution des unités descriptives se traduit directement par l'enregistrement et la gestion d'un grand nombre de structures ensemblistes. Il y a deux raisons de penser que cette introduction d'unités complexes n'est pas irréaliste sur un plan opérationnel. La première raison découle des principes qui régulent la construction des unités descriptives : l'unité descriptive par défaut est l'unité Simple, de complexité structurelle nulle ; et toute autre unité descriptive n'est pas définie uniquement sur sa « bonne constitution » elle n'est effectivement déclarée que si elle assure une économie descriptive. La deuxième raison qui limite les craintes de complexité, est le concept de profil d'approche, qui permet de dégrossir l'établissement des relations texte - texte sur la base de caractérisations représentatives minimales, légères et efficaces.

L'enjeu est donc à présent de poursuivre l'expérimentation progressivement. De fait, la modularité du système se prête à une évolution incrémentale. Deux étapes particulièrement significatives se profilent.

La première est la *construction de Communautés*, en intégrant le module de classification multiclasse non exhaustive (déjà réalisé). Ce serait un moment important pour l'expérimentation des hypothèses et propositions avancées dans cette thèse sur la pertinence et l'efficacité de ces structures pour la description textuelle. Bien sûr, l'utilisation de Communautés peut déjà s'éprouver « à la main » sur des textes attestés (les descriptions mises au point pour les intertitres des textes d'Action sont directement traduisibles en Relations). Mais l'enjeu de la construction automatique de Communautés est aussi de vérifier en quoi les principes d'opportunité et d'économie descriptives sont générateurs d'unités sémantiquement intéressantes. Cette hypothèse est réaliste (de nombreuses expérimentations montrent qu'une bonne application de statistiques à des corpus est fructueuse) mais demande à être qualifiée et validée.

La seconde étape importante à prévoir est le passage du moteur autonome à une utilisation souple d'outils extérieurs, en particulier un catégoriseur morpho-syntaxique ou un lemmatiseur (qui ajoute à l'information de catégorisation une réduction flexionnelle), et des outils de repérage de

relations syntagmatiques comme un extracteur de candidats termes et un module de repérage de segments répétés. Les expérimentations ainsi réalisées permettraient d'affiner nos pronostics sur le mode de *contributions d'outils de Traitement Automatique des Langues* à l'analyse textuelle. Il est essentiel que l'analyse textuelle sache tirer parti des investissements considérables de la recherche linguistique et informatique dans ce domaine, ou tout au moins que l'on évalue et comprenne la place effective que peuvent prendre ces traitements linguistiques locaux (au palier du mot ou de la phrase) dans une approche et des applications textuelles.

b) Révision de l'outil ADOC (Associations entre DOCuments)

Le chapitre sur la *Caractérisation d'un texte dans un corpus* propose une boîte à outils bien garnie, et un certain nombre d'observations pour l'usage de ces outils, pour la mise en œuvre de mesures quantitatives. La dernière partie du chapitre est consacrée à la présentation de pistes pour le cas de DECID.

Cela concerne la partie du traitement actuellement réalisée par le logiciel ADOC. ADOC dote d'une pondération chaque terme d'index (unité descriptive) affecté au texte : cette pondération doit être revue dans le cadre plus général des unités caractérisantes. ADOC opère ensuite le calcul de proximités texte - texte par un cosinus : on prévoit de remplacer cette mesure par une mesure plus qualitative, qui tienne compte du statut des unités en commun, et non de leur cumul indifférencié.

On conviendra sans peine que ces propositions ne sont encore qu'une esquisse, que l'on est face à un chantier ouvert et que le terrain d'expérimentation est vaste. Là encore, un point rassurant est la possibilité de travailler progressivement. Car d'ores et déjà, ADOC actuel peut être utilisé avec profit sur les nouvelles unités caractérisantes : en effet, comme elles sont plus contextuelles que des mots-clés obtenus par découpage, elles rendent compte d'interactions significatives entre unités élémentaires, et préparent ainsi des rapprochements motivés. L'ancienne formule, un cosinus directement sur des unités en fait élémentaires, faisait des cumuls aveugles et complètement incontrôlés d'unités élémentaires, non nécessairement valides.

Ce travail serait l'occasion de préciser des aspects méthodologiques propres aux calculs sur les textes. Dans ce domaine, la validité n'est pas au terme de la preuve, mais de l'épreuve. Il y a tout un cheminement : l'intuition initiale, l'expression d'une hypothèse, sa traduction concrète et opératoire dans les formules et le codage des données (les données étant davantage ce que l'on se donne, que ce qui est donné), le dépouillement critique des résultats en regard des choix initiaux (données, codages, mesures). La question des choix est particulièrement sensible quand elle s'apparente à une coupure : la condition d'arrêt d'itérations, le seuillage (élimination en deçà ou au delà d'une valeur fixée). Pour toutes ces étapes, les repères sont encore à rassembler et à préciser. Ces questions ne sont pas neuves, comme en témoignent les études de (Borillo & Virbel 1977), (Gardin 1974), (Gardin 1991). Pour autant, elles se posent toujours avec acuité aujourd'hui, au quotidien concret des expérimentations, facilitées par l'accès direct aux moyens de calcul, et par les ressources en puissance et en mémoire des machines actuelles. Les études sur données textuelles sont bien présentes, et sollicitent de ne pas laisser la réflexion critique et méthodologique s'éteindre.

2. Aspects du traitement des textes électroniques

a) Types et genres textuels

L'approche textuelle est inséparable d'une réflexion sur la manière d'établir, de décrire et de prendre en compte les genres textuels.

Notre étude du corpus des textes d'Action conduit à repérer trois pistes de travail, pour la mise en œuvre des genres dans des traitements automatiques.

La première piste concerne un mode de description de la structure interne de genres se divisant en parties conventionnelles. Ceci est moins restrictif qu'il y paraît : il s'agit des genres « à plan-type », ou de genres assimilables, sachant que chaque texte n'est pas en conformité formelle avec un unique plan-type précis, mais que le texte « s'inspire » d'une organisation interne qui elle-même n'est pas déterminée de façon univoque. La base acquise à travers l'étude des textes d'Action est une

typologie des relations globales des parties entre elles : succession, fusion / inclusion / exclusion, répétition et factorisation, présence, longueur. La conception d'un formalisme et d'un algorithme capables d'exploiter ces repères descriptifs n'est pas un problème informatiquement simple, et n'a vraisemblablement pas une résolution univoque. Cette recherche mérite cependant d'être tentée, car elle serait un pas significatif vers une reconnaissance globale de l'articulation d'un texte en parties, quand les algorithmes connus s'en tiennent à des indices locaux.

Une deuxième piste à explorer sur la question des genres concerne la mise au point d'un aide mémoire des caractéristiques linguistiques susceptibles de jouer un rôle actif dans la caractérisation des genres. Des études comme celles de (Bronckart 1985) ou (Biber 1988) sont de solides amorces. L'observation des textes d'ARD semble à son tour indiquer le rôle de caractéristiques comme : la reconnaissance de formulations figées et d'un vocabulaire de base du genre, qui fait office de « liant » ; différents modes d'utilisation de ponctuations ; le rôle des modes et des temps, qui se combinent ou se contrastent pour créer l'atmosphère d'un passage ou d'un partie ; un contraste possible aussi entre différents degrés de rédaction, avec d'un côté la prolifération de structures de listes et de constructions purement nominales, et de l'autres des phrases construites, plus proches de la grammaire traditionnelle. Le recensement de ce genre de caractéristiques, et la mise au point de méthodologies et d'outils pour en tirer parti dans l'analyse des textes –une analyse « genre-sensitive »–, serait une généralisation utile à large échelle.

Troisième piste décelée concernant les genres textuels : l'étude d'un mode de définition des parcours de textes, en fonction de leur genre et de l'application visée. Autrement dit, pour nos textes d'Action par exemple, qu'est-ce qui est lu, et de quelle manière ? Va-t-on se centrer sur certaines parties ? Va-t-on ne parcourir certaines qu'à la recherche d'éléments bien particuliers : des noms d'entreprises (attendus dans la partie *Partenariats*), des désignations synthétiques de l'activité (attendues par exemple dans la partie *But*, et peut-être dans les titres des *Documents de référence*) ? Derrière ces questions se dessinent déjà quelques types de parcours, qui peuvent être traduits dans une implémentation : un parcours partiel, qui ne prend pas en considération une part du texte ; un parcours focalisé sur la recherche de la présence de certains éléments, ou du contenu d'une certaine rubrique implicite.

La définition d'une typologie générale des textes nous paraît problématique. Car le corpus indéfini des textes existants se laisse-t-il quadriller ? La distribution des textes entre les types serait-elle sans litiges et sans restes ? Rien n'est moins sûr. Si typologie il y a, elle se laisse penser en termes de points de référence (non absolus), de familles de rattachement, et d'oppositions locales, plutôt que de contours (qui feraient la part entre ce qui est dans un type et ce qui est dehors) et de système universel. Autrement dit, l'indication de genre pour les textes d'un grand corpus suppose de convenir d'un référentiel de genres, documenté et interprétable par tous. La constitution d'un tel référentiel a tout intérêt à s'inspirer à la fois des contrastes entre les textes (notamment les contrastes linguistiques, cf. deuxième piste ci-dessus), et des catégories habituellement utilisées dans les pratiques d'exploitation du corpus. Chaque texte serait susceptible d'un rattachement qualifié, éventuellement multiple. Plusieurs référentiels de genres pourraient proposer plusieurs vues, complémentaires, sur un corpus. Dans les termes du modèle de structuration textuelle que nous avons proposé, un référentiel serait un *rangement*, dans lequel un genre serait une *boîte*.

b) Le contexte : par delà le monopole artefactuel des relations binaires

Il reste encore un point qui ouvre clairement une piste de recherche, et qui contribuerait à une mise en œuvre de la classification multiclasse non exhaustive plus fidèle aux principes que l'on veut ici modéliser. Rappelons l'importance de cette classification : elle sert à la construction des Communautés et à la représentation de la pertinence différentielle. Il s'agit en l'occurrence de procéder au regroupement d'unités non plus sur la base d'une mesure binaire (une distance) mais en étant capable de considérer directement la cohésion de regroupements de plus de deux unités. En effet, on peut concevoir des cas de figure où une série de paires (binaires) semblent ne rien présenter de remarquable, alors que les unités concernées présentent un comportement d'ensemble beaucoup plus remarquable et saillant.