CHAPITRE IV

Eléments pour une définition de la textualité

Aperçu

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. La recherche linguistique s'ouvre donc sur une analyse de la textualité, à savoir l'explicitation des propriétés constitutives des textes. A partir de cet inventaire très large, et dans le contexte de documents écrits à dominante scientifique et technique, quatre facettes textuelles sont définies pour guider la conception des traitements : *(i)* la matière linguistique du texte, *(ii)* son organisation interne, close et orientée, *(iii)* l'intertextualité (et notamment les formations que sont les genres), *(iv)* le rôle constitutif des lectures et la dynamique de l'interprétation (le sens est construit par et pour un lecteur).

Quelle place donner à la machine dans l'analyse des textes, quelle aide peut-elle apporter ? L'idée d'une *compréhension* automatique est ici rejetée. L'apport de l'ordinateur tient à ses capacités en termes de mémoire, de manipulation systématique et de vitesse de calcul. Le traitement suppose la définition d'une *représentation*, qui est déjà une interprétation : le sens de l'analyse effectuée est et reste du côté de l'utilisateur. A ce stade de l'exposé, un point sur les différentes conceptions de l'*interprétation* semble s'imposer, pour préciser la voie adoptée. Interpréter un texte, ce n'est ici ni établir sa vérité, ni expliciter le sens qu'il renfermerait ; mais c'est repérer des points d'appuis et des contraintes qui orientent la construction d'un sens. Dans le même esprit, le concept de *pertinence*, central pour les systèmes documentaires et les applications de recherche d'informations, est examiné à son tour. La diversité des facteurs de pertinence est rappelée et illustrée, mettant par là même en évidence l'insuffisance des modèles qui préenregistrent des jugements de pertinence pour des paires requête - document. Pour un système comme DECID, il s'agit aussi de choisir une forme de représentation de la pertinence. Parmi les cinq modèles identifiés, dont celui, le plus connu dans les moteurs de recherche, de la *pertinence linéaire* (la pertinence est évaluée par un score chiffré, et les documents sont présentés selon une liste ordonnée), le choix se porte sur une *pertinence différentielle*, qui permet une exploration thématique, méthodique et dynamique des propositions.

Table des matières du Chapitre IV

A. MULTIPLES VUES SUR LE TEXTE : L'INVENTAIRE DU COLLECTIONNEUR

1. Avertissement

Qu'est-ce qu'un texte ? La question éveille de multiples résonances. C'est même le titre littéral d'un essai de Paul Ricœur, d'un ouvrage collectif dirigé par Edmond Barbotin, et gageons que cette liste pourrait s'étendre.

D'inspiration les plus diverses, les contributions saisissent des propriétés de tous ordres. Chacune vaut d'être entendue. Les deux écueils seraient d'une part la censure, qui refuse arbitrairement un point de vue (nous pensons au contraire qu'il y a à trouver dans chacun un fondement de vérité), d'autre part croire définir la textualité par seulement un ou quelques-uns de ces aspects.

Aussi large que serait l'inventaire, il faut renoncer dès à présent à rassembler toutes les propriétés attribuable au texte, et même à en 'tenir' ne serait-ce qu'une seule ! La textualité, en un sens universel et intemporel, est un objet illusoire[1]. Ne nous trompons pas d'objectif : l'inventaire est destiné à recueillir les éléments, issus de l'expérience acquise dans notre culture actuelle. Cela fournira la base pour prendre en compte la dimension textuelle des documents rencontrés dans le cadre de l'application de diffusion ciblée qui nous concerne. On pourra ainsi proposer *une* définition, utile pour notre contexte, mais également établie non sans un certain recul.

Si, à ce stade, l'ensemble paraît bien dépareillé, dans un second temps une vision plus synthétique pourra être élaborée, intégrant et mettant l'accent sur les aspects qui apparaissent les plus pertinents dans notre contexte.

Pour l'instant, le parcours tous azimuts est un survol de repérage.

2. Du côté de l'informatique et d'autres supports d'inscription et d'enregistrement

a) Le codage alphabétique

Est dit *textuel* ce qui procède d'une langue, se transcrit, s'articule en lettres et en mots dans une écriture, par opposition à ce qui relève d'autres médias : les images, les sons. Le multimédia commence par distinguer ces différents modes d'expression, pour ensuite les mettre en relation et les intégrer en un tout plus riche. Le premier temps de la démarche induit une manière spécifique de considérer le texte. Le format du fichier informatique exploite directement la représentation en termes de chaînes de caractères, ce qui d'ailleurs occupe un espace mémoire notablement moindre que le codage des images en pixels. Les traitements s'appliquant au textuel *a minima* s'assimilent aux manipulations d'une suite de caractères : repérage et transformations d'expressions régulières[2], mise en forme et séquencement par l'intermédiaire de caractères spéciaux (notamment fin de ligne).

---

[1] « les structures textuelles sont essentiellement sémantiques. En tant que telles, elles relèvent donc plutôt de normes et de régularités que de règles –et échappent à une linguistique restreinte qui concevrait les règles conformément à la théorie des langages formels. Le caractère culturel de ces normes dissuade de considérer la textualité comme un invariant. Du moins, s'il existe au plan sémantique des formes générales voire universelles de la textualité, c'est à une sémantique comparée de les caractériser. Nous estimons donc que la textualité ne peut se définir en soi, et nous entendons seulement proposer le cadre conceptuel d'une typologie des textes. » (Rastier, Cavazza, Abeillé 1994, §VII.2, p. 172).

[2] Ce formalisme bien connu des informaticiens permet la définition de patrons décrivant un ensemble de chaînes de caractères. Des opérateurs fournissent une notation condensée de : *(i)* la gamme des caractères possibles à une position donnée, *(ii)* la présence facultative ou obligatoire d'une séquence de caractères, *(iii)* la réalisation unique ou la répétition d'une séquence de caractères.

Le jeu de caractères est un alphabet. L'alphabet de l'anglais fait partie des caractères de base de l'ordinateur ; les caractères diacritiques (avec accent, tréma, cédille, tilde,...) et les alphabets non latins ont obligé à concevoir des extensions pour les textes d'autres langues. Et bien sûr ce modèle « oublie » les écritures idéographiques.

Des propriétés statistiques sont observées sur les suites de caractères réalisées par les textes (cf. la loi de Zipf). Elles sont centrales pour les problématiques de la compression du volume de données enregistrées ou du cryptage (Salton 1989, §5 et 6). Ces considérations visent cependant le langage plus que la textualité.

b) L'expression libre dans le cadre d'une langue

Dans le jargon des bases de données, les champs textuels s'opposent aux champs factuels et numériques. Ce qui est factuel, c'est ce qui prend sa valeur parmi un ensemble donné d'alternatives (*vrai / faux*, codes départementaux, date, répertoire de noms d'auteurs, etc.). Pour ce qui est textuel, il n'y a pas de liste de possibilités prévues, la seule contrainte est en général une longueur maximale.

La particularité des champs textuels est alors leur extrême variabilité, qui fait que sur un très grand nombre d'enregistrements il peut n'y en avoir pas deux identiques. La recherche de l'identité cède le pas à la recherche du similaire. La langue, grâce aux descriptions qu'en donnent la morphologie, la syntaxe, la sémantique, met en relation des mots, des constructions. L'analyse d'un ensemble de champs textuels, par exemple l'ensemble des réponses à une question d'une enquête, s'appuie sur la linguistique pour opérer des transformations et des réductions, et ainsi forger des représentations confrontables, comparables.

c) La linéarité

Le texte se déroule, il se présente comme une suite de mots ordonnés. Si des effets de superposition ou d'échos sémantiques semblent échapper en partie à l'impératif de succession, et si une lecture peut prendre des libertés en consultant simplement quelques passages ici et là, l'expression, à travers l'écriture et la mise en page, propose un parcours systématique, qui mène du début à la fin du texte.

L'ordre joue un rôle au niveau sémantique[3], même s'il ne détermine le sens ni systématiquement, ni entièrement. Des artifices linguistiques et graphiques contribuent à exprimer un parallélisme –le caractère non significatif de l'ordre pourtant présent– là où le texte force un avant et un après[4].

Cette linéarité ne préjuge pas d'une identique linéarité de la représentation que peut se construire un lecteur : le texte est plutôt perçu comme un tout, une composition d'ensemble[5], synthétique, quand bien même la langue semble forcer le détour par une expression analytique[6].

d) Présentation, maquette, typographie

Un texte, dans sa réalisation matérielle, est mis en forme. Les logiciels de traitement de texte sont ainsi présentés comme des systèmes d'*édition*, et, pour les plus avancés, comme de la *publication assistée par ordinateur* (PAO). Il s'agit bien d'une dimension du texte lui-même : la présentation choisie n'est pas extérieure au texte, elle est en interrelation avec l'expression linguistique et concourt

[3] Cf. la composante *tactique*, proposée par François RASTIER pour la description sémantique des textes.

[4] En guise de remarque : la plupart de ces artifices linguistiques (comme la coordination et les connecteurs) ou graphiques (comme les listes) servent aussi bien à marquer l'absence d'ordre significatif qu'à souligner un ordre précis : enchaînement temporel, causal, logique, etc. Leur valeur est ainsi manifestement affaire d'interprétation.

[5] « Pour comprendre un texte, il faut être capable de passer de la séquence (lire-comprendre les propositions comme venant les unes après les autres conformément à la contrainte de la linéarité de la langue) à la figure. Il faut, comme P. Ricœur l'a montré, être capable de comprendre le texte comme faisant sens dans sa globalité configurationelle. » (Adam 1990, §I.1.4, p. 48)

[6] « on ne peut pas tout dire de quelque chose en même temps ; le langage oblige à présenter de manière analytique, point par point, des réalités synthétiques, globalisantes. L'arbre *décrit* par une succession de termes plus ou moins précis n'est pas l'arbre *perçu*. [...] Le langage contribue donc en cela à distordre la réalité. » (de Almeida, Bellamy, Kassai, p. 53)

à la construction de la signification[7]. Les choix de découpage des paragraphes comme des mises en relief se comportent comme des instructions de lecture, des indices d'une intention du rédacteur, des « traces d'actes de discours », à valeur performative[8].

3. Une linguistique qui s'aventure hors de la phrase pour aller vers le texte

a) La cohésion : liens de continuité

Le texte est ce qui déborde la phrase. Les liens syntaxiques décrivent l'unité de la phrase. Mais il y a aussi des articulations interphrastiques, c'est-à-dire d'une phrase à l'autre, ou dont la portée est de l'ordre du paragraphe.

Les anaphores (reprise par un pronom, l'élément repris est l'antécédent du pronom) et les ellipses (reprise partielle) procurent le suivi d'une notion avec moins de lourdeur qu'une répétition littérale.

Les connecteurs, en introduisant un élément, le positionnent par rapport à ce qui précède (*bref*, *par conséquent*). Il y a aussi des systèmes de connecteurs : *d'une part* / *d'autre part*, *premièrement* / *deuxièmement* / etc. Certains décrivent l'effet de ces liens d'enchaînement par le terme *connexité*, en réservant le terme cohésion aux liens de type reprise (anaphore etc.) (Charolles 1988).

L'impression de continuité sémantique du discours, par opposition à une succession de ruptures de type coq-à-l'âne, est une forme d'isotopie[9] : les unités lexicales ont des sens qui entrent en relation (Morris, Hirst 1991), elles partagent des sèmes.

Certains enchaînements locaux s'analysent en explicitant des présuppositions, et des polarités argumentatives[10], qui renforcent la plausibilité de telle succession de propositions et expliquent le caractère apparemment curieux de telle autre. Un point du discours n'est pas seulement l'aboutissement de ce qui précède, c'est aussi la préparation de ce qui suit ; le contexte local n'est pas simplement un contexte antécédent (en témoigne aussi le phénomène de cataphore).

---

[7] « les propriétés relatives de la réalisation typographique et de l'organisation spatiale de certains objets [textuels] participent à la composante sémantique du document : *l'architecture d'un texte, perceptible par le biais de ces propriétés de mise en forme matérielle, est directement partie prenante dans la construction du sens de ce texte*. » (Pascual 1991, §I.1, p. 46)
Cela devient le support de pratiques méthodiques d'« analyse de contenu » :
« Ainsi pour constater et mesurer les différences de présentation d'un événement dans la presse, on prend classiquement pour indicateurs : la surface totale (en centimètres carrés) de l'article d'information ; le nombre de mots ; la position dans le journal (première page ou pages intérieures) ; la position dans la page ; la surface totale de l'espace consacré au titre et aux sous-titres ; la taille des caractères du titre (s'il y a des majuscules, on prend la taille des caractères non majuscules) ; le nombre des illustrations ; la surface (en centimètres carrés) des ilustrations. » (Mucchielli 1974, §3.3, p. 60)

[8] « un 'chapitre', une 'section', ou un 'paragraphe', aussi 'naturels' et justifiables que soient les motifs que l'on peut avancer pour leur existence dans un texte donné, doivent d'abord leur existence au fait que par quelque moyen je performe l'existence de telle ou telle entité ayant tel statut dans mon texte. [...] Il n'est pas jusqu'à des énoncés dont les conditions d'établissement semblent très éloignées de la performativité, tels les théorèmes ou les démonstrations qui ne relèvent pas, à notre avis, de cette performativité textuelle. Ainsi, un théorème ou une démonstration dont on peut montrer qu'ils sont mal formulés ou qu'ils comportent une erreur demeurent un 'théorème' ou une 'démonstration' dans un texte donné s'ils ont été performés en tant que tels (ils seront dans ce cas un pseudo-théorème, ou une démonstration fausse, dans l'univers des mathématiques ou de la logique, mais un théorème ou une démonstration dans celui de la performativité textuelle). » (Virbel 1987, §2.2, pp. 86-87)

[9] C'est Greimas qui a introduit le concept d'isotopie sémantique (voir par exemple (Greimas 1966, §VI.1)). Ce concept est central chez lui, au point que le texte à étudier n'est plus défini que par l'intermédiaire d'une isotopie :
« Nous entendrons [...] par *texte* [...] l'ensemble des éléments de signification qui sont situés sur l'isotopie choisie et sont enfermés dans les limites du corpus. » (Greimas, §IX.1.c, p. 145)

[10] Ceci renvoie tout particulièrement aux travaux de Oswald DUCROT et de son école.

La recherche de points de cohésion moindre, voire de rupture ou de discontinuité de la cohésion sur un plan (thématique, temporel, etc.) est une tactique utilisée pour repérer automatiquement un découpage d'un flux textuel en textes, d'un document composite en sous-parties, ou d'un texte en passages successifs.

b) La progression

Dès le niveau interphrastique se manifeste la progression du texte. Un ordre est souligné par les connecteurs ; à ce qui est connu s'ajoute ce qui est nouveau (c'est la paire *thème* / *rhème*)[11], et le texte entrelace constamment le connu et le nouveau[12]. Chaque élément (personnage, idée,...) est introduit, c'est-à-dire placé dans le contexte existant, puis précisé, enrichi, peu à peu transformé[13]. L'étude de la progression introduit une dynamique dans la représentation du texte et de ses lectures[14].

Sauf convention de lecture particulière (lié à un genre), le texte qui patine, tourne en rond, se répète, n'est pas admis, ou n'est pas reçu comme tel : on lui prêtera un gain en précision, une insistance. Inversement, face à une apparente rupture du fil du discours, le lecteur soupçonne une altération du support (perte d'une page, trou dans un fichier), ou encore induit une transition avec ce qui précède ou à un niveau plus général.

Le plus souvent, le rédacteur (l'écrivain) ménage une *tension*[15], qui retienne l'attention du lecteur qui se demande où l'auteur veut en venir. La dynamique de lecture est également sensible à la

---

[11] Pour une introduction à ces notions, voir les dictionnaires linguistiques et autres ouvrages de référence, par exemple (Ducrot & Todorov 1972, § *Combinatoire sémantique*) ou (Pottier 1992, §XV.6.2).
Dans le cadre d'une analyse automatique de textes, (Hahn 1992) s'appuie sur le travail de F. Danes, qui aurait identifié trois types d'enchaînements (chaque succesion d'une phrase à l'autre relèverait de l'un de ces trois cas) : *(i)* la conservation du thème (développement autour d'un sujet donné), *(ii)* la transformation du rhème en thème (ce qui est commentaire à propos d'un thème devient le thème suivant), *(iii)* le lien à un même thème plus général (hyperonyme) implicite (les thèmes décrivent différents aspects, de façon parallèle, à un même niveau). L'utilisation de ces structures fournirait des moyens d'enrichissement de requêtes (indication de thèmes en relation), pour une recherche sur les textes analysés.
L'implémentation présentée par (Hahn 1992) fait appel à d'importantes ressources en termes d'outils de Traitement Automatique des Langues Naturelles et de bases de connaissances sur le domaine.

[12] « la textualité peut être définie comme un équilibre délicat entre une continuité-répétition, d'une part, et progression de l'information, d'autre part. Ainsi B. Combettes : « L'absence d'apport d'information entraînerait une paraphrase perpétuelle ; l'absence de points d'ancrage renvoyant à du 'déjà dit' amènerait à une suite de phrases qui, à plus ou moins long terme, n'auraient aucun rapport entre elles » [COMBETTES Bernard (1986) - « Introduction et reprise des éléments d'un texte », *Pratiques*, 49, Metz, p. 69]. » (Adam 1990, §I.1.3, pp. 45-46)
Voi aussi (Combettes & Tomassone 1988).

[13] Le travail de thèse de Jean-Philippe DUPUY (Dupuy 1993) est une foisonnante étude, d'inspiration lexicométrique, de la répétition à l'intérieur d'un texte, et de la manière dont elle fait sens dans un jeu d'identités et de différences, d'articulation entre l'unité et l'altérité, de tension entre le retour et l'évolution :
« répéter, c'est tracer un lien à la surface du discours, construire une relation, mettre en rapport deux occurrences ainsi que deux zones cotextuelles qui entrent en opposition ; [...] les répétitions, qui semblent scander le temps, ne font que mettre en évidence sa lente métamorphose. » [en l'occurrence, celle du personnage du texte] (Dupuy 1993, p. 24)
« [Dans cette étude,] le texte a été appréhendé successivement selon quatre niveaux structureaux, lexical, morphologique, temporel et sémantique [...]. A chaque fois, on a essayé de montrer que l'on peut accéder aux sens du texte en étudiant comment il organise sa répétitivité. Non que l'émergence du sens passe nécessairement et uniquement par l'itératif : c'est souvent la différence qui signifie ; détecter une répétition, c'est précisément s'offrir la possibilité d'observer la différence (diégèse en devenir, autre cotexte et autres relations), d'autant plus patente qu'elle se trouve comme pointée par l'invariant réitéré. » (*ibid*., p. 507)

[14] (Grau 1983) propose ainsi un traitement automatique pour suivre les développements thématiques d'un texte narratif. Le texte est parcouru phrase par phrase, et chaque phrase est intégrée dans le contexte formé par les phrases précédentes. Des principes de récence (locale, thème de la phrase précédente) et de dominance (globale, thème principal) guident le suivi des thèmes. Une limitation importante vient cependant du traitement purement séquentiel du texte, en particulier le modèle n'est compatible qu'avec un « texte où le thème principal est introduit dès les premières phrases » (p. 126).

[15] Dans le modèle qu'il propose, (Greimas 1966, §XI.2.c, p. 206) donne une explication de ce que l'on peut appeler *intrigue*, *suspense*, *ressort* ou *tension dramatique*.

progression, avec d'une part une assimilation rétrospective, qui récapitule ce qui a été lu, et une anticipation, prospective, élan en prolongement du point courant.

c) La cohérence : la construction d'un référentiel

Le texte s'ancre dans un univers qu'il présente et organise. Le lecteur se construit une représentation de cet univers, et cette représentation évolue dynamiquement au fil de sa lecture. Le texte présente alors une unité de sens. La cohérence –peut-être d'ailleurs plus comme une présomption que comme un fait–, assure ainsi l'intelligibilité du texte (Charolles 1988).

Au fil du texte, plusieurs désignations peuvent être interprétées comme faisant référence à un même objet dans l'univers sous-jacent au texte : c'est le phénomène de coréférence.

Un même texte peut présenter plusieurs foyers énonciatifs, c'est-à-dire plusieurs points de vue, qui ont pour effet de superposer plusieurs univers. Ces points de vue servent de référence pour le choix des temps et des modalités, et pour toutes les formes d'évaluation.

4. Structure et déploiement interne

a) Elasticité

Un texte s'inscrit dans une série de réécritures, comme entre deux infinis, du plus succinct au développement prolixe. Des opérations de condensation mènent aux résumés, à un intitulé, à un mot ou une proposition jugée fondatrice. Des opérations d'expansion conduisent aux commentaires, aux explicitations. La langue elle-même présente cette souplesse de passage d'une formulation synthétique à une formulation analytique et vice-versa, et qui devient le principe d'association entre le mot et sa définition dans les dictionnaires[16].

Certains[17] ont induit de cette propriété le fait que tout texte soit réductible à une macro-proposition, qui prend la forme d'une proposition au sens grammatical, et qui contient toute l'essence du texte. Cette macro-proposition se détermine par élagage et regroupement. Une telle conception doit être dénoncée comme doublement réductrice. Elle stipule le caractère accessoire d'une grande partie du texte, ce qui est pour le moins désobligeant vis-à-vis de l'auteur, dont l'effort aurait consisté en un magistral délayage. Elle donne pour aboutissement la proposition, et même une unique proposition fixée, s'interdisant donc tout à la fois de considérer d'autre paliers (le paragraphe, le mot) et la diversité des points de vue auxquels se prête un texte. On relèvera simplement que cet engouement pour la proposition est né dans un contexte encore fortement marqué par un intérêt dominant pour la syntaxe.

b) Une possible hétérogénéité de la forme : les séquences

Un même texte peut successivement décrire, raconter, argumenter... Les éléments linguistiques mobilisés et le mode de lecture diffèrent d'une séquence à l'autre. C'est un lieu d'hétérogénéité du texte. La pure narration, la pure explication, etc., sont des cas d'école (Adam 1992). On peut d'ailleurs s'interroger sur la possibilité de découper de telles séquences homogènes, juxtaposées et successives.

---

[16] « [Dans une définition,] le défini lexicalise de façon synthétique ce que le définissant lexicalise en général de façon analytique [...]. On peut appeler *expansivité* la propriété universelle des langues qui permet que des unités de sens soient expansées dans des unités de complexité plus grande : le rapport entre un titre et le texte qu'il introduit en illustre un cas limite. La propriété converse est la *rétractivité*, qui permet les pratiques de résumé. Expansivité et rétractivité sont des propriétés herméneutiques : c'est par convention locale soumise à conditions que l'on admet l'équivalence d'unités, quel que soit leur degré de complexité relatif. » (Rastier, Cavazza, Abeillé 1994, §III.2.1, pp. 48-49)

Comme l'avait déjà noté (Greimas 1996, §VI.2.a et b, p. 72 sq.), cette souplesse de passage d'un palier à l'autre est une motivation pour penser une sémantique unifiée.

[17] Teun VAN DIJK et Walter KINTSCH ont ouvert la voie.

c) Tabularité

Présenté sur une page, le texte prend un caractère spatial[18], et le regard peut le balayer sans se cantonner au conduit des lignes. La poésie joue manifestement sur cette disposition plane pour mettre en valeur la superposition des finales des vers (accompagnant le rapprochement des sonorités par un rapprochement graphique). L'acrostiche en fait même une clé de lecture.

Il y a ainsi deux modes de perception, complémentaires, du texte. D'une part, la linéarité : de celle-ci relèvent l'ordre et la successivité, la dynamique d'un cheminement, l'orientation temporelle (*avant* / *pendant* / *après*). La linéarité se dessine dans le déplacement d'un point de vue local. D'autre part, la tabularité, qui s'oppose méthodiquement, point par point, aux caractéristiques données pour la linéarité. La tabularité se situe dans le registre du global : c'est une perception globale, simultanée (ou du moins qui neutralise l'ordre successif par la coprésence d'une multiplicité d'ordres virtuels). Au dynamisme d'un parcours elle oppose la disposition respective d'éléments, considérés dans leurs interrelations[19].

d) Arborescence orientée

Le sommaire d'un ouvrage présente le texte par sa structure, un découpage qui rythme le texte. Cette structure présente quatre propriétés. *(i)* Elle est sans restes : toute partie du texte, sauf peut-être des pièces liminaires comme surajoutées et périphériques, prend place, entre deux bornes d'un découpage. *(ii)* La structure est orientée : chaque partie se situe entre un avant et un après, sauf les deux positions remarquables de première et dernière. *(iii)* La structure est hiérarchisée, au sens où elle s'organise en parties de niveaux successifs de généralité / spécificité. *(iv)* Enfin, la structure est emboîtée, si bien que l'intégrité de chaque partie n'est mise en cause par aucune autre partie : elle est soit entièrement incluse, soit analysée en sous-parties plus fines qui ne vont pas chercher d'éléments à l'extérieur de la partie.

Cette structure prépare deux axes de lecture : un axe « horizontal », qui suit la linéarité du texte et enchaîne les parties, et les parcourt systématiquement l'une après l'autre (c'est le parcours *en profondeur d'abord* des informaticiens) ; et un axe « vertical », qui s'appuie sur les intitulés et commence par la vision d'ensemble pour la détailler peu à peu et accéder au texte par « morceaux » (c'est cette fois-ci le parcours *en largeur d'abord*).

La structure peut connaître des réalisations minimales. Un roman peut s'en tenir à une simple division en chapitres, une nouvelle se présenter d'un seul tenant. En revanche, une documentation technique multiplie les niveaux de granularité et ménage ainsi des accès de consultation sur un point donné.

e) Délimitation

Le texte compte un nombre limité de pages, et son déroulement chemine d'un début à une fin. Les contraintes physiques imposent ce caractère fini (Bazin 1994). La rédaction s'accommode de ce cadre et le charge de signification : les genres dessinent les manières de commencer un texte et de le finir ; et tout ce qui est exprimé dans le texte doit se trouver dans l'espace des pages qui le constituent.

f) Cœur et périphérie

Des parties singulières ménagent la transition entre le propos du texte et les accès vers et depuis le texte. Ces parties se trouvent d'ailleurs aux marges de l'objet physique que constitue le livre, tel qu'il est manié : premières et dernières pages, couverture, mais aussi illustrations,

[18] Jack GOODY (Goody 1979) montre qu'ainsi, par son déploiement spatial, l'écriture a permis à d'autres formes de pensée d'advenir. Typiquement, la présentation sous forme de tableau conduit à se représenter simultanément le croisement de deux séries de modalités. Un tableau ne peut être linéarisé, exprimé oralement, sans perdre une partie de sa substance (la symétrie d'entrée par les lignes ou par les colonnes, la mise en évidence radicale des « trous » ou des surcharges dans les cases du tableau, etc.).

[19] L'articulation linéarité / tabularité traverse tout le travail de (Dupuy 1993).

commentaires marginaux. Passages entre le texte et la situation de lecture, ils orientent la façon d'aborder le texte ou de faire un écart (détour par une note et un renvoi par exemple). A la suite de (Genette 1987), on convient d'appeler ce rapport *paratextualité*.

5. Le texte et son entour

a) Liens et citations

De par l'ensemble des textes connus de l'auteur et présents à son esprit, et de ceux par rapport auxquels le texte se positionne pour ses lecteurs, aucun texte n'échappe à une multitude de liens qui le rattachent à d'autres textes.

Au sens large, une citation peut être explicite (annoncée, référencée) ou implicite (greffée sans démarcation dans le fil du texte), voulue ou fortuite, mise en valeur, allusive ou furtive, rejetée ou appropriée, littérale ou accommodée, conventionnelle ou inattendue[20]. Au point que tout texte puisse finalement être tenu pour un centon, à savoir tout entièrement forgé à partir de citations[21].

b) Situation et implicite

Le texte délimite un environnement intérieur, linguistique, et un environnement extérieur, situationnel, respectivement le cotexte et le contexte[22].

Ancré dans un contexte qui le situe, tout texte, dans sa finitude, comporte une part d'implicite. Il part d'une certaine connaissance commune et à partir de laquelle il se déploie. Cette connaissance commune peut être plus ou moins universelle et atemporelle (elle n'est jamais complètement l'un ou l'autre, de par son ancrage culturel) : renvoi à l'actualité, à des circonstances qui relèvent de la vie privée, etc.

La prise en compte des conditions de production, ou au contraire le travail sur la matérialité linguistique du texte détaché d'une situation particulière, est parfois traduit par l'opposition *discours vs texte*.[23]

---

[20](Maingueneau 1991, p. 139 sq.) distingue l'*intertexte*, à savoir l'ensemble des citations effectives, et l'*intertextualité*, qui est le type de citation que le genre (donc la pratique) dans lequel s'insère le texte autorise. L'intertextualité représente donc le domaine virtuel des citations. L'intertextualité se divise en *interne vs externe*, selon que le texte cité se trouve dans le même champ pratique que le texte étudié ou bien qu'il en sort.

[21] Roland BARTHES résume ainsi cette prégnance insoupçonnée et généralisée des citations :
« Le texte redistribue la langue (il est le champ de cette redistribution). L'une des voies de cette déconstruction - reconstruction est de *permuter* des textes, des lambeaux de textes qui ont existé ou existent autour du texte considéré, et finalement en lui : tout texte est un *intertexte* ; d'autres textes sont présents en lui, à des niveaux variables, sous des formes plus ou moins reconnaissables : les textes de la culture antérieure et ceux de la culture environnante ; tout texte est un tissu nouveau de citations révolues. Passent dans le texte, redistribués en lui, des morceaux de codes, des formules, des modèles rythmiques, des fragments de langages sociaux, etc., car il y a toujours du langage avant le texte et autour de lui. L'intertextualité, condition de tout texte, quel qu'il soit, ne se réduit évidemment pas à un problème de sources ou d'influences ; l'intertexte est un champ général de formules anonymes, dont l'origine est rarement repérable, de citations inconscientes ou automatiques, données sans guillemets. Epistémologiquement, le concept d'intertexte est ce qui apporte à la théorie du texte le volume de la socialité : c'est tout le langage, antérieur et contemporain, qui vient au texte, non selon la voie d'une filiation repérable, d'une imitation volontaire, mais selon celle d'une dissémination –image qui assure au texte le statut, non d'une *reproduction*, mais d'une *productivité*. » (Barthes 1973)
Pour une étude développée des processus de citation dans les textes, voir (Compagnon 1979).

[22] Cette terminologie n'est pas complètement stabilisée.
Bernard POTTIER distingue par exemple trois modalités de contexte (Bommier 1994a, p. 7) :
- l'*antétexte* est le contexte linguistique, c'est-à-dire le texte énoncé peu avant, qui est présent à la mémoire des interlocuteurs (ou du rédacteur / lecteur), et auquel peuvent notamment référer des anaphores.
- le *co-texte* est l'accompagnement du texte utilisant un support non verbal, par exemple une illustration.
- le *contexte* doit être compris au sens large : c'est tout ce qui peut caractériser la situation d'énonciation, et qui interagit (de façon plus ou moins marquée) avec le texte.

[23] « un discours est un énoncé caractérisable certes par des propriétés textuelles, mais surtout comme un acte de discours accompli dans une situation (participants, institutions, lieu, temps) [...]. Le texte, en revanche, est un

L'explicite, ce qui est dit dans le texte, peut servir de base d'application de déductions et d'inférences. Ces raisonnements servent à expliciter une part de ce que le texte porte en puissance sans l'exprimer.

c) L'autonomie

L'échange de paroles impose la coprésence des interlocuteurs. Le texte, par le biais de l'écriture, intercepte[24] le rapport auteur - lecteur et se détache d'une situation particulière, *hic* et *nunc*. S'il use de déictiques, c'est pour renvoyer à son propre univers interne qu'il élabore, qu'il s'agisse d'un développement théorique et abstrait, d'un monde fictionnel, ou de la vision d'une réalité.

L'autonomie n'est pas autarcie, et le texte ne prend sens que pour un lecteur, dans un contexte. D'où le rejet d'un principe d'immanence, qui voudrait que « tout » soit « dans » le texte. La lecture est une actualisation et une appropriation[25], elle prend sa matière de construction à la fois dans et hors du texte. Si l'on peut reconnaître un fonctionnement endogène au texte, qui instaure, crée et fait évoluer lui-même son univers et son maniement de la langue, cela n'élude pas l'interaction sans laquelle le texte, coupé de toute réalité, ne peut se situer et prendre sens.[26]

d) L'affiliation historico-culturelle et l'appartenance à un genre

Qu'il s'agisse de sa rédaction ou de ses lectures, le rapport au texte est médiatisé par la culture (Beacco 1992).

C'est cette dimension qui le constitue comme archive[27], non seulement trace historique, mais instance prenant place dans un tissu de rapports sociaux, et s'inscrivant –de façon significative– dans un mode de prise de parole.

---

objet abstrait résultant de la soustraction du contexte opérée sur l'objet concret (discours). » (Adam 1990, Introduction §3, p. 23)

[24] C'est le terme de Paul RICŒUR : on trouvera l'idée développée dans plusieurs essais de (Ricœur 1986).

[25] Nous suivons toujours (Ricœur 1986). Au fil de ces essais, Paul RICŒUR engage à dépasser l'opposition épistémologique entre une conception objective et une conception subjective du texte. Il s'applique à montrer l'alliance féconde de l'*explication* (qui s'ancre dans la matérialité du texte et dégage ses structures internes) et de la *compréhension* ou *interprétation* (qui ouvre dynamiquement le texte sur un sens personnel pour le lecteur, en reconfigurant sa manière de voir le monde).

[26] Une étude linguistique du texte peut ainsi procéder à une triple désontologisation, méthodique et raisonnée :
« *(i)* remplacer le problème de la référence par celui de l'impression référentielle ; *(ii)* celui de l'énonciateur, par celui du foyer énonciatif, tel qu'il est représenté dans le texte et/ou situé par les règles du genre ; *(iii)* et celui du destinataire par celui du foyer interprétatif, dans des conditions analogues. » (Rastier 1996b, §1.1, p. 16)
Plutôt que de se perdre dans une insaisissable réalité extratextuelle, l'analyse observe celle-ci depuis ses traces, ses points de contact avec le texte. Aux pôles extrinsèques du texte (l'auteur, le monde, les destinataires) répondent les pôles intrinsèques du texte, tels que les circonscrit le genre.

[27] Ce terme renvoie à l'école française d'Analyse du Discours. Dominique MAINGUENEAU en est un porte-parole :
« pour l'AD [l'école française d'analyse du discours] il ne saurait être question de traiter les matériaux verbaux comme de simples véhicules d'information ; elle veut les appréhender *comme des textes*. Si pour l'analyse de contenu ces textes sont en quelque sorte transparents aux représentations des sujets sociaux qu'ils sont censés refléter, l'AD prend acte de leur opacité, refusant de les projeter directement sur une réalité extradiscursive : l'interprétation doit prendre en compte le mode de fonctionnement des discours, les modalités de l'exercice de la parole dans un univers déterminé. [...]
Tel qu'il se détermine ici, l'objet de l'AD pourrait être dénommé une *archive*, laquelle regroupe un ensemble d'« inscriptions » référées à un même positionnement. [...]
Pour l'AD les soubassements sémantiques d'*archive* ne sont pas dénués d'intérêt. Son étymon latin, l'*archivum*, provient de l'*archeion* grec, lui-même dérivé de l'*archè* de l'*archéologie*. Lié à l'*archè*, « source », « principe » et à partir de là « commandement », « pouvoir », l'*archeion*, c'est le siège de l'autorité (un palais par exemple), un corps de magistrats, mais aussi les archives publiques. La fonction de mémorisation, de trésor textuel qui est celle de l'archive et dont participe l'AD elle-même en recueillant, en manipulant les énoncés déjà proférés, est ainsi systématiquement rapportée à la détermination d'une enceinte, d'un pouvoir qui est pouvoir de dire, à l'affirmation de la légitimité d'un corps d'énonciateurs consacrés. Or s'il est vrai que l'AD récuse l'idée d'un point d'origine du discours, l'imaginaire constitutif de l'archive suppose une relation à une source du sens, la

Rattaché, par l'usage, à une pratique particulière, il se rapporte nécessairement à un genre[28]. Le texte neutre[29], standard ou spontané, qui serait délié de tout genre, ne peut pas exister.

Le genre n'est pas un moule formel, préexistant, extérieur au texte, il est constitutif du texte[30]. Le genre a une incidence sur la composition du texte (thèmes abordés, vocabulaire, découpage en parties,...) et sur ses modes de lecture[31]. L'affiliation à un genre est porteuse de signification : le texte s'y positionne, il s'en réclame ou le subvertit, etc. (Maingueneau 1991, §5.1)

Des typologies sont proposées (Adam 1992) (Petitjean 1989ab) (Bronckart, Coste, Roulet 1991), renvoyant pour la plupart à des considérations fonctionnelles (raconter (narratif), convaincre (argumentatif), etc.). La discussion autour de ces propositions peut s'engager à partir des points suivants : *(i)* pour tout texte que j'ai là, maintenant, devant moi, concrètement, trouve-t-il sa place dans la typologie considérée ? et avec quelle clarté : l'attribution est-elle laborieuse ? *(ii)* La typologie place-t-elle mon texte dans la même classe que d'autres textes, avec lesquels il est pourtant en contraste évident ?[32] Le premier point épingle une typologie trop restrictive, le second une typologie trop accueillante (ces deux défauts pouvant se cumuler).

---

délimitation d'un espace fondateur, authentifiant. L'AD s'intéresse en effet surtout aux discours *autorisés* qui, au-delà de leur fonction immédiate, supposent un rapport aux fondements et aux valeurs. Considéré comme « archive », un ensemble de textes ne se définit pas seulement comme la réponse à un faisceau de contraintes pratiques, il permet aussi de légitimer un certain exercice de la parole pour un groupe donné. Dès lors, étudier des articles scientifiques ou les publications internes à une entreprise industrielle ne saurait se résoudre dans la seule prise en compte de leur utilité, dans la mise en rapport d'une structure et d'une fonction : c'est une certaine organisation de l'univers d'une collectivité qui se trouve impliquée. L'étude de l'archive joue aussi un rôle comparable à celle du mythe pour les sociétés primitives. Pour l'AD comme pour le mythologue, il s'agit de considérer des positions énonciatives qui nouent un fonctionnement textuel à l'identité d'un groupe. »
(Maingueneau 1991, §1.1, pp. 9, 22)

[28] « Un acte de communication n'est pas une simple transmission de messages entre deux interlocuteurs idéalisés, comme l'*Emetteur* et le *Récepteur* pour Saussure, *A* et *B* pour Jakobson, ou *Jill* et *Jack* pour Bloomfield. L'usage d'une langue est par excellence une activité sociale, si bien que toute situation de communication est déterminée par une pratique sociale qui l'instaure et la contraint.
Sur cette évidence se fondent nos affirmations sur l'omniprésence des genres. » (Rastier 1989, §I.3.III)
Les types de textes se répartissent alors en *discours* (qui correspondent aux domaines d'activités dans la société et à la division du travail : politique, religieux, médical, etc.), puis en *genres* (qui sont associés aux différentes pratiques ayant cours dans le domaine en question). (Adam 1990, Introduction §3, p. 20-21) (Rastier, Cavazza, Abeillé 1994, §VII.4.1)

[29] « A ceux qui demandent comment traiter les textes neutres ou ordinaires, qui ne seraient ni littéraires ni techniques, nous répondons qu'il n'en existe pas. Cette question est inspirée sans doute par la philosophie du langage ordinaire, et plus généralement par l'idée qu'il existe un emploi neutre du langage, littéral, à la fois d'usage général et simplement dénotatif. En fait, tous les usages linguistiques sont normés, relèvent d'un genre et d'une pratique sociale, et même ceux qui donnent l'impression de liberté, notamment les usages privés, n'échappent pas à ces déterminations. » (Rastier, Cavazza, Abeillé 1994, §VII.1.3)

[30] « Les genres du discours ne sont pas des catégories intemporelles mais des réalités historiques, inséparables des sociétés dans lesquelles ils émergent. A la lumière de la conception pragmatique du langage, on assiste à une modification de l'image traditionnelle qu'on s'en fait, celle d'un ensemble de « procédés », de « cadres », qui permettent de donner une certaine forme à un « contenu » qui en serait indépendant. On préfère y voir une activité sociale ritualisée, soumise à des conditions de réussite qui intègrent un ensemble diversifié de paramètres (statut des énonciateurs, du public, lieux d'énonciation, etc.). Dans cet ordre d'idées on sait par exemple quels progrès ont été réalisés dans la compréhension des Evangiles quand on a étudié leur texte en prenant en compte l'usage qui en était fait dans les communautés chrétiennes où ils se sont constitués, au lieu de ne voir dans ces dernières que des « circonstances » contingentes. » (Maingueneau 1991, §5.1, pp. 178-179)

[31] « le genre est une catégorie *instituante* qui prend la forme d'un 'horizon d'attente' au niveau de la lecture, d'un cadre discursif au niveau de l'écriture, et dans tous les cas d'instance de 'socialisation'. » (Petitjean 1989b, p. 120)

[32] Par exemple, que deviennent ces deux questions quand je considère les descriptifs d'activité des agents de la Direction des Etudes et Recherches d'EDF, et non un roman 'canonique'...
Un regard critique plus détaillé sur les typologies fonctionnelles pourra être trouvé dans (Rastier 1989, §I.3.IV.A).

La tendance est à la description des genres comme des hybridations des fonctions précitées, les textes « purs » apparaissant des cas d'école. Plus souple encore, la définition du genre d'un texte peut être obtenue par une panoplie de critères de caractérisation[33].

e) Une parole fixée, inscrite

Les paroles s'envolent, l'écrit reste, résume le dicton. La composition linguistique du texte est stable, le choix et l'ordre des mots sont fixés définitivement[34]. De plus, le codage alphabétique opéré par l'écriture assure une reproductibilité exacte et littérale, *ad libitum*, et donc une persistance du texte dans une démultiplication de ses exemplaires. Ceci suppose que l'on s'en tienne essentiellement à l'expression linguistique et aux segmentations marquées (vers, paragraphes, parties), et que l'on néglige la nature du support (choix du papier,...) et peut-être la disposition (mise en page et typographie). Le texte est à la fois constitué par la matérialité de son support qui l'institue, et caractérisé par sa dématérialisation, qui lui permet de persister par delà l'existence d'exemplaires concrets.

Dans une lecture, cette inscription sur un support permet des libertés qui échappent à l'oral[35]. En effet, la linéarité du texte n'engage pas celle de la lecture, à la différence de la chronologie de l'oral. Le texte peut être consulté ponctuellement, le lecteur peut revenir sur un point précédent[36], ou anticiper sur le déroulement linéaire ; il peut feuilleter, survoler, s'arrêter sur un point[37]. Ceci est facilité par l'évolution des supports : on est passé de l'accès séquentiel du codex (rouleau), à l'accès direct avec le livre (ouverture sur n'importe quelle page) ; et les nouveaux supports électroniques instrumentent et renforcent l'accès direct (liens statiques et dynamiques, recherche en texte intégral et navigation).

f) Une épaisseur temporelle

La génétique des textes étudie la formation d'un texte à travers les brouillons successifs. La philologie s'efforce de rétablir la version originelle d'un texte, dont les copistes du Moyen-Age ou les éditeurs, même réputés sérieux, se sont écartés. De nos jours, l'utilisation des traitements de texte

---

[33] François Rastier propose quatre composantes sémantiques : la *thématique*, la *dialectique*, la *dialogique* et la *tactique*. D'une façon très simplifiée : la thématique s'intéresse au repérage des éléments de contenu et à l'identification du sujet du texte ; la dialectique traite des intervalles temporels dans le temps représenté, de la structuration et des interactions des entités ; la dialogique étudie les points de vue (modalisation) ; et la tactique rend compte de la disposition linéaire des unités sémantiques, avec les effets d'ordre et de succession, tant au plan de l'expression que du contenu. Ces composantes concernent tous les paliers de l'analyse (mot, phrase, texte) et sont organisées en hétérarchie (aucune ne domine ni ne précède une autre).
« Les genres sont définis par des interactions normées entre les composantes [sémantiques]. [...]
[...] les interactions des composantes sémantiques n'ont pas à être explorées *in abstracto*. Elles sont codifiées par les discours, et les genres (dont chacun peut être défini –quant à son contenu– comme un type d'interaction entre elles) et en cela relèvent de normes, évidemment culturelles. Aussi n'entendions nous pas formuler une typologie, mais en définir les critères. » (Rastier, Cavazza, Abeillé 1994, §VII.4.6)

[34] « On ne dira jamais assez, par exemple, l'importance du 'bon à tirer' qui sépare nettement l'acte d'écrire, révisable et interminable, de l'œuvre elle-même ». (Bazin 1994)

[35] Le texte n'est donc pas seulement une expression linguistique fixée sur un support, mais il est *conçu* avec l'idée qu'il est fixé, se prêtant à une pratique interprétative de lecture et de relectures. Certains enregistrements d'interventions, à l'oral, satisferaient cette condition : la lecture préparée d'un texte, et peut-être un certain nombre d'émissions différées. En revanche, inclure dans l'étude des textes des retranscriptions d'échanges dans des situations où il est hors des préoccupations « normales » des locuteurs de garder et marquer une trace de l'échange tel qu'il se déroule, élargit par trop le champ de notre étude et nous ferait perdre des propriétés fortes de la textualité.

[36] Cette possibilité, de revenir plusieurs fois sur un passage complexe, fait que « l'écrit peut se permettre d'imposer à la mémoire du récepteur une charge supplémentaire » par rapport à celle acceptable à l'oral (de Almeida, Bellamy, Kassai, pp. 99-100).

[37] La structuration du texte guide la construction d'un parcours : le lecteur s'oriente en fonction du type de document et des traces (typographiques, linguistiques) de son organisation, il se repère par rapport à ses attentes sur les fonctionnalités des parties (Dillon 1991).

modifie les habitudes de rédaction (facilités d'insertion, de déplacement et de duplication d'une zone de texte), multiplie les versions. L'attention qui était apportée à la conception et à la planification initiale, dans une rédaction manuscrite, est reversée sur une phase de corrections et d'enrichissement (Piolat, Isnard, Della Valle 1993). Les 'marques de révision' sont proposées pour tracer les retouches.

Le rattachement à son auteur inscrit le texte dans une œuvre, peut-être comme étape dans la réalisation d'un projet (esthétique, scientifique), ou comme élément de réponse à une question première qui hante l'auteur. Tel texte est perçu comme une évolution, une révision, une annonce, d'un autre texte de l'auteur. Le choix d'un ordre de lecture n'est pas sans incidence interprétative (Tardieu 1987).

Quant à sa lecture, le texte s'enrichit des lectures qui ont marqué la vision que l'on adopte du texte. Ce processus de sédimentation (les lectures se superposent au fil du temps et des traditions) est aussi ce qui maintient une continuité entre un texte original, éloigné dans le temps et l'espace, et ses possibles lectures actuelles.

6. L'homme face au texte

a) Une existence motivée

Le texte apparaît comme le fruit de l'expression originale d'un auteur (éventuellement pluriel) à l'intention d'un lectorat. Il se pose comme un acte, acte d'écriture, acte qui intervient dans le cours des choses et dans l'histoire[38]. *A contrario*, une suite de mots, même « bien formée » (pour reprendre l'expression consacrée des logiciens), générée par une machine et non orientée par quelque choix humain (choix des éléments à présenter, choix du mode de construction du discours, etc.), peine à être reçue comme un texte (pour autant que l'on sache qu'il s'agit d'une production machinale) (Dumesnil 1992). Un texte est une intelligence, une sensibilité, qui se communique. Il est crédité d'un sens.

Par sa simple existence, le texte se pose comme légitime (Maingueneau 1991, p. 173). Son auteur s'en porte garant.

Le texte se justifie par son utilité (au sens large) et son originalité (non sans lien avec sa nouveauté) (Chabin 1997). Les typologies fonctionnelles s'efforcent ainsi de situer chaque texte par rapport à un usage général visé, un mode de relation de communication : décrire (représentation par les mots d'une scène statique), raconter (un enchaînement d'événements qui fait sens), argumenter, enseigner, distraire, procurer une émotion esthétique... Quant à son originalité, le texte se présente comme un apport ou un écart dans le contexte dans lequel il s'inscrit.

b) Le support de lectures et d'interprétations

Du texte à son lecteur humain, il y a un acte. En fonction de ses attentes, des contraintes linguistiques posées par le texte, de règles interprétatives, le lecteur parcourt le texte. Il se construit dynamiquement une représentation de ce qu'il a perçu, de ce qui a été saillant[39].

7. Le texte électronique : une autre textualité ?

a) L'incidence du support sur la nature du texte

Il y a une différence fondamentale entre des documents conçus, ou adaptés, pour une publication hypertexte[40] (et qui nous intéressent pour ce dernier ensemble de propriétés), et des

[38] (Ricœur 1986) trace un riche parallèle entre *texte* et *action*.

[39] Par exemple, une lecture professionnelle ne procède pas de la même manière, ni avec les mêmes objectifs, qu'une lecture « gratuite », « pour le plaisir » (Brouillette 1996).

[40] Pour un rapide tour d'horizon historique sur la formation du concept d'hypertexte et des premières réalisations (Vannevar BUSH, Paul OTLET, H.G. WELLS, Douglas C. ENGELBART, Theodor Holm NELSON, Bill ATKINSON, Tim BERNERS-LEE), voir (Teasdale 1995).

documents qui sont la reprise, sans réaménagement, de documents papiers existants (Oßwald 1995) (Amitay 1997)[41].

Les possibilités offertes par la forme électronique induisent de nouvelles formes textuelles, qui, par delà même l'apparition de nouveaux genres[42], bousculent certains des repères donnés précédemment pour la textualité.

Il ne faut pas croire pour autant à la disparition du papier, et à l'étouffement définitif des modes de lecture qu'il induit et des propriétés textuelles que nous avons vues précédemment. On pourrait plutôt esquisser une complémentarité :

Papier	Electronique
lire	apercevoir (signalement)
concentration (recueil)	dispersion (réseau)
stabilité, référence	dynamique
localisation (objet à sa disposition)	diffusion

b) Un document diffus et fragmenté : la clôture du texte en question

L'hypertexte rend problématique la clôture du document. Autant l'unité que constitue la page est clairement délimitée, autant une unité supérieure, qui rassemblerait des pages pour former un document, n'est pas toujours claire à cerner. Jusqu'où suivre les liens ? Trancher sur le statut, interne ou externe, d'un lien par rapport au document, devient parfois une véritable opération herméneutique.

Tout se passe un peu comme si l'ancienne évidence unitaire du texte devenait une évidence unitaire de la page, fragmentant de la sorte les unités de communication[43]. Mais la vision devient alors parcellaire, la page isolée ayant parfois un contenu indigent et non autonome. Et, enregistrés sous la même appellation de *page*, se trouvent des documents des « niveaux » les plus divers : page d'accueil générique, exemple, illustration, article de référence, courrier électronique, etc. (Koch 1996)

---

[41] La mise sous forme hypertextuelle de documents « classiques » électroniques existants donne lieu à une réflexion théorique et pratique très riche. Voir par exemple tout le travail de thèse de (Papy 1994), en particulier §4.1 (p. 84 sq.) –comment « découper » le document pour former les noeuds de l'hypertexte ; les textes techniques s'y prêtent mieux que les textes littéraires–, et §6.6 (p. 152 sq.) –la difficulté qu'il y a à éclater le texte en unités.

[42] Einat Amitay plaide haut et fort pour définir l'hypertexte comme un nouveau genre :
« The idea behind this dissertation is that hypertext is a new genre of expression and that it is systematically different from other communicative verbal means of expression like flat hierarchical text or speech. » (Amitay 1997, §6, p. 49)
Son travail sur les spécificités linguistiques et organisationnelles d'un corpus de pages Web, et notamment sur la forme et l'usage des ancres (zones actives pour un renvoi hypertexte), serait plutôt une excellente introduction à une analyse des *pages personnelles*. Si genre il y a, il se situe selon nous à ce niveau (un « type » de page) ou encore en deçà (par ex. les pages personnelles d'une certaine communauté). Vouloir décrire le Web dans son ensemble appauvrirait les régularités décelables (peu de choses sont communes à toutes les pages Web), sans non plus refléter une unité réelle et effective (en pratique on a affaire à un secteur du Web).
Marie-Anne CHABIN propose, comme première piste de travail à propos des archives numériques, de prendre acte de cette différence des documents électroniques, qui ne sont pas une simple retranscription des formes de textes connues :
« Etablir une typologie spécifique des documents numériques : à côté des natures de documents qui ont simplement changé de support, l'utilisation du numérique promeut de nouveaux types de documents tels que les documents collectifs issus du *workflow*, ou les très nombreuses mises à jour. L'analyse de leur raison d'être, de leur provenance, de leur mode de fabrication, etc., doit permettre d'esquisser des types de documents. » (Chabin 1997, pp. 215-216)

[43] Cette évolution vers une fragmentation de plus en plus marquée de l'information est analysée comme une source de surcharge cognitive : il faut en effet constamment passer d'un fragment à l'autre, et à chaque fois (re)constituer un contexte. Il faut aussi gérer simultanément une multiplicité d'informations autonomes, en percevant leur positionnement respectif et en organisant leurs priorités (voir les travaux de Saadi LAHLOU à EDF-DER et de Charles LENAY à l'UTC de Compiègne).

La page, ajustée par les contraintes de transfert et d'affichage, devient le nœud central, et le point intercalaire entre deux « zones » nouvelles. L'unité supérieure « physique » est le *site* (pour Internet), ou plus généralement un répertoire principal. Et, zone plus petite que la page, la *fenêtre* d'affichage, qui cadre le champ de vision et renforce la linéarité et le découpage du texte à l'intérieur de la page (le début est plus visible que la fin ; la présence nombreuse de frontières et de titres facilite le repérage, où que soit positionnée la fenêtre).

c) Perte de certains guides de parcours : butinage et désorientation

Les pages d'un livre, reliées dans un certain ordre, proposent d'emblée un parcours préparé et systématique. En suivant l'ordre du livre, le lecteur sait qu'il aura une vue complète de l'ouvrage. Il peut se reposer sur ce fil conducteur. Il a à tout moment une idée du chemin parcouru et de ce qui reste à parcourir. Grâce aux indications portées par la structuration interne du document (découpage en chapitres et sections et intitulés, paragraphes mis en valeur ou au contraire présentés comme subsidiaires), il adapte facilement son parcours de lecture à sa situation (intérêts, contraintes de temps, etc.). Le lecteur prend donc certaines libertés, mais sur une base linéaire.

Ces repères sont pour une bonne part perdus quand il s'agit d'un hypertexte[44], qui par essence offre des lectures *purement* non-linéaires. Il y a d'abord la question du point d'entrée : toutes les pages ne sont pas équivalentes pour donner un contexte introductif et ouvrir sur une lecture constructive. Ensuite, en ce qui concerne l'enchaînement des pages, une page propose couramment plusieurs liens, ce qui rompt la linéarité et l'évidence du parcours. Il n'y a plus vraiment de tactique systématique pour faire le tour d'un document. Bien sûr, l'algorithmique fournit deux modes de parcours d'un arbre hiérarchique, en largeur d'abord et en profondeur d'abord. Le réseau hypertexte peut en effet être vu comme une simple structure arborescente, en prenant la page de départ comme racine (à quelques exceptions prêt, qui peuvent compliquer la situation : cycles, etc.). La logique du parcours en largeur d'abord crée des discontinuités de contexte (on saute d'une branche à l'autre de l'arbre) : c'est tellement anti-naturel qu'elle est spontanément très peu pratiquée, ou seulement très localement (sans sortir du contexte d'une page). La logique en profondeur d'abord entraîne dans des dérives sans fin, faute de repères de clôture. Elle contribue à ces digressions et flâneries que d'aucuns stigmatisent. La démarche réelle est intermédiaire, et donc plus aléatoirement opportuniste qu'efficacement systématique. Une aide importante (mais qui ne résout pas tout) est l'indication qu'un lien mène sur une page qui a déjà été visitée ou non.

A cela s'ajoute que, au lieu d'une identification des « niveaux » d'information, tout est « page », et il est souvent difficile de savoir, en décidant de suivre un lien, si l'on trouve une information synthétique ou détaillée, une illustration ou une page d'accueil qui invite à explorer tout un nouveau site (Teasdale 1995).

Au mieux, l'hypertexte favorise une stratégie de découverte (« tomber sur » une page inattendue et intéressante) (Michel 1997, §2.5, p. 224), mais gêne la construction de visions intégrales et intégrées, et le repérage par rapport à des points de référence caractéristiques (réalité mouvante de l'Internet, uniformité du format 'page'). L'absence d'un texte principal, d'une unité textuelle, semble substituer la *lecture*, qui procède par enrichissement de l'interprétation au fil du texte, au *parcours*, qui n'est que déplacement d'un centre d'attention à un autre, sans capitalisation progressive[45]. Les québécois, en baptisant les logiciels de navigation des *butineurs*, ont trouvé une image parlante.

La difficulté pour cerner un ensemble qui forme un document et avoir la vision globale d'un tout a bien été identifiée –« myopie »– par les concepteurs d'interface, qui s'ingénient à fournir à l'utilisateur une vue d'ensemble de son parcours (éventuellement « aplatie », cf. le mécanisme de

---

[44] (Zizi 1995, §1.3.2) donne une bonne description des *Problèmes inhérents à la navigation hypertexte*, qu'elle développe en trois points : *myopie*, *désorientation*, et *digression*.
Sur la désorientation, voir aussi : (Papy 1994, §2.5.3, p. 56 sq.).

[45] (Bachimont 1999b).
L'usager des hypertextes exprime le besoin d'outils pour éviter l'inconsistance et la dispersion de ces parcours :
« Je veux naviguer facilement sans perdre le fil de ma pensée. Ce qui m'intéresse, c'est un historique de ma pensée. J'ai besoin de ne retenir qu'un lien fort. Trop rebondir peut distraire plus qu'enrichir. J'aimerais avoir des aides pour enrichir le rebondissement sans connaître la distraction. » (Merle, Fradin 1994, §9, p. 48)

retour arrière, qui mémorise le déroulement 'en profondeur' mais pas 'en largeur' depuis le début de la session). Soit dit en passant, ce diagnostic heuristique révèle bien l'importance de l'interdétermination du local et du global dans la construction de l'interprétation, point sur lequel nous aurons l'occasion de revenir.

Le support papier favorise une vue synoptique : pour travailler sur un dossier, on dispose l'ensemble des documents sur l'espace de travail, et on s'organise et s'oriente en fonction de la hauteur des piles, de leur ordre de succession, de la proximité (ce qu'on a placé près ou bien loin de soi). L'électronique ne reproduit pas naturellement ni efficacement ces dispositions d'ensemble. En revanche, la force (potentielle) du support numérique est dans le calcul de vues synthétiques, correspondants à un angle de lecture, qui embrassent un volume de documents d'un tout autre ordre de grandeur (Bachimont 1999c).

d) Sources diffuses : des documents multipliés et mal identifiés

La généralisation de l'usage des traitements de texte, et l'ouverture de moyens puissants de diffusion hors des circuits, contrôlés, des maisons d'édition, affaiblit la légitimité accordée au texte, et bouleverse la constitution des fonds d'archives (Chabin 1997) (Michel 1997). L'auteur d'un document n'est pas toujours clairement identifié ; et aucun comité de lecture n'a approuvé le texte et reconnu qu'il « méritait » d'être lu (Bazin 1994). En outre, le document, qui n'a pas une forme stable et unique, a pu être modifié subrepticement ; et à la surmultiplication des exemplaires (prolifération des copies pour information) se mêlent confusément les variantes de version.

Le document devient plus insaisissable, dans tous les sens du terme. Il n'y a peut-être pas à s'en étonner pour ce qui concerne Internet. L'origine du Web était justement une organisation en réseau telle qu'elle échappe à toute tentative de destruction. Ce qui est perdu localement peut être rétabli, retrouvé ailleurs. C'est une dispersion, une dilution tactique. Les sites miroirs, les liens multiples et sans systématique, empêchent de cerner une cible.

Au quotidien dans les bureaux, le stockage sous forme électronique relâche certaines contraintes d'encombrement et d'ordre (Chabin 1997), ce qui favorise une conservation plus 'quantitative' et moins 'qualitative', plus systématique et moins rigoureuse.

e) Liens organisateurs d'un espace

L'avènement de l'hypertexte a matérialisé le renvoi « point à point », d'une zone d'un texte à un point –ce qui ne préjuge pas du caractère ponctuel de la cible du renvoi : référence interprétée comme l'œuvre qu'elle désigne, début d'une partie (page, intitulé, paragraphe) associé au développement qui y est opéré, terme pris dans son contexte.

Il y a une topographie des relations de texte à texte[46]. La littérature tendrait à mettre en valeur un canon, un texte qui donne accès à tous les autres. Les lianes de l'Internet, ou les textes bien rangés dans une base de donnée, donneraient plutôt l'image d'une multitextualité, dans laquelle il n'y a pas de texte dominant.

Les liens tissent ainsi un espace, avec ses voisinages, ses chemins ; il ne s'agit (actuellement) que de parcours locaux, de linéarisations élémentaires, qui n'aiguillent que d'une page à une autre.

f) Documents vivants

Le support électronique favorise l'évolution et l'ajustement continu du document (Papy 1994, §2.2.4, p. 42 sq.). Par exemple sur Internet, un document apparaît à une adresse donnée. Il évolue, avec des discontinuités possibles, si l'on identifie le document par son adresse : remplacement complet d'un texte par un autre, « déménagement » de la page à une autre adresse. Puis il disparaît sans prévenir et sans laisser de traces... Cette caducité et cette fugacité se généralisent à la plupart des

[46] Des études récentes suivent par exemple les types de cheminement, de page à page, sur une portion du Web (Wexelblat, Maes 1997).
On peut aussi s'intéresser à la typologie des liens eux-mêmes ((Papy 1994, §2.5.2, p.55) expose cette problématique mais n'entreprend pas de la creuser).

documents conçus sous forme électronique, et gagne une large part des documents dans les entreprises[47].

Le document devient même intrinsèquement dynamique : des calculs composent une page en fonction du moment ou / et d'une indication introduite par celui qui la consulte. Le document est-il alors le 'moule' et ses remplissages virtuels, ou y a-t-il autant de documents que de réalisations de la page ?

La publication sur les réseaux électronique réduit la distance qui sépare le lecteur de l'auteur (contemporain), voire des autres lecteurs. Le texte peut évoluer en échos aux lectures qui en sont faites (Bazin 1994).

g) De l'empreinte à la matrice : un potentiel de réalisations multiples

Le texte électronique se prête à une diversité de formes de présentation : impression papier, réorganisation par des tris et des filtres, etc. Le texte électronique est générateur d'une multitude de textes donnés à la lecture. L'auteur se voit alors confronté à la nécessité de structurer son texte et ainsi de guider (contraindre) les modes d'appréhension et d'accès offerts par le calcul, en les anticipant. (Cotte 1999)

h) Bilan : réinvention de la lecture

> Plusieurs questions se sont posées [...] au fil de notre parcours [étudiant l'impact des nouveaux supports électroniques pour le texte]. Elles concernent la constitution et l'appropriation d'une mémoire collective, le rôle du témoignage, la fiabilité de l'information, la délocalisation du savoir.
>
> Toutes convergent, finalement, vers la question du « sens », c'est-à-dire ce qui donne consistance au fait de vivre en communauté. En effet, la sophistication croissante des dispositifs de traitement de l'information semble s'accompagner d'une évaporation des référents stables, clairement repérables et transmissibles, que produisait l'ordre du livre.
>
> [...] il ne faut pas perdre de vue que les enjeux se situeront, désormais, beaucoup plus du côté des processus de lecture que de la fixation des contenus.
>
> Autrement dit, il faudra veiller à ce que tous les citoyens disposent des outils adéquats et maîtrisent les nouvelles techniques de lecture. Plus profondément, il faudra favoriser le partage des mêmes pratiques [...] [pour] réinventer ensemble, dans le contexte du relativisme et de la virtualité, l'espace public du savoir, sans lequel la connaissance n'est pas culture.
>
> (Bazin 1994)

[47] « L'expérience de consultant de l'auteur de ces lignes lui permet d'avancer que 30 à 60 % des documents produits aujourd'hui par une entreprise n'existaient pas il y a 10 ans ou n'existeront plus dans 10 ans, soit qu'ils correspondent à une nouvelle procédure, soit que la fonction qui les produit ait été redéfinie et qu'ils aient changé de nom, soit que, agencés différemment dans des dossiers nouveaux, ils n'aient plus la même apparence. » (Chabin 1997, §2.6, p. 211)

B. PROPOSITION DE SYNTHÈSE : LES QUATRE FACETTES DU TEXTE

1. Le texte dans le contexte de l'application DECID : champ d'étude

a) De « vrais » textes

L'enjeu est de pouvoir prendre le document de travail le plus banal, tel qu'il est. Il ne s'agit ni d'imaginer des textes qui n'existent pas (textes attestés), ni de s'en tenir à des textes calibrés (issus soit d'une réécriture, soit d'une contrainte qui serait ajoutée à la rédaction). Ce ne sont pas les textes qui sont faits pour l'application, mais c'est l'application qui est au service des textes rencontrés.

Ancré dans la réalité, le corpus affirme l'inanité d'une vision universaliste en s'inscrivant dans un domaine et en y prenant sens.

Même pour le corpus destiné à fournir les caractérisations des destinataires, il n'est pas question de mettre en place un contrôle rédactionnel qui veillerait à une certaine qualité normée des textes. Bien sûr, il n'est pas mauvais que l'application de diffusion ciblée encourage les chercheurs d'EDF à faire une « bonne » rédaction des textes descriptifs de l'activité, c'est-à-dire à donner une description riche, détaillée, informative : cela ne peut que profiter à tout le monde (les destinataires sont mieux caractérisés et donc mieux servis, et l'application de diffusion ciblée est plus performante). Mais l'effort se porte prioritairement sur la conception d'un système suffisamment souple et puissant pour tirer un parti aussi intéressant que possible de l'existant.

b) Des documents scientifiques et techniques, à vocation informative

Les textes qui nous intéressent pour DECID sont ceux qui servent de support de mémorisation, de transcription de connaissance, dans la mesure où s'y applique un travail de compréhension. Ce sont des documents qui entrent dans une pratique professionnelle de constitution et de mise en œuvre d'un savoir scientifique et technique.

La manière d'aborder les textes ne serait pas la même si l'on avait affaire à des textes littéraires ou juridiques par exemple[48]. Il est difficile d'éluder la question de la qualité et des effets du style dans un texte littéraire, et d'y trouver un plan de lecture fortement présent. Une lecture qui s'en tient à l'intrigue est possible et légitime, mais n'est pas pleinement 'convaincante' : on attend de l'auteur littéraire une portée significative de son maniement de la langue, c'est une présomption qui oriente la lecture et la construction d'un sens. Le sens d'une œuvre littéraire n'est pas dans le vrai, pas nécessairement dans le vraisemblable ; il peut faire grande place à la musique des mots (phonétiques). Alors que, dans un contexte scientifique et technique, la tendance est à la normalisation de la formulation et du lexique (un concept est désigné par un terme précis, une pièce est identifiée par un identifiant consigné dans une nomenclature stricte), l'œuvre littéraire recherche des façons inédites de dire les choses.

Nous avons aussi évoqué les textes juridiques : les pratiques de lecture extrêmement nuancées et scrupuleuses, qui permettent, sur un mot, de basculer d'un univers dans un autre, ne correspondent pas au dégrossissement et aux approximations de l'approche adoptée ici. Tout au plus pourrait-on proposer une première lecture, mais là encore sans doute trop insuffisante.

---

[48] La différence que nous voulons souligner est celle de pratiques interprétatives contrastées. La question, dans l'absolu, de la (plus grande) facilité ou difficulté à traiter textes scientifiques ou littéraires, est sans doute une fausse question. Citons par exemple le témoignage suivant :
« une idée circule dans la communauté des Traitements Automatiques des Langues, selon laquelle les textes techniques seraient en général plus simples à traiter, grâce, notamment, à des constructions syntaxiques plus simples. Même si l'on observe effectivement sur nos corpus des caractéristiques linguistiques pouvant simplifier les traitements automatiques, nous montrons [par des extraits du corpus] [...] des constructions syntaxiques qui peuvent être complexes. [Ces extraits] montrent également la présence d'expressions imprécises. » (Assadi 1998, §I.5.2, p. 68)

Bien sûr, il s'agit ici de dominantes. On ne peut refuser à certains rédacteurs de notes techniques un certain art de la rédaction, un style personnel, une finesse extrême dans le choix des termes et des tournures, le recours à quelques figures de style ou procédés rhétoriques... Notre position n'est pas décréter l'insignifiance de tout cela, mais de considérer que ce n'est pas premier dans une application de diffusion ciblée. L'application choisit un point de vue sur les textes : les textes en tant que documents, reflets de connaissances et de compétences dans les domaines d'activité où ils circulent. A ce titre aussi, des documents que l'on classerait comme administratifs ou médiatiques rejoignent le corpus de DECID, tant que l'on considère que le point de vue de l'application apporte une lecture intéressante de ces documents.

c) L'écrit

Le texte est inscrit, fixé sur un support. Les définitions les plus larges envisagent toutes sortes de support, y compris la bande magnétique, qui enregistre un discours.

Nous voulons nous en tenir ici aux documents rédigés pour être lus. Cela a une incidence sur leur constitution. L'auteur ajuste en effet sa composition au mode de réception. Typiquement, la retranscription d'un échange oral n'est pas du même ordre qu'un texte conçu pour être diffusé sous forme de livre ou d'article. Ce qui est oral (à l'origine) oblige à limiter les développements (l'attention et la disponibilité des auditeurs ou interlocuteurs jouent fortement), permet de compter sur une certain interactivité (des questions sont l'occasion de revenir sur un point évoqué rapidement), peut faire grande référence au contexte qui réunit les participants, multiplie les modalités (ton de voix, gestes, projection de transparents), etc.

Un document écrit, pour assurer sa fonction de transmission d'un savoir ou d'informations, est conçu de façon à pouvoir être consulté dans diverses situations et sans contact avec l'auteur. Ceci suppose au contenu du document une certaine autonomie.

Le rôle des graphiques et illustrations peut être d'importance inégale selon les documents. Sans prétendre donner une lecture complète du document, l'analyse dans DECID s'en tient au texte (sans se priver des éléments textuels dans les tableaux, des légendes titrant et commentant les graphiques) et essaie d'en tirer le meilleur parti.

d) Le rapport au texte est celui de la lecture

Dans le contexte de l'application de diffusion ciblée, les documents sont visé sous l'angle de la lecture par opposition à celui de la rédaction. Le travail s'effectue sur les parcours de lecture et les informations construites à partir du document, et non sur le passage d'une idée à son expression linguistique, à la délimitation de ce qui est à dire, aux contraintes rédactionnelles et à la manière d'investir un genre, à la création d'un objet linguistique reçu et reconnu par un certain public. Côté traitements automatiques, c'est se situer parmi les outils d'analyse *vs* les outils de génération[49].

Nous ne demandons pas à la machine de 'produire' des textes, mais de partir d'un existant. Son rôle est de proposer des configurations, des présentations, qui renouvellent les modes d'accès aux textes, sans pour autant pouvoir se substituer à une interprétation humaine.

Nous sommes convaincus que l'ordinateur peut apporter une aide réelle pour aborder des volumes textuels de plus en plus présents (voire oppressants), et qu'il est moins souvent opportun pour porter des données codifiées dans le monde des textes. Les données codifiées ne sont-elles pas exploitables plus efficacement dans leur pureté, leur simplicité et leur acuité originelle ? La mise en texte n'est pas une transformation conservatrice, sans pertes ni gains. Les gains sont issus de l'explicitation d'une lecture d'un format conventionnel (autrement dit, comment déchiffrer telle série de mesures, tel diagramme), et l'introduction d'un point de vue sur les données (qu'il faudra attribuer soit au rédacteur humain, soit au concepteur qui a édicté le comportement de la machine, dans les limites de sa maîtrise de l'algorithme). Les pertes de la transformation des données en texte se

[49] Plus précisément, on pourrait opposer *analyse* vs *génération* au niveau syntaxique, *interprétation* vs *production* au niveau sémantique, et *compréhension* vs *énonciation* au niveau mental, cf. (Rastier, Cavazza, Abeillé 1994, §I.2.2, p. 16)

manifestent comme un effet d'enrobage, surtout vis-à-vis d'un format codifié qui s'est forgé au long d'une pratique et est le plus ajusté à la situation concrète.

e) Des textes en nombre

L'automatisation apporte un relais face à un volume de texte que l'homme n'est pas en mesure d'appréhender. La valeur ajoutée de l'outil n'existe que par cet effet d'échelle.

Très concrètement, il n'y a aucun intérêt à mettre en place un outil de diffusion ciblée et de repérage de destinataires, lorsque l'ensemble des destinataires potentiels est une petite équipe de personnes qui se connaissent bien ! La lecture humaine est évidemment supérieure au traitement mécanique que peut effectuer la machine.

2. Description des quatre facettes textuelles

a) Présentation

Organisation d'ensemble

Annonçons dès à présent les quatre facettes que nous proposons de retenir pour guider la suite de notre étude. La présentation en tableau est utilisée pour faire ressortir des regroupements et oppositions qui structurent les quatre facettes comme formant système. Les annotations dans les marges du tableau proposent une interprétation possible (et ouverte) pour décrire ce système.

	Vision interne : texte *objet unique (objectivité relative)* *paradigme logico-grammatical*	*Vision externe : document* *contextes pluriels (subjectivité)* *paradigme rhétorico-herméneutique*
Domaine : *situé,* *cotexte (culturel, situationnel,...)*	1. MATIERE LINGUISTIQUE	4. RÔLE CONSTITUTIF DE LA LECTURE
Domaine : *système,* *contexte (textuel)*	2. ORGANISATION INTERNE clôture et autonomie, linéarité, hiérarchie, orientation	3. INTERTEXTUALITE

L'ordre de parcours proposé par la numérotation des facettes concorde, au moins pour les trois premières, avec un élargissement des paliers concernés. La linguistique s'attacherait aux petites unités, l'organisation interne déborde la phrase et déploie le texte. L'intertextualité peut être pensée comme un palier encore supérieur (Kanellos, Thlivitis 1997).

Comparaison et discussion

Dans un contexte plus général, celui d'une réflexion sur un format général d'encodage des textes (y compris des parchemins) sous forme électronique, (Sperberg-McQueen 1991) aboutit à une analyse comparable de la textualité. Il énonce et commente neuf axiomes :

> Axiom 1 : Markup reflects a theory of the text.
> Axiom 2 : One's understanding of texts is worth sharing.
> Axiom 3 : No finite markup language can be complete.
> Axiom 4 : Texts are linguistic objects.
> Axiom 5 : Texts occur in / are realized by physical objects.
> Axiom 6 : Texts are both linear and hierarchical.
> Axiom 7 : Textual cross-references form a structure.
> Axiom 8 : Texts refer to objets in a real or fictive universe.
> Axiom 9 : Texts are cultural and therefore historical objects.
> (Sperberg-McQueen 1991)

Il est frappant de trouver de multiples parallèles : on rapportera sans hésitations l'axiome 4 à notre première facette, la seconde facette a son pendant dans l'axiome 6, la troisième facette trouve un écho dans l'axiome 7 mais aussi l'axiome 9 (où Sperberg-McQueen aborde les conventions de rédaction –ce que nous rattachons à la question des genres– et les faisceaux de versions d'un même texte). Notre quatrième facette n'a pas de correspondant direct, mais se retrouve sans peine derrière les réflexions sur le sens et la portée du codage (axiomes 1, 2 et 3).

Toutes nos quatre facettes sont confirmées chez Sperberg-McQueen. Il reste à considérer la correspondance inverse : deux axiomes (le 5 et le 8) semblent oubliés de notre grille.

L'axiome 5 est l'occasion pour Sperberg-McQueen de souligner les effets de mise en page et de disposition matérielle des blocs de textes. Cela peut généralement être réintégré dans notre deuxième facette, nous l'illustrerons dans notre chapitre sur le codage des textes dans DECID. Les cas extrêmes, de disposition élaborées (acrostiches complexes ou calligrammes) relèvent d'effets recherchés dans le domaine littéraire. L'autre incidence du support est celle de son altération possible : un passage illisible, un morceau manquant, une tache malencontreuse. Là encore, cette dimension, significative pour un corpus d'archives, ne paraît pas devoir être ajoutée à nos facettes pour le contexte que nous nous sommes fixé[50]. L'axiome 5 déborde en quelque sorte notre problématique, et serait à reconsidérer pour une adaptation des facettes à un contexte plus général.

En revanche, l'axiome 8 s'écarte des fondements de notre approche en soulignant la dimension référentielle ou dénotationnelle des textes. Notre perspective est de recourir à une sémantique différentielle. Considérons de plus près comment Sperberg-McQueen entend la chose. L'axiome 8 est celui qui fait l'objet du plus bref commentaire :

> Because texts refer to things, wether real or fictive, we need to be able to mark the objects referred to in texts, e.g. place names and personal names. Such markup may be required for stylistic study (to distinguish Mr. Brown from the color brown) or historical study (to see who knew the Paston family) or for subject indexing.

Les conséquences tirées de cet axiome, qui sont donc d'opérer un codage des « objets » du texte, seront critiquées (et rejetées) à l'occasion de la réflexion sur le codage des textes pour DECID. En effet, ce codage, par son aspect autoritaire et tranché, va à l'encontre de la dimension interprétative. Quant à l'exemple sur la confusion Mr. Brown / couleur *brown*, il procède d'une vision sémasiologique et non contextuelle. L'axiome 8 est donc le seul point de désaccord véritable entre la proposition de Sperberg-McQueen et la nôtre.

Dans l'ensemble, nous estimons que les quatre facettes sont un guide plus clair, à garder à l'esprit lors de la construction de la modélisation, que la série des neuf axiomes.

L'utilisation des facettes dans le système DECID

Des affinités particulières se tissent entre les différentes facettes et les chapitres qui présentent la réalisation du système DECID. Ainsi, la recherche d'un format de codage sera l'occasion de revenir sur les structures d'organisation interne d'un texte. Le repérage et la construction d'unités mobilise des connaissances sur la langue. La mise en contraste d'un texte dans un corpus, et la confrontation de texte à texte renvoient de façon évidente à l'intertextualité. Et la

[50] Le fait de ne pas prendre en compte ici la matérialité physique du texte est une approximation. Elle se légitime dans la mesure où, dans les pratiques professionnelles que nous considérons, les variations entre deux éditions (pagination, choix des caractères, etc.) ne sont généralement pas perçues comme significatives : il s'agit toujours du *même* texte.

Nous avons pourtant rencontré des situations où cette approximation est mise en défaut. Un chercheur à qui l'on présentait un document, en lui demandant d'expliciter les critères qui lui permettraient de savoir à qui le faire suivre (soit donc une problématique de diffusion ciblée de l'information, précisément), disait avoir accordé une importance significative aux perforations dans la marge gauche. Cela lui permettait le raisonnement suivant : il y a des trous, donc c'est un feuillet destiné à être rangé dans un classeur, donc c'est une information de travail, qui doit régulièrement être mise à jour –pas un ouvrage de référence, faisant état d'une connaissance stabilisée. C'est typiquement un document utilisé par une équipe dans la phase de réalisation d'un projet, et pas un document de synthèse qui pourrait intéresser un directeur. Une information sur l'aspect physique du document a donc été utilisée pour avoir une première interprétation générale du document.

réflexion sur la dimension interprétative et le rôle constitutif de la lecture trouve des prolongements dans la définition d'une interface, accompagnant l'interaction de l'utilisateur avec les textes.

Mais si l'une ou l'autre facette est plus en évidence selon un certain point de vue, les facettes restantes n'en restent pas moins présentes. Par exemple, l'idée même de construction des unités est une idée interprétative (facette 4), et sa mise en œuvre prendra appui sur des zones de localité dessinées par la structuration interne du texte (facette 3). Et la caractérisation d'un texte repose sur sa description externe (facette 2), mais aussi interne (facette 2).

Ces quatre facettes vont donc être tour à tour activement mobilisées dans la conception du système DECID. Cela montre la cohérence et l'efficacité que nous avons trouvées à ce résumé des dimensions textuelles, dans notre contexte. Cette synthèse ne prétend à aucun statut théorique, encore moins à l'universalité. Elle est dédiée à guider des réalisations pratiques, concernant des documents écrits. Elle trouverait toute sa justification en se révélant utile pour d'autres applications, dans ce domaine des systèmes documentaires : elle y servirait de repère pour l'introduction équilibrée d'une vision textuelle.

b) La langue comme matériau du texte

Une partition musicale transcrivant une symphonie, *(i)* a une structure interne, close et orientée, ligne mélodique et déploiement du contrepoint ; *(ii)* est en rapport de reprise et d'opposition avec d'autres œuvres, notamment celles de la même forme musicale, du même compositeur ou de la même période ; *(iii)* et est actualisée dans une interprétation musicale, qui lui donne sens. Mais nous n'y reconnaissons pas un texte (sinon par métaphore) : son matériau est la gamme et le système rythmique, alors que le texte se forge dans la langue.[51]

Langue naturelle, langage formel

Une langue n'est pas assimilable à un langage formel.

La langue présente des *régularités*, qui servent assurément de point d'appui pour l'interprétation : capacité à donner sens à un énoncé encore jamais rencontré, à un néologisme. Ces régularités jouent aussi un rôle pour réduire la charge cognitive liée au langage : elles permettent de « factoriser » des connaissances.

Certaines régularités peuvent s'apparenter à des lois, elles sont tellement intégrées à la langue et objectivées que le locuteur n'a plus de prise dessus. Mais ces lois (ou normes fortes) ne sont pas des lois logiques, et la langue ne se laisse pas décrire par une mécanique formelle qui détermine ce qui est dicible et ce qui ne l'est pas, et qui régisse le sens en termes de calcul. En effet, rien ne peut empêcher de parler ou d'écrire à sa guise ; et même plus : l'activité interprétative est irrépressible, on ne peut s'empêcher de comprendre « quelque chose » à ce que l'on entend ou lit[52].

---

[51] L'exemple peut être retourné, et d'autres préféreront au contraire une définition du texte qui s'étende à l'œuvre musicale comme à d'autres expressions dans d'autres registres sémiotiques (Barthes 1973). Etablir la réalisation dans une langue comme une caractéristique définitoire du texte est un choix, ici voulu, mais qui ne prétend pas l'invalidité d'autres conceptions (qui ont leur pertinence pour un autre point de vue).

[52] Ce que l'on désigne comme le « caractère *compulsif* de l'interprétation sémantique ». (Rastier 1991, §VIII.2, p. 212) poursuit : « Les linguistes ont beau faire, on sait que les phrases réputées absurdes, voire asémantiques peuvent toujours être interprétées : cela reflète sans doute un processus hautement complexe, comprenant des hypothèses sur un émetteur et une situation de communication fictive, des processus de « réécriture » interprétative (création d'acceptions figurées ou idiomatiques). Mais en deçà, ces opérations ne sont possibles que sur la base des simulacres associés aux lexies. Même des non-mots suscitent de tels simulacres, pour peu que leur formation respecte les règles morpho-phonologiques de la langue : c'est pourquoi on peut lire *Finnegan's Wake* même dans les passages où aucun des mots ne figure au dictionnaire. [...]
Bref, pour simplifier, de la même façon qu'on ne peut s'empêcher d'entendre, on ne peut s'empêcher de comprendre. (Note : C'est là une allégorie du péché originel, ou de moins de la condition humaine : nous sommes condamnés au sens.) »

Problèmes d'ontologies - l'autonomie de la linguistique

Il faut également souligner l'irréductibilité de la sémantique (linguistique) tant à un ensemble de référents identifiables dans le monde réel, qu'à des abstractions conceptuelles qui appartiendraient à un monde des idées, voire à des universaux de pensée. Ceci conduit à abandonner les conceptions référentielle et inférentielle de la linguistique, au profit d'une approche différentielle. Seul ce cadre fait place à la diversité fondamentale des langues –telle qu'il n'est jamais d'identité de sens entre une expression dans une langue et le meilleur équivalent que l'on puisse trouver dans une autre[53]– et à la pluralité des ontologies[54]. D'ailleurs, loin d'être soumise à la simple explicitation d'une réalité externe universelle, univoque et prédéterminée, bref à une distribution d'étiquettes, la langue modèle la vision du monde que se construit un locuteur[55].

Il en va également ainsi des jargons professionnels, ou du langage commun (sociolecte) qui a cours à l'intérieur d'une entreprise[56].

Que le lecteur ne nous en veuille pas pour le plaisir que nous avons à rappeler ici ces propos de Bergson –et pour notre incapacité à les raccourcir...[57]– :

> Quel est l'objet de l'art ? Si la réalité venait frapper directement nos sens et notre conscience, si nous pouvions entrer en communication immédiate avec les choses et avec nous-mêmes, je crois bien que l'art serait inutile, ou plutôt que nous serions tous artistes, car notre âme vibrerait continuellement à l'unisson de la nature. Nos yeux, aidés de notre mémoire, découperaient dans l'espace et fixeraient dans le temps des tableaux inimitables. Notre regard saisirait au passage, sculptés dans le marbre vivant du corps humain, des fragments de statue aussi beaux que ceux de la statuaire antique. Nous entendrions chanter au fond de nos âmes, comme une mélodie quelquefois gaie, plus souvent plaintive, toujours originale la mélodie ininterrompue de notre vie intérieure. Tout cela est autour de nous, tout cela est en nous, et pourtant rien de tout cela n'est perçu par nous distinctement. Entre la nature et nous, que dis-je ? entre nous et notre propre conscience, un voile s'interpose, voile épais pour le commun des hommes, voile léger, presque transparent pour l'artiste et le poète.
>
> Quelle fée a tissé ce voile ? Fut-ce par malice ou par amitié ? Il fallait vivre et la vie exigeait que nous appréhendions les choses dans le rapport qu'elles ont avec nos besoins. Vivre consiste à agir.

---

[53] Ce travail de *translation* (Rastier 1995b) fait le caractère éminemment humain et interprétatif de la traduction. Le traducteur (et interprète) est sans cesse tenue d'adopter un point de vue, fût-il par fidélité à la formulation d'origine (*sourcier*) ou à l'effet de sens obtenu (*cibliste*) (Ladmiral 1986).

[54] Une ontologie est l'ensemble des entités que l'on perçoit et discerne, et la manière dont on les organise. L'ontologie d'un animal se centre par exemple sur ses prédateurs, ses proies, et ses congénères. L'ontologie d'un expert dans un domaine est évidemment plus détaillée que celle du *quidam* dans ce même domaine. Face à un salon meublé, la femme de ménage distinguera d'abord les meubles sous lesquels passer le balais, les meubles à cirer, etc. alors que l'ébéniste pourrait percevoir surtout les différences de styles, les piqûres de vers récentes, etc. et le déménageur sera sensible aux volumes et aux fragilités.

[55] Telle langue nordique déclinera toute une gamme de vocabulaire pour qualifier les différents états de la neige, là où le français courant ne proposera que de trois ou quatre adjectifs. De même, le spectre des couleurs n'est pas découpé de la même manière dans toutes les langues (Hjelmslev 1968, §13, p. 71 sq.).

« Si chaque langue est une vision spécifique du monde, les relations internationales ont la lourde tâche d'« ajuster » non pas des langues, par un simple transvasement de contenu des unes dans les autres, mais bien des conceptualisations différentes, des manières de voir qui souvent ne sont même pas perçues comme telles. [...]

On peut dire que chaque langue est un prisme à travers lequel ses usagers sont contraints de voir le monde. Ce prisme ordonne le monde et l'expérience en catégories qui les rendent pensables. » (de Almeida, Bellamy, Kassai, pp. 51 et 61)

[56] « La vie en entreprise peut fournir un langage commun qui fonctionne par allusion au vécu collectif. C'est le langage élaboré à partir de l'expérience commune en entreprise, et par exemple véhiculé par ses publications internes. Ce sociolecte, qui permet la connivence, est donc facteur d'une communication plus économique et plus efficace sur le lieu de travail.

[...] le 'jargon' professionnel [...] se justifie par le besoin des professionnels d'affiner leur vision et leurs techniques.

Cependant, il n'est pas moins vrai que les langages professionnels [...] ont une fonction distinctive, voire démarcative. Ainsi, l'emploi de certains termes s'explique aussi bien par le besoin de se distinguer de ceux qui ne font pas partie de la profession que par celui de souligner l'appartenance au groupe. »

(de Almeida, bellamy, Kassai, pp. 62 et 89)

[57] C'est un texte qui a joué un rôle précurseur dans mon attrait pour la linguistique.

> Vivre c'est n'accepter des objets que l'impression utile pour y répondre par des actions appropriées : les autres impressions doivent s'obscurcir ou ne nous arriver que confusément. Je regarde et je crois voir, j'écoute et je crois entendre, je m'étudie et je crois lire dans le fond de mon cœur. Mais ce que je vois et ce que j'entends du monde extérieur, c'est simplement ce que mes sens en extraient pour éclairer ma conduite ; ce que je connais de moi-même, c'est ce qui affleure à la surface, qui prend part à l'action. Mes sens et ma conscience ne me livrent donc de la réalité qu'une simplification pratique. Dans la vision qu'ils me donnent des choses et de moi-même, les différences inutiles à l'homme sont effacées, les ressemblances utiles à l'homme sont accentuées, des routes me sont tracées à l'avance où mon action s'engagera. Ces routes sont celles où l'humanité toute entière a passé avant moi. Les choses ont été classées en vue du parti que j'en pourrai tirer. Et c'est cette classification que j'aperçois beaucoup plus que le contour et la forme des choses [...]. L'individualité des choses et des êtres nous échappe toutes les fois qu'il ne nous est pas matériellement utile de les apercevoir. Et là même où nous la remarquons (comme lorsque nous distinguons un homme d'un autre homme), ce n'est pas l'individualité même que notre œil saisit, c'est-à-dire une harmonie tout à fait originale des formes et des couleurs, mas seulement un ou deux traits qui faciliteront la reconnaissance pratique.
>
> Enfin pour tout dire, nous ne voyons pas les choses mêmes ; nous nous bornons, le plus souvent à lire des étiquettes collées sur elles. Cette tendance, issue du besoin, s'est encore accentuée sous l'effet du langage. Car les mots (à l'exception des noms propres) désignent des genres. Le mot, qui ne note de la chose que sa fonction la plus commune et son aspect banal, s'insinue entre elle et nous, et en masquerait la forme à nos yeux si cette forme ne se dissimulait déjà derrière les besoins qui ont créé le mot lui-même. Et ce ne sont pas seulement les objets extérieurs, ce sont aussi nos propres états d'âme qui se dérobent à nous dans ce qu'ils ont d'intime, de personnel, d'originalement vécu. Quand nous éprouvons de l'amour ou de la haine, quand nous nous sentons joyeux ou tristes, est-ce bien notre sentiment lui-même qui arrive à notre conscience avec les mille nuances fugitives et les mille résonances profondes qui en font quelque chose d'absolument nôtre ? Nous serions alors tous romanciers, tous poètes, tous musiciens. Mais le plus souvent nous n'apercevons de notre état d'âme que son déploiement extérieur. Nous ne saisissons de nos sentiments que leur aspect impersonnel, celui que le langage a pu noter une fois pour toutes parce qu'il est à peu près le même, dans les mêmes conditions, pour tous les hommes. Ainsi, jusque dans notre propre individu, l'individualité nous échappe. Nous nous mouvons parmi des généralités et des symboles [...]. Nous vivons dans une zone mitoyenne entre les choses et nous, extérieurement aux choses, extérieurement aussi à nous-mêmes.
>
> (Bergson 1900, §III.1)

Heureusement, la langue n'est pas enfermée et figée dans ses *mots*.[58] Et pour poursuivre le propos de Bergson, un *texte* peut être reçu telle une œuvre –ciselée par son auteur–, et le lecteur, se faisant interprète, est créateur de sens.

Du débat sur le rapport entre la langue et la pensée, retenons le caractère irréductible de l'une à de l'autre, et leur relativité culturelle.

> Il est de la nature du langage de prêter à deux illusions en sens opposé. Etant assimilable, consistant en un nombre toujours limité d'éléments, la langue donne l'impression de n'être qu'un des truchements possibles de la pensée, celle-ci, libre, autarcique, individuelle, employant la langue comme son instrument. En fait, essaie-t-on d'atteindre les cadres propres de la pensée, on ne ressaisit que les catégories de la langue. L'autre illusion est à l'inverse. Le fait que la langue est un ensemble ordonné, qu'elle révèle un plan, incite à chercher dans le système formel de la langue le décalque

[58] Quant à la critique linguistique de ce passage de Bergson, elle pourrait reprocher deux détails (?) : *(i)* il n'y a pas la langue de l'humanité, mais une diversité de langues, irréductibles les unes aux autres, ancrées dans une culture et modelées par l'histoire d'une société ; *(ii)* les « choses » ne préexistent pas indépendamment, leur identification et leur délimitation sont déjà un acte interprétatif.
Et l'accent mis sur un façonnement de la langue utilitariste mériterait discussion :
« Il faut affirmer nettement que le signe-outil s'oppose au véritable symbole, tout comme au signifiant linguistique. La pression de l'utilité, si elle était constante, empêcherait tout simplement *l'apparition* d'un fonctionnement linguistique ou symbolique, qui repose précisément sur la mise à *distance* de l'utilité [...]. Voilà pourquoi les scénarios utilitaristes de l'origine du langage ne sont pas crédibles [...]. Et s'il y a bien un *intérêt* de l'humain pour le langage, il faut reconnaître qu'il est *d'abord* d'un autre ordre. [...] [Cet intérêt] s'apparent-t-il au désir de classer, bricoler et cuisiner des mondes symboliques –soit à une façon d'aménager, d'habiter, de concilier le soi, le monde, les autres ? ou bien est-ce d'emblée la constitution du désir comme poursuite d'un enjeu, retracée par une intrigue certes finie, mais qu'il *faut* toujours recommencer (ce désir serait toujours celui d'une autre suite, d'une autre fin) ? » (Visetti 1999, p. 147)

> d'une « logique » qui serait inhérente à l'esprit, donc extérieure et antérieure à la langue. En fait, on ne construit ainsi que des naïvetés ou des tautologies. (Benveniste 1966, §6, p. 73)
>
> Nous nous en tenons [...] à cette position mesurée : les signifiés des langues et les représentations mentales sont les uns comme les autres des formations culturelles. Ils ne se confondent pas et se conditionnent mutuellement. Cependant leur unité est telle qu'une position dualiste qui admettrait une détermination unilatérale du signifié à la représentation, ou la détermination inverse, ne permettrait pas de saisir la complexité de leurs interrelations. (Rastier 1991, §III.4, p. 96)

La langue n'est ni un décalque du monde et de la réalité perceptible, ni une projection des catégories de la pensée et des concepts mentaux. Mais elle tient un rôle médiateur entre ces deux pôles.

> le rôle médiateur du monde sémiotique [entre le monde physique et le monde des représentations] [...] tient à la double nature des signes [...], qui relèvent du physique par leurs signifiants, et qui peuvent être associés à des représentations mentales par les signifiés qu'on leur attribue, directement ou non.
>
> Ce rôle s'entend de deux façons, puisque le biologique est inclus dans le physique. Relativement au physique (au sens très restreint de l'objectivité perçue), le sémiotique est le médiateur entre les « états de choses » et leurs représentations. En d'autres termes, le face-à-face millénaire et figé qui oppose le sujet à l'objet devrait s'effacer avec le dualisme dont il procède : car on ne passe pas directement d'une objectivité physique à une représentation subjective. [...]
>
> Touchant la médiation entre le représentationnel et le biologique, nous formulons l'hypothèse que le sémiotique constitue corrélativement l'instance médiatrice entre les états mentaux et les états cérébraux –indépendamment du fait que les échanges sémiotiques structurent une part des tissus cérébraux.
>
> (Rastier 1991, Epilogue, p. 243)

Le texte, objet linguistique, et l'objet de la linguistique

C'est récemment que le texte a été reconnu comme l'objet réel de la linguistique, ou autrement dit l'observable de la science étudiant la langue. Certaines théories linguistiques ont été bâties sur le mot, ou le plus souvent sur la phrase (ou énoncé, ou proposition). Affirmer que le texte est l'objet premier de la linguistique, c'est dire que les mots sont des unités non pas données et figées, mais construites à travers leurs usages en contexte. C'est également percevoir que la compréhension d'un texte n'est pas simple affaire de juxtaposition ou de composition de phrases, et que la syntaxe voit son rôle relativisé.

Incidence pratique pour DECID

Ces considérations liées à la nature linguistique du texte jouent un rôle fondamental dans la conception de l'outil de caractérisation des textes développé ici, tout particulièrement pour la définition d'unités élémentaires, descriptives et caractérisantes. Cela oriente la mise en œuvre des moyens apportés par les outils de traitement automatique du langage naturel.

c) La construction interne du texte, sa clôture et son orientation

Cette facette reprend déjà ce qui a trait à la structure à la fois arborescente et orientée d'un texte, telle qu'elle est présentée dans une table des matières, ainsi que ce qui a été dit sur la progression. Soit donc deux axes qui ordonnent matériellement le texte : un axe « horizontal », linéaire et séquentiel, et un axe « vertical », hiérarchique.

Avertissement : des propriétés situées, relatives

La définition et la pertinence de cette facette se conçoivent bien dans le domaine que nous nous sommes fixé. Il s'agit bien de textes conçus sous une forme écrite, et destinés à être utilisés dans une activité de lecture. Le support d'un livre ou d'un ensemble de feuillets matérialise nettement et d'emblée son caractère clos et délimité, ce qui est peut-être moins évident d'une conversation, qui peut glisser d'une préoccupation à une autre, sans qu'il y ait au final le sentiment d'une unité d'ensemble et d'une composition réfléchie. D'autre part, la structuration interne dont il est ici question est très présente et marquée dans les documents scientifiques et techniques, par opposition par exemple aux romans.

La façon de concevoir et organiser un texte n'est pas indépendante de son support. La facette que nous considérons ici est au tout premier plan concernée par les transformations du texte papier au texte électronique, du document imprimé à l'hypertexte. Sans s'arrêter aux modifications superficielles, il faut y percevoir des facteurs d'évolution qui influent sur les modes de pensée[59]. Les pages du Web n'ont des pages d'un livre que le nom : elles fragmentent le texte et rendent son contour plus diffus. Les liens hypertextes font éclater les rapports de proximité et de localité, et les repères du parcours linéaire se dérobent dans la mise en abyme des renvois illimités. Les grandes lignes que nous traçons pour décrire la facette de l'organisation interne du texte sont une base pour saisir des compositions porteuses de sens, mais aussi pour contraster les formes de mobilisation de ces structures.

Dimension horizontale

La linéarité orientée du texte non seulement souvent guide la lecture, mais joue un rôle interprétatif. Par exemple, dans la plupart des genres, il est de convention qu'un élément simplement évoqué en un point du texte, mais *a priori* non connu du lectorat (personnage, notion plus technique, etc.), a été introduit dans les pages précédentes. Et le lecteur qui a pris la liberté d'entrer directement en un point du texte sait que ce qu'il lit peut avoir des liens de dépendance avec ce qui précède et s'appuyer dessus.

Cette logique cumulative est tempérée par des phénomènes de proximité. Le lecteur peut retenir davantage dans le détail ce qu'il vient de lire, et garder une idée plus synthétique des premières parties. Il reste que, sur un support écrit, il y a toujours la possibilité latente de revenir au besoin sur un passage et de réactualiser un moment de lecture antérieur, comme de prendre du temps pour mieux mémoriser une partie du texte, –deux possibilités que n'offre pas la communication orale directe.

Le déroulement linéaire, ponctué par de multiples découpages (paragraphes, sections, etc.), crée des zones de localité. Ces zones ont un rôle de premier plan dans la construction et l'actualisation des unités sémantiques : par leur proximité et leur mise en relation, des éléments se renforcent, se propagent, d'autres sont virtualisés, inhibés. La dynamique interprétative est extrêmement sensible à ces interactions à différentes échelles de contexte.

Ce découpage est le lieu d'introduction de dénombrements mnémoniques : numérotation d'un ensemble de points, rythme qui équilibre l'ensemble du texte (par exemple : quatre parties, qui se divisent en trois sections chacune).

Dimension verticale

L'organisation hiérarchique du texte est le plus souvent étiquetée par des intitulés (chapitres, sections, etc.), qui, selon le genre, remplissent diverses fonctions : formulation synthétique du thème

[59] Il y a là tout un programme de recherche, tel celui mené par Bruno BACHIMONT, qui étudie le passage de la *raison graphique* (écriture traditionnelle) à la *raison computationnelle* (dynamique apportée par le support électronique) :

« La technique permet d'accroître et d'élargir les possibilités de donner un sens au monde en proposant des structures d'appréhension nouvelles. Autrement dit, la technique permet de constituer de nouvelles catégories conceptuelles pour penser le monde, c'est-à-dire de constituer de nouvelles rationalités. On peut alors penser que chaque type de système technique sera constitutif d'un type particulier de rationalité. Ainsi, on parlera de *raison graphique* pour les techniques d'écriture.

L'écriture est un exemple paradigmatique du rôle constitutif de la technique dans la genèse des connaissances. (Goody 1979) a montré comment l'apparition de l'écriture dans une culture s'accompagne de l'émergence de nouvelles catégories intellectuelles comme les listes, les tableaux, les formules. Ces structures conceptuelles sont des artefacts de l'écriture. Ses traducteurs ont proposé le terme de 'raison graphique' pour désigner le type de rationalité constituée par la technique de l'écriture.

L'apparition de supports d'inscription dynamiques comme l'ordinateur renouvelle la technique de l'écriture. Se pose alors la question de savoir dans quelle mesure ces nouvelles techniques vont reconfigurer la géographie du savoir et constituer des nouvelles structures conceptuelles. La 'raison computationnelle' correspond à la rationalité constituée par les supports dynamiques et le but de la recherche entreprise ici est de la mettre en évidence et d'en préciser les caractéristiques. »

(`http://www.biomath.jussieu.fr/~bb/FullRecherches.htm`, 20 octobre 1998)

abordé dans la partie, effet d'accroche, repérage dans un plan-type, fonction de la partie (ex. : introduction, glossaire), etc. La constante est le rapport global posé entre l'intitulé et la partie correspondante. On peut aussi noter la convention générale d'antéposition : le titre ou l'intertitre est placé avant le texte ou la partie à laquelle il est rattaché. D'où un effet d'annonce (qui joue sur les anticipations du lecteur), et une manière d'accompagner le mouvement naturel de détermination du local par le global. En effet, la formulation synthétique du titre pose un cadre interprétatif préparant l'entrée dans le détail du texte.

Le texte forme un tout, il s'affiche comme une unité close, du moins autonome, et complète. D'où une définition du texte comme « unité linguistique de taille maximale, appréhendable dans une perspective interne » (Bronckart & *al.* 1985, introduction à la première partie, p. 11). Il ne faut pas s'y tromper : que l'unité soit de taille maximale n'implique pas une grande étendue (au plan syntagmatique), la taille maximale signifie simplement qu'on ne se le représente pas comme un extrait mais comme un tout. Le texte n'est pas créé par le nombre de pages ni celui de phrases. Un panneau qui indique *Danger* ou *Interdiction de fumer* est déjà un texte. Nous rejoignons pleinement Todorov quand il écrit :

> La notion de *texte* ne se situe pas sur le même plan que celle de phrase (ou de proposition, syntagme, etc.) ; en ce sens, le texte doit être distingué du *paragraphe*, unité typographique de plusieurs phrases. *Le texte peut coïncider avec une phrase comme avec un livre entier* ; il se définit par son *autonomie* et par sa *clôture* (Ducrot, Todorov 1972, § *Texte*)

Formant ainsi une entité, le texte a un nom, qui permet de l'évoquer, de le citer : son titre.

Le texte n'est pas censé faire appel à des éléments extérieurs à lui et non conventionnellement connus du lectorat auquel il s'adresse (autonomie). Il peut bien sûr, au fil de l'exposé, renvoyer explicitement à d'autres textes, mais en termes de complément, pas de passage obligé pour poursuivre la lecture commencée. Le texte porte donc en lui-même tout le nécessaire pour construire un univers qu'il invite le lecteur à parcourir.

Définir le texte comme un tout c'est aussi lui supposer une certaine homogénéité, une cohérence d'ensemble. Un document qui rassemble plusieurs composants disparates sera plutôt présenté comme un recueil de textes (*textes* au pluriel), tout en considérant que l'acte de les avoir réunis dans un même document laisse présumer une cohérence d'ensemble.

Le fait que le texte soit à la fois une structure close et orientée lui confère des extrémités, qui sont autant de passages particuliers. Selon l'axe horizontal, ce sont le début et la fin, soient notamment, au palier du texte, toutes les pièces liminaires et annexes qui bordent le développement central. Le rôle singulier du début et de la fin peut encore se retrouver au niveau de chaque partie (délimitée comme un texte dans le texte), voire au palier du paragraphe[60]. A la zone de début sont associées les présentations générales, le contexte introductif ; à celle de fin, les reprises synthétiques et conclusives, et éventuellement l'annonce du début suivant, une transition.

Des considérations générales à la réalisation concrète

Cette présentation a incontestablement un caractère naïf et, prise à la lettre, se heurterait à de nombreux contre-exemples. Ce qu'il convient d'en retenir, ce sont les points d'attention qu'elle propose. Dans un contexte applicatif donné, pour un corpus particulier, ce sont autant d'aspects qui peuvent jouer un rôle particulier dans l'analyse : dans les textes que je considère, qu'induit la linéarité (cf. la *composante tactique* chez F. Rastier), quelle place prend-elle dans la lecture et la construction de l'interprétation ? et qu'en est-il de l'imbrication hiérarchique des parties ? et de leurs intitulés ? et des zones de début et de fin ? etc.

Cette facette de la structuration interne du texte invite à voir l'impact de la mise en page et des modes de lectures qu'elle soutient : découpage plus ou moins marqué, disposition, points d'accroche. Par exemple, le lecteur peut sauter certains passages, encouragé par une présentation qui en annonce le caractère marginal, ou qui le présente comme un complément plus technique facultatif dans le cadre du texte, ou encore qui unit une série d'alternatives parmi lesquelles une seule est pertinente pour le lecteur et a été repérée. On relève dans le texte d'un document scientifique et

[60] Dans ses analyses textuelles, (Dupuy 1993) accorde ainsi un « poids » particulier aux zones de début et de fin (première et dernière phrase du texte, début et fin des paragraphes).

technique des éléments qui se distinguent et ressortent du reste, mais avec des statuts très différents : exemple[61], résumé, atomes d'information (de type formules), éléments-clés de validation et d'évaluation (hypothèses explicitées, formulation de résultats), etc. Les références d'une bibliographie sont typiquement hétérogènes quant à leur signification.

D'une façon générale, l'organisation du texte met en relation tel et tel items, résume ou souligne une information centrale ; capter la teneur informative d'un texte, c'est bien repérer ce qui est présenté de façon à en être retenu (ce qui frappe l'attention, mobilise la mémoire).

Dans la conception d'un traitement automatique, tous ces points concernant la structuration interne du texte interviennent dès le choix d'un format de codage des textes. Ils peuvent ensuite alors entrer en ligne de compte pour l'étape de caractérisation des textes, qui prépare elle-même la présentation d'un texte vis-à-vis des autres textes et pour l'utilisateur.

Pour les genres privilégiés, qui font l'objet d'une étude approfondie, la description de la structuration interne peut être enrichie, notamment en ajoutant la caractérisation de parties conventionnelles (que nous appelons « rubriques »), leur organisation entre elles, et leur composition propre. Dans le cas de DECID, ce travail a été lancé pour le corpus des descriptifs d'activité, utilisé pour la constitution des profils de destinataires. Les efforts spécifiques pour ce corpus doivent cependant rester mesurés : ce corpus joue actuellement un rôle central, mais est amené à évoluer avec la politique de mise en œuvre de l'ordonnancement à la Direction des Etudes et Recherches d'EDF (les tendances actuelles sont : réduction du nombre de textes, modifications du plan-type, couverture plus générale de chaque texte).

d) L'intertextualité

Une facette qui s'impose

Un document n'est jamais perçu isolément[62]. Il s'entoure d'autres textes : textes par rapport auquel le rédacteur se positionne, textes que le lecteur utilise ou a rencontré dans la même pratique. Et plus généralement : textes que l'ouvrage côtoie sur l'étagère, textes présents à l'esprit ou prêts à refaire surface du fond de la mémoire, textes dont l'analogie est plus ou moins « rationnelle » (du rattachement au même sujet à la ressemblance de la couverture...).

Intertextualité et pertinence

Perçu parmi une multitude d'autres textes, le texte reçoit une valeur relative, il prend sens par rapport aux autres[63]. Il y a une « attente » intertextuelle du lecteur : tout texte semble devoir se justifier par ce qu'il comporte de différent, de novateur, d'original par rapport aux autres, tout en s'inscrivant dans un existant et se calant sur des repères connus. Le texte respecte un canon, tout en

[61] « Le rôle et la structure des exemples ne peuvent être décrits que par rapport au contexte dans lequel ils s'insèrent. En premier lieu, ils diffèrent selon les discours et les genres. Ainsi, dans le discours philosophique, ils paraissent avoir un rôle de problématisation ; dans le discours scientifique, un rôle d'objectivation ; dans le discours technique, un rôle de typification. » (Rastier, Cavazza, Abeillé 1994, §VII.7.2, p. 195)

[62] Et plus généralement, son contexte (autres textes, origine du document, situations de lecture prévues) contribue directement à lui donner sens (Poitou, Ballay, Saintive 1997, p. 12 sq.).
Voir aussi (Thlivitis 1998), qui élargit la *Sémantique Interprétative* de François Rastier à une *Sémantique Interprétative Intertextuelle*, notamment par une généralisation de ses structures de classes sémantiques (formation de classes de textes).

[63] La situation d'un document dans un ensemble de documents accessibles, dans un fonds qui a été rassemblé, est chargée de sens, et participe directement à la valeur qu'il reçoit : « Citons comme exemple le fait de savoir pour une lettre de réclamation d'un client qu'elle est isolée ou qu'elle fait partie d'une série de cinquante lettres du même type, et si elle est isolée, qu'elle l'a toujours été ou qu'elle l'est aujourd'hui par suite de la destruction des autres ; le fait de savoir que ce compte rendu de séance d'une page est comparable à tous les autres comptes rendus de ce comité ou, au contraire, que c'est la seule séance qui a bénéficié d'un compte rendu, ou encore que d'ordinaire les comptes rendus font 20 pages ; le fait de savoir que tel projet de plan a été étudié et par qui, qu'il a été approuvé ou non, que le plan a été réalisé ou qu'il est resté sans suite. » (Chabin 1997)

s'en démarquant, ce qui lui donne da raison d'être[64]. D'où deux formes de saillance, c'est-à-dire de « relief » des éléments du texte au fil de la lecture : la saillance de ce qui est très visiblement en commun avec d'autres textes (une citation), et la saillance de ce qui se démarque de l'expression commune (un terme particulier, une idée originale).

Enfin, la pertinence d'un document à traiter d'une question ne s'évalue pas tant comme une adéquation du document en lui-même à cette question que comme prenant place dans une composition équilibrée de plusieurs documents complémentaires : si je vais rechercher de l'information sur une question dans un centre de documentation, il est vraisemblable que je reparte avec deux ou trois ouvrages, par exemple l'un qui fait référence, l'autre plus d'actualité, le troisième qui apporte un point de vue original et stimulant sur mon sujet, sans qu'aucun des trois ne puisse être considéré comme meilleur que les autres.

Un document se situe non seulement par le sujet qu'il aborde, l'information ou les données qu'il apporte, mais aussi par le point de vue adopté. Un même texte peut présenter plusieurs points de vue, dont l'attribution est significative : ce qui est assumé par l'auteur, caution, éviction. Autrement dit, l'intertextualité ne se centre pas sur la seule composante thématique, généralement considérée via des statistiques distributionnelles de mots-clés. La composante dialogique aurait aussi un rôle de structuration intertextuelle, mais à un niveau plus fin, et demande par exemple la prise en compte des modalités et de leur association avec les éléments thématiques. Les composantes thématique et dialogique interviennent de façon différente, étant donné qu'il est difficile de confronter des documents selon leur composante dialogique indépendamment de leur composante thématique, et de trouver un sens à cette confrontation. Ce qui se conçoit plutôt, c'est de consulter des documents présentant des positions contrastées au sein d'un même débat, tout en cherchant à cerner les principaux pôles d'opposition.

Une communauté intertextuelle remarquable : le genre

Les genres (ou types) textuels sont tout à fait concernés par la facette de l'intertextualité. Chaque genre constitue lui-même un ensemble intertextuel, car tout texte renvoie implicitement aux textes du même genre. Un document, écrit à l'intention d'une certaine utilisation, se rallie à un genre, en en adoptant les conventions plus ou moins codifiées. (Même pour s'en démarquer, il doit encore garder un lien au genre visé.) Techniquement, réunir un corpus de documents d'un même genre permet d'étudier les régularités et conventions du genre. Si ensuite l'on est amené à considérer des documents de genres différents, les caractéristiques d'un genre sont des paramètres particuliers, que l'on peut choisir de faire ressortir (le genre devient une dimension de contraste fort) ou au contraire d'estomper (par exemple, on souhaite trouver des similarités thématiques, même entre des documents de genres différents). En revanche, l'absence de caractérisation préalable des genres favorise des rapprochements inadéquats, telle formulation conventionnelle et banalisée dans un genre étant remotivée par rapprochement avec une expression relevant de la thématique d'un texte d'un autre genre.

La reconnaissance et la prise en compte des genres est une étape nécessaire dans la réalisation d'une application qui opère des calculs sur des textes[65]. L'étude et la caractérisation des genres

[64] D'où le sentiment paradoxal, l'absurde latent –et même avoué–, dans la nouvelle de Jorge Luis BORGES, qui raconte la laborieuse réécriture *littérale* du *Don Quichotte* par un certain Ménard (recueil *Fictions*). Et pourtant, on y joue de la différence essentielle des deux œuvres, qui tient justement au contraste entre les contextes sur lesquelles elles se profilent. Il nous faut donc compléter notre formule : le texte prend sens par rapport aux autres, mais aussi par les autres qui lui sont rapportés. Quant à Borges, il conclut : « Ménard (peut-être sans le vouloir) a enrichi l'art figé et rudimentaire de la lecture par une technique nouvelle : la technique de l'anachronisme délibéré et des attributions erronées. [...] Attribuer l'*Imitation de Jésus-Christ* à Louis-Ferdinand Céline ou à James Joyce, n'est-ce pas renouveler suffisamment les minces conseils spirituels de cet ouvrage ? ».

[65] « La typologie des genres textuels paraît indispensable pour les traitements automatiques. Soit en général, car l'analyse des corpus en situation montre que le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres. [...] Soit en particulier, car les genres sont déterminés par des pratiques sociales spécifiques, dans lesquelles les applications informatiques prennent place. Elles doivent donc tenir compte de ces contraintes propres à ces pratiques où elles s'insèrent. » (Rastier, Cavazza, Abeillé 1994, §VII.4.1)

dominants attendus ne couvre cependant pas nécessairement tout le registre des textes effectivement soumis au système, outre qu'elle suppose un investissement (temps et expertise) qui dépasse aisément les moyens consentis. L'exigence de robustesse d'une application opérationnelle conduit à concevoir un fonctionnement acceptable malgré l'absence d'information de genre sur un texte.

Le corpus, esquisse matérielle de l'intertexte

L'intertextualité conduit à la question du corpus, de sa constitution et de son étude. Suffisamment large et représentatif[66], le corpus constitue un univers de référence pour le traitement, un univers textuel[67], qui permet déjà des caractérisations justifiées et significatives. Le corpus est exploré selon l'approche différentielle, mettant en valeur les ressemblances et les différences, ce qui est uniforme et ce qui est contrasté. Les études de statistiques lexicales et textuelles se penchent sur ce domaine.

Evoquer le corpus annonce déjà la facette suivante : l'association de textes est un acte herméneutique, c'est une clé de lecture.

e) Le rôle constitutif des lectures

Multiples déterminations

Le rapport du texte au(x) lecteur(s) est double : un texte est écrit *pour* des lecteurs, et prend sens dans son appropriation *par* un lecteur.

En étant destiné à un certain lectorat, le texte rejoint d'autres textes rentrant dans une même pratique et adressés à la même communauté. En cela, la facette du texte comme objet de lectures rejoint celle de l'intertextualité. Tout ce que l'on ajoutera ici, c'est que le concept de *genre textuel* s'en trouve renforcé, puisqu'un genre est justement cet ensemble de textes rassemblés par une même pratique de lecture.

Un même texte se prête généralement à plusieurs types de lectures, qui sont autant de points de vue effectifs sur lui. Par exemple, face à un article scientifique, et de prime abord, le chercheur novice peut être particulièrement sensible au titre, aux mots-clés ; l'expert du domaine voit immédiatement l'auteur, la revue, et 'décode' la bibliographie. Qui plus est, l'objectif de lecture ne

---

[66] Le corpus est toujours en deçà de l'intertextualité effective : qui saurait cerner même un seul des ensembles de textes, plus ou moins présents dans la mémoire d'un lecteur lorsqu'il se trouve face à un texte donné ?
« The interpretive act by which we make sense of these presuppositions does not simply rely on receiving signs and recognizing their signifieds. Instead, we insert these signifiers into the network of discourses always already present but never fully elaborated during our reading of the text. Intertextual interpretation is therefore the survey of a set of possible meanings that readers attempt to disentangle from a textthat is nothing more than fragments from countless other texts knitted together.
Investigating a discursive space can never reach any sort of ultimate mapping. No database can be constructed that would permit researchers to explore every discourse that resonates in a text, especially since cultural, social, and political discourses are not fully transcribed and machine-readable. Nevertheless, databases such as ARTFL enable us to explore intertextuality in ways that did not exist before computers. »
(Wolff 1994)

[67] (Thlivitis 1998) défend l'idée qu'il est possible de rendre compte de toutes sortes de contextes, y compris des contextes « non linguistiques », par le biais de l'ajout de textes ; le texte étudié peut alors être pleinement considéré au sein d'un tel *univers textuel*, au sens le plus fort du terme :
« Nous faisons une hypothèse principale pour la suite de ce travail, inspirés en partie d'un constat empirique : dans les analyses de textes il est toujours possible d'exprimer à l'aide d'un texte (*e.g*. commentaire, critique littéraire, exposé pédagogique d'une analyse littéraire) toutes sortes de connaissances utilisées pour l'analyse. [...] cette hypothèse ne supprime pas la nécessité d'un *entour* mais affime la possibilité de l'internaliser dans un *univers textuel*. [...] [Nous présentons donc] la notion d'*intertextualité*, [...] de façon intuitive, comme un moyen de 'capter' l'entour de manière textuelle. » (Thlivitis 1998, §1.1.3, pp. 17 & 19)
Pour notre part, nous nous en tiendrons au fait que l'intertexte est un environnement significatif, sans statuer sur son éventuelle complétude. Quant à la possibilité et la validité d'une description centrée sur le texte comme objet linguistique, nous pensons qu'elles sont effectives, grâce à la notion de *pôles intrinsèques* du texte, cf. (Rastier 1998).

répond pas nécessairement au contrat de lecture, implicitement déclaré par l'appartenance à un genre : le texte prépare des lectures privilégiées, mais ne peut pas les déterminer entièrement.

Le texte ne prend sens et valeur que dans une lecture : c'est d'une certaine manière ce qui lui donne d'exister, d'être présent[68]. L'adjectif « constitutif » choisi pour désigner cette quatrième facette textuelle garde toute sa force.

L'acte interprétatif

La question de la relation d'un lecteur à un texte est celle de l'interprétation. Nous proposerons plus loin un point sur les diverses manières dont est entendue l'activité interprétative, point qui accuse la divergence de vues sur la question. Sans entrer dès à présent dans le débat, nous allons mentionner ici les principales propriétés que nous voulons retenir dans notre étude de la textualité.

L'interprétation a une dynamique, elle recherche et construit un sens fondé sur le texte et pertinent pour le lecteur. Le texte ne détient pas une signification pleine et unique qu'il s'agit d'extraire (un *contenu*, à tirer d'un *contenant*). Le lecteur aborde le texte avec des attentes (en fonction de la situation qui l'amène à rencontrer ou à remarquer ce texte) et des présomptions (de cohérence, d'intérêt, de facilité / difficulté, etc.), qui déjà orientent sa manière de percevoir et de parcourir le texte, lui lecteur, à ce moment-là. Le texte prend un certain relief : points saillants, points latents (perçus mais non encore considérés) ; des points se font proches, d'autres se répondent tout en se contrastant. Ce parcours 'inégalitaire' laisse quelquefois des traces : marque-page, annotations, surlignages.

En un résumé imagé, le *texte* donne un ensemble de points de repères (des éléments linguistiques, typographiques, un positionnement intertextuel, etc.) ; l'*interprétation* est un parcours circulant parmi ces points de repères, où l'image de la circulation n'interdit ni une lecture linéaire, progressant régulièrement du début à la fin du texte, ni une lecture plus intermittente ou partielle, qui scrute, saute, revient, annote, etc. ; la *pertinence* enfin se définit par l'intégration des fruits de ce parcours dans l'univers personnel du lecteur, étant pertinent ce qui n'apparaît ni redondant ou superflu dans cet univers, ni trop étranger et sans ralliement significatif.

Pas de texte sans lecture

Le texte renvoie, étymologiquement et de façon suggestive, au textile, au tissage. L'image est souvent reprise pour rappeler le croisement, en chaque point du texte, des axes syntagmatiques et paradigmatiques. Les vocabulaires ont quelques connivences : la trame du tissu, et la trame de l'intrigue...

Le texte pourrait être aussi cette texture, dans l'entrelacs des mots de la langue et du vécu du lecteur. L'un apporterait la chaîne, l'autre la trame ; si bien que le texte sans le lecteur s'effiloche et perd toute consistance. Image qui rappelle aussi qu'il y a mille manière de passer la navette : quels fils sont saisis, avec quelle alternance ; choix de couleurs, choix de matières. Chaque lecteur, en recevant le texte, s'y implique, pour l'investir de sens[69].

Orientations pour DECID

Identifier un texte à une représentation formelle particulière nie la dynamique de l'interprétation –car toute lecture se construit, évolue, et c'est un processus fondamental pour l'appropriation d'une connaissance dans une communication écrite. Dans le cas de DECID, en caractérisant un point de vue d'un lecteur par son profil, puis en confrontant un profil et un document, l'enjeu est bien de rendre compte de la pluralité des lectures auxquelles se prête un même document (y compris un document technique).

---

[68] (Adam 1990, Introduction §4, p.28) cite ainsi RUTTEN F. (1980) - « Sur les notions de texte et de lecture dans une théorie de la réception », *Revue des sciences humaines*, 177, Université de Lille III, p. 83 : « On ne lit pas un texte, il y a texte parce qu'il y a eu lecture ».

[69] La lecture peut en partie se matérialiser au sein du texte sous forme d'annotations. Pour une étude des *actes annotatifs*, voir (Virbel 1994).

Au plan du traitement automatique envisagé, il est clair que la prise en compte de la facette interprétative du texte a de multiples incidences techniques, particulièrement pour tout le processus de caractérisation (filtrages et sélections, pondérations et évaluations, ajouts, réductions), comme pour la conception de l'interface, qui détermine les interactions possibles (lectures) entre l'utilisateur humain et les textes (texte soumis en requête, textes formant la base des profils, textes présentés en réponse). Les résultats issus du calcul prennent le statut d'une base suggestive, à partir de laquelle l'utilisateur construit lui-même la réponse qu'il recherche, plutôt que le statut de sortie du système (*sortie* dans tous les sens du terme), sélection déterminée et *a priori* bonne sur laquelle il n'est pas prévu de revenir.

La discipline qui s'est penchée de longue date sur l'acte de lecture et d'interprétation est l'herméneutique : elle pourra être consultée comme guide. Moins intimidantes peut-être, la kyrielle de méthodes de lectures[70] montre l'application de certains principes (herméneutiques sans doute, mais rendus familiers) à des textes et dans des pratiques rencontrés dans l'entreprise. La facette interprétative peut (doit) être prise en compte dès la description linguistique : la Sémantique Interprétative de François Rastier fait le lien entre des unités linguistiques et une dynamique interprétative.

f) Epilogue : résonances de l'image du texte comme tissu

TISSAGE	TEXTE
Le tissage n'est pas scindable, il est d'un seul tenant ; ses lisières sont bien définies, elles font même l'objet d'un traitement spécial (retour de la navette, point d'arrêt).	Le texte est pensé avec un début et une fin, avec des hypothèses et une conclusion, avec un tenant et un aboutissant, avec sa clôture, son sujet ; il pose également son cadre.
Le tissu n'est pas tant la somme matérielle des fils que leur agencement étudié ;	La « forme » (le style, l'organisation des parties) n'a pas moins d'importance que le « fond » ; elle est généralement à son service pour mieux le révéler.
si bien que d'une certaine manière, l'air et le vide entre les croisées des fils réalisent aussi le tissu (c'est indéniable qu'ils contribuent notamment à sa souplesse, à ses propriétés thermiques, etc.).	La trame et la chaîne du tissu sont classiquement associées respectivement aux axes paradigmatique et syntagmatique. Le texte emprunte sa forme au matériau linguistique, mais le sens n'apparaît qu'au détour des mots.
Il serait par conséquent délicat de définir le tissu uniquement par sa matérialité ; il est plutôt appréhendé pour sa fonctionnalité, toute liée à son étendue couvrante.	Aussi paraît-il vain de considérer le texte comme une série de caractères ; ce sont plutôt les propriétés de l'objet de communication et d'expression humaine qu'il s'agit de repérer.
Le tissu joue des effets de motifs et de texture : ce sont des effets globaux bien que créés par des contributions locales (insignifiantes à elles seules).	L'interprétation du texte se nourrit d'informations locales et globales, simultanément : on ne peut avoir une compréhension juste des unes sans une connaissance des autres et réciproquement (cf. l'isotopie)

(parallèle repris de (Bommier 1994a, p. 22))

[70] Les manuels et formations de lecture rapide ne se comptent plus. Certaines méthodes de lecture se dotent d'acronymes mnémoniques comme SQL2R (Survoler, Questionner, Lire, Réfléchir / Raisonner, Répondre).

C. TEXTES ET TRAITEMENTS AUTOMATIQUES : OBSERVATIONS QUANT AU STATUT DU TEXTE DANS LES PÔLES DE RECHERCHE ACTUELS

1. Linguistique

a) Texte et lexique

La recherche et les réalisations en Traitement Automatique du Langage Naturel accordent une place de plus en plus dominante au lexique. Les formalismes syntaxiques s'ancrent dans le lexique et l'investissent (LFG, TAG[71]). Le(s) crédit(s) accordé(s) à la constitution et à l'entretien des ressources lexicales occupe(nt) une place dominante : l'éventail s'étend des ressources les plus « linguistiques » (dictionnaires) aux plus conceptuelles (thesaurus et autres ontologies), en passant par les terminologies, les vocabulaires multilingues et les réseaux sémantiques (cf. le très grand succès de *WordNet*[72]). Dans ce cadre, le statut du texte apparaît ambivalent.

D'une part, le texte est opposé au terme, comme dans le parallèle suivant :

texte	**terme (terminologie)**
occurrence	*type*
sens (description/interprétation en contexte)	*signification (construction normative)*
linguistique (morphosyntaxique)	*conceptuel (ontologies)*

Notons que cette opposition n'apparaît pas comme une opération d'exclusion, mais de complémentarité (alternance de deux points de vue, distincts et compatibles). Cette perspective prête à se focaliser sur la problématique de la terminologie, sans confusion avec une *autre* problématique qui serait celle du texte.

D'autre part, sauf pour le cas de *termes* (vocabulaire conventionnel, technique, de spécialité) pris dans leur domaine, l'incidence de l'environnement d'un mot est nettement affirmée, selon l'idée que le mot ne prend sens qu'en contexte[73]. Dans cette mouvance se situent des recherches sur la polysémie, visant à expliciter les mécanismes permettant d'identifier « le » « bon » sens d'un mot. Les guillemets précédents indiquent déjà notre hésitation : peut-on inventorier les sens d'un mot ?[74] Peu de linguistes se refuseraient pourtant à reconnaître l'aspect continu des effets de sens. En mettant le contexte à l'honneur pour la désambiguïsation, c'est peut-être encore une fois aller à l'encontre de la textualité, puisque l'entour du mot n'est considéré que pour mieux l'isoler ensuite. Il s'agirait à

---

[71] LFG : *Lexical Functional Grammar*, en français la *grammaire Lexicale Fonctionnelle*, conçue à la fin des années soixante-dix par Joan BRESNAN et Ronald KAPLAN.
TAG : *Tree Adjoining Grammar*, ou *grammaire d'Arbres Adjoints*. Lancée par A. JOSHI au milieu des années soixante-dix, elle est toujours l'objet de développements actifs, notamment autour d'Anne ABEILLÉ en France.
L'ouvrage français de référence sur les formalismes syntaxiques actuels est :
ABEILLÉ Anne (1993) - *Les nouvelles syntaxes –Grammaires d'unification et analyse du français*, Armand Colin, coll. Linguistique, 327 pages.

[72] On trouvera une présentation de *WordNet* dans (Habert, Nazarenko, Salem 1997, §III.4, p. 85 sq.).

[73] Dans la pratique, cependant, les contextes cités débordent rarement la phrase, il n'est donc pas du tout évident que l'on s'intéresse directement à une dimension textuelle (une forme d'isotopie par exemple).

[74] Tout dépend aussi de la finesse recherchée (l'homographie n'est pas du même ordre que la pluralité d'acceptions... Pour une description systématique, voir par exemple (Martin 1983, §II.II, pp. 75-95) ou (Pottier 1987, §V.3)). En outre, avivé par une approche sémasiologique, le problème de l'ambiguïté peut souvent paraître artificiel à plusieurs égards : en convoquant parallèlement des sens complètement étrangers au contexte du corpus ; en décrétant que l'auteur *a priori* ne recourt pas à ce mécanisme (alors qu'il peut être mobilisé délibérément, par jeu, par effet de style fondé sur la richesse sémantique du mot, pour éviter une prise de position précoce ou dangereuse...).
Pour un argumentaire plus complet, dense et illustré, voir par exemple (Gayral 1998, §1.1).

l'inverse de pouvoir faire le lien entre les occurrences et le référentiel terminologique préexistant, sans filtrer inconsidérément l'enrichissement sémantique émanant du contexte en présence (effets de cooccurrence, saillance,...). En définitive, ces approches courent le risque de dévier vers une vison éclatée, morcelée, du texte. Les expériences de traitements inter-lingual sont là pour souligner les limites des correspondances mot-à-mot, voire terme à terme.

b) Texte et phrases

Si le texte se construit dans l'enchaînement des phrases, alors les études sur l'anaphore, les connecteurs, rejoignent les études sur la textualité (Fuchs & *al*. 1993, §8.1.1). Néanmoins, *le textuel est d'une autre nature que le transphrastique* : le texte ne se réduit pas à une suite de phrases, ni même à la composition d'éléments locaux[75]. Si macro-syntaxe il y a, le jeu des préfixes ne doit pas occulter qu'elle n'est pas dans la stricte continuité de la (méso- ?) syntaxe : en témoigne l'échec reconnu des « grammaires de discours » pour l'analyse des textes (Charolles 1988) (Vandendorpe 1994) (Rastier, Cavazza, Abeillé 1994, §VII.2). Le niveau textuel ne saurait être totalement pris en charge par une extension directe des outils existants pour la morpho-syntaxe[76], ce serait mésuser de

[75] Un titre comme « Au-delà de la phrase » (article de Christos CLAIRIS dans *Modèles Linguistiques*, X (2), pp. 79-82) est symptomatique d'une linguistique qui souligne les limites de la phrase, qui déclare son intention de sortir de ce cadre, mais qui ne peut s'empêcher de penser en termes de phrases (cette fois-ci au pluriel), sans basculer dans l'univers d'une autre nature qu'est le texte. Même difficulté pour Gérard SABAH (introduction à la partie consacrée à la *Structuration du discours*, cf. extrait ci-dessous) et, dans une moindre mesure, pour Catherine FUCHS et Bernard VICTORRI (Fuchs & *al.* 1993, chapitre *Compréhension automatique de textes*) : l'adoption de la conception phrastique dans ces ouvrages de synthèse est à la fois symptôme et vecteur de son retentissement dans la communauté du Traitement Automatique du Langage Naturel. Sans compromettre la possibilité d'une sémantique unifiée, il faut réaffirmer que le palier de la phrase et le palier du texte sont irréductibles l'un à l'autre.
« Les recherches sur les langues se sont longtemps concentrées sur l'étude de la structure et du sens de phrases isolées. Cette étape est bien sûr nécessaire pour traiter du discours, mais il faut également préciser comment ces contenus se combinent pour former des ensembles plus importants : après avoir construit des représentations internes des diverses phrases, on doit les intégrer dans une structure qui dépasse le niveau de la phrase et représente une compréhension à un niveau plus global, montrant qu'un discours est plus qu'une simple succession de phrases. Il s'agit donc principalement de mettre en évidence l'unité d'un texte ou d'un dialogue, en effectuant des raisonnements permettant de découvrir les liens qui existent entre les différents éléments qui le composent. De ce point de vue, il est clair qu'une approche purement linguistique ne peut construire le sens global d'un texte qu'à partir de ses constituants et de la déclaration explicite des relations qui existent entre eux. » (Sabah 1989, introduction à la deuxième partie, p. 187).

[76] « Au cours des trois dernières décennies, le texte s'est de plus en plus affirmé comme objet d'étude autonome tout en résistant aux diverses tentatives de formalisation qu'on a tenté de lui appliquer. Parmi les pistes empruntées par la recherche, deux grandes orientations se sont partagé les faveurs. L'une se concentre sur l'organisation de la signification et s'intéresse au niveau profond du texte en étudiant son articulation paradigmatique : c'est la sémiotique. L'autre, s'inscrivant dans la ligne des études sur la phrase, a d'abord privilégié l'axe syntagmatique pour tenter de déboucher sur une linguistique transphrastique.
[...] Comme le note de Beaugrande [DE BEAUGRANDE Robert (1990) - « Text linguistics through the years », *Text*, 10 (1/2), pp. 9-17], une telle ambition reposait sur l'hypothèse fondamentale qu'il n'y avait entre la phrase et le texte que des différences d'ordre quantitatif dont on pourrait ultimement rendre compte en renforçant les systèmes de règles. C'était ignorer radicalement, comme le note le même auteur, que 'ce qui fait qu'un texte est un texte n'est pas sa *grammaticalité* mais sa *textualité*' (p. 11). Plus globalement, on pourrait aussi reprocher à cette approche de n'avoir pas perçu que, du système phonologique au texte, en passant par la morphologie et la syntaxe, l'emprise des contraintes diminue progressivement, pour faire place à une liberté croissante à mesure qu'on monte dans la hiérarchie des composantes du langage. Comme je l'ai montré ailleurs, la 'grammaire de récit' n'a pu s'établir qu'en occultant cette spécificité du texte, et en effectuant sur son corpus des opérations de sélection et de réécriture garantes des 'découvertes' qu'on voulait y faire [VANDENDORPE Christian (1989) - *Apprendre à lire des fables : une approche sémio-cognitive*, Montréal, Le Préambule / Balzac, pp. 87-98]. » (Vandendorpe 1994, pp. 331-332).

ceux-ci. Le risque ici serait de miser sur un opportunisme de mauvais aloi, orienté par les moyens au détriment des buts[77].

Les modélisations du texte (Sabah 1988, 1989), quand elles procèdent par instanciation dynamique d'une représentation statique, en intégrant les résultats de l'analyse successive des phrases une à une, n'accèdent pas à la dimension globale du texte. Ainsi, la DRT[78], avec ses représentations cumulatives, s'écarte indéniablement des représentations suscitées par une lecture humaine (vision synthétique plutôt qu'exhaustive et littérale, par exemple, ne serait-ce que pour des considérations de mémoire).[79]

Alors que la phrase (ou la proposition) reste traditionnellement au centre de nombre de travaux linguistiques (cf. le choix des exemples de travail), le texte se fait peu à peu reconnaître comme *l'objet réel* de la linguistique (Rastier 1993c)[80]. Mais peu de systèmes implémentés considèrent ces questions de textualité[81]. Sans doute sont-elles hors de propos pour certaines finalités

---

[77] Dit de façon brutale, ce n'est pas parce que l'on sait faire certaines analyses au niveau de la phrase, qu'il serait bon de les transposer au niveau du texte, sans s'inquiéter de leur utilité ni de leur validité.

[78] *Discourse Representation Theory*, élaborée par H. KAMP.

« Kamp (84) se propose de construire une représentation dynamique des divers éléments du discours constituée d'espaces imbriqués ou indépendants. Cette représentation (appelée structure de représentation discursive) est construite progressivement : une phrase nouvelle du discours est intégrée dans la représentation existante et provoque l'expansion de la représentation discursive. Un « espace » est construit pour représenter une phrase ; il contient la mention des éléments qui interviennent dans cette phrase et les relations que la phrase explicite entre ces constantes. » (Sabah 1988, §10.3.2.1)

Le modèle repose sur des principes compositionnels, et une sémantique dénotationnelle : les objets sont identifiés par des constantes et variables formelles, leurs relations codées par des fonctions (prédicats logiques). Toute l'attention est focalisée sur la description des phénomènes de portée (matérialisée par des espaces, cadrant ce qui est « accessible » ou non, et représentés par des boîtes), dans la phrase (quantificateurs) ou d'une phrase à l'autre (anaphores).

« Une nouvelle phrase du discours provoquera la création d'un espace englobant l'espace ancien et construit de façon analogue. Les relations ensemblistes entre les divers espaces construits permettent alors d'expliquer les références possibles d'un élément à l'autre (anaphores). Outre les contraintes usuelles (genre, nombre,...) des règles précisent les possibilités d'accès d'un espace à l'autre et un ordre de préférence dans le parcours des diverses constantes, par exemple dans la recherche des antécédents des pronoms. » (Sabah 1988, §10.3.2.1)

« Les deux formalismes 'généralistes' les plus utilisés pour analyser les textes [...] [sont] la théorie de la représentation du discours (DRT) de H. Kamp et [...] la théorie des graphes conceptuels de J. Sowa. En effet, l'ambition de ces deux formalismes, que nous avons présentés à propos de l'analyse sémantique de la phrase, dépassent largement ce cadre : l'une de leur principales qualités est justement de permettre de traiter des phénomènes inter-phrastiques comme l'anaphore, en se donnant les moyens de représenter des cadres de référence qui permettent le calcul des co-références et, dans une certaine mesure, des relations spatio-temporelles. » (Fuchs & *al*. 1993, §8.2.3, p. 238)

Mais, même sur le terrain de ses spécialités (*donkey-sentences* pour les quantificateurs, et anaphores), la théorie n'est pas à l'abri des critiques (Bourigault 1990).

[79] Cette remarque porte sur l'utilisabilité des représentations construites du texte, et non directement sur leur mode de construction.

[80] Et l'intuition de (Hérault 1981), qui prévoit un module d'*hyperanalyse* dans son système :

« notre unité d'analyse ne saurait être ni le mot, ni la phrase, mais [doit] englober de longues fractions du texte. » (Hérault 1981, p. 95)

[81] Voici le constat émis dans l'introduction d'un des principaux ouvrages de référence et de synthèse, en français, dans le domaine du Traitement Automatique du Langage Naturel :

« Que signifie 'comprendre le langage' ? La réponse n'est pas claire pour l'homme, mais elle est encore plus complexe si l'on se demande comment montrer qu'un système automatique a compris. Les recherches se sont longtemps limitées à la seule phrase or, il est clair que le contexte dans lequel une phrase apparaît doit être pris en compte et que le sens d'un texte n'est pas la simple juxtaposition du sens des phrases qui le composent. Les différents niveaux de compréhension possibles montrent que les inférences (raisonnements) nécessaires peuvent être très variés, allant de l'extraction du sens du texte à ses diverses interprétations possibles. De plus, on peut reconnaître des unités de sens dans des parties plus importantes : dialogues, descriptions de scènes, explications, récits, etc... Nous manquons encore de méthodes de représentation et d'analyse efficaces du sens de toutes ces unités, bien que des termes comme linguistique du discours ou linguistique du texte commencent à avoir droit de cité. » (Sabah 1988, §1.1, p. 20)

(ex.: analyse syntaxique pour elle-même). Surtout, elles soulèvent d'importantes difficultés sur le plan de la faisabilité, car elles vont à l'encontre de l'architecture calculatoire des moyens informatiques et de son sémantisme sous-jacent[82] : compositionnalité (Nazarenko 1998), conception d'une signification fonctionnant par dénotation, décontextualisation[83]. Nonobstant, la linguistique n'a pas à être bridée par les contingences techniques : on peut viser à élaborer une théorie rationnelle du texte, quitte ensuite à l'appauvrir sciemment (en fonction du contexte applicatif), à l'approximer, dans une modélisation formelle pour l'implémentation (Rastier, Cavazza, Abeillé 1994, §II.5, note)[84].

c) Texte et statistiques sur corpus

La linguistique à base de corpus fait face aux problèmes d'échelle (Jacob 1994) : il s'agit de réaliser une industrialisation robuste de systèmes de Traitement Automatique du Langage Naturel, d'acquérir automatiquement des informations (notamment linguistiques) sur un grand ensemble de textes à traiter. La magie de promesses formulées sur la base de quelques exemples ponctuels s'évanouit pour laisser place à plus de modestie[85]. Une conception booléenne des résultats décline : ici, pas de conclusion sur le registre du *vrai /faux* ou *valide / invalide*. Y a-t-il pour autant un regrettable compromis, mettant en balance robustesse (donc utilisabilité) et formalisation (garante de rigueur) ? La lexicométrie, la construction d'indicateurs statistiques, l'application de l'analyse des données, ont ouvert la voie d'une rationalité conférant la primauté à une vue d'ensemble, plus souple, et réintroduisant la tâche interprétative[86].

L'approche à base de corpus est-elle une approche textuelle ? Cette perspective est en tout cas favorable à faire jouer l'intertextualité, les rapports qu'entretiennent les textes les uns par rapport aux autres ; souvent, aussi, le prétraitement du corpus (en faisant appel à des instruments de Traitement Automatique du Langage Naturel classiques) s'efforce de respecter sa nature linguistique. Enfin, dans le soin apporté à la constitution d'un corpus homogène il y a déjà des considérations typologiques. Plusieurs des facettes que nous avons proposées sont donc présentes.

2. Autour de l'informatique

a) Texte et cognition (en Intelligence Artificielle)

Dans la lignée de la tradition philosophique, se refusant à dissocier langage et pensée, la linguistique est comptée dans les disciplines majeures des sciences cognitives, et le texte a pu être considéré comme manifestation directe de la compréhension[87]. Ce succès ne semble pas avoir profité

Cette formulation de la problématique reste tributaire d'une conception qui vise au calcul d'une représentation du sens d'un texte, par opposition à une conception plus herméneutique.

[82] « Les composantes sémantiques ne sont ni ordonnées ni hiérarchisées *a priori*. [...]
Ces propriétés rompent avec le modularisme et la séquentialité qui ont dominé bien des théories linguistiques [...]. [Celles-ci décrivaient] l'action successive de modules autonomes déclenchés dans un ordre strict, la sortie du premier devenant l'entrée du suivant, etc. Ainsi, ces modules n'interagissent pas en cours de traitement [note : Ces présupposés appartiennent au sens commun de l'Intelligence artificielle, et plus généralement de l'informatique. Ils y sont nécessaires pour éviter l'explosion combinatoire et l'élaboration des algorithmes très complexes qui régissent les processus parallèles.] » (Rastier 1989, §I.8.A)

[83] L'objectif de généralité et de portabilité ont pu aussi contribuer au succès de la syntaxe (et donc des analyses phrastiques), selon l'idée (fausse, cf. les études de corpus contrastant les genres) que le système grammatical est constant à travers tous les usages de la langue, par opposition à la sémantique, jugée coûteuse car dépendante du domaine d'application.

[84] Nous reprendrons ce débat dans le chapitre sur la définition des unités pour DECID.

[85] Ainsi, dans certains acronymes (TAO par ex.), le A glisse de « automatique » à « assisté(e) ».

[86] L'interprétation intervient ici aussi bien *en amont du calcul*, dans la définition du corpus, de son découpage en unités, qu'*en aval*, dans le commentaire éclairant les résultats chiffrés et leur usage.

[87] Notamment, en psychologie, la production d'un résumé a beaucoup servi de test de compréhension.

à la notion de texte : ne voit-on pas le texte réduit à une production *linguistique quelconque*, elle-même *imparfait véhicule* de transmission d'information[88] ?

Par exemple, le recours à un formalisme externe, pour la « représentation des connaissances » « extraites » d'un texte, pourrait[89] s'interpréter comme une double négation de la textualité. En premier lieu, dissocier information (référentiel/objectif) et point de vue (énonciatif/subjectif), postuler l'indépendance du fond (à extraire) par rapport à la forme (à neutraliser), faire strictement la part entre syntaxe et sémantique, bref isoler une connaissance de son support et de sa manifestation, procèdent sans doute (à différents niveaux) d'un même mouvement qui demanderait légitimation[90], puisque gommant l'ancrage sémiotique (linguistique, textuel) de l'objet[91]. En second lieu, déterminer « le » contenu d'un texte, c'est faire fi de sa dimension herméneutique[92], à savoir du travail interprétatif que suppose la construction et l'appropriation d'une connaissance. Car la connaissance (ou la signification) n'est pas immanente au texte. Dans la triade langage / cognition / interprétation (ou lire / savoir / comprendre), les membres entretiennent des rapports complexes et aucun n'est réductible aux autres.

Les Systèmes de Consultation de Documentation Technique, tels celui présenté par (Assadi 1998), redonnent aux textes leur place. Il s'agit d'un hypertexte contenant quatre modes d'accès à l'information : une table des matières, une recherche en texte intégral et deux index, l'un représentant les concepts du domaine et l'autre les tâches de l'utilisateur. Chaque concept est relié à ses occurrences[93], et la construction même du réseau (l'*ontologie régionale*) est basée sur le corpus[94]. Le modèle des tâches de l'utilisateur rend compte du contexte de consultation, par des ingénieurs et techniciens dans le cadre d'une activité bien déterminée.

---

[88] Ainsi, François RASTIER s'écarte d'une conception utilitariste du texte, et oppose, au *paradigme* (positiviste et réducteur) *de la communication*, où le langage est vu comme code, le *paradigme de la transmission*, rendant compte de la « réélaboration interprétative » à l'œuvre dans le commentaire, la tradition, et la traduction. « Où la communication transmet le signifiant, la transmission communique le signifié, [...] non par un transport d'information, mais par création et recréation. » (Rastier 1995b, p.166).
« A la conception instrumentale du langage qui prévaut notamment chez les cognitivistes orthodoxes, nous opposerons d'une part que la langue n'est pas un instrument, mais une condition historique *a priori*, un *milieu*. D'autre part, que si elle est certes utilisée pour communiquer, elle ne se réduit pas à cette fonction. Seul un instrument est déterminé par sa fonction. » (Rastier 1991, §III.5, p. 102, note 1)

[89] Bien sûr, le traitement automatique supposera la mobilisation d'une modélisation. En revanche, l'excès ici condamné est de tenir la représentation construite à partir du texte comme équivalente, voire « supérieure » à celui-ci ! Nous pensons que le texte et sa représentation sont incommensurables : la représentation perd nécessairement une part de la richesse du texte (par ex. autres interprétations possibles, euphonie, etc.), mais (si elle est conçue astucieusement) c'est pour mieux se prêter au traitement voulu en mettant en évidence les éléments requis.

[90] Des nuances seraient à introduire : ainsi, un document écrit d'information scientifique et technique a une vocation affichée de transmission de connaissances de travail, et vise à une relative autonomie par rapport à l'auteur (il est destiné au public le plus large, moyennant un certain niveau de connaissances).

[91] Avec de vertigineux corollaires : à la limite, si le texte doit être considéré comme un simple vecteur de connaissances, l'étude du texte en tant que tel est-elle du ressort de la linguistique ?

[92] Bien qu'au cœur de la problématique de l'*information retrieval*, dont la conception très réductrice de la pertinence est dénoncée dans cette thèse, Gerard SALTON fait écho de cette critique dans la section qu'il consacre aux systèmes experts (Salton 1989, §11.4.2).

[93] Et même mieux : une relation ternaire, indécomposable, unit un (ou plusieurs, ou aucun) concept, un (ou plusieurs, ou aucun) texte, et une (ou plusieurs, ou aucune) expression (terme) (Assadi 1998, §1.4.2.1, p. 55 sq., et §3.3.1, p. 156)

[94] La méthodologie, baptisée *analyse conceptuelle interactive* (ACI), adopte des principes issus de la sémantique différentielle de François Rastier. L'ACI comporte deux phases : une phase d'amorçage, l'analyse macroscopique, et une phase itérative de raffinement, l'analyse microscopique (on trouvera une présentation synthétique dans (Assadi 1996)). La partie interactive intègre pleinement la nécessaire intervention d'une compétence interprétative, celle de l'expert humain.

b) Texte et hypertexte

La définition d'une norme comme SGML[95] pour la structuration des documents électroniques a ouvert une réflexion sur les « fonctionalités » du document, les attentes du lecteur vis-à-vis des usages possibles. Pour ce qui concerne le texte, l'accent est mis sur son *découpage* et sa *structuration*. L'hypermédia invite aussi à caractériser la place spécifique du texte, et son degré (variable) d'autonomie, dans des documents faisant conjointement appel à d'autres registres sémiotiques : image, graphique, son.

Les applications (livre électronique, Internet), avec la matérialisation de *liens* entre les documents, sont également l'occasion d'étudier les parcours de lecture. Il s'agit de trouver des modes pertinents d'ancrage, de typage, de gestion des liens ; leur activation dans un processus de *navigation* met au jour des problématiques de mémorisation et de repérage, au sein d'un texte et /ou d'un ensemble de documents.

c) Texte et ergonomie des interfaces

La promotion du langage naturel, pour les interfaces homme-machine, va de pair avec le souci de rejoindre le plus grand nombre d'utilisateurs dans leur pratique courante. Au delà de la reconnaissance frustre de mots-clef prédéfinis (filtres), le dialogue est apparu comme le centre des préoccupations. La généralisation et la réutilisabilité de résultats sur l'enchaînement des répliques apparaissent d'autant plus délicates que le dialogue ne constitue pas (linguistiquement) un type de texte[96], et que les régularités que l'on décèle doivent être attribuées au genre sous-jacent (à cerner et à caractériser). Il n'est alors pas surprenant de voir liée la faisabilité et la réussite d'un système avec la bonne délimitation d'un domaine fermé d'application, puisque ce sont justement les pratiques, dans leur cadre, qui induisent les genres.

L'utilisation d'un système informatique étant elle-même une pratique particulière, on remarque effectivement que le dialogue prend une tournure spéciale quand l'interlocuteur est une machine (explicitation, style télégraphique, dépersonnalisation) (Fuchs & *al*. 1993, §10.1.3.1). Par ailleurs, la motivation généreuse d'exotérisme cache peut-être une illusion démagogique (Rastier 1991, §VI.4), à savoir que dans certains cas l'ergonomie réside moins dans une convivialité apparente que dans l'usage d'un langage approprié, plus efficient[97], et plus clair (car ne reportant pas le travail d'interprétation sur une machine dont on ne connaît avec précision les rouages internes).[98]

3. Systèmes documentaires et recherche d'information : le modèle vectoriel

a) Une approche tout naturellement textuelle

Dans les systèmes documentaires, ce qui est soumis au calcul, ce sont d'abord des textes : texte d'un document, texte de la requête. D'où une sage heuristique : rien n'oblige à entrer dans des

[95] *Standard Generalized Markup Language*, pour les formats d'échange entre documents électroniques.

[96] La conversation ne s'inscrit pas dans une typologie des textes. Si on voulait lui trouver une unité, ce serait plutôt un type de structure « dialogique » (une *séquence*, au sens de (Adam 1992)), qui apparaît dans plusieurs genres.

« Même les échanges linguistiques qui paraissent les plus spontanés sont réglés par les pratiques sociales dans lesquelles ils prennent place, et relèvent donc d'un discours et d'un genre. La conversation, par exemple, n'est pas un genre ni un discours –malgré certains théoriciens de l'analyse conversationnelle. Nous disposons tous de plusieurs genres conversationnels, liés à des pratiques différentes, de l'entretien à la conversation de cantine, et dont chacun a ses spécificités. » (Rastier, Cavazza, Abeillé 1994, §VII.4.1)

[97] cf. d'ailleurs ce que nous observons sur les requêtes adressées à DECID : quand il n'a pas à sa disposition la forme électronique (fichier) qui lui permettrait de procéder par copier / coller, l'utilisateur préfère souvent taper quelques mots-clés, plutôt que de soumettre un texte dans son entier. L'influence des pratiques de recherche documentaire n'est pas non plus négligeable ici.

[98] L'utilisateur doit être un tant soit peu en intelligence avec les traitements effectués, dans leur principe : cf. B. BACHIMONT (1992) - *Le contrôle dans les systèmes à base de connaissances*, Hermès.

considérations sur la signification des mots pris isolément. En cela, l'approche est résolument textuelle.

> It is well known that most text words and expressions are highly ambiguous when considered out of context. On the other hand, principles are also at work that limit the potential for ambiguous interpretation, at least in ordinary discourse and nonliterary writing. The need for writers and readers to communicate regularizes the language to some extent [...]. Consequently, much of the language ambiguity disappears in ordinary discourse when the wider linguistic, social, and temporal contexts are taken into account.
>
> The environment in which text units and complete texts are embedded also plays a major role in the influential *use theory* of meaning proposed by Wittgenstein and others. The following quotation is reflective of this point of view (Wittgenstein 1953) :
>
> « For a large class of cases –though not for all– in which we employ the word 'meaning' it can be defined thus : the meaning of a word is its use in the language ».
>
> The use theory seems especially appropriate in information retrieval in which the major concern is not with the intrinsic meaning of the words and text units in isolation but with the global meaning of complete text entities. In the retrieval environment it is not normally necessary to assign specific semantic interpretation to individual text words. Instead, it suffices to determine whether different texts –for example, a query text and the texts of stored documents– are close enough to be relatable, that is, whether text use and text environments are congruent.
>
> In practice, it is not always easy to determine the precise purpose and social environment in which a given text is placed. It is, however, normally possible to study the linguistic context in which the words occur. [...] one assumes that word meanings are related when text words and expressions appear in similar local contexts [...].
>
> (Salton, Allan, Buckley 1994, pp. 98-99)

b) Et pourtant : l'oubli du texte

Les conférences américaines comme TREC ou SIGIR ont centré l'attention sur les performances des systèmes documentaires, performances mesurées comme une adéquation entre les résultats des calculs et une pertinence considérée comme donnée. Il s'agit pour le système d'être au plus près d'une collection de réponses préenregistrées, « pour telle requête, tel document est pertinent, tel autre ne l'est pas ».

Le travail de recherche se concentre alors sur l'ajustement de formules mathématiques, pour obtenir le meilleur accord avec les corpus d'association requête - document. La question de la textualité est bien souvent éludée : quelles propriétés du texte sont significatives pour l'application de recherche documentaire ? comment la modélisation choisie en rend-elle compte ? Tout ceci n'est évoqué qu'évasivement, et n'évolue pas beaucoup : fréquence comme indicateur d'importance, discriminance comme indicateur de significativité. En revanche, c'est la discussion de tout un bataillon de formules qui est exposé, pour conclure sur « la meilleure ». Le choix de l'introduction d'une certaine mesure, ou de l'utilisation de tel type de fonction, est justifié parce que « c'est ce qui marche le mieux », après tâtonnements expérimentaux : « voyez la courbe précision / rappel, elle est au-dessus de toutes les autres... » Bien que TREC se défende d'être une pure compétition de systèmes, pour être avant tout un lieu de débat scientifique sur les qualités et limites des différentes techniques, la discussion critique se situe davantage au niveau des heuristiques avantageuses, que de la conception du texte sous-jacente[99].

En fait, la question du texte s'identifie ici très souvent au problème des « textes longs », dès que l'on se harsarde à quitter les corpus fondateurs que sont les résumés des notices bibliographiques[100], les dépêches et les artciles de presse : nous allons voir ce qu'il en ressort dans ce qui suit.

---

[99] Notre analyse n'est pas isolée : « Du côté de la prise en compte des phénomènes linguistiques, les sciences de l'information les envisagent souvent dans la perspective de l'interrogation, en termes de taux de rappel et de précision. On cherche rarement des explications dans les textes sources. » (Bertrand-Gastaldy 1993)

[100] On peut défendre la pertinence de s'en tenir à un corpus de résumés, mais les arguments sont inégaux :
« La performance [des] méthodes [mises au point dans ce travail] n'est envisagée ici que sur des corpus de textes condensés, à savoir des formes textuelles réduites dans lesquelles on s'attache à mettre en valeur les notions

c) La normalisation homothétique

Si le texte est représenté comme un ensemble de mots (éventuellement avec des pondérations), la comparaison directe des ensembles est avantageuse pour les textes les plus longs : ils ont plus de mots, donc ont plus facilement des mots en communs avec d'autres textes.

Or bien sûr un texte long n'a pas nécessairement à « peser » plus lourd qu'un texte court. Chaque texte représente une unité, *a priori* tout aussi achevée et complète quelle que soit sa longueur effective. La parade la plus classique, pour rééquilibrer l'influence des différents textes, est la normalisation : un texte long a beaucoup de mot, mais chaque mot en commun n'apporte qu'une petite contribution à la ressemblance. Ce procédé ne donne pas entière satisfaction, car les textes de plus de quelques pages sont alors mal représentés : leur grand nombre de mots les pénalise.

Plus fondamentalement, un modèle fonctionnant par homothétie pour norm(alis)er tous les textes du point de vue de leur longueur ignore les particularités stylistiques actives chez les uns et les autres (impact ou évitement des répétitions, texte complet ou extrait,...). Si transformation il y a, pour passer d'un texte court à un texte long, elle ne se laisse pas penser en termes de dilution, et le texte long n'est pas un texte court « gonflé ». La réduction homothétique est égalisante, au lieu d'être (s)élective. Elle fait disparaître les saillances locales (une notion importante, si elle n'est abordée que passagèrement, est très vraisemblablement fortement dévaluée par la réduction). Elle maintient toute la diversité du vocabulaire du texte le plus long, au prix d'une dévaluation générale, d'une miniaturisation artificielle et fragilisante.

d) L'échantillon

Le début

Une autre stratégie, rudimentaire et risquée (mais pratiquée par certains moteurs de recherche du Web), consiste à ne considérer que le début des documents : on représente par exemple le texte par les 100 premiers mots rencontrés. La qualité d'une telle représentation peut être très variable selon le type de texte : les règles de rédaction pour les journaux (règle des W) ou les pages Web (la page s'affiche dans une fenêtre de hauteur limitée) voudraient qu'effectivement le tout début du texte soit un bon résumé ou donne une bonne idée de l'ensemble[101], mais toute page Web ne se conforme pas nécessairement à ce principe d'ergonomie, et le monde des documents ne se limite pas aux articles et aux pages Web.

En fait, tout est affaire de nuance, entre simplification aveugle et heuristique bien pensée et maîtrisée. Rejetable d'un point de vue général[102], la stratégie de s'en tenir au début des textes peut se

---

essentielles, c'est-à-dire dans lesquelles le phénomène de répétition est lié à la volonté de mettre en valeur les termes significatifs. Dans ce travail, cette performance n'est envisagée que sur des résumés d'articles scientifiques et techniques (résumés de publications et de brevets) disponibles directement via l'accès à un serveur de bases et banques de données. En effet sur les serveurs, les documents circulent sous une forme 'titre et résumé', les techniques développées trouvant ainsi un vaste champ d'application. » (Chartron 1988, §II.2.4, p. 24)

Que les bases de résumés soient un objet d'étude important, pourquoi pas. Mais que les fréquences des mots dans les résumés soient plus directement interprétables... le texte, résumé ou développé, introduit toujours une distance par rapport aux dénombrements que l'on peut faire sur lui.

[101] (Hérault 1981) lui se servirait des débuts des textes comme d'un extrait représentatif, sur lequel ajuster ses outils pour le traitement :

« La mise en place d'un autre module, appelé *Adaptation*, serait certainement très utile. Nous avons, en effet, constaté que, pour un texte non littéraire, l'Auteur (qui apparaît impersonnellement) donne à son lecteur dans les premières pages (moins de 5 000 mots, d'après les quelques travaux que nous avons déjà faits) toutes les 'clefs' qui lui permettront ensuite de suivre convenablement l'architecture du texte. [...] Tout ceci concourt à la création d'un module qui, pour un texte donné et compte-tenu d'une pré-analyse sur les 5 000 premiers mots, adapterait les autres modules à ce texte, c'est-à-dire réduirait considérablement les fichiers de données à partir desquels ils fonctionnent. » (Hérault 1981, p. 121)

[102] Ce rejet tient à la fois à son inefficacité au plan pratique, et plus fondamentalement à l'éradication qu'il promeut. Le principe de ne systématiquement lire que les débuts de documents est un principe discriminateur, qui

justifier parfaitement dans le cadre d'un traitement sur un genre bien défini. Pour un tel genre, le « début » correspond à une partie, dont le rôle correspond à la représentation que l'on veut obtenir.[103]

Les phrases à concentration de vocabulaire caractéristique

Des systèmes de résumé automatiques sont réalisés pour réduire le volume du texte en vue des traitements ultérieurs (Salton 1989, §12.3.1). Ce traitement, qui peut être assez élaboré, est rarement mis en place pour un seul usage de calcul interne : il sert à générer et afficher des versions abrégées des textes de la base, à la demande de l'utilisateur. Des moteurs de recherche Web offrent par exemple ce genre de résumés, pour donner un aperçu des pages retenues dans la liste des résultats d'une interrogation.

Les systèmes non basés sur une modélisation approfondie du domaine des textes à résumer fonctionnent par sélection et recomposition d'extraits du texte. Des phrases sont repérées en fonction de leur place dans la structure, et de la présence de mots fortement pondérés (*i.e.* discriminants sur le corpus et fréquents dans le texte). On veille à la complémentarité du vocabulaire des phrases retenues. Pour éviter que le résultat ne ressemble à un patchwork de phrases sans continuité, des critères linguistiques sont ajoutés (expressions-clés, anaphores, place), ou encore l'extraction s'oriente vers une extraction de paragraphes. Dans ce dernier cas, le résumé retient les paragraphes les plus centraux du point de vue du vocabulaire, éventuellement après les avoir organisés selon une classification automatique.

> The human being is capable of reading a text and summarizing its message in the form of a new, shorter text. The computer cannot yet do this, and any abstract that it creates at the moment has to be made up of words and sentences drawn exclusively from the original text. This type of abstract is better termed an 'abridgement'.
>
> [...] Based on ideas of Dr. Michael Hoey, of Birmingham University, our systems variously trace the patterns of lexical repetition in a text and use this information to select key sentences. Sentences found to be most heavily cohesive are deemed to be core information bearers.
>
> [...] Although our system does not apply a weighting to any particular section of the text, it tends to select initial sentences in journalistic articles because they are lexically rich and so achieve the required threshold in terms of repetition. This accurately reflects journalistic practice, where the essence of the text is typically summarized in the opening sentence or sentences. [...]
>
> Automatically-generated products, whilst being fast and excellent for some purposes, are not yet all readable or reader-friendly. This is partly because the computer can only represent the writer's model of text, whereas the human agent, as abstractor or indexer, adopts the reader's perspective. I think that the kind of software described in this paper, in addition to being used to present finished products to the user, will serve a very useful function as an intermediary in the information chain. For example, the automatic abridgements could be used to find other relevant texts in databases ».
>
> (Renouf 1993b)

---

organise la disparition des documents ne suivant pas cette « norme », pour un monde où on ne lirait que les gros titres. Une telle vision, qui institue une hiérarchie informationnelle, linéaire, est inacceptable.

[103] Des expérimentations sur le corpus des notes internes (qui nous intéresse au premier chef pour DECID) plaideraient en faveur d'une bonne représentation par les pages introductives :

« L'indexation automatique est légèrement meilleure pour le corpus 4 [page de garde, page de synthèse, sommaire, et les 3 pages suivantes] que pour le corpus 3 [texte intégral], le gain intervenant sur la précision apportée aux indexations T (+ 1 %) [Thesaurus] et N (+ 3 %) [Nouvelle Terminologie], le silence restant constant, ce qui accrédite l'hypothèse de n'indexer que le début des documents (premières pages) où est ciblé le 'sujet' traité. » (Monteil 1993, p. 17)

Quelques précisions doivent être ajoutées :

- d'une manière générale, l'application d'indexation automatique peut être gênée par la longueur des textes pris dans leur entier, qu'elle ne maîtrise qu'en filtrant les mots-clés une fois extraits du texte (donc sans indication de leur contexte et de leur position) ; cela peut engendrer un biais, à l'avantage des premières pages sur le texte intégral.
- le corpus des notes internes n'est pas homogène du point de vue du genre (compte-rendu de réunion, article de congrès, rapport de stage, etc.) ; l'étude du rôle du début de la note ne peut aboutir à une conclusion globale (sur l'ensemble du corpus) que si tous les genres en présence ont été examinés, et concordent –avec une marge d'approximation acceptable– quant au rôle de leurs premières pages.

Quelle que soit la forme du résumé, il reste que ce n'est qu'une autre vue, partielle, du texte intégral, et que les thèmes qui ne sont pas traités de façon majeure dans le texte sont éludés. En ce sens, une recherche qui sache prendre en compte un texte dans sa totalité, et à plus forte raison s'il est long et développé, vaut d'être mise au point (Hearst, Plaunt 1993).

e) La scission en passages

Nouvelle définition des unités de recherche

On a proposé de raisonner au niveau des « passages » plutôt que des documents entiers. Cela homogénéise les longueurs donc permet un bon fonctionnement du calcul. D'autre part, dans le cadre d'une recherche d'information, le repérage d'un passage pertinent au sein même du document apparaît comme un plus, puisqu'on localise ainsi immédiatement l'information cherchée sans avoir à parcourir le document[104].

De fait, l'objet d'un système documentaire n'est pas le document en tant que tel, mais l'unité de recherche[105]. Parler d'*unités documentaires* ou d'*unités textuelles* (au lieu de *textes* et de

[104] C'est la position de Christian FLUHR, concernant le système SPIRIT :
« pour la comparaison des documents, il est préférable que leur longueur soit assez homogène. Il peut être difficile d'indexer par un procédé commun des résumés contenant quelques centaines de mots et des livres renfermant quelques centaines de milliers de mots. Dans ce cas, on pourra subdiviser les ouvrages longs en chapitres ou même en paragraphes, afin d'avoir un fond de documents homogènes en longueur. Le fait de n'avoir que des documents assez courts permet une meilleure efficacité de réponse documentaire. Ceci, par ailleurs, facilite la validation des réponses. » (Fluhr 1977, §III.5, pp. 174-175)
En l'occurrence, les extraits de résultats figurant en annexe semblent montrer que les documents utilisés faisaient entre 15 et 80 mots environ (« mots pleins », c'est-à-dire à l'exclusion des conjonctions, déterminants, etc.).
La plaquette de présentation du système, maintenant commercialisé, présente le découpage des documents comme une des opérations de constitution de la base. Les passages sont déterminés manuellement, par l'administrateur du système, qui par son choix (re)définit la notion de document pour sa base. (Quant à la requête, qui peut être textuelle, ce qui excède une certaine longueur est tronqué, du moins pour ce que l'on peut percevoir du fonctionnement de SPIRIT-W3).
« L'information contenue dans une base de données est découpée en unités documentaires (ou documents) qui représentent l'unité de recherche. A une question posée, SPIRIT propose un certain nombre de documents que l'utilisateur peut visualiser. La notion de document est variable d'une base de données à l'autre. Pour une base bibliographique ou un catalogue de produits, chaque notice ou chaque description de produit constitue un document différent. Dans d'autres cas, le découpage à opérer est moins évident. Une base contenant par exemple l'ensemble du code des impôts doit-elle être découpée par chapitre ou par article de loi ? L'administrateur de la base doit effectuer le découpage en tenant compte des règles générales suivantes :
- chaque document doit avoir un contenu homogène ;
- les documents trop courts dispersent l'information dans la base et la rendent difficile à retrouver ;
- les documents trop longs ne permettent pas d'utiliser au mieux les mécanismes d'optimisation de la recherche. »
(plaquette de présentation de SPIRIT, société T.GID, 1993).
On retrouve le même principe pour le calcul dynamique de liens effectué par Similidoc :
« Le système développé a pour objectif la recherche de similitudes entres parties de documents. A chaque recherche, il y a création de liens dynamiques entre les parties de documents suivant leur degré de similitude au sens du système. [...]
Cet outil a été utilisé pour effectuer des rapprochements entre des textes relatifs à l'assurance qualité. La granularité choisie est celle du paragraphe. Une baisse de précision est observable pour des paragraphes courts dont le contexte est limité. En contrepartie, la délivrance directe des paragraphes supposés pertinents est un atout important par rapport à celle des textes complets. »
(Betaille, Massotte, Joubert 1998, pp. 136 et 142)

[105] Le basculement de la recherche *documentaire* à la recherche d'*informations* dans des *bases de données textuelles* trahit le peu de cas que l'on fait de l'unité que constitue le texte :
« les clés d'accès [fournies par de nouvelles techniques de gestion documentaire et de recherche rétrospective dans un fonds] visent essentiellement à répondre à une demande d'information liée à la fourniture du document. Parallèlement à cette problématique du *document*, un autre type de demande d'information s'est profilé, s'inscrivant dans une problématique d'*ensemble de connaissances* stockées dans les *corpus documentaires*. [...]

documents) rappelle qu'il s'agit d'entités construites pour l'application. Faisons-nous fausse route en examinant la prise en compte du texte dans les systèmes documentaires en général ? C'est peut-être au contraire souligner la décision herméneutique qui intervient lorsque l'on choisit de reconnaître tel ensemble de paragraphes comme un texte –il n'y a pas de texte en soi, il n'y a de texte que par décision d'un lecteur.

La critique majeure que l'on peut formuler à l'encontre de cette redéfinition est qu'elle perd l'unité du document, tel qu'il était perçu initialement. Le texte est éclaté en « morceaux », considérés ensuite indépendamment les uns des autres lors du calcul qui les sélectionne.

L'articulation global / local

Une manière de profiter des avantages du calcul sur les passages, sans sacrifier la cohésion d'ensemble du document, est de procéder en deux temps : un premier calcul sélectionne des documents (avec éventuellement des scores inégaux, certains très faibles), et un second calcul explore les passages des documents sélectionnés.

On conjugue ainsi un point de vue global et un point de vue local, et fait jouer des zones de localité signifiantes comme la phrase ou le paragraphe : en effet, quelques termes en commun sont plus probablement pertinents s'ils sont en relation de proximité que s'ils sont dispersés aux confins du texte. On fait donc d'une pierre deux coups : la longueur du texte n'est plus irrémédiablement pénalisante, et l'on compense les faiblesses du seul contexte textuel pour saisir les interrelations entre les mots. En effet, on s'efforce ainsi d'écarter les rapprochements injustifiés par un seul mot de forte pondération, ou par un ensemble de mots « dépareillés ». Dans (Salton, Allan, Buckley 1994), les conditions de similarité entre deux textes prennent ainsi la forme suivante : *(i)* la similarité (globale) entre les deux textes est supérieure à un certain seuil ; *(ii)* il existe *n* paires de phrases, l'une dans le premier texte, l'autre dans le second, telles que la similarité (locale) entre les deux phrases de la paire soit supérieure à un seuil fixé, et que plusieurs mots contribuent significativement à la similarité (par exemple, si la similarité entre les deux phrases est due à une seul mot pour plus de 90 % de la valeur calculée, alors la paire n'est pas comptée comme similarité locale valable).

La discussion porte alors sur la manière la plus adaptée de définir des passages, et le mode de composition des deux calculs de similarité. Les deux cas sensibles sont ceux où la similarité globale est très faible et la similarité locale significative, et la réciproque. Première voie : calcul global puis local, mais le principe même de cascade peut en effet être trop sélectif s'il élimine en amont ce qui aurait en définitive pu être retenu à l'issue du calcul complet. Deuxième voie, garder dans la même base et les textes, et les passages, comme autant d'unités autonomes (ce qui évite le filtrage des uns pour accéder aux autres), augmente fortement le volume de la base et renvoie au problème de la perte de la contextualisation des passages et de leur regroupement en un tout.

Troisième voie, combiner les valeurs des rapprochements locaux sélectionnés avec la similarité globale sur le document ; en effet, dans un extrait du texte qui aborde un sujet spécifique, il peut n'être pas fait mention de la problématique générale du document, qui reste implicite. La combinaison des rapprochements locaux et globaux est une manière de prendre en compte à la fois des thèmes mineurs, développés seulement très localement dans le document, tout en les rapportant éventuellement à un contexte d'ensemble (« A, dans le contexte de B ») (Hearst, Plaunt 1993).

Vers une décomposition automatique du texte : segments et thèmes

Le découpage d'un document en passages est quelquefois pris en charge par le système lui-même. D'abord envisagée à des fins de redéfinition d'unités de recherche, la question de la délimitation de parties à l'intérieur du document s'est élargi à une forme d'analyse de la structuration interne du texte.

Dans le cas de figure le moins « linguistique », les contextes locaux sont définis comme des segments de longueur fixe (en nombre de mots). Technique fruste au premier abord, elle se révèle efficace par plusieurs aspects (Callan 1994). Premièrement, elle forme des zones régulières et pas trop

---

La croissance et la multiplication des documents sur support magnétique amènent divers utilisateurs [...] à considérer les bases de données textuelles comme des réservoirs de connaissances ». (Chartron 1988, §I.2, p. 11)

courtes. En effet, les paragraphes s'avèrent dans certain cas des contextes trop restreints ; la longueur optimale des fenêtres serait d'ailleurs de 200 à 250 mots selon (Callan 1994). Deuxièmement, la mise en œuvre de fenêtres complètement chevauchantes (l'extrémité d'une fenêtre est le milieu de la fenêtre voisine) évite un cloisonnement des contextes et s'adapte mieux au fait qu'un texte ne présente pas nécessairement un unique découpage significatif.

La façon la plus naturelle de définir les passages est, quand on dispose de l'information, de reprendre la structuration logique du texte, à savoir le découpage en paragraphes ou en sections. Il semble ainsi que l'on prenne en compte une sémantique apportée par la forme du document, chaque contexte ainsi marqué reflétant une intention de l'auteur. Pour autant, il n'est pas assuré que l'auteur fasse un emploi canonique et toujours également motivé de cette structuration. De plus, les documents électroniques ont rarement des indications univoques de la structure, il s'agit de traces (comme les retours à la ligne, les sauts de ligne, etc.) qui doivent être utilisés, non sans risque d'erreur.

Pour (Hearst, Plaunt 1993), les frontières où trancher entre un passage et le passage suivant sont repérées comme des points de discontinuité thématique[106]. Autrement dit, on quitte un ensemble de mots, qui expriment un certain ton, un certain sujet, pour entrer dans un vocabulaire différent, qui aborde un autre sujet ou / et adopte un autre ton. L'algorithme de leur outil *TextTiling* calcule toutes les similarités entre segments adjacents (un segment est une suite de 3 à 5 phrases, la mesure de similarité est une mesure vectorielle de type cosinus, avec une pondération interne, caractérisant le vocabulaire du segment par rapport à celui du texte). Ensuite, la courbe des similarités obtenue pour le texte est lissée, et les coupures du texte sont placées selon les 'creux', les 'vallées' de la courbes. Contrairement aux attentes, l'expérimentation ne révèle pas de supériorité significative de cette technique par rapport à l'utilisation du découpage selon les paragraphes.

C'est avec (Salton & *al*. 1996) que l'on bascule clairement de l'optique de découpage dans l'étude de la structuration interne des textes. Les techniques de calcul de similarité, connues pour caractériser les rapprochements entre textes à l'intérieur d'un corpus, sont cette fois-ci déployées à l'intérieur d'un texte, pour caractériser les liens entre paragraphes. La base est le calcul des similarités entre les paragraphes pris deux à deux, ce qui, si l'on ne retient que les similarités suffisantes (seuil), se traduit par un graphe. Chaque paragraphe est un noeud du graphe, et chaque similarité significative est un arc qui relie les deux paragraphes concernés. Deux formes de structuration sont recherchées, les *segments* et les *thèmes*.

Les segments sont les composantes connexes du graphe, une fois effacés tous les liens concernant des paragraphes distants (par exemple séparés par plus de trois paragraphes). Les segments reflètent l'organisation linéaire du texte, sans faire de sauts qui uniraient des parties distantes dans le texte. Ils sont analogues à la conception du passage dans *TextTiling* : cohésion lexicale interne, et délimitations correspondant aux changements de vocabulaire.

Les thèmes, eux, groupent les paragraphes en fonction de leur similarité, indépendamment de la distance qui peut les séparer dans le texte. Le regroupement s'opère par une classification automatique par agrégation de trios de paragraphes (triangles dans le graphe), chaque trio étant représenté par le vecteur somme des vecteurs paragraphes. Les thèmes sont donc censés représenter des composantes thématiques du texte, et séparer les différents aspects qui sont traités de façon plus ou moins intriquée dans le document.

L'étude du graphe permet encore de repérer les paragraphes les plus centraux, caractérisés par le grand nombre de liens qu'ils entretiennent avec les autres paragraphes. Cela fournit matière à des formes de 'résumés automatiques' (cf. la recherche d'échantillons, ci-dessus).

[106] (Nakhimovsky, Rapaport 1989) déclinent cinq classes de discontinuités (pour le cas de récits narratifs) : inversion entre premier plan et arrière plan, changement de lieu ou de moment, changement de point de vue, rupture thématique. Ils remarquent notamment que la reprise d'un syntagme nominal complet (au lieu d'une ellipse ou d'une anaphore possible) constitue une marque de discontinuité.
Les discontinuités qu'ils se proposent de repérer (sans en montrer une implémentation) sont en deçà du paragraphe. Ils n'utilisent pas l'information du découpage en paragraphes, pour mettre au point leur algorithme. Ces chercheurs américains précisent néanmoins que les paragraphes sont davantage que des indicateurs de discontinuité : ils instaurent une discontinuité par leur simple présence.

Dans le cadre d'une application de recherche d'information, les segments et thèmes sont mobilisés pour faire des calculs de similarités globaux et locaux (les thèmes remplaçant les passages pour les textes qui ne sont pas organisés de façon linéaire). Ils sont aussi proposés pour proposer des parcours dans les documents : traversée générale (en prenant les paragraphes centraux des différents thèmes), ou traversée thématique (en sélectionnant les paragraphes centraux à l'intérieur du thème retenu).

f) Que penser de tout cela ?

La question des documents longs est celle de la construction d'une représentation synthétique. Les modèles utilisés par les moteurs de recherche ont une logique cumulative : plus il y a de mots dans le document, plus sa représentation est grande, avec un effet de dilution. La solution n'est peut-être pas de réécrire ou de redéfinir le texte lui-même. Une autre voie est de respecter le texte dans son intégrité, et de construire des unités descriptives synthétiques et des représentations entrant dans le calcul de façon souple.

Avec la décomposition du texte en segments et thèmes, on assiste à un recyclage magistral des techniques intertextuelles en techniques intratextuelles. Le fait majeur, est la reconnaissance de zones de localité du point de vue de la sémantique (thématique), et de leurs interrelations possibles à l'échelle du texte. La description reste cependant dans l'ordre de la simplification : le 'grain' est défini à l'avance (par exemple le paragraphe), et chacun est un atome univoque. Autrement dit, il ne semble pas prévu de rendre compte de chevauchements thématiques par exemple.

4. Lexicométrie intratextuelle : l'étude des rythmes

Il s'agit ici de rendre compte de travaux complémentaires à ceux évoqués dans les paragraphes précédents, et donc, parmi les travaux de lexicométrie, ceux qui optent pour une approche intratextuelle plutôt qu'intertextuelle de la distribution d'unités dans les textes. L'accent est donc mis ici sur notre deuxième facette (la structuration interne du texte et notamment sa linéarité), en reléguant pour un temps au second plan la troisième facette (intertextualité), dont la présence est habituellement dominante dans les études un tant soit peu quantitatives.

Le texte n'est pas une réalité ponctuelle et uniforme : son déroulement est l'occasion de contrastes entre régularité et rafales[107], dispersion étale ou accumulation localisée.

Le texte réunit les conditions d'une étude du rythme :

> Le rythme ne peut apparaître, selon nous, qu'à certaines conditions :
>
> - présence d'une *linéarité* (linéarité du temps, mais aussi par exemple linéarité du scriptural ou de l'oral).
>
> - présence d'éléments *discontinus*, distincts les uns des autres tels que des notes ou des mots. Un son continu, de même hauteur et intensité, comme celui que produit un klaxon, ne saurait créer un rythme.
>
> - présence d'une *récurrence* : le rythme repose sur le retour, la réapparition, régulière ou non, d'éléments identiques (une même note, un même silence, une même durée).
>
> - présence d'une *différence*, d'un écart. Dans le domaine musical, par exemple, il faut distinguer l'isochronie (la goutte d'eau qui tombe régulièrement) du rythme, qui naît ou bien d'une différenciation qualitative, produite par l'accentuation d'un temps ou la heuteur du son, ou bien d'une différenciation quantitative, créée par une opposition de durée. De même qu'elle fonde le sens, la différence apparaît ainsi à la base du rythme.
>
> Le texte réunit ces quatre conditions, puisqu'il dispose linéairement des éléments discontinus – les mots–, qui présentent un caractère itératif (itérations lxicales, morphologiques, sémantiques, etc. qui contribuent à la cohérence) mais manifestent simultanément des différences (autre entour cotextuel, autres relations, ce qui assure la progression) : il apparaît ainsi fondé de chercher à mettre en évidence les rythmes textuels. »
>
> (Dupuy 1993, pp. 509-510)

L'étude du rythme fait place à plusieurs facettes textuelles, outre la deuxième, directement concernée (par le biais de la linéarité). Le caractère linguistique du texte (première facette) est le

[107] C'est Pierre LAFON qui a introduit le concept de *rafale* et en a proposé une mesure, voir par exemple (Lafon 1981a).

principal point d'appui pour la définition des unités et de leurs variantes. La linguistique fournit des systèmes d'identités et de différences, lexicales ou morphologiques par exemple. Et la place de l'interprétation (quatrième facette) peut être explicitement reconnue, en soulignant le caractère relatif des choix de modélisation (le choix de ce dont on suit la répétition), et en s'en tenant à des résultats nuancés et graduels, jamais définitivement établis ni universels.

Cette voie est encore relativement peu suivie[108], et mérite d'être reprise et poursuivie, en complémentarité avec d'autres approches (plus intertextuelles).

[108] Les travaux de Pierre LAFON et ceux de Jean-Philippe DUPUY ne sont toutefois pas les seuls.
(Lessard & Hamm 1991) ont par exemple une analyse des formes de répétitions, pour des séquences de plusieurs unités. Leur problématique est très proche de celle des segments répétés (cf. André SALEM). Ils identifient plusieurs modes de répétition (selon l'écart croissant entre les occurrences : répétition rhétorique, répétition d'insistance, répétition idiolectale), et s'intéressent également à la relation entre les variantes pour les reprises non littérales.

D. RECEVOIR UN TEXTE

1. Compréhension

a) Que saisir de la compréhension d'un texte ?

Repères généraux

Il est difficile de souscrire à l'idée de systèmes réalisant une 'compréhension automatique'. La discussion doit au moins commencer par clarifier ce que l'on met derrière le mot *compréhension*. (Sabatier & *al*. 1997, §4).

Une conception extrême et fortement réductrice est celle qui établirait le résultat de l'analyse automatique du texte comme « la » compréhension unique, ultime et universelle d'un texte. Une autre version des faits, plus réaliste et plus modeste, est de se donner une grille de lecture appropriée au type de textes à traiter. La compréhension est alors le remplissage convenable de la grille, pour chaque texte, à partir des informations que l'on y trouve. Dans ces deux cas, la compréhension consiste en la transcription d'un texte dans un certain formalisme, –nécessairement non équivalent : une (large ?) part du texte échappe à la dite compréhension. Il semblerait préférable de considérer là qu'il s'agit d'*une* lecture, quelque peu mécanique, calibrée pour un certain type de textes et un certain usage[109], répétitive et fortement déterminée *a priori* (peu prête à percevoir le texte dans sa singularité et sa texture propre).

---

[109] La thèse de Laurent Doré (Doré 1992) est un exemple de travail dans cette optique :
« En se proposant de comprendre de façon automatique un compte-rendu [médical], on doit aboutir à une représentation informatique qui reflète le contenu du texte et réponde de la même façon aux objectifs de la communication dégagés précédemment », à savoir la transmission d'informations sur « l'évolution de l'état du patient » et « l'enchaînement chronologique des actions réalisées » (Doré 1992, p. 37).
La « connaissance pragmatique spécifique [du protocole de traitement et de surveillance du cancer de la thyroïde (TSCT)] fournit un shéma d'intégration qui va être instancié par les actions du texte ». (*ibid*., p. 42)
Laurent Doré fonde alors l'algorithme du traitement sur les observations suivantes, faites sur un corpus de 80 compte-rendus d'hospitalisation : premièrement, l'ordre du texte suit l'ordre chronologique des actes médicaux et des observations ; deuxièmement, on reste dans le domaine médical et hospitalier, si bien que les termes ne présentent pas d'ambiguïté. Le traitement se fait phrase par phrase ; la question de la prise en compte de phénomènes interphrastiques (anaphores) est soulevée, celle (plus générale) de la coréférence est résolue dans le cadre du modèle, qui délimite les références possibles.
La valeur pratique de ce traitement automatique trouve une justification dans un contexte défini :
« demander aux médecins de s'exprimer dans un formalisme rigide utilisant des codes prédéfinis [comme cela a été fait dans d'autres travaux], se heurte d'une part au manque de disponibilité voire à l'hostilité des médecins et d'autre part rencontre des contraintes pratiques dues à la rigidité intrinsèque de tous les formalismes *a priori*. » (*ibid.*, p. 153) (le texte reste donc un mode d'expression irréductible ; le traitement est acceptable parce qu'il en donne une vue, mais ne se substitue pas au texte).
« Après l'étude des différentes questions possibles [huit jugées intéressantes, proposées par des médecins], il ressort que les informations recherchées (un résultat, une date, un nombre, etc.) font dans la plupart des acs directement référence aux événements réalisés ou prévus. Or la représentation de l'histoire que nous construisons est précisément centrée sur des concepts d'action correspondant aux événements mentionnés dans le texte. » (*ibid*., p. 155)
Les limites du systèmes ne sont pas masquées :
« notre approche est effectivement apprpriée pour l'analyse de textes narratifs dans des domaines techniques sous-tendus par un modèle de fonctionnement (resp. de dysfonctionnement) auquel correspondent des plans d'intervention préétablis. » (*ibid*., p. 150)
« le contexte pertinent n'est pas forcément équivalent *a priori* à la représentation courante de l'histoire. Celle-ci reste un contexte par défaut valable uniquement dans le cas de texte relatant des faits de façon strictement chronologique comme nous l'avons estimé pour les compte-rendus d'hospitalisation. » (*ibid*., p. 156)

D'une manière générale, la classe des *systèmes de compréhension de textes* (Sabatier & *al*. 1997) porte un titre trop lourd. La compréhension réalisée consiste en la capacité à extraire, et éventuellement reformuler, un élément du texte, ou en la transcription correcte du texte dans le formalisme qu'on s'est choisi. Première objection : la compréhension est tout entière « dans » le texte, éventuellement étendu par une base de connaissances (fournisseur d'inférences), celle-ci n'étant jamais que partielle et partiale (ce peut être néanmoins une bonne représentation d'un arrière plan conventionnel). Autrement dit, il ne s'agit pas d'une compréhension au sens d'une appropriation personnelle d'un texte, et de la manière dont il fait écho en soi et motive une action originale, innovante, créatrice. Il y a encore une seconde objection à l'étiquette *systèmes de compréhension de textes* : en fait de textes, la plupart de ces systèmes (*grosso modo*, ceux qui n'utilisent pas de grille de lecture) s'en tiennent à un cumul d'analyses ponctuelles, phrastiques. A proprement parler, ce ne sont pas des textes qu'ils considèrent, mais du texte[110]. Ils n'acquièrent pas une vue intégrée de l'unité textuelle, qui fasse sens, mais en obtiennent une vue morcellée en une série d'informations élémentaires, éventuellement recombinables, certes, mais sans réelle perspective globale et synthétique de l'unité que forme le texte.

Il faut ainsi dénoncer une modélisation de la compréhension dans le prolongement direct d'analyses linguistiques lexicales et morpho-syntaxiques. Bien que le texte soit rédigé dans une certaine langue, et qu'il puisse y avoir une pertinence à effectuer une analyse des structures linguistiques qu'il réalise, la compréhension ne saurait s'arrêter à une vue de type étiquetage des mots, des liens syntaxiques, ou des « sens ». En tant que lecteur, ce que je retiens d'un texte, ce n'est pas le détail précis de ses mots ou de ses constructions linguistiques, mais une idée d'ensemble. Les outils classiques de Traitement Automatique du Langage Naturel ne sont pas pour autant à exclure : ils peuvent être mis au service d'une vision textuelle.

Une proposition linguistique : la sémantique interprétative

La compréhension est plutôt une forme d'interprétation, d'actualisation du sens d'un texte par et pour un lecteur donné, à un moment donné. En ce qui concerne le texte, on peut donc s'efforcer de repérer les contraintes linguistiques qu'il instaure, les indices qu'il fournit, qui, sans déterminer l'interprétation / compréhension, participent à son élaboration.

> dans les termes de la sémantique linguistique [, la] compréhension, déliée des réquisits psychologiques, est une interprétation : elle consiste à stipuler, sous la forme de paraphrases intralinguistiques, *(i)* quels traits sémantiques sont actualisés dans un texte, *(ii)* quelles sont les relations qui les structurent, et *(iii)* quels indices et/ou prescriptions permettent d'actualiser ces traits et d'établir ces relations, qui sont autant de chemins élémentaires pour des parcours interprétatifs. La première stipulation suppose une analyse componentielle ; la seconde, structurale ; la troisième,

[110] La citation suivante, d'autant plus significative qu'elle est centrale dans le document fédérateur (Sabatier & *al*. 1997), trahit cette conception non textuelle :
« Dans le cadre d'une évaluation qualitative, nous pensons que la meilleure forme de tests pour évaluer des systèmes de compréhension de textes est celle du type DQR où :
- D est un ensemble de phrases déclaratives (ou de données) ;
- Q est une question ;
- R est la réponse attendue à la question Q, réponse qui peut être déduite de D.
Tout système se prêtant à des tests DQR peut être considéré comme un système *complet* de compréhension du langage naturel. Nous qualifions de complet un système qui analyse (D et Q) et qui synthétise (R) du langage naturel. La synthèse est une réaction appropriée aux propos qui lui sont adressés : ce peut être une réponse à une question (Q + R) ou bien une réaction du système (R sans Q) sur la consistance, sur l'ambiguïté, sur la redondance des propos tenus, etc. Ce peut-être une demande d'information du système, etc. » (Sabatier & *al*., 1997, §6)
Les textes ne sont ici que pré-textes à interrogations : ils ne sont vus qu'à travers le prisme de questions ponctuelles et factuelles. Il s'agit de retrouver des faits objectifs, d'y accéder, dans une collection de faits objectifs, prédéfinis (*i.e.* dont les valeurs possibles sont connues à l'avance). D'ailleurs dans l'espace de ces quelques lignes, les rédacteurs glissent non sans raison de la *compréhension de textes* à la *compréhension du langage naturel*. Evidemment, il serait absurde de demander à la machine : qu'est-ce qui fait sens *pour vous* dans ce texte ? qu'en retiendriez-vous et pourquoi ? Tout au plus le calcul est-il capable de fournir des représentations suggestives, support pour l'interprétation d'un utilisateur humain.

> interprétative et herméneutique. Il en résulte non une traduction, mais une explicitation, qui généralise les principes de la définition, en les réfléchissant pour assurer la pertinence de leur application.
>
> En tout cas, le terme de *compréhension* est sans doute trop fort pour une telle conception qui ne recourt pas à un sujet psychologique ou philosophique. Mais cette insuffisance devient une vertu dès lors qu'il s'agit de proposer une méthode d'interprétation explicite et au moins partiellement automatisable.
>
> (Rastier, Cavazza, Abeillé 1994, §I.2.2, pp. 11-12)

Appropriation et construction : l'image de l'interpolation

La compréhension est une interprétation, c'est-à-dire la construction et l'appropriation d'un sens[111].

Le texte apporte des points d'appui, qui orientent et contraignent la construction d'un parcours interprétatif ; le lecteur apporte lui aussi ses propres repères. Un sens peut alors se dessiner comme un chemin (mieux : un cheminement)[112], une interpolation, qui fait se rejoindre et interagir le monde du lecteur et celui du texte. Quand ces deux mondes s'inscrivent dans des espaces trop étrangers l'un à l'autre, le tracé d'un parcours se fait laborieux et fragile. Quand au contraire le monde du texte est déjà comme trop bien intégré au mode du lecteur, Il y a à peine à modifier des tracés existants –il est difficile d'échapper aux ornières bien creusées–, et l'excursion perd de son attrait. La compréhension la plus fructueuse est donc celle qui peut se développer sur des bases textuelles et personnelles suffisantes, et vient renouveler le paysage intérieur du lecteur en lui ouvrant de nouvelles perspectives, de nouvelles pistes, plus prometteuses qu'hasardeuses. D'où une remotivation du terme lui-même : com-prendre, prendre avec soi, garder quelque chose de la rencontre avec le texte, incorporer une part de la réalité du texte dans sa propre réalité.

Du point de vue de la mémoire, cette image d'une compréhension comme interpolation rejoint une expérience commune. A partir de quelques points que l'on se remémore d'abord, se recompose une pensée cohérente, un tout intégré.

En ce qui concerne les situations de travail dans l'entreprise, chacun sait la difficulté qu'il y a à reprendre le dossier d'un collègue, à « adopter » l'armoire de documents laissée par son prédécesseur[113]. L'image du dépôt, d'un gisement ou d'une mine d'informations montre cruellement

---

[111] Ce n'est pas une nouveauté, mais mérite toujours d'être réaffirmé ! Sur ce point, voir par exemple (Dumesnil 1992) (la compréhension comme construction), (Poitou, Ballay, Saintive 1997) (les savoirs, comme appropriation des connaissances).

[112] Ou une *trajectoire* à travers des *attracteurs* :

[Les] formations sémiotiques ont des structures propres qui, à défaut de leur conférer une objectivité, contraignent les parcours interprétatifs, sans les déterminer pour autant. Par exemple, un tiret inhibe la propagation des traits sémantiques entre les syntagmes qu'il sépare, alors que les deux points la favorisent. A grande échelle, ce type de contraintes, auxquelles s'ajoutent des contraintes situationnelles, dessinent des parcours préférentiels. Plus généralement, on pourra définir les sens d'un texte comme des parcours entre des comportements sémantiques stabilisés (ou *attracteurs*, dans la terminologie des systèmes dynamiques). Le 'mouvement' du texte, qui le rend irréductible à une suite de phrases, serait alors une trajectoire dans un paysage d'attracteurs, le passage d'un attracteur à un autre dépendant des objectifs de la pratique interprétative en cours. » (Rastier 1994, §2, pp. 334-335).

[113] Sans compter l'effet « ticket de métro », fort justement perçu et nommé par Simone Joseph-Waterlot :

« [dans le métro,] on doit garder son ticket pendant le voyage en cas de contrôle, et à la sortie, on ne fouille pas ses poches pour le jeter. Et si jamais on a l'idée de s'en débarrasser au cours d'un voyage suivant, on en retrouve plusieurs et ne sachant plus lequel est le bon, celui qu'il faut garder, on les remet tous illico dans sa poche.

La gestion d'un dossier peut s'y apparenter. Sauf que l'on ne sait jamais quand finira le voyage. Il est sûr qu'il y a une phase de démarrage, puis de vie active d'un dossier, suivi d'une veille plus ou moins profonde. Le passage d'une étape à l'autre s'effectuant progressivement. C'est à la mise en veille qu'il serait bon de faire un tri sérieux, et de jeter les brouillons et les pièces en double chez un collègue. [...]

Mais si la période est passée, et qu'au terme de l'étude, le dossier n'ait pas été rangé, le coche est raté, les choses devenant de moins en moins lisibles au fil du temps. [...] On garde alors tout, sans prendre le temps de trier. Ce qui explique que les armoires soient pleines et en nombre toujours insuffisant. [...]

[De plus,] le rangement, la partie visible du tri, n'étant pas une activité gratifiante, rares sont ceux qui prennent le temps de l'effectuer avant de se lancer dans le travail suivant.

[...] les successeurs se retrouvent avec un passé inutile ou inutilisable ».

ses limites : l'information n'est pas dans les documents, où il suffirait de la piocher ; elle est à construire par et avec une appropriation. D'ailleurs, l'idée que l'on se fait d'un document et la manière dont on l'aborde évoluent.

Vouloir décrire le monde et expliciter l'ensemble des connaissances, et penser inculquer ainsi une forme de compréhension à la machine qui lui permette de « faire le tour » d'un texte, c'est oublier l'interprétation à l'œuvre dans l'appropriation, la constitution, et la transmission d'un savoir. Les connaissances préenregistrées sont figées, normatives, partielles et partiales, et désespérément incomplètes...

Discussion : affinités et écarts avec la pertinence selon Sperber & Wilson

La conception de la compréhension comme d'une forme de rencontre fructueuse entre le texte et le lecteur, dans laquelle le monde du texte et celui du lecteur se rejoignent sans se superposer, partage des points forts avec la théorie de la pertinence élaborée par (Sperber & Wilson 1986).

Sperber et Wilson décrivent une situation de communication à partir de de la transmission d'un message (typiquement une parole, qui peut être modulée par l'intonation, le geste) entre deux interlocuteurs. Ils récusent bien l'idée selon laquelle la langue se réduirait à un codage, empaquetant un contenu informationnel[114] : la langue (l'expression verbale) fournit des points d'appui à partir desquels l'interprète / destinataire construit un sens. Et cette élaboration d'un sens ne naît que s'il y a effectivement rencontre du monde de l'interprète et de celui proposé par le message : il y a une activité « productive » de sens que Sperber et Wilson décrivent à travers le concept d'*effet contextuel* (p. 187). Le contexte lui-même n'est pas donné et fixé *a priori* (p. 215), il s'ajuste et se reconfigure dans l'activité même d'interprétation.

Mais il est difficile de suivre Sperber et Wilson dans le développement complet de leur théorie, et plus encore dans la modélisation d'inspiration logique[115] qu'ils échaffaudent. L'être humain serait un dispositif efficace de traitement de l'information (p. 76), identifiant *la* signification (au sens de *meaning*, le « vouloir dire ») la plus rentable et s'arrêtant à elle (p. 256). Il y a une réduction imposée et hâtive des possiblités de sens (une phase de *désambiguation*, cf. p. 267, p. 306)[116], calculées à partir d'une sémantique compositionnelle et dénotationnelle[117]. La mécanique de description du monde mental de l'interprète est réduite à des processus inférentiels sur des expressions logiques[118]. Le choix de l'interprétation est régi par les critères du moindre *effort* et de l'*effet* maximal (p. 188 sq.), mais la manière dont se concilient ces deux facteurs pour évaluer une rentabilité (ou une productivité, un rendement) n'est pas claire[119]. Enfin, rien n'est moins sûr que l'universalité d'une herméneutique « intéressée », tout entière régie par des considérations de rentabilité immédiate. Le modèle que développent Sperber et Wilson en se référant à des situations

---

(Joseph-Waterlot, Lahlou 1995, §III.3, pp. 28-29)

[114] La langue n'encode pas une pensée (*ibid*., p. 345), et la communication n'est pas le tranfert d'une pensée (*ibid*., pp. 287-288).

[115] L'activité de l'esprit humain ne se calque pas entièrement sur des règles logiques, concèdent les auteurs –voir par exemple (*ibid*., p. 109).

[116] Le modèle chasse également toutes formes de contradictions (logiques) : si la représentation du monde que se fait un individu vient à avoir deux éléments contradictoires, alors tout est mis en œuvre pour éliminer la contradiction, avec l'idée qu'il faut trancher (effacement pur et simple de l'alternative la plus faible, par exemple, cf. p. 176).

[117] La signification explicite d'un énoncé est déterminée par l'intermédiaire d'associations entre les mots et des *concepts* (c'est l'*entrée lexicale* du concept) (p. 141), l'*entrée encyclopédique* du concept rassemblant les informations sur son extension ou sa dénotation (p. 135), et l'*entrée logique* assurant le relais avec la représentation du monde dont dispose une personne. Le module linguistique travaille mot à mot, séquentiellement (p. 278), avec des anticipations globales possibles au niveau de l'énoncé (p. 306).

[118] La représentation du monde est assimilée à un ensemble d'hypothèses, sur lesquelles s'appliquent des règles logiques (inférence, élimination), chaque hypothèse étant modulée par sa force (établissement initial), son degré de confirmation et son accessibilité (réactivations nombreuses ou/et récentes).

[119] Un test systématique de « toutes » les possibilités ? Mais en s'arrêtant à la première « satisfaisante » ? (cf. 256)
Une maximisation absolue semble trop coûteuse, et une maximisation relative trop arbitraire.

d'échanges verbaux brefs, ne se laisse pas généraliser à la diversité des pratiques de lecture et d'interprétation.

Le passage à une utilisation opérationnelle de l'idée de rencontre et d'interpolation entre le monde du texte et celui du lecteur n'impose pas de recourir à un appareillage logique lourd, rigide, et *in fine* trop étriqué pour rendre compte du sens. Par exemple, le modèle vectoriel en recherche documentaire, et utilisé dans un premier temps par l'application DECID de diffusion ciblée, rend compte de façon grossière mais robuste des « recouvrements » significatifs entre les pôles d'intérêts du destinataire et les sujets abordés dans le document. Nous gardons donc à l'esprit une partie de la théorie de (Sperber & Wilson 1986), sans adopter leur modèle.

Modélisation : points d'appui plutôt que contenu

Les systèmes qui visent une forme de compréhension par la machine sont amenés équiper la machine d'un 'monde'. Le premier constat est la nécessité de se restreindre à une réalité cernée et très limitée, et fonctionner dans ce monde clos. La deuxième est, malgré cette délimitation, le travail énorme pour rassembler et constituer les connaissances dont doit disposer le système pour mener à bien la construction de représentations utiles. Qu'il s'agisse de réseaux d'inférences, de systèmes experts, de raisonnements par cas, on est toujours face à des systèmes très lourds (Jacob 1994).

Un troisième constat pointe un écart fondamental de comportement dans les conditions aux limites. Ces systèmes, qui sont bornés par les limites même des connaissances qu'ils emmagasinent, échouent plus ou moins élégamment face à un texte un peu inattendu. Le lecteur humain a au contraire une tendance irrépressible à donner du sens, trouver un sens à un texte, à 'broder' à partir de quelques éléments qui stimulent son imagination (à défaut peut-être de son savoir rationnel) : l'interprétation est compulsive. Plutôt que de déclarer forfait devant une matière textuelle insuffisante, le système devrait pouvoir partir des moindres traces qu'il reconnaît, même si elles ne forment pas une structure bien formée reconnue, et proposer un résultat. Un indicateur de fiabilité (en termes de traitements automatiques), ou une indication sur la faiblesse de l'ancrage au texte (en termes plus interprétatifs), serait opportun pour guider l'utilisateur dans la manière de considérer les résultats.

Plutôt qu'une représentation qui renfermerait le sens du texte, ce qu'il est intéressant de passer à la machine, ce sont des points d'appui pour la construction d'une interprétation[120]. C'est s'en

[120] (Hérault 1981) propose l'opposition entre *compréhension explicite* et *compréhension implicite* :
« il y a *compréhension explicite* toutes les fois où un système non humain obéit aux instructions qui lui sont données, sous forme écrite ou orale et dans un langage 'ordinaire'. [...] il est parfaitement envisageable de commander un système mécanique (ou, ce qui revient au même, électronique), même très complexe, à partir d'un texte à propos duquel le système ne dispose que de très sommaires informations. Cette situation est, nous semble-t-il, la seule qui permette de porter un jugement sur la qualité de la compréhension, que nous qualifions d'explicite car il lui correspond une manifestation physique. Cependant, à une compréhension de ce type est, selon toute vraisemblance, nécessairement associée une 'sémantique fermée'. Autrement dit plus vulgairement, on sait à l'avance 'de quoi on parlera'.
Que se passe-t-il maintenant, si l'on ne dispose pas d'informations sur 'ce dont on parlera' ? Remarquons tout d'abord qu'il s'agit là de la situation normale, celle du lecteur qui entre, pour la première fois, en contact avec un document, à propos duquel il ne possède que de vagues indications, dérivant, par exemple, du titre, du nom des auteurs ou de son volume. Notons ensuite que l'analyse de la compréhension dans cette 'situation normale' a essentiellement donné lieu à un seul type de recherche, où il s'est agi de savoir si l'on pouvait correctement traduire le phénomène compréhensif en terme d'assemblages de 'traits sémantiques distinctifs'. Disons tout de suite que notre opinion à propos de ces travaux est fortement négative : toutes les fois où ils ont été engagés sérieusement, nous sommes contraints de constater que, pour le moment, ils ont abouti à un échec ou à une impasse, compte tenu du nombre gigantesque des *traits* qu'il faut manipuler, même pour décrire une situation très simple, et compte tenu, aussi, de la complexité de la combinatoire qu'ils engendrent. Nous demeurons donc sur la réserve en ce qui concerne cette partie linguistique de l'Intelligence Artificielle.
L'analyse directe de la compréhension en 'sémantique ouverte' semblant utopique, à quoi peut correspondre l'*approche implicite* que nous avons mentionnée ci-avant ? Il s'agira, pour l'essentiel, d'extraire, par des procédés automatiques, d'un texte donné suffisamment d'informations pour que tout lecteur potentiel, ayant une solide connaissance du domaine abordé, puisse en déduire une exacte description du 'ce dont il parle'. Dans cette approche, la fermeture sémantique est réalisée grâce à la compétence du lecteur. Et il est clair, par exemple, qu'un très grand savant, spécialiste des problèmes génétiques, ne saurait 'comprendre', sauf cas exceptionnel, un

tenir à une réalité du texte, sans s'aventurer vers ce qui ressemble trop à des mirages, qui s'évanouissent ou reculent. C'est aussi respecter la dynamique et la pluralité des interprétations possibles.

> One way to use computers effectively without claiming universality is to explore discursive formations that underlie *how* we read texts. The emphasis of such an approach would be to locate words, phrases, and syntactic constructions that produce meanings in a particular instance of reading. (Wolff 1994)

b) Place de la compréhension dans les traitements automatiques

Conception et interface : singer n'est pas la (seule) solution

Les approches de modélisation diffèrent quant à leurs choix fondamentaux :

- modéliser en s'efforçant de reproduire les processus naturels (avec donc une possibilité de validation théorique d'un modèle descriptif) : c'est de cas de l'*IA* (Intelligence Artificielle) dite *forte* ;
- modéliser en visant à obtenir le même comportement, les mêmes réponses, sans nécessairement mimer le moyen d'y parvenir (en quelque sorte, on se donne une 'obligation de résultat', mais pas une 'obligation de moyens') : *IA faible*[121].

La même division se retrouve pour la conception de l'ergonomie des systèmes :

- voiler le fonctionnement du système, et dans l'idée que l'utilisateur n'aie besoin d'aucune connaissance supplémentaire pour s'en servir, qu'il n'aie pas à changer ses façons de faire naturelles parce qu'il se trouve devant une machine : *IA simulatrice* ;
- adapter le système aux moyens disponibles et aux tâches à effectuer, en donnant à l'utilisateur la possibilité d'interagir en appréhendant le fonctionnement du système, en s'adaptant, éventuellement en jouant sur certains paramètres : *IA opératoire*.

En raison de la complexité des processus d'interprétation et du peu de connaissances en la matière, une solution anthropomorphique n'est guère souhaitable. De plus, rien n'assure qu'elle aurait des performances (qualité du traitement) supérieures[122]. Notre objectif est l'efficacité de l'application, pas une investigation du fonctionnement cognitif pour elle-même. La voie de l'IA faible est plus souple et pas moins puissante : c'est elle qui est choisie ici.

> Au lieu de voir dans la formalisation des connaissances un modèle permettant de *reproduire* le comportement cognitif d'un être humain possédant ces connaissances, il s'agit à présent de considérer qu'une telle formalisation permet de construire un système dont le comportement, une fois *interprété*,

---

ouvrage traitant des problèmes technologiques liés à la miniaturisation des ordinateurs. En d'autres termes, c'est le lecteur lui-même qui reconstruira le contenu du texte à partir des éléments qui lui seront proposés. Dès lors, deux types de problèmes sont immédiatement soulevés : cette reconstruction étant, par définition, cohérente, comment peut-on évaluer son degré de fidélité par rapport au texte ? Est-il même envisageable qu'une reconstruction complètement inexacte (qui serait en quelque sorte un contresens global) soit créée ? En second lieu, on doit se demander comment sera traité le passage de la langue du texte à la langue du lecteur, ces deux langues étant presque toujours différentes. »

Ces quelques paragraphes ont le mérite d'évoquer plusieurs idées importantes : le concept de « sémantique fermée », la quête illusoire de représentation exhaustive du sens, la compétence irremplaçable du lecteur. Pour autant, parmi les présupposés contestables, celui sur *le* sens du texte, que le lecteur peut retrouver de façon plus ou moins *exacte*. Quant au problème du passage de la langue du texte à celle du lecteur (problème que nous avons au reste du mal à cerner), nous le suspendrons ici en le réservant aux systèmes fortement multilingues (c'est-à-dire où l'utilisateur peut être confronté à des textes d'une langue qu'il ne connaît pas, ce que nous n'envisageons pas pour DECID).

[121] Par exemple, s'agissant de « la compréhension de textes rédigés dans la langue naturelle [...] il ne s'agit pas de fonctionner comme un humain, ni de comprendre comme un humain, mais de fonctionner de telle manière qu'un humain interprète le comportement du programme comme une compréhension. » (Bachimont 1992, §1.5.3.2, pp. 25-26).

[122] Croire à la supériorité intrinsèque de la simulation est se fourvoyer sur le statut de la modélisation, car il n'y a pas de correspondance scientifique entre d'une part les mécanismes sous-jacents d'un phénomène, et d'autre part la valeur de signification de ce phénomène pour une personne, cf. (Bachimont 1992), notamment §1.5.3.2.

> *approxime* le comportement cognitif d'un être humain possédant ces connaissances. (Bachimont 1992, §1.5.3.2, p. 25)

Dans le cas de DECID, cela consiste à se tenir au plus près de ce dont on dispose : les textes, tels qu'ils peuvent être fournis à la machine.

> In our approach, since the computer does not move away from the text, we must discover ways in which it can be made to work with what is available in the text itself. Accordingly, the basic units of information are words, singly and in combination, word frequencies, and the positions of words in relation to each other. 'Sticking to the text' in this way leads us to develop systems that do things differently from the way a human would ». (Renouf 1993b)

Contrôle et suspens de l'interprétation

Calculer n'est pas comprendre. La machine reçoit des données (une suite de symboles), y applique (aveuglément) un certain nombre de réécritures, et conclut sa tâche sur un critère externe, logiciel (sorties prévues dans l'algorithme) ou matériel (panne). De décision, elle n'a guère : elle suit ce qui a été prévu dans la conception du traitement. Même un choix aléatoire n'est que l'appel d'une certaine fonction, présentant certaines propriétés mathématiques.

En soumettant un texte, dans un certain format qu'il a préparé, et en connaissant les principes du traitement qui lui sont appliqués, l'utilisateur est en mesure de poursuivre l'interprétation et de revenir au texte en s'appuyant sur les indicateurs apportés par le calcul. La phase de manipulation symbolique opérée par la machine est comme une restructuration d'entités, indiquées et déposées au départ, et retrouvées et réappropriées à l'arrivée. On peut parler d'un suspens, d'un report de l'interprétation.

La machine ne comprend pas. Il y a intelligence du traitement, non que la machine soit dotée d'une intelligence, mais que l'utilisateur soit en intelligence avec les propriétés du traitement, que les fondements, objectifs et limites du traitement lui soient intelligibles.

> Les signes sur l'écran sont des symboles ininterprétés, des codes que manipule le programme. Ils sont lus cependant par l'utilisateur qui les interprète et leur attache du sens. Il faut que ce sens soit conforme à ce qui est attendu [du traitement]. (Bachimont 1992, §1.5.3.2, p. 26)

c) La dimension applicative : des contextes favorables

Certaines applications d'analyse de textes visent à extraire certains types d'information (prédéterminés), pour les enregistrer dans une base de données. L'utilisation escomptée est alors l'interrogation sur la base des informations recueillies. Le système est évalué en fonction de sa capacité à répondre à des questions à propos de certains éléments des textes.

La diffusion ciblée ne travaille pas sur l'information elle-même : elle travaille sur sa transmission.

L'observation de situations courantes

Dans un courrier à propos de DECID, Pierre Dumesnil (chercheur à l'INT, Evry) remarque :

> En tant que destinataire potentiel, il me plairait qu'un texte me parvienne qui ne soit pas explicitement dans mon domaine (selon les descripteurs) [*i.e.* les indices apportés par l'auteur ou la collection sont insuffisants, il a fallu considérer le texte lui-même], mais que je puisse lire avec profit. C'est la situation que je vis avec mon libraire qui ne comprend pas trop ce que je lis, mais qui, néanmoins, m'indique ce qui *pourrait m'intéresser*. Souvent, il a raison.

Le libraire pratique une forme de lecture professionnelle, de perception du texte, qui ne mobilise pas une appropriation de connaissances apportées par le texte. Ces types de lectures ont déjà été la source d'inspiration de systèmes informatiques, comme l'exemple qui suit.

Un exemple : de la lecture d'analyse documentaire à la conception d'une application automatique

Le principe de départ du système SERAPHIN (*Système Expert de Repérage Automatique des Phrases Importantes d'un texte et de leur Normalisation*) (Le Roux, Monteil 1993) est de repérer dans les textes des indices de surface pour évaluer le caractère central ou non d'un propos dans le document. Par exemple, une formule comme « il est essentiel de noter que... » manifeste l'importance

accordée par l'auteur à ce qui suit immédiatement. Cette approche revendique le modèle fourni par les pratiques des documentalistes, et plus précisément le cas d'une personne qui connaît bien le fonctionnement des documents et de la langue, même si elle ne possède pas tout le bagage technique pour comprendre un texte de spécialité. La connaissance convoquée par une telle documentaliste, pour analyser le document (pour le classer, le conseiller), serait moins une connaissance encyclopédique qu'une connaissance linguistique et documentaire : tels sont les présupposés sur lesquels s'est fondé le système SERAPHIN. Bref, on pourrait donc analyser un texte et en repérer les principales idées sans pour autant passer par une étape de compréhension.

La recherche documentaire

L'ouvrage de référence du courant de l'*information retrieval* souligne que la tâche de recherche documentaire n'implique pas nécessairement une modélisation fine du contenu du document, ni un objectif de compréhension automatique[123].

> On the one hand, some individuals are convinced that to retrieve items « about » certain subjects, it is necessary to use all available facts pertaining to these items. This operation necessarily requires an analysis of meaning which is not substantially different in information retrieval from other areas of language understanding. In particular, a desirable indexing, or content analysis, approach would then consist of translating the document or query into some formal language consisting of concepts and relationships between the concepts. This introduces the notion of a semantic network and of translations from one language (the input) to another (the formalized index descriptions). [...]
>
> The opposite view about the importance of language analysis [and understanding] in retrieval comes to very different conclusions. [...] The reason may be that a fundamental difference exists between information retrieval on the one hand and certain other language processing tasks on the other. In retrieval one needs to render a document retrievable, rather than to convey the exact meaning of the text. Thus, two items dealing with the same subject matter but coming to different conclusions are treated identically in retrieval, that is, either they are both retrieved or they are both rejected. In a question-answering or language translation situation, these documents would of course be treated differently. This amounts to a qualitative difference between document retrieval on the one hand and question-answering or language translation systems on the other. For example, to answer a specific question about an apple it is helpful to have some detailed knowledge about apples. To retrieve documents about apples, it may be unnecessary to understand precisely what the concept of apple actually entails. Instead, it may be sufficient to detect rough similarities between documents and concepts –for example, it might be enough to know that an apple is more similar to a pear than to an elephant. [...]
>
> This view of information retrieval rejects the notion that information retrieval is simply an early stage of more refined question answering. »
>
> (Salton, McGill 1983, §7.2)

Dans l'application de diffusion ciblée, le système n'a pas à *comprendre* le document, pas plus qu'il n'a à dresser un *portrait* de chaque destinataire. C'est une caractérisation qui est visée : caractérisation des documents qui permet de les positionner les uns par rapport aux autres, caractérisation de l'activité des destinataires, qui reflète leurs intérêts et compétences professionnels.

2. Représentation

a) De justes rapports

La primauté du texte

Le texte est accusé de fournir une représentation partielle (à cause de l'implicite), redondante (synonymies), imprécise (polysémies). La traduction dans un formalisme est alors présentée comme

[123] Mail il y a des tenants des deux écoles, et par exemple (Fox 1987) collectionne les formes de contribution possibles de l'Intelligence Artificielle à la recherche d'informations : représentation des connaissances avec des *frames*, représentation du temps, systèmes experts, etc.

un bénéfice : là, l'expression est rigoureuse et explicite[124]. Cette vision doit retournée, pour rétablir la primauté du texte et reconnaître sa juste place à la représentation formalisée. La formalisation est effectivement nécessaire au traitement, et une formalisation appropriée donne toute son efficacité au traitement automatique. Le sens et la valeur d'une représentation formalisée sont relatives à une application. La langue, à l'œuvre dans le texte, reste la représentation la plus riche et la plus expressive, c'est bien une forme privilégiée de la communication humaine[125]. La représentation formelle ne saisit que certains aspects d'un texte.

Même si l'application informatique n'opère que sur une représentation formelle, la vision sous-jacente peut n'être pas soumise aux contingences techniques. Prenons un jugement désabusé comme celui-ci : « si l'on décide de recourir à l'informatique, alors on adopte une vision du texte en termes de décomptes et de positions de mots ». C'est placer le texte en dépendance par rapport à l'informatique. Cette perspective est en fait à renverser. L'objectif est de se donner une description du texte, dont on tire une modélisation implémentable. La description initiale fonde l'ensemble, et assure une signification qui déborde le cadre d'une manipulation de symboles. L'informatique intervient comme étape de l'application, en tant qu'outil. Elle n'a pas à dicter *a priori* une vision réductrice du problème.

Ce qui revient à la machine

Les atouts de la machine se définissent en regard des limites naturelles des capacités cognitives humaines :

---

[124] « Language, a mere go-between in our communicative intentions, creates certain severe shortcomings which are unacceptable from a logical point of view : homonymy (various things called by the same name) ; synonymy (various names for one thing) ; extensional indeterminacy ; and indistinction among levels. In trying to solve these obstacles, we traditionally fall back on logic, a discipline that studies the structure, foundation and use of cognoscitive expressions, allowing, in short, a meticulous analysis of thought. [...]
The importance of formal logic applied to operations of content analysis comes from the fact that logical symbols, unlike linguistic ones, have a perfectly accurate meaning. One of the most important discoveries of contemporary methodology is having realised that, using language in its syntactic plane (and therefore disregarding the other two) makes the intellectual work much easier. [...] [The logical primitive elements and rules] do not form a language, a means of communication, but rather a truly syntactic framework : their elements are opaque entities, though there is always the possibility of transforming a calculation into a language by interpreting its symbols and giving them a meaning ». (Pinto Molina 1994)

[125] « les langues naturelles, elles, n'ont pas les propriétés d'un code ; elles évoluent dans le temps, elles comportent nécessairement de l'implicite, et elles ne connaissent pas de correspondance bi-univoque entre forme et sens : c'est précisément cette non-biunivocité constitutive sur le plan des signifiants (marqueurs et structures de marqueurs) et le plan des signifiés (valeurs sémantiques) –sources des phénomènes d'ambiguïté, de polysémie, de synonymie, de paraphrase– qui donne aux langues cette marge de jeu, cette labilité leur permettant d'être des instruments de communication (et pas seulement des moyens de consigner l'information). » (Fuchs & *al.* 1993, introduction §3.2, p. 25)
« Il est [...] communément entendu que la langue ne serait pas suffisamment précise ou qu'elle serait ambiguë. [...] [Pourtant], dans son effectivité, la langue ne fonctionne pas comme un assemblage codifié d'éléments aux propriétés préalablement fixées, mais comme un *système* dont la cohérence est testée à la fois de manière interne (cohésion) et de manière externe en référence avec un monde, imaginaire ou réel, jugé possible. Cette propriété, *si elle est maniée avec suffisamment de virtuosité*, permet en particulier de s'affranchir des significations disponibles, des règles grammaticales ou syntaxiques sans que la construction finale soit privée de sens, sans qu'elle soit ambiguë et sans qu'elle soit déclarée « illégale ». Cette « torsion » des règles et des « valeurs » des éléments à assembler serait destructrice pour un langage, elle ne l'est pas pour la langue [...].
[...] dans la vie la plus quotidienne, sans être poètes, locuteurs et scripteurs d'un côté, auditeurs et lecteurs de l'autre, manifestent la capacité à dire et à entendre l'inouï, à écrire et à lire l'inédit, non pas comme simple assemblage, combinatoire ou enchaînement de ce qui avait déjà été dit ou écrit, mais comme *vraie nouveauté* ou mieux, comme *création*, non logiquement déductible des traces externes antérieures de la langue. Cette capacité à énoncer et à communiquer efficacement le nouveau –ou, de manière infiniment plus rapide que le langage, le non immédiatement déductible– constitue à nos yeux ce qui rend inexpugnable la position de la langue. »
(Dumesnil 1995, pp. 13-14 et 16-17)
Voir aussi (Rastier 1995c, §I.C), sur les imperfections dont a été accusée la langue et sur les entreprises de construction d'une langue parfaite.

- sa capacité à faire un balayage intégral, *systématique*, exhaustif, selon le point de vue qui lui a été prescrit. La machine ne perd rien, n'oublie rien, ne néglige rien des entités qu'on lui a demandé de repérer et d'enregistrer.
- sa capacité à *embrasser dans leur ensemble* d'énormes volumes de données, et d'accéder avec une égale facilité à une multiplicité d'informations. L'ordinateur a ainsi réellement permis de mettre en œuvre les techniques d'analyse des données (analyse factorielle, classifications), qui font appel à des structures matricielles de grande taille qui, sauf approximation ou cas particulier, ne se décomposent pas en structures de taille plus réduite où répartir la tâche et mener des petits calculs indépendants. La machine est ainsi capable de suggérer et de mettre en évidence des rapports qui passeraient inaperçus dans la « masse ».
- sa *rapidité pour effectuer des calculs*. Calcul doit être pris dans un sens très large, qui inclut non seulement les opérations arithmétiques, mais aussi des opérations élémentaires sur divers types de données (troncature d'une chaîne de caractères, etc.), ainsi que toutes les *fonctions* et *procédures*, combinaisons élaborées de transformations que les programmes informatiques permettent de définir.
- ses bases de codage *explicites, déterministes, et discrètes*. Il y a de multiples manières de définir l'égalité (même enregistrement en mémoire, même valeur), mais pour une définition donnée le résultat est tranché : l'ordinateur n'« hésite » pas[126], n'a pas de problème perceptif et interprétatif. Cela le rend particulièrement apte dans des fonctions de contrôle et de vérification systématique de contraintes.

Ces propositions rejoignent celles qui ont pu être exprimées sur la conception de systèmes *anthropocentrés* :

> c'est la machine qui assiste l'homme, non pas l'homme la machine [...] [,] en lui proposant ses services en matière d'organisation et de gestion des ressources, de calculs symboliques, de comparaisons, bref de services de contrôle de cohérence et de suggestion. (Kanellos, Thlivitis 1997)

Si l'ordinateur n'a rien de la créativité et de l'intelligence humaine –il ne fournit rien qu'il n'ait reçu, sous forme de données ou d'algorithme[127]–, il est en revanche capable d'épauler l'homme dans certains traitements. Il démultiplie[128] les possibilités sur les aspects que nous avons énumérés. Cette aide est comme celle d'un outil, qui prolonge les possibilités d'action sous une forme différente : le *calcul* de la machine ne peut être l'image que d'une infime partie des mécanismes de *raisonnement* humain ; l'*enregistrement* de la machine est un reflet, déformé et appauvri à l'extrême, de la *mémoire* humaine.

Il y a bien une place pour la machine dans la sémantique des textes, aux côtés de l'homme qui seul donne du sens :

> Yves-Marie Visetti (Visetti 1991) annonçait un renversement de tendance, au vu des échecs des tentatives d'artificialiser et de copier l'humain, et proposait la coopération entre l'homme et la machine. Qui remettrait en cause une telle coopération pour le langage ? Sûrement pas nous, qui utilisons un traitement de texte pour cet article, et un correcteur d'orthographe pour vérifier sa place dans une norme langagière. Mais nous allons ici proposer un nouveau type d'outils, que l'on pourrait

[126] Des oscillations éventuelles ne relèvent pas d'un comportement intrinsèque de la machine, mais d'un algorithme dont la conception est de la responsabilité de l'informaticien.

[127] « L'expression dans les langages formels des connaissances réduit la connaissance à sa formalisation : l'approche cognitiviste, formaliste, ne trouve alors dans les données qu'elle même, ce qu'elle y a donné. L'IA, dans cette mesure, ne traite les problèmes non tels qu'ils se présentent, mais tels qu'elle les re-présente, les formalise. L'IA devient ainsi tautologique en ne traitant plus que les problèmes qu'elle sait résoudre. » (Bachimont 1992, §8, p. 309).

[128] Ou « amplifie » : « [Engelbart (dans GREIF Irene (ed.) (1988) - *Computer supported cooperative work : a book of readings*, San Mateo, CA, Morgan Kaufmann) parle des machines comme servant à amplifier l'activité mentale.] Cette idée d'amplifier est tout à fait intéressante parce qu'elle procède d'une orientation vraiment différente de celle de l'intelligence artificielle, les systèmes experts, etc. ; il ne s'agit pas de substituer quoi que ce soit à l'activité mentale, il s'agit de lui donner des amplificateurs, c'est-à-dire un outillage qui lui permette de tirer davantage de son activité mentale grâce à un outil plus puissant. Engelbart dit d'ailleurs, qu'il n'y a pas d'intelligence humaine sans outillage, et l'intelligence artificielle c'est un outil particulier, mais ce n'est pas une autre espèce d'intelligence, c'est toujours le même mode de fonctionnement, une activité mentale qui tire ses connaissances de son outillage. » (Poitou, Ballay, Saintive p. 11)

> grossièrement appeler des outils de traitement de sens. Tout comme les textes que nous tapons ne proviennent pas de la machine sur laquelle ils sont simplement mis en forme, une interprétation, ou une attribution de sens à un texte relève uniquement de l'humain, avec tout ce que cette notion peut supporter de psychologique et de social. Dans le type d'outils que nous proposons, la machine ne sert donc qu'à mettre en forme une interprétation, et surtout à guider cette opération dans un ensemble de contraintes et d'étapes. Nous verrons également comment cet ensemble d'opérations peut se révéler source de création, en compensant l'aspect exhaustif de ce processus par la proposition de nouvelles directions dans l'exploration du sens. Enfin, la « standardisation » des données interprétatives permet également la qualification d'une interprétation, et propose des voies vers l'appréhension de l'intertextualité, toujours en utilisant les facultés de calcul de la machine. (Tanguy, Thlivitis 1996)

La machine prend place dans une stratégie d'*aide* au traitement d'un grand volume de textes[129], et non sous la forme du *remplacement* d'une compréhension humaine.

> Even though we can read more text more systematically with computers, we must still contend with our 'horizon of expectations', a preexisting frame of reference that governs how we interpret texts according to subjective perceptions.
>
> Perhaps one way to use computers effectively in textual analysis is to see how the text is able to manipulate how we read. Intertextuality from this angle would not be a static system of fixed signifiers but rather openings in the text that compel the reader to participate in the production of meaning. [...]
>
> [Although the analysis is never complete,] databases such as ARTFL[130] enable us to explore intertextuality in ways that did not exist before computers. [...]
>
> Anxiety about computers in the humanities may finally have less to do with methodologies and more to do with our expectations for 'science'. The technology at our disposal leads many individuals to expect that literature will finally be explained scientifically. Unless we make it clear what computers can and cannot do, the myth of science will obscure rather than enhance literary studies.
>
> (Wolff 1994)

Soulignons en particulier combien l'introduction du calcul est un apport majeur dans le cadre de l'inter-détermination global - local, dès lors que le global excède les capacités cognitives de présentation synoptique. En effet, pour *comprendre* –prendre ensemble– il faut pouvoir embrasser un ensemble pour situer et interpréter chaque élément. Le support numérique et ses possibilités de traitement (bien pensées) prend avantageusement le relais des dossiers papier, lorsque ceux-ci ne peuvent plus être disposés, agencés, de façon significative, sur la surface d'un bureau. Accordons-nous une illustration complète et récapitulative de ce propos, dans le cadre de l'étude de la mise sous forme hypertextuelle des dossiers patient dans un hôpital :

> Le dossier papier [un « dossier patient » dans un hôpital] [...] autorise [...] une lecture rapide et efficace en fonction des objectifs de lecture fixés dans la pratique hospitalière. En effet, [...] [il] peut s'étaler ([sur] une table par exemple) et [...] la position des documents dans cet espace conditionne la signification des informations contenues dans ces documents. Par ailleurs, outre la position dans l'espace, la nature physique du support papier conditionne l'interprétation des informations. Par exemple, si le dossier est étalé en paquets correspondant chacun à une hospitalisation passée, un paquet peu épais renverra à une hospitalisation de routine dont la consultation n'est que de peu d'intérêt ; en revanche, un paquet plus épais correspond à une hospitalisation au cours de laquelle des complications sont survenues et par conséquent elle mérite le détour. Par ailleurs, la couleur plus ou moins jaune du papier indique l'ancienneté de l'hospitalisation : une hospitalisation ancienne étant consultée en dernier [...].
>
> Quand on dématérialise le dossier [en en faisant un document électronique, sous forme hypertexte], on perd les aides matérielles à la navigation / consultation apportées par le support papier. [...] En effet, la consultation [d'une collection de documents] se structure à partir du moment où, étalé sur un espace, [la collection de documents] peut s'appréhender globalement comme un tout : on embrasse sa finitude d'un seul regard et c'est dans ce cadre fini que l'on instrumente la consultation en interprétant la position spatiale dans l'espace comme une prescription interprétative sur le contenu. Puisque la signification est une position dans un réseau de valeurs sémantiques, il est indispensable

[129] (Michel 1997, §2.4, pp. 223-224) voit dans les nouvelles procédures d'investigations sur des grands volumes de textes, l'apparition d'une *macro-information*, par opposition à la *micro-documentation* traditionnelle, où il s'agit de sélectionner quelques références de textes à lire.

[130] ARTFL : *American and French Research on the Treasury of the French Language*. Collaboration entre l'Université de Chicago et l'Institut National de la Langue Française (INaLF / CNRS), à partir du corpus de littérature française qui a été constitué pour la réalisation du dictionnaire *Trésor de la Langue Française* (TLF).

d'embrasser le réseau dans sa globalité pour attribuer une position et donc une signification au document consulté. Or, il est bien clair que la synopsis globale de [la collection de documents] est perdue lors du passage au support informatique. [...]

Pour pallier cette désorientation inhérente à l'informatisation des hyperdocuments, on s'attache, dans certaines expérimentations, à reproduire sur l'écran des équivalents iconiques des aides matérielles liées au support papier (indices de couleur, taille analogique des dossiers représentant une hospitalisation, reproduction du fait de tourner les pages, etc.). Egalement on tente de suggérer à l'utilisateur une vision d'ensemble de l'hypertexte pour qu'il sache où il en est : par exemple on lui soumet un damier de rectangles colorés, chaque rectangle correspondant à un document, la couleur indiquant qu'on l'a déjà consulté ou non.

Mais, par ailleurs, on gagne la possibilité de déléguer au système le calcul de la navigation ; au lieu qu'il s'agisse de se repérer en fonction de la contiguïté spatiale, il s'agit de se repérer dans la consultation du dossier informatisé en fonction de liens calculés. La *finitude synoptique* du dossier, qui s'offre au regard, doit laisser place à la *finitude computationnelle* du dossier. Le support informatique peut calculer sur l'ensemble du dossier et proposer via les liens le point de vue synoptique dont a besoin l'utilisateur pour s'orienter. On gagnerait ainsi la possibilité d'appréhender des [collections de documents] dont le volume matériel interdit toute synopsis. Il en est ainsi des [collections de documents] portant sur des systèmes techniques complexes : par exemple, la documentation technique d'un Airbus est aussi volumineuse que l'Airbus lui-même [...].

Il n'est pas question de lire l'hyperdocument à la place du lecteur : le lien calculé propose un sens de parcours, il ne l'impose pas. Pour deux raisons : la première renvoyant aux considérations [...] sur le fait que les actes interprétatifs d'un lecteur humain ne sont pas d'ordre calculatoire et qu'il n'est par conséquent pas possible de calculer une lecture ; la seconde tenant au fait que, pour que lecture il y ait, il faut qu'il y ait une actualisation active par le lecteur du sens proposé par le système. [...]

Si l'on se souvient que la meilleure manière de traiter [une collection de documents] papier est de l'étaler sur un plan matériel pour que sa vision globale permette d'interpréter les positions des documents vis-à-vis de l'ensemble, on constate que l'on conserve la même idée à un niveau local désormais, l'écran. [...] [Le calcul permet de projeter] le global dans le local. Par exemple, lorsque l'on calcule une table des matières [...], c'est bien de cela dont il s'agit : l'appréhension de la totalité du texte, tâche difficile, fastidieuse voire impossible à l'échelle d'un individu, pour en déduire un document de synthèse accessible dans son unité et sa globalité.

(Bachimont 1999c, pp. 21-27)

Sans interprète, pas de sens

La machine n'opère que des réécritures, selon des opérations d'ordre syntaxique. Or la syntaxe ne détermine pas la sémantique. C'est l'utilisateur qui se fait interprète et confère un sens aux calculs de la machine.[131]

Plus généralement, les objets ne préexistent pas extérieurement à un sujet (une personne), et indépendamment. Ils sont délimités, identifiées, construits, constitués, dans une activité herméneutique. Ceci renvoie à l'approche phénoménologique.

Une réflexion sur la constitution ontologique *et* épistémique des sciences du langage fait apparaître en leur sein le travail d'une *double herméneutique* : les structures langagières sont à la fois la *condition de possibilité* et le *résultat* de l'activité interprétante des sujets parlants. [...]

Les textes ne sont [...] en aucune manière des représentations renvoyées sur un réel supposé objectif.

[131] « [...] nous savons que l'ordinateur est une machine à manipuler des signes [:] [...] les algorithmes ne transforment les codes qu'en fonction de leur forme syntaxique, et non en fonction de leur interprétation sémantique. [...]

[Or] une connaissance ou représentation interprétée ne contient pas dans sa forme ce qui fait d'elle une connaissance, *i.e.* les principes de son interprétation. [...]

[Donc] le principe de calquer la syntaxe sur la sémantique [sic] est intenable en son fondement : le sens et la forme ne sont pas en relation biunivoque, et les programmes de l'IA [Intelligence Artificielle] ne sont pas des connaissances.

[...] Par conséquent, il faut confier à l'utilisateur la tâche d'interprétation, qui devient par là même une tâche de validation.

[...] Pour parler du sens il faut un interprétant [c'est-à-dire quelqu'un qui interprète.] »

(Bachimont 1992, §1.5.2, pp. 20, 24-25)

(Havelange 1995, pp. 136-137)

La subjectivité originelle et ultime de tout traitement sur les textes peut être tempérée par la recherche et la mise en évidence de régularités et de convergences, subsumant les variations d'un traitement à l'autre :

> Si l'on peut raisonnablement espérer conduire une *analyse* rigoureuse et exhaustive, la phase d'*interprétation* reste inévitablement subjective, incertaine mais cependant nécessaire : il faut en assumer le risque, en tentant de le limiter par une recherche constante de convergences (multiplication des analyses, croisement des résultats, etc.). (Dupuy 1993, p. 34)

b) Une heureuse fatalité

Représenter, c'est réduire

Le texte est irréductible, toute transposition « perd » quelque chose de lui. C'est le cas de ses descriptions secondaires (mots-clés, résumé), aussi bien que de tout encodage formel (arbres syntaxiques, étiquettes « sémantiques »).Toute modélisation appauvrit. Il n'y a pas 'dans' le texte un 'contenu' qu'on peut capter et transporter dans une représentation[132].

> La réduction apparaît, sous sa forme la plus simple, comme la suppression de la redondance. [...] [Elle] ne peut se faire qu'au prix d'un certain appauvrissement de la signification : le niveau de généralité une fois choisi, la description ne peut apparaître que comme la sélection des éléments de contenu pertinents et comme le rejet (ou la suspension provisoire) d'autres éléments, considérés comme stylistiques et non pertinents pour la construction du modèle. (Greimas 1966, §IX.3.b, p. 159)

En explicitant ce qu'elle retient, la représentation perd la dimension, en perspective infinie, de déploiements implicites du texte.

Pour autant, élaborer un traitement automatique suppose bien à un moment de construire une représentation du texte, qui inévitablement en gomme certains aspects mais pour mieux se focaliser sur d'autres : tout l'intérêt de la recherche est dans le choix, le repérage et la modélisation de la dimension du texte que l'on s'efforce de capter, sans prétendre qu'il s'agisse encore du texte dans sa plénitude.

> L'utilisation de la théorie sémantique pour des applications informatiques consiste en premier lieu à transposer une lecture descriptive en lecture réductive, c'est-à-dire à sélectionner les unités sémantiques pertinentes pour la tâche. (Rastier, Cavazza, Abeillé 1994, §I.2.2, p. 16)

La modélisation a à la fois un rôle descriptif –rendre compte de ce que sur quoi l'on veut baser l'analyse–, et un rôle normatif –ici, permettre la comparaison, rendre commensurable.

Réduire, c'est commencer à interpréter

La réduction procède d'un choix, qui met en relief les caractères « intéressants ». La réduction est intrinsèquement subjective et relative, mais, ainsi orientée, elle gagne en pouvoir de signification ce qu'elle perd en universalité.

En centrant la représentation sur ce qui est utile dans le traitement, la réduction concourt à l'*efficacité* du traitement. En organisant la description et en en donnant les contours, la réduction répond aussi à des critères d'*ergonomie* : elle aide à saisir, à percevoir, la réalité décrite dans sa diversité.

L'indexation, dans les pratiques documentaires, devrait ainsi être guidée par des préoccupations herméneutiques. L'enjeu est de mettre en rapport le texte et le lecteur. L'analyse documentaire procède de l'élucidation (avec un objectif de « fidélité » de la représentation) et de la contextualisation (limiter l'émiettement des documents, des œuvres). La représentation documentaire est bien appelée à concilier « l'ouverture herméneutique potentiellement infinie et la nécessaire réduction documentaire ». (Richardot 1996)

Donner la représentation d'un texte, dans DECID, c'est expliciter les ingrédients de lectures du texte. Un texte, y compris un texte technique, est susceptible de multiples lectures, en fonction des préoccupations du lecteur, de ce qui motive la lecture, etc. La représentation, conçue pour une

[132] C'est une des raisons de récuser la possibilité de définir un langage pivot, qui servirait d'intermédiaire pour passer d'une langue à une autre, sans déformation de sens (ni perte, ni ajout !).

certaine application, s'inscrit elle-même dans un mode d'approche du texte. Pour DECID par exemple, il ne semble pas prioritaire de se focaliser sur les actes de langage, même si c'est une question par ailleurs très étudiée. Il y a, dans le choix d'une représentation, un opportunisme bien dosé.

c) *Les voies de réduction*

La projection

C'est le choix initial d'un espace de description.

Pour un texte : va-t-on considérer la police de caractère utilisée, l'ordre des mots, le changement de page, etc. Le choix le plus courant, en lexicométrie, est de considérer le texte comme une suite de caractères, en distinguant des caractères constitutifs de mots (lettres), des caractères séparateurs de mots (espace, apostrophe), et éventuellement des caractères particuliers (ponctuations).

Le texte est en fait toujours tributaire d'une projection, sur le plan de sa réalisation matérielle (codage, mise en page,...) et sur le plan de son rapport à un lecteur, lors d'une lecture (attention portée sur certains aspects et pas sur d'autres, sciemment, consciemment, et inconsciemment). La projection est la réification de la *perception* : elle explicite ce que capte la machine, ce qu'aperçoit le lecteur.

Les projections possibles d'un texte sont *multiples*, et constituent autant de représentations pertinentes pour des points de vue et des usages différents. Le texte est dans chaque cas considéré sous un certain angle. Chaque projection s'opère suivant un *axe*, qui organise la façon dont le texte se « rabat » sur le plan de projection, et lui donne donc sa direction (son sens ?) générale, assurant la cohérence d'ensemble de la représentation vis-à-vis de la matière initiale.

(Morizet-Mahoudeaux, Terray, Brunié, Kassel 1998) font de la projection un processus fondamental dans les systèmes hypertextes : chaque manière d'utiliser les données du fichier électronique enregistrant le texte est une projection[133]. La définition qu'ils proposent résume la plupart des points que nous venons de voir (c'est nous qui soulignons) :

> Definition : a projection is any systematic operation, which gives a perceptible output, from a determined format. (Morizet-Mahoudeaux, Terray, Brunié, Kassel 1998)

Ajoutons un dernier point, qui annonce le paragraphe suivant : la projection est *globale*, au sens où elle considère tout le texte, et où elle retient sa réalisation entière sur la ou les dimensions qu'elle retient. C'est ainsi qu'elle se distingue d'une opération de sélection.

La sélection et l'élimination

La seconde voie de réduction est pratiquée sous ses deux faces : sélection et élimination. Dans les deux cas, il s'agit de délimiter un sous-ensemble dans un ensemble plus vaste. On obtient ainsi deux parties (le sous-ensemble *vs* le complémentaire), que l'on fait correspondre aux deux alternatives, garder *vs* laisser.

Les deux opérations, de sélection et d'élimination, seraient donc formellement complètement équivalentes, si ne s'introduisait un facteur de dissymétrie. Tout dépend du caractère clos et bien déterminé de l'ensemble initial vis-à-vis des valeurs que peut prendre le critère. Soit le partage de l'ensemble de départ se fait sur le mode une partie *vs* le reste (vision dissymétrique), soit on a deux parties qui ont chacune leur consistance propre (vision symétrique).

Quand on se donne une liste de mots-outils, ou un référentiel terminologique, on opère un découpage dissymétrique du vocabulaire de la langue. On ne sait pas définir le reste du vocabulaire de

---

[133] Extraits de (Morizet-Mahoudeaux, Terray, Brunié, Kassel 1998) : We can « not only improve our understanding of the nature of digital documents, but, more precisely our understanding of how we are accessing them. Effectively, we do not have access to digital documents directly, since they always are an abstraction of the electronic state of the computer, which we model through a bunch of numbers. This calls for defining the operation that makes the document readable. We will call it a *projection* of the digital document. »
« The table of contents from a structured text is a projection of the document, even if it omits a great part of the file, since it is still a view of the same file. »
« Each projection is an interpretation by the software of the semantics of the format. More precisely, the software is a tool used by the author of the programs for building interpretations of this semantics. »

la langue de façon positive, c'est-à-dire autrement que « ce qui n'est pas dans la liste (ou le référentiel) ».

Faire une sélection à partir d'un référentiel, c'est savoir ce que l'on garde, et ignorer ce que l'on laisse. Autrement dit, c'est retrouver à la sortie ce que l'on a mis en entrée, et ne pas accorder d'attention à ce qui ne correspond pas à ce qui est prévu. Le traitement est déterminé par le format du résultat que l'on veut obtenir. C'est l'attitude que l'on prend pour remplir une grille d'analyse (avec des rubriques précodées), pour repérer des exemples d'usage d'un mot dans un corpus, pour effectuer une indexation contrôlée (toutes les notions à traduire par un mot-clé sont fixées).

Faire une élimination à partir d'une liste de mots-outils, c'est savoir ce que l'on veut laisser, et garder ce dont la preuve de l'inutilité n'est pas (encore) faite. Le rejet ne se fait que *a posteriori*, en connaissance de cause. Cette attitude prévaut dans les dispositifs de veille, dans lesquels une nouveauté doit être repérée. C'est aussi l'approche générale de DECID : l'outil n'est un apport, pour la diffusion de l'information, que s'il est capable de percevoir des destinataires mal représentés par la grille de l'organigramme, et pas toujours connus des collègues : activité marginale par rapport à l'équipe de rattachement, récent embauché, personne sur un site éloigné, etc.

Le regroupement, la synthèse

Le regroupement procède par fusion : ce qui était distingué, et constituait plusieurs unités, est finalement saisi en une seule unité. Une information sur la nature de la fusion peut enrichir la nouvelle unité. L'unité joue alors le rôle d'une formulation synthétique de ses composantes.

L'illustration la plus évidente est celle des regroupements en classes. Les procédures ascendantes opèrent par regroupements successifs d'éléments ou de classes. Cette approche adopte généralement un point de vue local : la constitution des classes s'organise à partir des voisinages locaux, des proximités des éléments deux à deux. L'attention est portée sur la constitution interne de chaque classe, considérée indépendamment. La démarche inverse se pratique également. Les procédures descendantes raisonnent par divisions successives de l'ensemble initial. Elles considèrent donc à chaque étape une partie et cherchent une scission optimale du point de vue de l'ensemble de ses éléments. D'autres méthodes sont globales en cherchant à optimiser une partition en fonction de critères portant sur la structure d'ensemble. *In fine*, ce sont les rapports inter-classes qui ont une importance dominante dans la représentation obtenue. (Quatrain, Béguinet 1996, §3.3)

Quand (Greimas 1966, §V.3.c, p.68) annonce trois étapes pour la description, à savoir, successivement, l'*inventaire*, la *réduction*, et la *structuration*, il emploie le terme *réduction* dans un sens plus restreint qu'ici, car la réduction se traduit essentiellement chez lui par des opérations de regroupement (peut-être un peu aussi d'élimination). L'*inventaire* quant à lui a son pendant dans la projection (et la sélection) ; la *structuration* pourrait être une forme encore différente de réduction, la réduction par analyse, qu'il nous reste maintenant à présenter.

L'analyse et la description par des lois

Un ensemble d'unités « primitives », muni de lois de composition, permet de représenter en puissance un beaucoup plus grand nombre d'unités « complexes ». Les primitives et les lois constituent un résumé d'un ensemble de possibilités virtuelles. La réduction n'est satisfaisante que si le décalage entre les unités prévues par les lois et les unités effectivement réalisées est négligeable ou ne perturbe pas la description. D'autre part, pour qu'il y ait à proprement parler réduction, le nombre des primitives doit être d'un ordre de grandeur inférieur à celui des unités à décrire, et les lois doivent être simples et peu nombreuses (donc globalement productives).

Les langages d'indexation pré- ou postcoordonnés s'appuient sur cette propriété. Avec les langages précoordonnés, la description du référentiel d'indexation est relativement concise, structurée et systématique, et les *termes* d'index effectifs sont nombreux et précis. Pour les langages postcoordonnés, les possibilités de représentation d'un document sont démultipliées par les possibilités combinatoires des termes d'index[134].

[134] Voir par exemple (Lefèvre 1997, §4.2.1) pour une explication plus développée de la coordination dans les langages documentaires.

Au plan des représentations graphiques, la recherche de *variétés* (au sens mathématique du terme) s'apparente à ce type de réduction : on ajuste un type de forme connu aux données, qui, plus général et plus souple, promet d'être mieux adapté que les seules formes linéaires (droites, plans)[135].

d) Repères pour la mise en œuvre

Démarche méthodologique

Il y a trois étapes-clés auxquelles se jouent la qualité de la représentation pour l'application : la conception, la réalisation informatique, et la validation.

> Le schéma régulateur [que nous proposons pour la construction d'un système dans le domaine de l'Intelligence Artificielle] comprend trois étapes clés : *(i)* spécification en termes sémiotiques du comportement que le futur système doit posséder, c'est une étape de description sémiotique ; *(ii)* modélisation scientifique et construction technique d'un système respectant les descriptions sémiotiques, c'est une étape scientifique ; *(iii)* évaluation au niveau phénoménologique de l'adéquation entre les descriptions sémiotiques et le comportement produit par le système, c'est une étape d'évaluation sémiotique. (Bachimont 1992, §8.2, p. 313)

La dernière étape est rendue nécessaire parce que la description sémiotique donne les conditions nécessaires pour que la machine fonctionne conformément à ce qu'on en attend, mais non des conditions suffisantes (Bachimont 1992, §8.2.3, p. 316).

Des critères pour qualifier la représentation

Bruno Bachimont distingue trois dimensions d'évaluation de l'adéquation de la représentation :

> Les critères d'évaluation découlent du triptyque nature / représentation / utilisation. Un premier critère consiste à évaluer la facilité d'expression des connaissances dans le formalisme (adéquation nature / représentation), un second à évaluer la pertinence des représentations dans le système (adéquation représentation / utilisation), un troisième à évaluer le comportement global de l'architecture (adéquation nature / utilisation). (Bachimont 1992, §8.5.2, p. 339)

Chacun de ces critères reçoit un nom. Le premier (adéquation nature / représentation) évalue la capacité du formalisme choisi à rendre compte de ce qui est pertinent pour le traitement, c'est un critère d'*expressivité*. Le second (adéquation représentation / utilisation) évalue si l'utilisation des représentations par la machine est explicite et formulée en termes pertinents pour le concepteur, si les principes de fonctionnement sont clairs, c'est un critère de *transparence*. Le dernier (adéquation nature / utilisation) évalue si le traitement correspond aux spécifications qui ont pu être posées, s'il est conforme à ce qu'on attend du système, c'est un critère de *correction* (Bachimont 1992, §2.5, p. 72).

3. Interprétation : huit conceptions

a) Introduction au parcours proposé

L'examen des différentes conceptions de l'interprétation suit les jalons posés par (Rastier 1987), ce dont nous ne nous cacherons pas. Cette reprise se justifie par un changement de fil conducteur. La progression ici aménagée va par « nombre de sens croissant ». Elle part de la conception la plus restrictive (quasi booléenne) à celle la plus largement ouverte (infinité), en passant graduellement par les conceptions intermédiaires.

Cet ordre assez systématique nous paraît intéressant pour explorer méthodiquement le domaine[136].

---

[135] L'équipe de Martin RAJMAN (EPFL, Lausanne) explore ainsi l'application l'analyse curvilinéaire pour construire une représentation optimale (déformation minimale et minimum d'espace « perdu ») de l'ensemble des textes d'une base. (`http://liawww.epfl.ch/~lnmain`)

[136] On ne suit pas ici de près les travaux d'Umberto ECO, qui a eu un rayonnement majeur dans les études sur l'acte de lecture, notamment à travers ses essais (Eco 1979) ou (Eco 1990).
Il faut également signaler la parution récente de plusieurs ouvrages de synthèse sur la question de la lecture, par exemple (Jouve 1993), (Dufays 1994).

b) Véricondition

La logique formelle dispose d'un concept d'interprétation. Or la logique est quelquefois choisie comme langage de représentations de textes[137] (ce choix relève du débat sur *Texte et phrases*, car le niveau de travail de la logique est la proposition).

La logique piège le sens dans une alternative *vrai / faux*. Elle suppose un monde (ou modèle), représentation univoque et complète de la réalité concernée, et l'association biunivoque des objets du monde et des unités d'expression. Elle statue alors sur la correspondance (isomorphe) de la présence et l'agencement des entités langagières avec le monde. Cette conception fait du monde est un décalque du lexique et de la syntaxe du langage, le premier fixant ses objets, la seconde instituant les relations possibles entre eux. Il en résulte une sémantique statique et atomiste (Rastier 1994, §1.2.c, p. 329 sq.). L'extension opérée par la logique modale ne fait qu'introduire et gérer une multiplicité de mondes.

Tout le traitement est polarisé sur les objets de travail de la logique : les quantifications (existence, universalité), la négation, la portée des différents opérateurs.

Cette conception se heurte à de sévères impasses dès lors que l'on veut s'en servir pour une sémantique des textes[138]. La notion de monde, qui serait porteuse de la sémantique du texte, ne fait que déporter la question du sens sans la résoudre[139]. Pire, cette représentation perd tout ancrage à la réalité : on abandonne la réalité du texte pour partir à la quête d'une description artificielle, appauvrie, et dont le mode de définition est non résolu. La traduction logique opère de façon myope, proposition à proposition, et s'arrange mal des souplesses de construction et d'expression de la langue[140]. Enfin, décrire le déploiement d'une interprétation comme une fonction d'appariement d'un terme à un objet, et d'une proposition à une valeur de vérité, est contraire à l'expérience de lecture[141]. Ce que l'on retient d'un texte est différent d'un moment à l'autre, d'une personne à l'autre, et ne se laisse pas réduire à une collection d'assertions validées ou invalidées.

c) Extraction et univocité

Sens hors-contexte

Devant la montée en volume de l'information textuelle disponible, d'aucuns élaborent des systèmes d'extraction de l'information des textes, de synthèse automatique, de filtrage, qui seraient une digestion appropriée, directement assimilable. La critique à leur encontre est évidente : l'expérience interprétative de chacun montre l'incidence déterminante du contexte dans le sens trouvé au texte.

> The decision to use poststructuralist theories must imply a prior decision to at least suspend the belief that texts have unique meanings that can be extracted with the right tools. (Wolff 1994)

---

[137] La DRT fait une place importante au formalisme logique. On trouve aussi des initiatives, qui appellent la logique de leurs vœux, mais laissent dubitatif quant à la forme concrète que cela pourrait prendre et à la faisabilité de l'ensemble.
« It is often the case that human knowledge materialised in scientific texts is a 'world' of logical facts (universal concepts) arranged in a logical structure [...]. If we add the contrastable reality that any analytical process is, above all, a logical process, then any doubts about the possible contributions of logic to WTDCA [Written Text Documentary Content Analysis] are fully dissipated. » (Pinto Molina 1994)

[138] Autant le formalisme logique est un outil puissant, précis et rigoureux, pour structurer des *significations* (lexicales), autant il s'avère inadapté à la description du *sens* (d'un texte).

[139] « [Pour une approche logiciste du langage,] la vérité des phrases est relative à un modèle, d'où la nécessité de construire des modèles (partiels ou non) de l'univers sémantique décrit avant de pouvoir leur assigner des valeurs de vérité. » (Rastier 1987, §IV.2.6.3, p. 102)

[140] Un écart qui n'est pas des moindres : les primitives de la logique (les objets, propositions et arguments, sur lesquels s'appliquent ses opérateurs), ne sont pas contextuelles (Bachimont 1999c).

[141] La représentation des tautologies par exemple n'est pas satisfaisante : sur ce point voir (Rastier 1987, §VII.1).

Détermination par optimalité

Pour résoudre l'implicite et les ambiguïtés, qui empêchent de trouver le sens du texte, des principes régissant la communication ont été proposés. Les maximes de Grice[142], conçues pour décrire ce qui est compris au cours d'une conversation, découlent du postulat que les interlocuteurs cherchent à échanger de l'information de façon coopérative et efficace. (Sperber, Wilson 1986) poursuivent dans cette voie, en concluant que le sens effectif d'un message est celui dont l'intégration à ce qui précède et à la situation est la plus directe. L'interprétation pertinente serait celle qui s'écarte le moins du contexte, celle qui s'ancre dans le plus d'éléments présents, et qui s'avère productive (elle apporte quelque chose à l'interprète).

Ces propositions sont intéressantes en ce qu'elles permettent de qualifier et de ne pas traiter uniformément divers cheminements interprétatifs, en prenant réellement appui sur le contexte (texte, mais surtout situation de communication). Elles deviennent réductrices dans leur acception universalisante (le principe d'économie ne règle pas toutes les pratiques interprétatives, loin de là) et éliminatrice (plusieurs interprétations peuvent coexister, même si l'une domine).

Première critique : le régime de la clarté

L'interprétation, conçue comme l'obtention du sens du texte, se place sous le régime de la clarté. Le travail interprétatif se cantonnerait à résoudre des difficultés ponctuelles, le reste du sens du texte relevant de l'évidence.

Aussi le cas général serait tout simple : le sens est là, donné, immédiat. Il n'y a pas à proprement parler d'interprétation, sinon quand la formulation choisie dans la langue gêne la reconnaissance du sens. Pourtant, pour peu qu'on l'examine, la notion de *sens littéral* est problématique (Rastier 1994, §1.2.a, p. 328) : par quel procédé se présenterait-il, sinon lui aussi par une interprétation, qui le relativise ?

> On admet ordinairement qu'en règle générale un texte a un sens littéral [...].
>
> Quand, exceptionnellement, ce sens immédiat ne répond pas aux attentes ou, pire, quand aucun sens littéral n'est immédiatement saisissable, on a recours au sens caché [...]. On sauve ainsi la possibilité d'identifier un seul et véritable sens.
>
> Une sémantique générative, ou une sémiotique générative, partent inévitablement d'un sens *ab quo*, qui serait le contenu à transmettre ; et les adjonctions qu'il reçoit dans le parcours génératif sont réputées inessentielles.
>
> Conformément à la perspective interprétative adoptée ici, on ne considérera pas le sens comme un donné. On constate en revanche que la polysémie des signes, l'ambiguïté des phrases, la plurivocité des textes sont des phénomènes –peut-être fondamentaux– de la sémantique des langues naturelles.
>
> [...]
>
> Aucun sens n'est donné immédiatement ; même celui de l'énoncé le plus simple est le résultat d'un parcours interprétatif complexe. [En outre, les textes plurivoques] [...] ne sont aucunement des aberrations vicieuses ou des exceptions déviantes. Toute normativité écartée, ils appartiennent à l'objet de la linguistique, tout aussi bien que *Max sliced the salami with a knife*, dont ils ne diffèrent sémantiquement que par le degré de complexité des parcours interprétatifs qui leurs sont associés. (Rastier 1987, §VIII.5.1, pp. 210-211)

> L'immanentisme et le littéralisme [...] témoignent de deux gestes d'objectivation [...] [selon] deux voies complémentaires, générative ou 'interprétative', respectivement : soit en considérant que le sens a été déposé dans le texte par l'esprit et/ou le monde, et qu'il reflète leur cours [...]. Soit en estimant, conformément au postulat réaliste qui fait le fond de toute la tradition occidentale que les

[142] Elles sont par exemple présentées dans (Sabah 1988, §10.2) :
Maximes de quantité : l'intervention doit apporter suffisamment d'information ; elle ne doit pas apporter plus d'information que ce qui est nécessaire.
Maximes de qualité : ne rien dire que l'on croit faux ; ne rien dire que l'on ne puisse démontrer.
Maxime de relation : l'information donnée doit être pertinente.
Maximes de manière : éviter d'utiliser des expressions obscures ; éviter d'utiliser des expressions ambiguës ; être bref ; donner les informations dans le bon ordre.

> textes sont des représentations plus ou moins transparentes du monde ou de l'esprit. (Rastier 1994, §1.2.d, p. 331)

Deuxième critique : une unicité arbitraire

Postuler qu'un texte ne prend qu'un seul sens est une attitude éliminatrice : il semble plus juste, même quand un sens s'impose davantage que d'autres, de hiérarchiser les interprétations en présence, sans exclusion.

La conception du texte comme une richesse de sens, à laquelle concourt également le lecteur, plutôt que comme inscription inaltérable, dépositaire et garant d'une vérité[143], est relativement moderne.

> Le traitement des ambiguïtés, réelles ou prétendues, soulève le problème du choix entre les diverses interprétations envisageables. Or, en I.A. comme en linguistique, on a généralement pris le parti d'éliminer les interprétations jugées impropres, en postulant l'univocité du texte traité. Nous préférons une autre approche. Pour une sémantique interprétative, l'équivocité est une donnée fondamentale. En règle générale, on a affaire à plusieurs interprétations. Dans le meilleur des cas, on peut établir qu'une interprétation est préférable à toutes les autres. En d'autres termes, et bien que toute notre tradition herméneutique milite contre cette conclusion, le sens d'un texte n'est pas de l'ordre du vrai, mais du plausible Plutôt que de révoquer les interprétations jugées impropres, il convient donc de les hiérarchiser, en graduant leur plausibilité relativement à une stratégie donnée. (Rastier 1991, §V.4, p. 160)

d) Explicitation totale

Complétude et ajustement

Le texte est par nature, nécessairement incomplet. L'ampleur du sens qu'il peut prendre pour un lecteur ne se mesure pas à son nombre de pages, et ne se laisse pas enfermer dans l'ensemble de ses mots. S'inscrivant dans une pratique, il s'appuie sur des 'acquis' implicites : les notions courantes dans telles domaine, la référence à tel ouvrage ou telle personnalité, sans compter la manière même dont le parcourent les lecteurs.

La compréhension d'un texte a alors pu être considérée comme l'établissement du sens « complet », permettant une « représentation exacte » et entière de son contenu, et pouvant être testée par la capacité à répondre à des questions à propos de l'information constituée à partir du texte (Fuchs & *al.* 1993, §8). Le « sens littéral » du texte est prolongé par la prise en compte de la situation et des conditions d'interprétation, permettant certaines actualisations et certaines inférences.

Représenter l'implicite

Des conceptions de l'analyse automatique des textes et de l'intelligence artificielle illustrent deux manières opposées de viser la dissipation de l'implicite du texte.

La première est de partir de l'explicite que représente la matière du texte initial, puis, par inférences et déductions logiques réitérées, d'ajouter progressivement toutes les propositions supplémentaires qui peuvent être construites en s'appuyant sur les données explicites du texte ou les informations déjà acquises par cheminement logique[144]. La logique donne le cadre du mécanisme qui

---

[143] Le texte est ainsi lié aux grandes institutions (Barthes 1973), épine dorsale de l'organisation en société : religion, comptabilité (finances et commerce), administration (Etat et pouvoirs centralisateurs), droit (la loi, les contrats) (Goody 1986). Il est quelque peu sacralisé, pour reprendre les termes de (Rastier 1996b).
Dans cette logique, la raison d'être de la philologie est de scientifiquement *restituer* le texte, l'exactitude littérale de l'écrit assurant alors la préservation de son sens, canonique (Barthes 1973).

[144] On peut se rallier à cette conception de l'interprétation tout en y reconnaissant un artefact de la modélisation choisie. C'est le traitement qui séparerait en étapes successives d'abord l'obtention d'un sens explicite et non contextuel, puis son enrichissement et sa transformation à partir de connaissances supplémentaires.
« Nous supposerons tout d'abord l'existence du sens littéral d'un énoncé, construit à l'aide de connaissances générales, syntaxiques et sémantiques (c'est principalement sur cette existence que portent l'essentiel des discussions théoriques). Nous supposerons ensuite une étape postérieure à cette première « compréhension », l'*interprétation*, visant à expliciter divers éléments portés par l'énonciation réalisée. Nous distinguerons alors [...]

permet de calculer l'implicite qui prolonge le texte. Le passage d'éléments explicites à l'ajout d'un élément implicite repose sur des règles, qui sont clairement une limitation pratique de la mise en œuvre. Il n'est pas simple d'avoir un jeu de règles cohérentes, et la description des opérations possibles est toujours inachevée.

L'idée des scripts et scénarios est de disposer d'une modélisation complète d'enchaînements d'événements typiques, de sorte que lorsqu'un texte évoque un des enchaînements d'événements prévus, la machine soit capable de rétablir l'ensemble des chaînons manquants. L'exemple canonique est celui du repas au restaurant : si le texte fait une simple allusion à un repas au restaurant, alors la machine est en mesure de représenter les entités 'sous-entendues' et le déroulement 'par défaut' (l'installation à une table libre, la commande de plats, le paiement de l'addition, etc.). Ici, l'ajout d'informations ne se fait pas de façon progressive, et unité d'information par unité d'information, mais au contraire saisit d'un seul tenant tout un ensemble d'événements qui fait système.

Une quête sans limites

L'idée est d'épuiser le texte, en procédant à toutes les inférences autorisées, en explicitant tous les présupposés et les non-dits. Mais le texte ne 'contient' pas un sens qu'il délimite (le 'contenu', pris dans un 'contenant'), et la thésaurisation du sens est vaine et sans fin.

> « nous ne prétendons pas à l'exhaustivité, largement illusoire, même pour l'étude d'un texte bref. (Note : Son nom indique assez qu'elle est épuisante, pour l'auteur comme le lecteur. Le principe d'exhaustivité, énoncé par Hjelmslev, et repris par la traditions sémiotique qui s'en réclame, repose sur un immanentisme que nous récusons) ». (Rastier 1989, Introduction §C, p.10)
>
> « En somme, la notion de présupposition existentielle peut amener à conclure que toute phrase, voire tout mot, présuppose l'existence de tout l'univers. » (Rastier 1987, §IX.2.3.2, p.226)

Focalisation et pertinence

L'explicitation d'un implicite pourrait s'en tenir à ce qui semble requis par le contexte d'interprétation. Si beaucoup d'extensions de la représentation sont possibles, toutes ne sont pas également valables[145]. Une lecture n'est pas une réception uniforme, elle construit un sens en fonction de points d'attention, de lignes de force, attentes vis-à-vis du texte ou éléments suscités par lui. La démarche d'explicitation exhaustive est anti-naturelle, *machinale*.

La pertinence de cette démarche est également discutable pour notre application. Dans un système qui confronte des textes entre eux et calcule des rapprochements, il n'y a pas forcément à chercher à résoudre le non-dit. L'implicite est de degré d'évidence et de pertinence varié. Deux textes sont proches *également* parce qu'ils partagent le même non-dit. A l'inverse, un ouvrage très spécialisé, qui ne revient pas sur les notions élémentaires du domaine, et un ouvrage d'initiation, qui les présente de façon très complète, ne connaissent pas les mêmes usages, et ne s'adressent pas aux mêmes lecteurs. Bien qu'ils s'inscrivent dans un même domaine de connaissances, ils ne sont pas « proches » pour autant.

---

deux types d'interprétations : la première consiste à étudier comment l'information apportée par l'énoncé s'intègre dans les connaissances que l'on a sur le monde de référence. Il s'agira ici de réaliser des inférences en vue d'expliciter les connaissances communes aux interlocuteurs, connaissances qui restent implicites. La seconde interprétation consistera à étudier dans quelle mesure le contexte d'énonciation influe sur le sens de la phrase prononcée, et comment il convient de modifier le sens littéral pour « calculer » une signification dépendant de la situation. Encore une fois, il ne s'agit là ni d'un modèle linguistique ni d'un modèle psychologique du langage, mais d'options que les contraintes de modularité ont conduit les informaticiens à adopter. » (Sabah 1988, introduction à la troisième partie, p. 259-260).

Le schéma qui suit montre les deux processus d'interprétation, indépendants l'un de l'autre, qui partent du sens littéral : d'une part celui qui mobilise des *connaissances sur le monde*, et aboutit au *sens « complet »* ; d'autre part celui qui prend en paramètre la *situation d'énonciation*, et aboutit à la *signification*.

[145] (Sperber & Wilson 1986) font ainsi la distinction entre tout ce qui est présent à l'esprit d'une personne à un moment donné, et tout ce qu'elle serait capable de percevoir ou d'inférer (ce qui lui est *manifeste*). Ce qui est manifeste est ce qui est accessible, mais n'est à expliciter que si c'est opportun.

e) Double sens

Une orientation a priori

Les deux sens que recèlerait le texte sont en fait toujours pensés comme les deux pôles qui s'opposent sur un axe orienté. L'axe choisi reflète les préoccupations et l'échelle de valeurs de la discipline de l'interprète : s'agit-il d'un exégète ou d'un psychanalyste, la perspective n'est pas tout à fait identique. En revanche, la structure d'articulation en deux pôles se retrouve partout. Il est donc possible de dresser un petit tableau qui présente, dans ses lignes, les variantes de réalisation de la structure bi-pôlaire :

axe	-	+
distance, dérivation	surface	profond
secret (dévoilement)	apparent	caché
travail interprétatif	manifeste	latent
subjectivité	dénoté	connoté
abstraction	littéral	figuré, spirituel
chronologie	premier, ancien	second, nouveau
histoire religieuse (exégèse patristique)	judaïque	chrétien

L'interprétation, qui se réduit à identifier les deux sens du texte, rejoint en fait une interprétation univoque, car le sens valorisé est généralement retenu comme seul véritable sens (Rastier 1987, §VIII.5.1, p. 210).

Une herméneutique convaincue

Le pôle positif, valorisé comme aboutissement de l'interprétation, est ce qui correspond à la norme que l'on s'est donnée et que l'on veut trouver dans le texte : grammaticalité en syntaxe, principe de charité (qui crédite au texte un sens édificateur) et théologie pour l'exégèse, interprétation libidinale, etc. (Rastier 1989, §I.2.I, p.22). L'interprétation n'est pas à l'écoute du texte. La signification ultime (voulue) est connue par avance, et la tâche interprétative est d'expliciter les règles ou normes qui justifient l'attribution de ce sens (Rastier 1987, §IX.4.1, pp. 247-250).

Des principes aux conditions linguistiques

La sémantique se refuse à ce jeu de l'herméneutique, et remplace l'interprétation fermée (le sens est fixé à l'avance) par une interprétation ouverte.

> L'image suggérée par l'étymologie même du mot « texte » [est celle d'] un *tissu* ; mais alors que précédemment la critique (seule forme connue en France d'une théorie de la littérature) mettait unanimement l'accent sur le « tissu » fini (le texte étant un « voile » derrière lequel il fallait aller chercher la vérité, le message réel, bref le *sens*), la théorie actuelle du texte se détourne du texte-voile et cherche à percevoir le tissu dans sa texture, dans l'entrelacs des codes, des formules, des signifiants, au sein duquel le sujet se place et se défait (Barthes 1973).

> une herméneutique sait toujours quel sens elle doit trouver. [...] [Le] sens caché n'est pas découvert mais retrouvé dans l'interprétation, car il se présentait d'emblée, épiphanique, dans le moment antérieur de la compréhension. [...]
>
> La sémantique textuelle demeure en deçà de toute herméneutique. Elle définit les conditions linguistiques de l'interpétation. Elle peut décrire des interprétations et les évaluer relativement à ces conditions, mais elle ne produit pas à strictement parler d'interprétation. En bref, elle ne recherche pas un ou plusieurs sens cachés ; dans le cas d'une pluralité de sens, elle décrit leur accessibilité relative, et évalue leur degré de plausibilité ; et surtout, elle ne sait pas quel(s) type(s) de sens elle doit trouver. [...] le texte apparaît comme une série de contraintes qui dessinent des parcours interprétatifs. (Rastier 1989, §I.1.C, p.18)

f) Plusieurs sens formant système

Ce cas s'apparente au précédent, sauf que le système n'oppose plus deux pôles sur un axe, mais prévoit *n* pôles. De la même manière, l'interprétation ne fait qu'établir les chemins qui relient le texte à chacun de ces pôles, déterminés à l'avance.

L'exégèse chrétienne médiévale illustre ce type d'interprétation multi-pôlaire, qui n'apparaît en fait que comme un raffinement de la théorie du double-sens :

> cette distinction devenue traditionnelle entre sens littéral (dit aussi *historique* ou *somatique*) et sens spirituel (dit aussi *allégorique* ou *figuré*) a été élaborée par la suite pour donner lieu à la théorie médiévale des quatre sens. Thomas d'Aquin la résume ainsi : « La première signification, à savoir celle par laquelle les mots employés expriment certaines choses, correspond au sens premier, qui est le sens *historique* ou *littéral*. La signification seconde, par laquelle les choses exprimées par les mots signifient, de nouveau, d'autres choses, c'est ce qu'on appelle le sens *spirituel*, qui se fonde ainsi sur le premier et le présuppose. A son tour, le sens *spirituel* se divise en trois sens distincts. En effet l'Apôtre dit : « la loi ancienne est une figure de la loi à venir » ; enfin, dans la nouvelle loi, ce qui a eu lieu dans le Chef est le signe de ce que nous-mêmes nous devons faire. Quand donc les choses de l'ancienne loi signifient celles de la loi nouvelle, on a le sens *allégorique* ; quand les choses réalisées dans le Christ ou concernant les figures du Christ sont les signes de ce que nous devons faire, on a le sens *moral* ; enfin si l'on considère que ces mêmes choses signifient ce qui est de l'éternelle gloire, on a le sens *anagogique* » (*Somme théologique*, question I, article 10, conclusion).
>
> (Rastier 1987, §VIII.1.1.1.B, p. 168)

Postuler un sens, ou deux, ou une série plus complexe mais tout autant limitée, c'est fermer l'interprétation : or le travail interprétatif, même en suivant la « trace » du texte, peut toujours aller plus loin. On n'a jamais tout dit, ni donc définitivement cerné ce qui fait sens à partir du texte.

> la critique cherche en général à découvrir le *sens* de l'œuvre, sens plus ou moins caché et qui est assigné à des niveaux divers, selon les critiques ; l'analyse textuelle récuse l'idée d'un signifié dernier : l'œuvre ne s'arrête pas, ne se ferme pas (Barthes 1973)

g) Equivocité et indétermination

Une conception non extrémiste

Il y a un point d'équilibre à trouver, pour rendre compte à la fois de la compulsivité de l'interprétation (on ne peut s'empêcher de donner du sens, même si l'expression linguistique dévie de l'usage –agrammaticalité, néologismes, etc.), et de la notion d'absurde (tout sens n'est pas également recevable, et on ne peut admettre de faire dire à un texte n'importe quoi).

Les lignes directrices ne sont pas dans des a priori...

Le sens d'un texte n'est pas déterminé *a priori* : s'il y a détermination, c'est celle d'*un* sens (sans exclure d'autres possibles) pour un lecteur, à un moment donné.

> On ne peut [...] attendre d'une sémantique interprétative qu'elle énonce *le* sens qui constituerait la vérité du texte. Ce serait là répéter l'erreur de la philologie, quand elle postulait qu'un texte a un et un seul sens. D'une part il n'existe pas *a priori* de Sens unique et ultime ; et de plus les sens d'un texte ne doivent pas être considérés comme immanents : nous souhaitons avoir montré que tout sens, et même tout sème, était le produit d'opérations interprétatives qui l'actualisent. (Rastier 1987, §IX.4.3.5, p. 263)

En particulier, ne sont fixés à l'avance ni le nombre de sens, ni leur organisation d'ensemble.

> à des théories des deux sens hiérarchisés *a priori*, nous souhaitons substituer une théorie des isotopies multiples non hiérarchisées *a priori*. (Rastier 1987, §VIII.1.1.3, p. 175)

Les différents sens qui coexistent sont évalués *a posteriori*. La hiérarchisation des isotopies procède d'un travail interprétatif, faisant intervenir le contexte, et de normes qui ne sont pas universelles (Rastier 1987, §VIII.4.3.1, pp. 202-203).

...les contraintes linguistiques fournissent des lignes directrices

La linguistique a son mot à dire dans l'interprétation, en ce que l'expression dans la langue donne des points d'appui et des contours dans lesquels ancrer l'interprétation. Celle-ci n'est pas purement détachée du texte, et ne peut ignorer sa matière linguistique.

On peut donner un exemple de telles contraintes pour l'établissement d'une isotopie, c'est-à-dire en quelque sorte d'un thème :

> Une première recommandation, formulée jadis (l'auteur, 1972, p.93), conseille, avant d'établir une isotopie générique, d'« identifier au moins un sémème appartenant sans équivoque » au domaine sémantique considéré, c'est-à-dire pourvu d'un sème générique inhérent, actualisé en contexte, et qui l'indexe dans ce domaine. (Rastier 1987, §IX.3.3, p. 240)

Autrement dit, l'expression linguistique ne contient (fixe) pas le sens, mais elle participe, en posant certains points de repères et certaines limites, à la construction d'un sens. C'est ainsi que le texte fonctionne pour faire mémoire, c'est ainsi qu'il retient et transmet :

> On peut [...] distinguer deux grands types de techniques : les techniques qui pro-gramment le geste, les techniques qui pro-gramment la reformulation. Dans le premier cas, la structure matérielle des outils conditionne le geste qui se saisit de l'outil. L'outil est une mémoire qui mémorise dans sa structure le geste à accomplir [...]. Dans le second cas, le but de l'outil n'est pas de commander le geste, mais de mémoriser une parole, un savoir, pour le transmettre et le diffuser. L'outil ne commande pas le geste, mais la parole ou la réécriture. L'apparition de la technologie de l'écriture correspond ainsi à l'émergence de ces techniques de la mémoire. [...]
>
> Puisque la mémorisation par la technique est d'emblée une restitution, elle comprend nécessairement une sélection. On ne peut en effet se rappeler de tout, ni tout mémoriser [en raison de la finitude rétentionnelle] [...]. Puisque l'enregistrement s'effectue en fonction d'une restitution future qu'il prescrit, la mémorisation par le technique sélectionne en fonction de l'usage qu'elle prescrit. Ainsi, on écrit pour être lu, c'est-à-dire pour être réécrit [car toute interprétation est une forme de réécriture]. Au lieu de chercher à s'exprimer, on cherche à contraindre la réécriture du lecteur. L'instrument technique sélectionne pour l'usage qu'il contraint en même temps qu'il constitue : une machine, un marteau, permettent de nouveaux gestes tout en contraignant la manière de les effectuer. Le marteau mémorise une certaine manière de frapper, de frapper « comme un marteau »
>
> (Bachimont 1999b, pp. 2-3 et 23)

h) Multiplicité artificielle

Combinatoire artéfactuelle

Les systèmes de traitement automatique de la langue qui cherchent à produire une représentation sémantique se heurtent souvent à une multiplication artificielle du sens. En effet, plusieurs significations peuvent être attachées à chaque unité lexicale (polysémie), et la combinaison des unités dans le texte se traduit par une combinatoire, plus ou moins contrôlée, des significations unitaires.

Ce type d'ambiguïté est clairement artificiel : la machine dénombre des dizaines de sens, parmi lesquels elle ne sait choisir, là où le lecteur humain n'en perçoit spontanément qu'un seul. Un premier diagnostic montre qu'une bonne part de ces ambiguïtés disparaîtraient avec une meilleure prise en compte du contexte (Rastier 1989, §I.1.B, p. 16).

Droit à l'existence d'un sens non fixable

Les systèmes rencontrent également une multiplication artificielle des sens quand il s'agit de choisir entre des alternatives de sens précises, là où le texte ne permet pas de trancher ou s'en tient à des considérations plus générales. L'interprétation n'a pas le devoir d'en dire plus que le texte, d'évoquer à la place du texte les prolongements que lui-même n'a pas tracés. Il peut tout à fait faire partie de la stratégie de l'auteur de maintenir une indétermination qui laisse ouvert un jeu de perspectives.

> Ces prétendues ambiguïtés reposent sur l'illusion édénique que chaque phrase pourvue de sens décrit complètement une partie de la réalité. Non seulement le sens est identifié à une désignation, mais encore à une désignation exhaustive.
>
> (Rastier 1987, §IX.2.3.2.C, p. 226)

i) Infinité

Le déconstructionnisme illustre cette conception de l'interprétation. Ici, tout peut être dit à partir de tout, et le texte n'oppose aucune limite aux sens dont on l'affuble.

Outre que cette approche tourne court pour apporter des éléments à un traitement automatique, elle contrevient à l'expérience commune de lecteur de tout un chacun. Dire que le texte se pourvoit d'une infinité de sens, sans restriction, c'est admettre qu'il n'a aucune incidence sur la construction du sens. Le sens ne serait que du côté du lecteur, voire dans la situation. C'est finalement dénier que la lecture puisse être l'expérience d'une rencontre d'une personne avec l'expression d'une autre personne, à travers la médiation du texte, et l'ouverture sur une altérité enrichissante. Qu'elle soit évasion, mise en question, information,... la lecture est une invitation à une réalité autre que soi.

> Quand on a convenu que les faits sémantiques, comme les autres, sont construits, s'ouvrent alors les voies d'un faux dilemme [...] : ou bien le récepteur découvre par les procédures appropriées le sens immanent au texte ; ou bien il le constitue, et ce sens éclate en une pluralité indéfinie, celle des lecteurs.
>
> La première thèse a été soutenue par un courant structuraliste : en appliquant au *texte seul* (isolé de son entour linguistique et « pragmatique ») des procédures universelles de décodage, un lecteur quelconque, armé de la bonne méthode, pouvait mettre en évidence son sens. [...] L'immanentisme en la matière est issu d'une longue tradition, antérieure à tout projet de description scientifique du sens, celle de l'herméneutique religieuse, fondée sur la révélation. Le sens serait immanent au texte parce qu'il y a été déposé -par Dieu ou par un homme, qu'importe. D'où les stratégies de *dévoilement*, de *mise en évidence*, etc.
>
> Une autre façon de méconnaître le type d'objectivité du sens (tout aussi unilatérale que la précédente, mais pour des raisons opposées) consiste à postuler la pluralité indéfinie du sens, situé alors dans le sujet dont l'inconscient, structuré comme un langage, parle au lieu du texte. Le sens devient alors transcendant au texte [...].
>
> On aurait pu espérer que cette théorie « désirante » du sens, à défaut de décrire les textes, produise au moins une agréable variété de lectures. Il n'en a rien été, car les lectures « pulsionnelles » n'ont vu dans les textes que les drames les plus œudipiens, les symboles les plus lourdement sexualisés. [...]
>
> Affirmer l'objectivité sans nuances du sens, ou sa subjectivité absolue, cela ne résout rien.
>
> (Rastier 1989, §I.1.A, pp. 13-15)

E. LA QUESTION DE LA PERTINENCE

1. Les expressions de la pertinence : examen des modèles rencontrés dans les applications documentaires

a) Pertinence binaire

Le modèle de la pertinence mis en œuvre dans le courant de l'*information retrieval* tient de la pertinence binaire, en ce qui concerne les méthodes d'évaluation. En effet, pour une requête donnée, chaque document reçoit l'appréciation *pertinent* ou bien *non pertinent*. La pertinence est donc représentée par une valeur parmi deux possibles, l'une positive, l'autre négative.

C'est un peu l'idée que l'utilisateur à la recherche d'informations, à l'issue de sa recherche, a finalement séparé l'ensemble des documents présentés en deux catégories : ceux qui ne l'intéressent pas et qu'il laisse, et ceux qu'il décide d'emmener avec lui pour s'en servir. On conviendra qu'il y a bien ce choix (emmener *vs* laisser), mais que l'interprétation sous-jacente relève d'une réalité plus complexe. Autrement dit, il serait simpliste de dire que les documents emportés correspondent exactement aux documents intéressants pour toute personne qui aurait formulé une requête identique. Tel document n'est pas retenu, bien que au cœur du sujet, parce que l'utilisateur le connaît déjà. Tel autre est sélectionné, mais il y avait un autre document équivalent, qui aurait tout aussi bien pu le remplacer. Tel autre encore n'est pas vraiment dans le sujet tel qu'il a été exprimé, mais s'est présenté au détour d'un rayon et correspond à une extension de la problématique qui concerne l'utilisateur. Etc.

Dans une curieuse étude, (Janes 1993) fait apparaître que, dans de multiples expériences, les personnes, à qui l'on a demandé d'évaluer la pertinence de documents sur une échelle de plusieurs valeurs, tendent à choisir préférentiellement les valeurs extrêmes. Ceci va à l'encontre du sens commun : on s'attendrait à ce que le jugement soit nuancé, et que les valeurs limites soient réservées à l'expression d'exceptions. En fait, il semble que les personnes visent à écarter ou à sélectionner, de façon décisive, les documents présentés, et que l'utilisation des valeurs intermédiaires est réservé aux documents pour lesquels on a l'impression de manquer d'informations pour juger (trancher). Autrement dit, ce qui reste entre les deux extrêmes est ce qui a du mal a être évalué, ce qui prendrait du temps à préciser, bref des « cas » difficiles de jugement de pertinence, voire des échecs (momentanés), enfin un résidu indécis et inconfortable que l'on souhaite minimal. Mais Janes s'interroge lui-même : cette attitude n'est-elle pas un artefact des conditions expérimentales relativement artificielles, où l'on demande à des personnes de juger la pertinence de documents, de s'exprimer à travers une échelle de pertinence, sans qu'elles soient elles-mêmes véritablement engagées dans une situation réelle et personnelle de besoin d'informations et de recherche bibliographique, dans laquelle leur attitude pourrait prendre d'autres formes d'expression ?

> We propose the following hypothesis. It could be that indeed *relevance* is, in part, [a] bipolar [concept]. It has been recognized for decades that decisions about really good and really bad documents are easy to make and that middling documents are less clear. We have found that people appear to judge that way, too : they see lots of bad documents, a few good ones, and a fair number scattered in between. Perhaps the process people go through is a two- (or multiple-) stage one :
>
> *(1)* Determine, very quickly, if the document is really good or really bad. If so, say so and the data appears to show that they don't much care exactly how really good or bad it is).
>
> *(2)* If not, then more time and effort must be taken to determine how much of it is good, whether or not it is from a trustworthy source, addresses the right issues, is in the right language, is available and accessible, etc.
>
> The first of these proceses is quick, relatively easy, and is done with confidennce. The second is slower, less certain, and done with more difficulty. [...]
>
> However, this could [...] be artifactual. People have made these decisions and producted these judgments *because they have been asked to do so as part of a research study*. We have no garantee whatsoever that this is reflective of what people really do when evaluating information in response to

> their information needs. Do they worry about how good documents are, or do they just dig in and find what's good in each one (if anything) and get on with whatever work they are doing ?
>
> (Janes 1993, p. 113)

b) Pertinence n-aire

La pertinence, de la forme tranchée pertinent / non pertinent, peut être déclinée en appréciations progressives : *a priori* très pertinent, assez pertinent, peu pertinent, hors sujet...

Sans rentrer dans une telle grille, somme toute encre assez rigide, le système SPIRIT, dans ses versions récentes (Fluhr 1994) (SPIRIT-W3), propose une présentation intéressante des résultats d'une recherche sur une base documentaire. Ce qui l'apparente à une pertinence *n*-aire, c'est le fait que les documents sélectionnés soient regroupés en classes, et que ces classes sont elles-mêmes présentées par ordre de pertinence présumée décroissante.[146]

Le principe est le suivant : la requête consiste habituellement en quelques mots. La première classe est formée par les documents présentant tous les mots de la requête. La deuxième, par les documents présentant tous les mots sauf un (le moins « significatif » au sens d'indicateurs statistiques). Pour la troisième classe, les documents ont encore tous les mots sauf un, mais cette fois-ci le terme manquant était un peu plus important. Et ainsi de suite, jusqu'aux classes correspondant aux documents retournés en raison d'un seul terme.[147]

Ce mode de sélection et de présentation est connu sous le nom de *quorum-level search* (Salton 1988). Ses limites les plus sensibles sont sa dépendance à la forme des requêtes, et la difficulté à maîtriser le volume des résultats.

En ce qui concerne la dépendance à la forme des requêtes, il est évident que cette heuristique, bien adaptée à un très petit nombre de termes, engendre une combinatoire peu gérable dès que la requête comporte cinq six termes ou plus. Le nombre de classes est alors démultiplié, et leur ordre est peu intuitif. Par exemple, si, en connaissant le fonctionnement du système, on sait que la classe des documents qui comporte tous les termes est en tête, et que l'on peut deviner que telle combinaison de termes, où il manque un terme subsidiaire (i.e. d'usage assez général dans le domaine de la base), sera vraisemblablement au deuxième ou troisième rang, que dire d'une certaine classe qui comporte la moitié des termes : à quel rang la trouver ?

Quand bien même on disposerait d'un index sophistiqué, qui permettrait de se positionner immédiatement sur la classe de résultats correspondants à une certaine combinaison de termes présents, l'exploration des résultats reste peu intuitive. Les documents sont dispersés dans un grand nombre de classes, et des propositions qui en définitive seraient globalement assez proches sont assignées à des classes séparées, uniquement en raison de petites variations de vocabulaire. Dès que le nombre de propositions est assez grand, le dépouillement est particulièrement austère : il faut examiner chaque classe l'une après l'autre, sans savoir à l'avance quelles combinaisons de termes se sont avérées les plus efficaces.

Il faut cependant retenir de cette représentation de la pertinence son expression relative à la requête soumise, dont on retrouve les termes[148]. C'est reconnaître qu'il y a toute une variété de mises en correspondances entre la requête et un document, et aussi qu'il ne revient pas à la machine de trancher, en éliminant d'emblée les documents qui ne sont sélectionnés que sur les indices

---

[146] On retrouve encore un tel modèle de pertinence chez (Denos 1997), qui introduit simplement en plus la distinction entre des mots (ou critères) obligatoires, qui doivent se trouver dans les documents de toutes les classes, et les mots optionnels, dont la combinatoire de réalisation induit les classes de présentation des résultats (*ibid*., §III.2.5, p. 88).

[147] Cette présentation est légèrement simplifiée : SPIRIT fait aussi intervenir la proximité des termes. Ainsi, lorsque des termes forment un syntagme dans la requête et sont retrouvés comme tels dans des documents, ces documents forment une classe distincte, qui obtient un meilleur rang que celle des documents où figurent les même termes, mais sans la relation syntagmatique.

[148] « Les classes de pertinence organisent l'ensemble des documents retrouvés en fonction du schéma de pertinence [*i.e.* la requête]. Elles fournissent à l'utilisateur une vue du corpus structurée en fonction du shéma de pertinence qu'il a formulé. Ainsi l'interface constitue le lieu de la confrontation entre le sens que l'utilisateur veut exprimer (son schéma individuel de pertinence) et le sens que le système est capable de produire à partir de l'expression du schéma de pertinence. » (Denos 1997, §I.2.2.1, p. 32)

apparemment les plus faibles. Enfin, l'organisation par regroupement, qui rencontre ses limites lorsqu'il faut parcourir séquentiellement un grand nombre de classes, contribue à une certaine efficacité du parcours : en effet, une sélection sur un motif lexical inadéquat peut être « sautée » dans son ensemble. Au raisonnement document par document se substitue dans certains cas un raisonnement collectif, groupe de documents par groupe de documents.

c) Pertinence linéaire

Avec la pertinence linéaire, on revient du côté des systèmes de recherche d'information, mais cette fois-ci des sorties « brutes », avant alignement sur les formats des campagnes d'évaluation.

Les techniques mis en oeuvres, qu'elles soient vectorielles ou probabilistes, fournissent une valeur numérique pour caractériser la relation entre une requête et un document. En général, il s'agit d'un réel positif, quelquefois plafonné (normé à 1 ou à 100 par exemple), et qui est une fonction croissante de l'adéquation du document à la requête. (La variation peut être inverse, si la valeur traduit non pas une similarité ou proximité, mais une distance).

L'image est spatiale : il y a ce qui est proche, et ce qui est moins proche, du thème d'intérêt. Cela se conçoit bien, si ce n'est l'organisation unidimensionnelle de l'ensemble. En effet, on pressent bien que deux documents peuvent être assez proches de la requête, sans l'être de la même façon. Et que les proposer *ex æquo*, ou préférer l'un par rapport à l'autre, n'a pas vraiment de sens. Ils sont complémentaires, et non en compétition pour se classer au meilleur *rang*. Si je me rends à la bibliothèque, les documents que je peux choisir ne m'apparaissent pas un à un, indépendamment et l'un après l'autre ; il y a plutôt des configurations, des regroupement et des oppositions, des alternatives et des complémentarités, qui me font repartir avec un ensemble de documents dont l'intérêt n'est pas forcément hiérarchisable, surtout de façon univoque.

Pour autant, les utilisateurs réclament *a priori* ce type de représentation. La mise en ordre des résultats, par valeur décroissante d'un indice de pertinence, est effectivement ce qui est affiché comme une fonction évoluée, par la plupart des applications documentaires. Les utilisateurs seraient donc convaincus qu'il s'agit là de ce qu'il y a de mieux, en matière de présentation des résultats. On peut penser aussi que les valeurs numériques (de l'indice de pertinence) rassurent, et sont perçues de façon très positive dans une culture où l'exactitude de la mesure chiffrée et le caractère scientifique sont assimilés et valorisés.

Concrètement, la diffusion ciblée a expérimenté ce type de pertinence. Nous en avons évoqué les inconvénients pratique (cf. introduction). Premièrement, cette présentation oblige à examiner les documents un à un. Si un terme inapproprié génère des rapprochements incongrus, la gêne occasionnée est massive, car il n'y a aucun moyen de mettre de côté, en une seule opération, les documents correspondants à cette anomalie : l'erreur est « diffuse ». Deuxièmement, autant l'organisation linéaire est parfaite pour un parcours systématique (on est sûr d'avoir tout examiné), autant le continuum des valeurs de proximité rend difficile la construction d'un parcours complet. Soit l'on parcourt toute la liste : cela risque d'être très long, pour une proportion de documents trop éloignés du sujet de plus en plus grande –soit une efficacité quasi nulle sur la fin de la recherche. Soit on choisit de s'arrêter à un point. Cela revient dans ce contexte à choisir un nombre de documents à examiner (mais il est difficile d'évaluer avec assurance l'ordre de grandeur du nombre de documents intéressants), ou bien à choisir une valeur seuil du score de similarité. C'est là que la monodimensionalité, qui est un artefact de ce modèle, est gênante. En effet, si tel aspect devient inintéressant approximativement à telle valeur du score, tel autre aspect, moins bien noté, continue à sélectionner des documents que l'utilisateur juge intéressants mais qui obtiennent des valeurs de similarité plus basses. Si bien que sur une même échelle sont confondus et mêlés des rapprochements selon des perspectives différentes, et fixer un seuil qui départage documents pertinents et documents non pertinents est impossible.

La pertinence linéaire est fille d'une réalisation numérique. Si elle rend compte, de façon synthétique, du comportement du système, elle est en fait étrangère à la notion humaine de pertinence, qui ne se laisse pas réduire à un alignement ordonné.

d) Pertinence différentielle

A vrai dire, nous n'avons pas connaissance de systèmes documentaires qui adopteraient cette conception de la pertinence, sinon ce que nous avons conçu dans le cadre de DECID.

La pertinence différentielle fonctionne par regroupements et oppositions, et traduit ainsi les interrelations entre documents proposés en résultat de la recherche. Les regroupements traduisent les familles, qui peuvent être traitées collectivement (notamment pour mettre de côté des documents tous sélectionnés sur un aspect qui n'intéresse pas l'utilisateur). Les oppositions servent à contraster les documents les uns par rapport aux autres, pour mettre en valeur la singularité de chacun dans le contexte de cette requête et de ce fonds documentaire. C'est en effet une combinaison de jugements d'équivalence et de caractérisations spécifiques qui construisent le résultat effectif de la recherche, et le choix motivé d'un ensemble de documents.

La forme que cela prend dans DECID est exposée en détail plus loin (chapitre consacré à l'interface et au parcours des résultats). Il s'agit d'une organisation arborescente, qui permet de descendre progressivement dans les niveaux de détail, et donc à chaque étape de se repérer par rapport à un nombre de propositions raisonnable. Le premier niveau présenté répartit les documents par domaine (cela peut correspondre à différentes disciplines, qui ont chacune de leur manière affaire au thème de recherche) : l'utilisateur a ainsi une vue globale (grossière mais complète) de l'ensemble des résultats. Il peut tout de suite écarter certaines branches et se focaliser sur les pistes les plus prometteuses. Au niveau de chaque piste, les différences internes font ressortir les intérêts propres des différents types de propositions possibles. Cette organisation respecte donc tout à fait la multiplicité de points de vue auxquels se prête un sujet, sans s'obliger à hiérarchiser ce qui n'est pas en soi comparable.

L'utilisateur - interprète des résultats, construit son propre parcours, son propre cheminement, en se repérant par rapport à cette organisation pistes / originalités. Autrement dit, il est impliqué et actif dans l'établissement de la pertinence, celle-ci n'est pas fixée pour lui, à sa place. En jouant sur les mots, à la suite de (Bachimont 1999a), la pertinence différentielle est également une pertinence différ*a*ntielle, à savoir qui diffère « le » sens (ultime), chaque interprétation étant force de proposition d'une vision nouvelle et personnelle sur un texte (ici sur les résultats, la sélection opérée par le calcul).

Au sens strict, la pertinence différentielle se traduit par une structure de partition (l'ensemble des documents est entièrement organisé en familles), ou une classification hiérarchique (système de classes emboîtées). Une forme plus souple est mise en œuvre dans DECID.

e) Pertinence polaire

La pertinence polaire correspond à une représentation spatiale, dans laquelle peuvent se dessiner plusieurs pôles d'attraction significatifs. Les documents se concentrent au niveau des différents pôles. Certains peuvent se positionner de façon intermédiaire, comme sous l'influence de différents pôles. Selon leur proximité relative à ces pôles traduit leur « attirance », potentiellement inégale.[149]

Selon les représentations, les pôles sont de nature diverse : pôles ponctuels, axes, zones. Le cas général est le choix d'une variété géométrique, qui s'adapte à la forme des résultats et reflète les regroupements « naturels » qu'opère l'interprétation.

La pertinence polaire ressemble à la pertinence différentielle en ce qu'elle donne une représentation globale, et multidimensionnelle (il y a plusieurs manières d'être proche de la

[149] Voir par exemple l'interface, assez spectaculaire, de *Websom*, qui fournit des cartes de documents avec des zones colorées décrivant comme des courbes de niveaux. Le site Internet de *Websom* a le mérite de fournir, outre une illustration / démonstration de l'interface, une documentation scientifique abondante explicant les principes de construction de ces cartes :
`http://websom.hut.fi/websom`
L'idée de cartographie d'un espace de documents suscite par ailleurs des techniques de calcul et des représentations diversifiées : (Appel 1991) et (Lelu & François 1992), (Chalmers 1993), (Zizi 1995).

requête)[150]. Elle s'en écarte par son caractère continu[151]. La pertinence différentielle délimite nettement des regroupements, les emboîte en un nombre limité de niveaux. La pertinence polaire permet tous les intermédiaires, toutes les positions singulières. Cela peut donner l'impression d'une plus grande fidélité à la réalité, infiniment nuancée. Cependant, sous cette forme, il n'y a plus de moyen simple et systématique de parcourir les résultats. Et, sans nier le fait que chaque texte et chaque document soit unique, la pertinence n'est peut-être pas tant la perception de cette irréductibilité d'un texte à un autre, que celle d'une structure qui, en résumant l'ensemble des documents embrassés, rend intelligible leur type d'adéquation au thème de recherche.

2. Etude pour la diffusion ciblée

a) Paramètres des choix de lecture professionnelle : qui lit quoi

La question est ici celle des choix de lecture, et non du comment la lecture est effectuée. Le mode de pertinence accordée au document oriente bien sûr l'adoption de telle ou telle stratégie de lecture, sans la déterminer entièrement. Ainsi, tel document que l'on met à son programme de lecture « pour information » sera, selon le contexte, lu attentivement, parcouru en diagonale, ou sédimenté dans une de ces fameuses piles étudiées par les experts du travail de bureau.

Les publications sur la pertinence dans le domaine des systèmes documentaires ont une insistance toute particulière sur le fait que la thématique, considérée seule, ne suffit pas à établir la pertinence. Autrement dit, ce n'est pas parce que le sujet d'un document concorde avec le thème de la recherche que le dit document répond aux attentes du chercheur. De multiples facteurs se combinent, dont on s'efforce de percevoir la trace et l'incidence, pour ne pas s'en tenir à une conception par trop simpliste et réductrice de l'accord requête - document.

> The Cranfield tests have, with some variations, been the primary model of experimental Information Retrieval research during the past twenty-five years. The research design of the Cranfield project provides a conceptual framework for reexamining the underlying assumptions of the traditional Information Retrieval model and the implications of using « relevance judgments » in Information Retrieval evaluation. Cranfield's experimental design involved four major steps : *(1)* building a test collection, *(2)* gathering users' questions, *(3)* obtaining relevance judgments, and *(4)* conducting tests of retrieval. [...]
>
> Examining the underlying dimensions of relevance in Information Retrieval experimentation, the Cranfield model views relevance as « on the topic », the relationship between the topic of a question and that of a document. The nature of the relationship between a docuument and a user's question is very precise and fixed. There are no concerns with the individuality of users. The model also assumes that users' information needs are conceptually well defined and that they know how to express them. [...]
>
> [Actually,] relevance is not a simple relationship between a document retrieved and a user's question but, rather, is psychological and contextual, involving an individual's cognitive states, perceptions, experiences, and knowledge about the problem at hand. It goes much deeper than simple topical relevance. [...] Other aspects of user-based relevance demonstrate its nature as

[150] Mais d'un point de vue strictement géométrique, la représentation plane n'exprime jamais que deux dimensions (2D) ; une animation de l'image (rotations, translations) –dont il faut veiller au coût en termes de puissance de calcul, de flux de transmission de données et d'espace mémoire– permet de simuler un déploiement spatial, en trois dimensions (3D). Or la pertinence, qu'il s'agit ici de représenter, ne se laisse pas caractériser par deux ou trois facteurs. La visualisation (2D ou 3D) impose de fait une réduction contingente, et facilement illusoire. L'habileté d'un calcul de représentation plane tient notamment à sa capacité à minimiser les (inévitables) déformations, de sorte à ce que l'interprétation de l'utilisateur ait le moins de risques de s'égarer.
Sur les manières optimales de projeter des données sur un espace plan, voir principalement les travaux en analyse factorielle. La couverture de (Fénelon 1981) résume le principe de l'approche en présentant côte à côte l'ombre chinoise d'un dromadaire vu de face, et l'ombre chinoise du même dromadaire de profil : l'analyse factorielle choisit le second. (Wismath, Soong, Akl 1981) proposent une approche originale et plus légère, la *triangulation* : les points sont placés un à un, de telle sorte que ses distances à deux autres points soient conservées. On respecte donc au total, pour n points, $(2n - 3)$ distances.

[151] La réflexion sur la signification et l'adéquation d'une représentation continue serait à poursuivre et approfondir, par exemple à partir de (Salanskis 1996).

> multidimensional and dynamic [...]. A user's information need state may be changed as he or she encounters relevant citations. [...]
>
> The individual and interpretative nature of user-based relevance [vs simple topical relevance] also demonstrates a serious problem in the interpretation of relevance judgments that are made by others (for example, subject experts or search intermediaries) than the actual users themselves.
>
> (Park 1993, pp. 322-323, 344, 346)

Les considérations qui suivent visent donc à rassembler, décrire et organiser un éventail aussi complet que possible des facteurs qui entrent en ligne de compte. C'est le fruit d'une synthèse des résultats publiés par les spécialistes de la pertinence dans le domaine de l'*information retrieval*[152], d'observations lors d'enquêtes réalisées à EDF sur les besoins et utilisations de l'information[153], et de la réflexion et de l'expérience personnelle de l'auteur de ces lignes. Ceci ne saurait donc être tenu pour un aboutissement, mais comme une proposition de cadre, pour orienter dès à présent les développements des applications documentaires, et si possible servir de base à une étude plus systématique et à une validation expérimentale.

Le lecteur en tant qu'individu

Sa personnalité

Chaque personne a un rapport à l'écrit et à la lecture qui est un trait de son caractère. Tout le monde n'est pas un grand lecteur, ou un lecteur rapide, un lecteur naturellement assidu[154] ou bien épisodique, un lecteur épanoui ou bien qui n'arrive pas à faire face à ce qu'il veut ou doit lire. Les bibliothécaires, vis-à-vis de leurs habitués, tiennent spontanément compte de ce facteur pour réguler le « volume » de leurs propositions.

Dans un contexte de recherche d'informations, la patience du chercheur est mise à l'épreuve : pour évaluer une première série de propositions, ajuster ses critères de recherche, dépouiller à nouveau des suggestions en grand nombre, etc. Il y a aussi, en interaction avec d'autres facteurs, une attitude plus ou moins tolérante, ou inversement, exigeante : des documents ne correspondant pas tout à fait aux attentes conviennent ; une recherche est poussée jusqu'à ses limites avant de se décider à la clore.

Son histoire de lecteur

La pertinence d'un document n'est pas perçue de la même façon si le document est déjà connu du lecteur, ou non. Si le document lui est connu, sa vision est encore influencée par : l'idée claire et précise qu'il a du document (entre avoir lu de façon approfondie un document vs en avoir entendu vaguement parler), l'impression qu'il en a gardé (favorable ou défavorable).

Le document lui-même peut être inconnu du lecteur sans lui être totalement étranger. Le lecteur n'a pas la même attitude vis-à-vis d'un document sans lien visible avec ses lectures antérieures, et un document d'un auteur qu'il connaît, ou publié dans une revue qu'il fréquente ou une collection qu'il affectionne, ou diffusé par un éditeur qu'il méprise...

Son humeur (du moment)

L'humeur du moment n'est pas sans lien avec le tempérament général de la personne. Son incidence sur les décisions de lecture est manifeste : que l'on contraste les situations de stress et celles vécue avec aisance, et l'attitude face à un document, l'intérêt qu'on lui trouve, peut varier du tout au tout. On reconnaît l'expérience, banale, de lectures redécouvertes, et de lectures désenchantement.

---

[152] Les références citées en bibliographie sont : (Barry 1993), (Cool, Belkin, Kantor 1993), (Harter 1992) (inspiré par (Sperber & Wilson 1986)), (Klobas 1995), (Park 1993), (Wang, Soergel 1993).

[153] Essentiellement les travaux à la Direction des Etudes et Recherches d'EDF, ceux coordonnés par Xavier SOINARD (centrés sur les systèmes documentaires et la documentation électronique) et ceux coordonnés par Saadi LAHLOU (concernant le traitement de l'information dans les bureaux).
Voir notamment : (Merle, Fradin, Soinard 1994, pp. 45-47, 65-71, 75-90).

[154] « Je suis un boulimique de l'information sur mon sujet de recherche mais on n'a pas assez de concurrence qui publie sur le sujet », déplore un chercheur EDF passionné... (Merle, Fradin 1994, §7.2, p. 42)

L'objectif : comment la lecture prend place dans le travail

Lien à la phase de recherche

On peut avoir à mener l'enquête sur un domaine dans son ensemble :

- ceci va de la première prise de connaissance et de la découverte des réalisations dans le domaine, à l'établissement d'un état de l'art.
- Si l'on a commencé soi-même à contribuer au domaine, il est utile de faire le point sur son positionnement, son degré d'originalité.
- D'un point de vue stratégique, on cherche à connaître les besoins, la mode, ce qui est au goût du jour[155]. Avec une visée prospective, on veut pressentir et repérer des tendances.

Un enrichissement intellectuel personnel dans le domaine est quelquefois recherché, pour mieux fonder et mener à bien un projet :

- formation (documents didactiques), cumul d'expérience (prendre connaissance des expériences marquantes et des principaux résultats acquis).
- regard sur les activités amont et aval, pour une bonne intégration du projet dans son environnement de mise en œuvre.
- regard sur des disciplines voisines, qui partagent une même méthodologie ou un même objet d'étude, voire renfort et complémentarité pluridisciplinaires.

Selon ses compétences dans le domaine, et éventuellement si l'on a en vue une tâche de communication (enseignement, publication), on est amené à s'orienter vers des documents généraux (éventuellement de vulgarisation), ou bien vers des documents pointus, spécialisés. La forme d'esprit du lecteur (analytique, synthétique) et sa formation (qui l'ont rendu familier avec certaines approches) ont également une influence, ainsi que le délai dont on dispose pour mener à bien la tâche (court terme, long terme) (Mainguenaud 1994, §2.1, p. 9).

Un projet en cours de définition peut être à l'affût de ce qui peut lui permettre d'innover :

- connaissance de nouveaux moyens (outils, méthodes) ;
- avancement de la recherche, nouvelles idées. Les idées novatrices ne sont pas forcément des idées neuves, ce sont non seulement des idées récentes, mais aussi des idées anciennes mais oubliées, et ayant de nouvelles potentialités dans le contexte présent, et également les pistes suggérées par des projets en perspectives.

Suivant la nature de la tâche, l'orientation choisie est soit théorique, soit appliquée : c'est ce qui fait préférer telle revue à telle autre par exemple.

Lorsque l'on se trouve au cœur de la mise en œuvre d'un projet, la motivation de la recherche peut être plus ciblée :

- réponse à un problème, aide à la résolution d'une difficulté rencontrée ;
- être conforté dans un choix (approuvé par un document de référence, un expert reconnu du domaine) ;
- remise en cause (évaluer les faiblesses d'un travail au regard de résultats obtenus ailleurs).

Selon les cas, le besoin est plus ou moins bien défini, plus ou mois aigu. La formulation de ce qui est cherché s'affine avec la connaissance que l'on a du domaine, et donc le nouveau venu sur un sujet sera moins bien armé pour prédire ce qui l'intéresse. On touche là un des paradoxes connus de la recherche documentaire, qui demande au chercheur d'exprimer ce qu'il cherche alors que dans bien des cas il ne sait ce qu'il va trouver.

Une lecture peut être jugée intéressante, alors qu'elle sort du besoin explicité lors de la recherche d'information. La lecture n'est pas non plus toujours en rapport direct avec l'activité en cours : il y a une pertinence à court terme, à moyen terme, à long terme.

La lecture en tant que tâche à part entière : le devoir explicite de lecture

Premier cas, les documents en relecture : en tant que responsable ou qu'expert, il est demandé un avis sur le document. Le travail demandé est une contribution à la qualité et à la fiabilité du

[155] « Etudes Gartner, publications hebdos : plus pour voir ce que retiennent les journalistes, sentir la mode sur des sujets que je connais déjà. » (Merle, Fradin 1994, §7.2, p. 43)

document. La lecture vise à repérer les points à préciser ou à corriger, et peut être à suggérer certains compléments.

Deuxième cas, un document qui parvient par un circuit de diffusion « officiel », émanant de la hiérarchie, et dont le destinataire est tenu de prendre connaissance. La lecture est imposée, en ce sens que le destinataire est censé être informé de son contenu et en tenir compte.

Vue générale : quatre types de rapport à l'information

(Lahlou 1994, §3.3, pp. 39-40) résume les résultats de son analyse lexicale du concept information en proposant quatre types de propagation de l'information : *acquérir*, *aviser*, *instituer*, *apprendre*.

> - *acquérir* (le sujet récupère activement de l'information). [...]
> - *aviser* (le sujet envoie de façon active de l'information) [...]
> - *apprendre* (un couple de sujets se transmet de l'information de façon coopérative et volontariste, situation dont l'enseignement scolaire est l'archétype) [...]
> - et enfin *instituer* (un groupe de sujets explicite officiellement un état de choses, le valide pour la collectivité). La circulaire, le jugement, sont des exemples types de cette activité [...]
>
> (Fischler, Lahlou 1995, §2, pp. 8-9)

Les caractéristiques du document

Appartenance à un genre (correspondant à un lectorat et à une pratique)

Le type de document suffit parfois à rejeter le document, comme ces surabondantes publicités, reçues au courrier, visuellement immédiatement identifiées, et que le chercheur jette à la poubelle sans même les ouvrir[156].

Le genre définit en partie :

- la couverture du document, son exigence de complétude : on opposera typiquement l'ouvrage de synthèse et une annonce de presse, un livre relié et un extrait (photocopie réalisée par un collègue) ou un classeur hérité de multiples prédécesseurs.
- la précision : de ce point de vue la présentation indicative, la vulgarisation, ou la communication d'expert n'ont pas le même niveau de détail ni le même degré de technicité.
- les attendus concernant les connaissances du lecteur : une note interne en diffusion restreinte peut compter sur la familiarité des structures et rouages internes à l'entreprise ; un public de techniciens est capable de comprendre des notions qu'un public de dirigeant ne saisit pas, et vice-versa.
- le temps et le travail de lecture estimés : nombre de pages, attention nécessaire à porter au moindre détail, composition d'un seul tenant ou organisée en sections autonomes, etc.

Là encore, il n'y a pas de genre supérieur aux autres, ou de caractéristique intrinsèquement bonne ou mauvaise. Si par exemple le besoin d'information est une réponse factuelle à une question précise, point n'est besoin d'avoir un document complet décrivant toute la discipline.

Autres repères bibliographiques, qui manifestent certaines « garanties »

Un document non publié n'est pas utilisable de la même manière qu'un document publié : sa fiabilité n'est pas assurée (par un comité de lecture), sa lisibilité peut être moins bonne (document de travail très spécifique et technique, document non finalisé). Hors des circuits de diffusion officiels, son accessibilité est problématique. Il ne peut contribuer que de façon incidente à une bibliographie, et n'est en général pas considéré comme une citation pleinement valable.

La publication est un premier repère, général, pour le lecteur. Il peut être sensible à reconnaître plus particulièrement : tel auteur (dont il apprécie les écrits), tel éditeur (représentant une certaine exigence de qualité), telle revue (appartenant à tel courant de recherche).

Les termes consacrés du domaine, désignant des concepts porteurs par rapport à la problématique de recherche, sont de bons points d'accroche, dès le titre.

---

[156] Avec de malencontreuses erreurs d'identification possibles. C'est ainsi qu'un conditionnement trop coloré et plastifié a conduit à de dommageables échecs de communication, depuis les très officiels carnets de santé jusqu'à des faire-part personnels originaux.

Récence

Selon les genres, et même à l'intérieur d'un genre, il y a de fortes variations de « durée de vie » d'un document. Par exemple, les articles de fond sont destinés à perdurer sensiblement plus que les articles d'actualité (la presse est classée comme denrée hautement périssable, le marchand de journaux n'emmagasine pas de stocks). L'ouvrage de référence vieillit beaucoup moins rapidement que l'article qui fait part de la dernière nouveauté.

Le récent n'est pas meilleur en soi. Tout dépend de la perspective adoptée[157]. La nouveauté est une valeur cruciale dans le cadre d'une veille (technologique, stratégique). Elle est un facteur moins aigu quand il s'agit de faire la synthèse des travaux sur une question (resituer la progression de la réflexion, tracer l'historique de divergences ou de convergences, contribue à l'intelligence de la situation actuelle).

La date du document est généralement un paramètre significatif. Selon le contexte, son interprétation est plus ou moins précise (on la considère au jour près (journal), à l'année près, à la décennie près). Sa valeur en dépend aussi : visée prospective ou rétrospective, chronologiquement focalisée ou large.

L'air du temps flotte autour du lecteur : le document qui est au goût du jour, reflète les tendances pressenties, se réclame d'un sujet à la mode, est écrit par un auteur à succès... est plus facilement considéré avec un œil favorable.

Dynamique de la confrontation lecteur / document

La pertinence n'est pas dans le document, elle est événement

La pertinence ne peut être attribuée entièrement au document, soit par une analyse de son « contenu », soit par une caractérisation orientée utilisateur, qui prévoie et reflète les attentes auxquelles un document peut répondre[158]. Le texte n'explicite pas toutes ses richesses, et d'autre part on ne saurait en épuiser toutes les lectures possibles. Comme il a été ci-dessus débattu, à propos de la compréhension et de l'interprétation, la lecture est la rencontre d'un texte et d'un point de vue, elle passe par une démarche personnelle d'appropriation et d'intégration.

> Un bon système d'information doit à la fois prendre en compte, évidemment, les contenus (pour bien archiver, indexer, éviter les duplications, etc.) ; mais aussi les usages. Ceux-ci sont étroitement liés aux positions des acteurs et à leurs besoins d'information. Or, chaque acteur a sa propre vision naturelle du monde, et c'est à partir de celle-ci qu'il définit ses besoins. Si une « même » information est archivée dans le système d'un unique point de vue, les acteurs devront adopter le point de vue du système pour y accéder, ce qui ne leur est pas naturel et engendre des dysfonctionnements, des sous-utilisations, des frustrations. (Lahlou 1994, §2.6, p. 18)

Le système documentaire fait figure de relais. Lorsqu'il sélectionne certains documents, il les dote d'une garantie tacite de pertinence (Harter 1992) : un document présenté est considéré avec une présomption de pertinence (« il doit y avoir une raison qui justifie de le proposer, cherchons à la comprendre »). D'une certaine manière, le document est vu sous un jour favorable ; mais sa pertinence n'est pas acquise.

La pertinence naît à la rencontre d'une attente personnelle et de propositions de lecture. Aussi n'y a-t-il pas à se focaliser sur des suggestions statiques (arrêt du calcul) : celles-ci sont envisagées de façon dynamique. La pertinence n'est pas prédéterminée, elle est construite par le lecteur concerné, en situation, dans sa manière de balayer et de s'approprier des propositions.

---

[157] Dans l'enquête prospective sur une bibliothèque électronique : « Il est faux de dire que l'âge du document est critère d'obsolescence et de destruction. Ce sont les mêmes questions que l'on se pose lorsqu'on se met à ranger nos armoires. Par exemple quand une note *annule et remplace*, j'aimerais qu'automatiquement la note qui est annulée et remplacée soit marquée par le système comme *remplacée*. Mais faut-il la détruire pour autant ? Je peux en avoir besoin pour comparer l'évolution. Peut-être est-ce un changement de lieu ou de rangement. Le document est gardé pour une autre raison que précédemment et son lieu de rangement l'indique (historique alors qu'auparavant document de référence « vivant »). » (Merle, Fradin, Soinard 1994, p. 95)

[158] Une saine réaction, dans le contexte d'une proposition d'un surlignage *a priori* (le lecteur ne sachant pas sur quels principes il a été fait) : « Cela ne me plairait pas d'avoir un texte pré-stabyloté ou pré-pointé. Qui l'a fait pour moi ? Je dois rester seul maître de mes centres d'intérêt quand je lis. » (Merle, Fradin, Soinard 1994, p. 67)

Dans l'application DECID, la mesure de rapprochement profil-document matérialise en fait le parcours de lecture et l'opération d'interprétation, qui ne sont ni « dans » le texte, ni « dans » le lecteur, mais bien au niveau de leur confrontation. Le sens est construit et est lié au lecteur, il n'existe pas indépendamment de celui-ci. En quelque sorte, c'est parce que le rédacteur écrit pour un certain public qu'il transmet une information : il « dépose » dans le texte les éléments qui permettent au lecteur averti de reconstruire un message significatif.

Multiplicité des manières de recevoir un texte

Sans entrer dans le détail ici, rappelons que la pertinence peut s'établir selon des modes d'appréhension du document très divers : lecture superficielle ou approfondie, recherche d'informations directement utilisables ou capacité du texte à susciter la créativité.

Certains documents ne seront que feuilletés : l'information est simplement de « savoir que cela existe ».

Similarité mais pas identité

Dans les systèmes de recherche d'information, la pertinence est généralement exprimée sous forme d'une similarité. Pour autant, la pertinence « maximale » n'est pas la similarité complète, au sens de l'identité. Le lecteur doit trouver des points communs avec son activité, ses centres d'intérêt ou ses compétences, pour se sentir concerné, mais il n'est réellement intéressé que si le document comporte aussi certains écarts, certaines différences, une certaine part d'inattendu[159], qui puissent être sources de connaissance[160]. Ainsi en est-il de la lecture d'une bibliographie : les références qui donneraient envie de voir le document correspondant sont celles dont l'auteur est connu mais portant sur un thème différent de ce qu'on connaît déjà pour cet auteur, ou bien celles dont le titre promet un développement intéressant, et dont l'auteur, inconnu, peut renouveler et enrichir la vision du lecteur sur le domaine. En somme, il n'y a de superposition intéressante que partielle.

Par exemple, dans le cadre du « *Qui Fait Quoi ?* » (calcul des similarités entre les différents projets de recherche à la Direction des Etudes et Recherches à EDF), un chef de groupe trouve intéressant de signaler des projets peu liés à ceux de son groupe, mais qui pourraient être amenés à l'être davantage.

C'est aussi ce que l'on peut appeler « l'effet étagère » : le document avec lequel on repart est le document voisin de celui que l'on était venu chercher. Un abonnement joue de cet effet, en présentant des articles non explicitement recherchés mais qui ouvrent sur diverses considérations en lien avec une thématique et une forme d'approche.

Enfin, l'intérêt d'un document n'est pas forcément dans son contenu immédiat, d'où une lecture que nous pourrions qualifier d'oblique.

---

[159] Avis recueilli lors d'une enquête sur les systèmes documentaires :
« Souhait de naviguer de façon autonome dans l'information pour se laisser toucher par des informations non recherchées, imprévues mais connectables et utiles.
Je suis preneur d'informations quitte à m'embêter un peu à lire. J'aime naviguer, me laisser capter.
Refus d'être enfermé dans une démarche trop rationnelle autour d'un thème (l'apport méthodologique des [professionnels de l'information] est important mais il faut être avec pour saisir l'imprévu). » (Merle, Fradin 1994, §11.3, p. 56)
Egalement, les réponses suivantes à un questionnaire, lors de l'enquête PUBE (Merle, Fradin, Soinard 1995, p. 17) :
5ième question : « Je préfère recevoir ce qui est strictement dans mes préoccupations »
Réponses : Vrai : 7 personnes ; Faux : 5 personnes (avis partagé).
6ième question : « J'aime recevoir des informations hors de mon domaine, ça peut me servir »
Réponses : Vrai : 11 personnes ; Faux : 0 ; Ne sait pas : 1 personne (quasi unanimité).

[160] C'est ainsi que (Harter 1992) lit (Sperber & Wilson 1986) et l'applique aux systèmes documentaires. Le document qui n'apporte rien qui ne soit déjà présent à l'esprit de celui qui le considère, n'est pas jugé pertinent : être pertinent, c'est avoir des *effets contextuels*, autrement dit s'insérer dans un contexte et y apporter des modifications.

Le document, situé par un cheminement qui le rapporte à soi

La décision de lecture suppose une démarche, un investissement personnel, un cheminement, qui explicite le sens de cette lecture pour le lecteur en question. Le document est par exemple obtenu au terme d'une navigation dans une base documentaire, qui le situe dans une logique de recherche et l'a identifié en le cernant. Un document « anonyme » (c'est-à-dire associé à rien de connu : auteur, collègue, laboratoire, revue, etc.) et impersonnel (dont on ne voit pas le rapport avec ses activités propres et qui n'est pas introduit par un intermédiaire) est d'emblée non pertinent.

Composer ses lectures : préférences et compromis

Attentes et aperçu

Le même lecteur peut être dans un cas satisfait avec un document, et une autre fois repartir avec dix avec l'intention de poursuivre sa recherche. Cela dépend évidemment de l'objectif de lecture : document prédéterminé, document quelconque qui comporte tel renseignement, recherche d'exhaustivité (bibliographique, approches, etc.). Par exemple, la recherche d'une citation connue se satisfait de l'ouvrage qui l'enferme, mais un état de l'art touche de multiples documents et n'est pour ainsi dire jamais clos. Le choix global d'un ensemble de lectures relève aussi de l'idée que l'on se fait de la largeur du domaine de recherche, et de la quantité d'information optimale pour être « assimilable » et suffisante.

Curiosité et plaisir de la nouveauté

Les gens demandent ce sur quoi ils ne travaillent pas, remarquent certains. Ouverture et culture générale, renouvellement par rapport à une problématique par ailleurs cultivée et bien connue, justifieront éventuellement ces choix aux yeux de l'entreprise ou du centre de recherches qui appelle à toujours plus de souplesse, de mobilité, d'innovation.

Singularité et spécificité

Ce qui est vague est déprécié. Le lecteur est d'autant plus motivé qu'il perçoit ce que tel document peut lui apporter d'unique.

Un document qui serait identifié comme étant *du ressort* d'un chercheur (un ouvrage générique sur son domaine : mécanique, acoustique,...) est rarement ce qui peut l'*intéresser*, qui est beaucoup plus spécifique.

Multidimensionnalité

Il est difficile d'indiquer un ordre sur les documents retenus, qui sont complémentaires. Différents critères de sélection ont été mobilisés, et ne sont pas naturellement commensurables : celui-ci est choisi pour sa récence, celui-là par qu'il fait autorité, etc.

Gestion du temps

Pris dans des contraintes de délais, le lecteur s'oriente vers ce qu'il juge le plus urgent ou le plus important, le premier aspect occultant parfois le second.

Les jugements et choix de priorité se concrétisent d'ailleurs dans l'organisation des documents sur le bureau par exemple :

> Pour maîtriser ou contrôler l'ensemble de l'information qu'ils ont à traiter, les interviewés ont recours à certaines stratégies sensori-cognitives dont certaines sont rapidement repérables dans les entretiens. Deux locutions sont constamment utilisées par les répondants, de façon très significative : « sous la main » et « sous les yeux ».
>
> [...] l'espace [est utilisé] comme métaphore de l'urgence.
>
> [...] Il y a ainsi équivalence pour certains sujets entre, d'une part, le niveau d'urgence ou d'importance et la proximité spatiale. Ce qui est important doit être à portée de la main et / ou à portée de regard.
>
> [...]
>
> La métaphore d'une information plus ou moins « vivante » revient souvent, comme si un dossier urgent ou non clos était plus « animé » qu'un dossier traité ou non urgent. Ceci renvoie sans doute à

> une sélection écologique. Dans la vie de l'homme primitif, un objet vivant, qui bouge, attire l'attention parce qu'il est potentiellement « à surveiller », parce qu'il est potentiellement dangereux... tandis que le paysage immobile devient un fond qui s'estompe. Il semble que les sujets aient transposé à l'information cette distinction entre ce qui est « vivant » (et qu'il faut surveiller) et ce qui ne l'est pas, et requiert moins d'attention. C'est l'idée de la saillance (étymologiquement de *saillir* : sauter, pour un animal), caractéristique qui fait qu'un objet s'impose à l'attention. D'ailleurs on sait que les humains (et de nombreux autres animaux) ont développé des capteurs différents (bâtonnets et cônes) dans la rétine pour les contours et pour le mouvement. Peut-être y a-t-il là une piste à creuser pour les métaphores d'interface documentaire ? Un dossier « urgent » gigoterait plus qu'un autre sur l'écran du micro...
>
> (Fischler, Lahlou 1995, §4, pp. 18-19)

Dans la pratique, la conduite n'obéit pas à la seule rationalité de l'urgence ou de l'importance. Car sont souvent considérées en priorité les informations qui peuvent se traiter rapidement et immédiatement, donnant mieux l'impression (gratifiante) d'avancer dans son travail (Fischler, Lahlou 1995, §5.2.4, p. 37).

Tentations de satisfaction

Le lecteur aurait l'esprit tranquille s'il avait la certitude d'avoir tout lu, ou du moins lu tout ce qui est essentiel. Ce sont là deux tentations (et de fait deux illusions) qui guettent l'utilisateur d'un système de recherche documentaire, et se manifestent par deux souhaits :

- disposer d'un jugement conclusif, arrêté, sur la qualité et l'intérêt des documents disponibles ;
- avoir l'indication de ce qui est essentiel –d'où aussi l'attrait pour certains systèmes qui se vantent de filtrage et d'écrémage, et fourniraient la substantifique moelle d'un texte, accusé de longueurs indues.

Dans les deux cas, le lecteur ne peut pourtant s'en remettre qu'à lui-même en dernière instance, et en tout cas pas à un système automatique. Le document ne peut rien lui apporter s'il ne s'implique dans sa lecture et cherche activement à en tirer profit. Quant à la détermination de l'essentiel, il est généralement raisonnable de présumer l'auteur innocent de toute volonté d'écrire sans intention de sens, ce qui fait qu'il n'y aurait aucune partie *a priori* vaine. Et réciproquement, la lecture est partielle (partiale) et retient tel ou tel point qui a mobilisé l'attention : il n'y a donc pas non plus de partie en soi toujours importante. On ne peut donc fixer ni ce qui serait toujours essentiel, ni ce qui ne le serait jamais. A chaque lecture revient sa part de discernement et d'engagement personnel.

La question « où s'arrêter ? », dans le dépouillement de listes de propositions (ou autrement) reste donc sans réponse formelle. Comment cela se résout-il dans la pratique ? Par des propositions à géométrie variable, des déploiements progressifs sur les aspects les plus intéressants (des « filons »).

La société du lecteur (communauté scientifique, collègues)

Intégration

Avoir lu un document rare et convoité, ou au contraire un document incontournable et que « tout le monde a lu », participe à l'intégration et à la reconnaissance de l'individu dans sa communauté de travail. C'est une équilibre entre ce qui valorise et distingue (avoir un plus, en tirer un certain prestige), et ce qui assure que l'on est bien dans la norme (ne pas avoir de manque, être à même de participer aux discussions).

Une lecture peut être motivée (ou à l'inverse écartée) en fonction du jugement que l'on prête à autrui : faire telle lecture risque de plaire ou déplaire à telle ou telle personne proche, dont l'avis importe (Klobas 1995).

Rapports hiérarchiques

La lecture fournie ou l'information envoyée par un supérieur mêle, au jugement que l'on peut avoir sur le contenu du document, des considérations de devoir plus ou moins claires. Une consigne explicite accompagnant l'envoi peut tenir d'un ordre (« j'aimerais avoir votre opinion sur ceci »), ou dégager de toute contrainte (« j'ai reçu cela, voyez si cela vous intéresse »). Dans tous les cas, avec ou

sans message clair sur le motif de l'envoi, la personne se sent plus tenue de prêter attention et d'accorder un certain égard aux propositions de son chef.

Les circonstances

Disponibilité

Disponibilité de la personne, qui se trouve plus ou moins de temps pour ses lectures.

Accessibilité du document : attendre des mois pour l'obtenir, le payer à prix d'or, devoir multiplier les démarches et tracas administratifs, devoir se déplacer pour consulter sur place, avoir un temps de prêt court, etc., peuvent décourager le lecteur, qui au besoin trouve un document alternatif plus simple à obtenir.

Egalement, s'il n'y a qu'un ou deux documents disponibles sur un sujet, le lecteur s'en contente en général plus volontiers et limite ses exigences, alors qu'il doit au contraire exacerber celles-ci lorsque, face à un très grand nombre de documents proposés, il doit faire un tri et n'en sélectionner que quelques-uns. Autrement dit, la pertinence accordée à un document dépend de l'intertexte effectif dans lequel il se positionne lors de la recherche d'information.

Relations interpersonnelles

La présentation d'un document à un lecteur se double d'une relation interpersonnelle quand par exemple le document est recommandé par une personne de sa connaissance. L'estime que l'on a pour cette personne a généralement une influence majeure sur la décision de lecture. La personne qui recommande fait effet d'un garant, de la même manière que l'on fait confiance à un auteur ou au comité de rédaction d'une revue, sauf que c'est ici personnalisé.

Déresponsabilisation par dilution collective

La revue, reçue sur abonnement, et qui circule systématiquement entre tous les membres d'une unité en suivant une liste de diffusion standardisée, peut être perçue de deux façons totalement opposée : ou bien, un devoir de lecture (document de travail officiel, visa à apposer) ; ou bien, une revue qui n'est pas vraiment destinée à toutes les personnes nommées (si la liste ne varie pas selon les documents en circulation), soit une situation telle que, si l'un omet de la parcourir, un autre l'aura lue et signalera à son collègue une information le concernant, au cas où il l'aurait manquée.

L'environnement agréable et efficace de proposition de documents

Pour un système automatique, une ergonomie insuffisante pénalise l'utilisateur : utilisation rébarbative parce que trop spécialisée, difficulté à obtenir des propositions qui semblent adaptées, sentiment de ne pas pouvoir maîtriser l'outil en sorte d'être en mesure d'arriver à ses fins et dans un temps acceptable, dépouillement des résultats laborieux et insatisfaisant s'il manque des éléments pour se faire une idée des documents proposés. Cela peut diminuer les motivations de lecture, et engendrer des déconvenues (tel document qui pouvait sembler pertinent se révèle décevant une fois l'exemplaire en main).

Vers la construction d'indicateurs de pertinence

L'observation des pratiques de recherche documentaire conduit alors à proposer des « règles » heuristiques, qui explicitent certaines combinaisons favorables de facteurs. Il s'agit en fait de mettre au jour de véritables stratégies interprétatives, l'esquisse d'une herméneutique documentaire. Les propositions suivantes nous semblent un exemple intéressant de ce type de recherche, en montrant notamment bien l'intertextualité à l'œuvre dans l'activité d'évaluation :

1. *Elimination rule*. Users tend to look for aspects of the document and the things which are obviously not what they look for to reject a document :

« This is a dissertation. It doesn't really say very much. I will tend to pass it. »

« ... Grossman, I don't know. I don't see any economist's writing at all. No. »

2. *Multi-criteria rule*. Users may feel more comfortable to accept a document by using more than one criterion :

« ... OK, the next one should be good. I know all the authors. The title sounds good. The authors are very competent and it's published in *American Journal of Agricultural Economics*. »

3. *Dominance rule*. Of similar documents, users will select the one which excels in at least one aspect and not worse on the other aspects, especially if only a few documents are wanted :

« ... This is the same as te previous one, almost the same. But, I like the other one better because this one is 1980 and that one is 1982. »

4. *Scarcity rule*. Users tend to select all seemingly relevant documents, if many documents are wanted and only few retrieved :

« I am going to put this, even though it's on Canada, the abstract sounds very good. I am going to put this high. Also, deals with soil erosion, which I don't think that we have seen too many [in the search outcome]. We have seen a lot of pest management, not much about soil. »

5. *Chain rule*. Users tend to select chained documents. Chained documents are critiques with original and papers from the same book. If one of such documents is selected, most likely the chained documents are also wanted. Vise versa.

« This is the article that led to the discussion above [previous article has the same title plus the subtitle *Discussion*]. This is the main article. I will put a check next to it also. »

(Wang & Soergel 1993, p. 90)

b) Le point de vue, réciproque, de l'expéditeur d'un document (notamment par diffusion ciblée)

Dans le cadre de la diffusion ciblée, à la question de la pertinence d'un document pour un lecteur (pertinence du destinataire, à qui est proposé un document), s'ajoute la question de la pertinence d'un envoi (pertinence de l'émetteur), qui n'apparaît pas dans la problématique documentaire classique.

Interprétations des propositions du système

Responsabilité et risque

Prendre l'initiative d'envoyer ou de faire suivre un document est perçu comme une responsabilité : le risque est de s'être fourvoyé sur les thèmes d'intérêt du destinataire et que le document paraisse hors-sujet, ou à l'inverse que le destinataire soit trop compétent ou déjà informé. Aussi l'exigence prioritaire semble être la précision de l'envoi (que l'on veut exempt d'erreurs, pour ne pas subir de rejet et de mécontentement), plutôt que l'assurance d'avoir trouvé toutes les personnes potentiellement intéressées (si la diffusion n'est pas « complète », elle est déjà un plus par rapport à l'absence de diffusion (sans l'aide de l'outil) ; et les contacts personnels pourront prendre le relais et prolonger, compléter et affiner la circulation de l'information ; le premier envoi n'aura été qu'une amorce).

Les propositions de destinataires pourraient donc être accompagnées par un indicateur de fiabilité.

En alternative à l'envoi, mettre à disposition et signaler (par voie d'affichage) peut être gratifiant. Il est naturel de ne pas vouloir imposer une information en présupposant sa pertinence pour la personne : c'est un signe de plus, qui témoigne que la pertinence ne se réduit pas à une simple adéquation de thématiques et de sujet traité.

Il y a donc des degrés d'action : l'envoi du document est un acte plus « fort » que l'envoi partiel ou le signalement.

La pertinence n'implique pas l'envoi

Quand l'utilisateur de la diffusion ciblée dépouille un ensemble de destinataires proposés par le système, il y a plusieurs cas de figure où la mention d'une personne est jugée positive, sans pour autant conclure à un envoi.

La situation la plus évidente est de trouver le nom d'un collègue proche (voire son propre nom). On sait que cette personne a déjà l'information, donc on n'utilise pas le système de diffusion ciblée pour lui faire parvenir. Pour autant, l'indication de cette personne par le système est bénéfique pour deux raisons. D'une part, cela rassure sur l'efficacité du système, et cela conforte le crédit qu'on lui accorde pour retrouver des personnes concernées : le système répond à la fonction que l'on attend de lui. D'autre part, les propositions du système peuvent ainsi former un ensemble cohérent et équilibré, qui permet par exemple de se rendre compte de la proportion (forte ou faible) de personnes que l'on était en mesure de trouver sans l'aide du système. La mention systématique de toutes les personnes potentiellement concernées permet de faire un point complet avant diffusion de l'information : il peut notamment y avoir des personnes connues de l'utilisateur du système de diffusion ciblée, mais qui avaient été oubliées, et que le système rappelle avec opportunité.

Il y a également des signalements (de personnes) jugés intéressants, mais qui n'ont pas à donner lieu à un envoi immédiatement. L'utilisateur du système de diffusion ciblée conserve leur nom pour une autre diffusion ou pour un contact ultérieur éventuel.

Les jugements de pertinence ne se juxtaposent donc pas avec les décisions d'envoi.

Les destinataires, collègues dans une même entreprise

Couverture

Dans le cadre de la diffusion ciblée, il y a une pertinence d'ordre global. Il est préférable que l'information soit bien répartie dans le centre de recherche et touche des entités qui ont peu ou pas de contacts entre elles (d'où un gain en matière de liens de communication), plutôt que de destiner tous ses envois à différents membres d'une seule petite équipe. Dans ce dernier cas en effet, on court-circuite le relais du bouche-à-oreille, de personne à personne, plus efficace et nécessaire à la vie de l'équipe.

L'apport du système est aussi de proposer des personnes moins évidentes à trouver par d'autres moyens : personnes dont l'activité est assez atypique par rapport à leur rattachement et à leur équipe (non retrouvées avec un organigramme), personnes dont l'activité est décrite avec un point de vue différent et un vocabulaire inattendu (non retrouvées par une recherche par mots-clés).

Focalisation

Multiplier les exemplaires et les copies ne multiplie pas d'autant l'impact, il peut même le diminuer, s'il n'y a pas un destinataire nominatif et privilégié. En effet, lorsque le destinataire est collectif, et que le document n'est pas spécialement attrayant, chacun se dit qu'un autre prendra bien la peine de le lire, et de repérer pour lui ce qu'il peut y avoir d'important. Assurer une diffusion focalisée évite une dilution de l'impact et responsabilise le destinataire.

Convivialité

Les documents se passent et circulent en empruntant le réseau des connaissances. Cet échange le renforce : c'est l'occasion de reprendre contact avec untel, d'inviter à collaborer plus étroitement, etc.

> Se faire parvenir des documents, c'est aussi une façon de communiquer entre amis et collègues. C'est une conséquence de l'existence [de ces] réseaux [de relation] (Joseph-Waterlot, Lahlou 1995, §II.4.1.d, p. 20)