CHAPITRE VI

Détermination d'unités de traitement

Aperçu

La confrontation des textes entre eux suppose la définition d'unités d'analyse, qui sont fournies par une opération d'indexation. L'indexation procède par *dotation* (l'analyse vise à reconnaître dans le texte des unités déjà définies, et les unités trouvées sont attribuées à la représentation du texte) ou bien par *érosion* (il s'agit à l'inverse de reconnaître ce que l'on élimine) ; ces deux approches sont complémentaires. Pour DECID, l'indexation par érosion a une importance certaine, étant donné que les textes considérés sont par nature riches de néologismes, de désignations spécifiques, de concepts nouveaux, toutes choses qui ne peuvent être prévues dans les unités d'une approche par dotation. Le point de vue textuel avertit aussi d'une question fondamentale qu'il faut affronter : la détermination du local par le global, exprimée autrement par la non-compositionnalité de la langue.

Pour commencer, il est nécessaire de voir les outils dont on dispose, grâce aux travaux de la communauté des Traitements Automatiques des Langues. En effet, le texte étant par nature linguistique (facette 1), il y a tout intérêt à se donner les moyens de prendre en compte cette propriété. On examine donc, tour à tour, comment chaque type d'outil peut concourir à une application qui doit calculer, entièrement automatiquement, des rapprochements texte - texte. Par exemple, il y aurait diverses manières d'utiliser les résultats d'un catégoriseur morpho-syntaxique, et toutes ne sont pas aussi intéressantes. Expérimentalement, DECID a initialement profité d'un outil d'indexation automatique avancé, mais auquel un traitement très fruste a succédé : les raisons de cette décision sont analysées.

Cet état des lieux des outils existants prépare la conception d'une architecture pour le traitement des textes, qui intègre les apports des traitements morphosyntaxiques (au palier du mot ou de la phrase), tout en se plaçant dans une perspective proprement textuelle. Une première innovation consiste à considérer l'indexation non plus comme une procédure qui affecte directement à chaque texte, au fil de sa lecture, des termes d'index, mais comme un traitement qui enchaîne plusieurs phases, *construction* des unités d'abord, et, ensuite, *élection* des unités pour caractériser chaque texte. Le mouvement est en harmonie avec l'interaction global / local désirée. Le traitement part d'éléments locaux (*unités élémentaires*), les reconsidère d'un point de vue global, et détermine alors les *unités descriptives* qui pourront être utilisées pour représenter les textes. Les textes reçoivent finalement des *unités caractérisantes*, sélectionnées parmi les unités descriptives disponibles comme représentatives du texte. Les unités descriptives sont ainsi dynamiquement adaptées au corpus ; elles peuvent être complétées par des unités qui apportent une certaine information sur les textes, par exemple relativement aux genres (par exemple, le genre des textes descriptifs des programmes de recherche à EDF-DER).

La mise en œuvre d'un tel modèle suppose une définition opérationnelle des trois catégories d'unités introduites : unités élémentaires, unités descriptives, unités caractérisantes. Les unités élémentaires font l'objet d'une première implémentation simple, sachant que l'essentiel se joue avec les unités descriptives, qui ont une certaine latitude pour remodeler le matériau de départ (les unités élémentaires).

Les unités descriptives se déclinent en huit types, que l'on peut regrouper en trois familles de « niveau » successif. Première famille, les unités initiales, qui ont pour rôle de rectifier les erreurs du découpage local en unités élémentaires, et de proposer une base descriptive valide après examen global. Certaines unités élémentaires sont confirmées et reprises telles quelles (*unités Simples*) ; d'autres sont ajustées au plan syntagmatique (*Solidarités*) ou au plan paradigmatique (*Assimilations*). Cette manière de faire correspond à une forme d'apprentissage endogène, au sens de Bourigault. La deuxième famille d'unités descriptives fournit des structures d'association syntagmatique (*Séquences*)

et paradigmatique (*Associations*) plus souples et plus nuancées. La troisième famille d'unités descriptive est celle des *Communautés*. Les types d'unités descriptives précédents reprenaient (avec quelques libertés) des perfectionnements connus dans le domaine des systèmes documentaires : termes composés, reformulation, etc. –on vérifie d'ailleurs que le modèle ici défini pour DECID n'a rien à envier aux systèmes documentaires traditionnels, même relativement évolués comme TOPIC. Les Communautés font entrer des moyens nouveaux, empruntés à la Sémantique Différentielle Unifiée (Rastier) : les isotopies sémantiques. L'astuce consiste à « retourner » la démarche habituelle, et dispense du même coup de constituer un (problématique) dictionnaire de sèmes : en résumé, ce n'est pas le repérage préalable de sèmes réitérés qui conduit à reconnaître l'isotopie, mais c'est la manifestation de l'isotopie qui traduit la récurrence d'un sème. L'isotopie est associée à la définition de zones de localité. Estimant que les interactions sont d'un ordre différent aux paliers de la période, du paragraphe, et du texte, trois types d'unités descriptives sont distingués en conséquence (respectivement, les *Relations*, les *Voisinages*, les *Arrière-Plan*). Techniquement, la construction automatique des Communautés appelle un algorithme de classification multiclasse non exhaustive ; comme ce cas n'est pas traité par les procédures connues de classification, un algorithme adéquat est élaboré, par adaptation et extension des Nuées Dynamiques (Diday).

Table des matières du Chapitre VI

A. REPÈRES PRÉLIMINAIRES

1. Rôle des unités

a) Justification théorique : pourquoi introduire des unités

La diffusion ciblée demande de savoir confronter un document à un profil, soit concrètement deux textes l'un avec l'autre. Or deux textes sont deux entités différentes : tels quels, ils sont incommensurables. L'analyse d'un texte en unités permet de repousser l'irréductibilité d'un degré, en ajoutant un nouveau niveau de granularité : ce ne sont plus les textes qui sont incommensurables, mais les unités, puisque c'est par définition à elles que s'arrête la décomposition possible du texte.

Les unités constituent le matériau à la base de la représentation de chaque texte : la représentation est une combinaison de ces unités. Le mécanisme de comparaison peut alors se fonder sur l'identité des unités et l'équivalence des structures de combinaison.

Pour prendre le vocabulaire de la linguistique, les unités sont le *lexique* des représentations des textes ; leur mode de combinaison, pour former une représentation, est régi par une *syntaxe*. D'aucuns, à l'instar de l'équipe québécoise précisant la Clé des procédés littéraires[1], trouveront plus claire une terminologie mathématique, selon laquelle les unités seront appelées *opérandes*, et leurs articulations possibles données par des *opérateurs*.

b) La relativité (multiple) des unités

Les unités sont relatives à la nature de la représentation et au niveau de détail voulus.

Nature de la représentation

La nature de la représentation dépend de l'usage pour lequel la représentation est conçue. C'est une question de pertinence, qui ne se tranche ni de façon uniforme, ni nécessairement de façon évidente.

Quelquefois, l'utilisation désigne directement les unités qui lui sont utiles : par exemple, pour une étude de métrique latine, un système d'unités permettant une représentation du texte donnant la longueur des syllabes successives est une base naturellement bonne pour poursuivre l'analyse (identification des dactyles, des spondées, etc.). Ou encore, s'il s'agit d'extraire certaines informations données d'un texte, pour compléter les champs d'un formulaire codifié (exemple : analyse de Curriculum Vitæ pour enrichir une base de données de candidats), la représentation se fera directement dans les termes de l'information recherchée (dans le cas du CV, les unités peuvent être : les noms et prénoms des candidats, les différents niveaux d'étude que l'on souhaite distinguer, les valeurs possibles de l'âge du candidat, .les types de postes, les employeurs répertoriés, etc.).

Dans d'autres cas, les unités sont choisies en faisant l'hypothèse (éventuellement implicite) qu'elles sont significativement corrélées à ce que l'on veut utiliser. Ainsi, une analyse lexicométrique peut choisir de se baser sur les formes graphiques des mots, pour tenter de repérer les enchaînements stéréotypés qui relèvent de la phraséologie du type de textes décrit.

Prenons un exemple volontairement trivial, pour illustrer l'importance de la nature des données : s'il s'agit, pour un ensemble de feuillets rédigés en français, de faire le tri entre ceux qui peuvent être dupliqués par une photocopieuse standard et ceux qui demandent du matériel plus évolué, la taille du papier, l'impression recto/verso, la couleur de l'encre, sont des unités de

[1] Consultable sous forme hypertexte à l'adresse :
`http://tornade.ere.umontreal.ca/~dupriez/cle.html`
la *Clé* donne accès au nom, à la définition et à la description complète d'une figure de style, à partir des opérandes auxquelles elle s'applique (prononciation, lettres,...) et des opérateurs qui les lient (répétition, inversion,...). L'idée de départ était que la consultation d'un dictionnaire supposait déjà connu le nom de la figure recherchée : la *Clé* propose alors un autre mode d'accès, non plus par le nom, mais par les composants d'une analyse descriptive de la figure.

représentation pertinentes ; en revanche, il y a peu de chances que les lettres de l'alphabet utilisées ou la police de caractères soient des unités adéquates.

Pourtant, certains partisans des réseaux neuronaux tendent à ignorer cette question de la nature des unités, en survalorisant le choix du niveau de détail. A les croire, pour peu que l'on fournisse une représentation assez fine de l'objet en entrée, l'apprentissage permet au réseau de se forger des unités internes pertinentes par combinaison des entrées, et après cette adaptation le traitement s'effectue au bon niveau, vers lequel le réseau a convergé. Le maillon délicat est évidemment l'adéquation de la nature des données en entrée pour la définition des unités internes. Autrement dit, si les réseaux neuronaux s'inspirent de l'organisation du système nerveux, ils ne peuvent éluder l'étape de la perception, qui conditionne, en tant que première étape, les traitements effectifs.

Une échappatoire est de mettre « tout » en entrée du réseau de neurones, en s'appuyant sur sa capacité à neutraliser les variables non pertinentes pour la tâche sur laquelle s'effectue l'apprentissage. Si elle n'est pas un tant soit peu guidée par une hypothèse sur les données potentiellement pertinentes, cette fuite en avant aiguillée par des mobiles purement quantitatifs peut se solder en errance (résultats non significatifs car il n'y a pas de données pertinentes captées) puis noyade (sous le volume des données brassées). En effet, si la description peut réunir de multiples points de vue, elle n'est jamais totalement exhaustive.

Niveau de détail

Le niveau de détail est borné d'un côté par le nombre d'entités à décrire et à distinguer (grosso modo, pour une syntaxe de combinaison des unités donnée, le nombre de représentations différentes croît avec le nombre d'unités), de l'autre pour éviter une dispersion (nuisible à la comparaison) ou/et une redondance (pesant sur les ressources à mobiliser).

L'exemple des sèmes

Les *sèmes*, composants définis et utilisés par l'analyse sémantique, illustrent ces propriétés des unités. En effet, en suivant (Rastier, 1987, § I.1.), ils sont :

- non universaux (ils sont à tout le moins relatifs à chaque langue),
- en nombre modéré (position équilibrée entre un petit nombre, qui calquerait des catégories fondamentales, et un nombre infini, à la mesure des variations du monde référentiel à décrire, ces deux extrêmes ignorant la spécificité linguistique en rabattant la langue sur une réalité d'un autre ordre, conceptuelle ou référentielle),
- de niveau de détail ajusté aux besoins de l'analyse - ce ne sont pas des composants ultimes ou primitifs, car une analyse plus fine peut décomposer un sème donné en unités de contenu plus petites.

Les sèmes ne sont en rien une réalité absolue, ils ne s'imposent pas à une description qui ne pourrait alors être qu'unique. Ils reçoivent leur valeur dans un jeu de relations et de différences :

> Cette unité minimale, cependant, que nous avons dénommée *sème*, n'a pas d'existence propre, et ne peut être imaginée et décrite qu'en relation avec quelque chose qui n'est pas elle, que dans la mesure où elle fait partie d'une structure de signification. (Greimas 1966, §VII.1.b, p. 103)

Conséquence : prévoir des mécanismes d'ajustement

Pour l'application de diffusion ciblée, les unités pertinentes pour la représentation des textes n'ont rien d'évident. Pour autant, il faut bien disposer d'un mode de représentation. Il faut donc :

- partir d'unités *a priori* assez fortement corrélées aux facettes (des profils, des documents) qui peuvent motiver un envoi : thème d'intérêt, type d'information,...
- prévoir des mécanismes d'ajustement, par la construction de nouveaux types d'unités mieux adaptées, et par la variation possible du niveau de détail.

2. Deux approches : dotation vs érosion

a) Faire le lien entre les unités et le texte à représenter

S'agissant de construire la représentation d'un texte, il faut définir la manière de trouver les unités à utiliser. Il y a alors deux démarches opposées[2].

Pour la première, la représentation initiale est vide ; on dispose d'un dictionnaire qui associe à des unités prédéfinies toutes les manifestations qu'elles peuvent prendre. La représentation regroupe alors toutes les unités qui sont reconnues dans le texte. C'est ce que nous avons appelé l'approche par dotation.

La démarche inverse consiste à partir d'une représentation grossière du texte lui-même, et à l'affiner peu à peu, par décompositions et éliminations successives. C'est la matière initiale qui est transformée pour devenir une représentation acceptable, conforme à l'usage attendu. Le nom d'approche par érosion voudrait évoquer le travail du sculpteur, qui, partant d'un bloc qui lui inspire un sujet, dégrossit la pierre par petits éclats et fait apparaître les grandes lignes puis la forme pressentie[3].

Paijmans (1992, 1993) fait la même distinction, et parle respectivement d'assignation et de dérivation (*assigned indexing* vs *derived indexing*). Greimas rencontre, dans la procédure de description du corpus, une question analogue : *Elimination ou extraction ?*. Il constate la relation duale entre ces deux alternatives, à la nuance près que « l'extraction paraît, à première vue, plus sujette à l'appréciation subjective du descripteur » (Greimas 1966, §IX.1.d, p. 146).

Ce serait encore à rapprocher de l'opposition entre *stratégie de conquête* et *stratégie d'apprivoisement*, que nous avions proposée comme lecture des pratiques en linguistique de corpus (Pincemin, Assadi, Lemesle, 1996, § 7.1). La stratégie de conquête consiste à aborder le corpus à travers une grille d'analyse (fondée par la théorie). On ne retient du corpus que ce qu'on y trouve comme éléments correspondants à la grille d'analyse. Un certaine part du corpus échappe donc à l'analyse, et l'effort est mis pour affiner et enrichir le modèle, pour obtenir des représentations de plus en plus complètes. On vise à capter, à enrégimenter, les données selon les vues du modèle théorique. La stratégie d'apprivoisement part d'une représentation grossière mais couvrant à sa manière l'ensemble du corpus. On la fait ensuite évoluer en l'affinant progressivement, à partir des régularités observées dans le corpus. Le modèle qui se dégage est alors par construction en affinité avec le corpus.

Les diverses formes d'indexation classiques se laissent alors classer dans ces deux catégories :

- l'indexation contrôlée par un thesaurus (précisant les termes vedette), l'indexation utilisant un dictionnaire et écartant tous les termes inconnus, relèvent de l'approche par dotation.
- l'indexation libre, plein-texte, en texte intégral, à savoir celle qui retient un ensemble de mots ou d'expressions du texte et ne prévoit que des éléments qui peuvent être éliminés (et non ce qui peut être sélectionné) correspond à une approche par érosion.

---

[2] Ces démarches opèrent dans deux sens opposés, pour autant elles ne sont pas incompatibles.

[3] L'analogie avec la sculpture ne s'arrête peut-être pas là. Michel-Ange, dans ses réflexions sur la sculpture, aurait opposé l'approche par moulage, à celle qui dégage peu à peu la forme de la masse. Suite à cette analyse, il aurait abandonné le bronze et fait le choix du marbre.
Cette démarche s'interprète comme le refus d'une *intrusion*, d'une *intervention autoritaire*, au profit d'une *découverte*, d'une *mise au jour* de ce qui est *déjà présent* au cœur de la masse. On *dégage progressivement l'excès* de matière qui recouvre encore la statue. L'artiste est humble devant la réalité de la pierre initiale. Le métier de sculpteur ne s'improvise pas, l'erreur est irréparable : dans la pierre on ne rajoute pas comme dans la cire ou la glaise. Dans le bloc, on n'efface pas, on ne recouvre pas. Le choix de ce matériau impose une certaine rectitude prudente de l'ouvrage.
Rapportée à une masse textuelle, la démarche devient celle qui considère que tout peut être pertinent. Ensuite, ce que l'on choisit de ne pas retenir est désigné explicitement : on sait ce dont on se débarrasse, en refusant d'éluder le problème herméneutique que pointe (Rastier, Cavazza, Abeillé 1994, épilogue, §2.4) : « à quelles conditions une unité est-elle considérée comme inessentielle ? ». On cherche ce faisant, progressivement, à mettre au jour des motifs significatifs déjà présents dans le texte, plutôt que de projeter sur le texte une grille qui ne permettrait que de trouver ce que l'on s'est déjà donné.

La même répartition se retrouve pour les outils d'extraction de termes :

- la stratégie la plus commune commence par la sélection de candidats-termes conformes à un des patrons prédéfinis (ces patrons décrivent les conditions lexicales, morpologiques, syntaxiques ou/et sémantiques que doivent respecter les éléments successifs) ; elle s'interprète comme une approche par dotation, puisque les candidats termes sont reconnus à partir de formes données *a priori*.
- la démarche inverse, inaugurée et promue par LEXTER[4], consiste à définir plutôt ce qui ne peut pas être un candidat terme, et à repérer ce qui se comporte comme frontière, pour ciseler et isoler peu à peu les candidats terme, sans préjuger de l'ensemble des formes qu'ils peuvent prendre. L'approche est bien une approche par érosion.

b) Caractéristiques

L'approche par dotation se fonde sur un référentiel qui perdure, autonome vis-à-vis des analyses successives. L'intérêt est de pouvoir le rendre très détaillé, construit et cohérent par rapport à un cadre théorique, et de le modeler en profitant de l'expérience acquise.

D'autre part, ce référentiel garantit une description cohérente et fournit les éléments de comparaison d'un texte à l'autre : c'est en quelque sorte un langage pivôt[5]. Une approche purement par érosion fournit des représentations locales des textes, qui s'ignorent les unes les autres.

Mais un référentiel est nécessairement fini et limité[6]. Cela convient bien lorsque les unités d'analyse s'inscrivent par nature dans une grille connue à l'avance. En revanche, lorsque les unités de représentation ne peuvent être cernées *a priori*, le référentiel ne fournit pas une couverture régulière : il donne une vue partielle, et même partiale, car restreinte au point de vue qu'il traduit. Le référentiel peut ainsi être de moins en moins satisfaisant au fil des ans, compte-tenu de l'évolution des centres d'intérêt (nouveaux champs d'investigation et sujets qui ne sont plus à la mode), des connaissances (nouvelles notions et notions périmées), de la langue (formulations obsolètes de descripteurs) (Sta 1997, §3.2.2).

Le coût de la constitution et de l'entretien d'un référentiel est inévitablement important. Des outils peuvent suggérer les modifications à apporter (Sta 1997), mais le travail de validation demande un temps notoire et la mobilisation de compétences terminologiques, documentaires, scientifiques et techniques.

---

[4] LEXTER est un logiciel d'extraction de candidats termes, conçu et développé par Didier BOURIGAULT dans le cadre de sa thèse à la Direction des Etudes et Recherches d'EDF.

[5] Même les mérites de la normalisation peuvent être mis en question :
« Une des limitations actuelles des systèmes informatiques de recherche en texte intégral est une certaine intolérance à la polysémie. Dans cet exemple, on essaie de structurer le champ documentaire de façon à s'y retrouver à peu près bien et l'administrateur de la base est amené à décider que comme le mot *classeur* est équivoque, il faut arbitrairement le normaliser. Il décide qu'on ne parlera de classeur que dans un des champs où le terme se rencontre dans l'entreprise et pas dans l'autre. C'est ce genre de contrainte technique qu'à mon avis il faut essayer de lever parce que ces normes ne sont utiles que si on les connaît. Du moment où on ne les connaît plus, elles font obstacle à l'exploration des bases de connaissances. Il faudrait avoir des systèmes capables non seulement de tolérer, mais d'exploiter la polysémie. » (Poitou, Ballay, Saintive 1997, p. 15)

[6] Quelles que soient l'ambition et la mobilisation autour d'un réseau sémantique comme *WordNet*, ou de son petit cousin *EuroWordNet*. (Pour une présentation de *WordNet*, voir par exemple (Habert, Nazarenko, Salem 1997, §III.4, p. 85 sq.))
Les choix faits pour la constitution de *WordNet* appellent trois remarques :
- sa structure se moule sur celle des taxonomies classiques (reprises en Intelligence Artificielle) ; le typage des relations est hérité sans discussion, aussi la nature des liens enregistrés est-elle très pauvre.
- on commence par isoler, dans des réseaux séparés, les catégories lexicales (un sous-réseau pour les noms, un sous-réseau pour les verbes, etc.) : c'est une option résolument *sémasiologique* (liée à la forme apparente des items lexicaux), alors que le projet s'annonce sémantique, et devrait donc prendre une base *onomasiologique* (liée à la signification, au « vouloir dire »). (Pour une critique de l'approche sémasiologique, voir (Rastier 1991, §III.5, p. 104 sq.))
- le point le plus intéressant dans la définition de *WordNet* est le concept de *SynSet*, et la place qui lui est reconnue dans le réseau par son usage systématique : il rend bien compte du fait qu'un sens se détermine en contexte, ici avec la donnée d'un ensemble de mots associés (donc un contexte plutôt paradigmatique).

Quand les unités sont lexicales, le référentiel prend très rapidement des proportions très importantes, alors que l'approche par érosion commence par s'intéresser aux mots grammaticaux, dont les paradigmes sont fermés (liste des déterminants, des conjonctions, etc.).

La voie choisie pour DECID combine donc les deux approches, en se fondant préférentiellement sur l'approche par érosion. De petits référentiels fournissent les éléments pour des analyses selon certains points de vue précis (notamment les éléments de description de genres textuels privilégiés). Un référentiel adapté à la base des profils est construit périodiquement, à partir d'une approche par érosion qui permet de prendre en compte les innovations. En caractérisant des activités de recherche actuelles et en projet, DECID est en effet voisin de la problématique de la veille technologique, pour laquelle

> le niveau de technicité, tout comme la grande diversité et le caractère neuf des thèmes éventuels de veille, rendent improbable l'existence d'un thesaurus approprié. (Quatrain & Béguinet 1996, p. 11)

3. Rencontre du cercle herméneutique : quelques défis pour le calcul

a) Les unités ne sont pas données, elles sont construites

Ce qui est premier, c'est le texte, non la description systématique des unités de la langue. C'est à partir de l'observation des textes et des autres pratiques langagières que le linguiste bâtit des modèles et que le lexicologue établit des dictionnaires.

Le texte n'est donc pas un assemblage univoque d'unités prédéfinies, qu'il suffirait de traiter séquentiellement, l'une après l'autre. L'activité du lecteur consiste plutôt à construire des unités, dans une interprétation en contexte.

> Toute interprétation suppose une stratégie d'analyse qui précise les tactiques à employer, et garantit la pertinence des éléments retenus.
>
> [...] En soulignant cela, nous nous écartons du paradigme positiviste encore dominant dans les sciences sociales, et *a fortiori* dans les domaines techniques : il voudrait que les faits s'imposent d'eux-mêmes par une simple évidence, alors que nous avons à les constituer. Les signes linguistiques ne sont que le support de l'interprétation, ils n'en sont pas l'objet. Seuls des signifiants, sons ou caractères, sont transmis : tout le reste est à reconstruire. En d'autres termes, l'interprétation ne s'appuie pas sur des signes déjà donnés, elle reconstitue les signes en identifiant leurs signifiants et en les associant à des signifiés. L'identification des signes comme tels *résulte* donc de parcours interprétatifs. On voit que ces parcours diffèrent des procédures (au sens informatique du terme) qui opèrent sur des symboles déjà donnés, et, en tant qu'elles sont formalisées, peuvent le faire indépendamment de la signification de ces symboles.
>
> (Rastier, Cavazza, Abeillé 1994, §I.2.2, p.12)

Des choix interprétatifs s'opèrent : saisie en bloc d'une locution ou remotivation de ses constituants, non perception des formes homonymiques non pertinentes (ou au contraire reconstitution de leur gamme, dans un jeu de mots).

Dès lors, l'application d'un *principe de compositionnalité*, qui suppose *donnés* les composants à partir desquels calculer le sens du tout, fait problème : car comment sont délimités les constituants, et le tout ? et comment est déterminée la nature de leur ordonnancement ? Il y a *a minima* une *opération* d'individualisation, qui reconnaisse les (ou des ?) unités, et une *opération* d'identification de la (ou d'une ?) structure de composition.

> [Une étude historique du principe de compositionnalité fait apparaître] que rien [–aucune des notions sur lesquelles il s'appuie–] n'était évident ni donné au départ : ni l'identification des unités (que l'on pense à la *scriptio continua* qui rendait difficile l'identification et le repérage des constituants graphiques), ni la délimitation du 'tout' (proposition, phrase, énoncé ?), ni la nature des propriétés engagées (lexicales, sémantiques, syntaxiques, grammaticales). (Godart-Wendling 1998)

C'est encore un principe, le *principe atomiste*, qui érige en évidence que la signification se distribue sans reste sur les unités, et donc que chaque unité possède une signification fixée indépendante. Cela conduit immédiatement à une multitude de difficultés pour la description des homonymes et des termes polysémiques, qu'il faut morceler en unités indépendantes[7].

[7] Un bon aperçu de ces difficultés est donné dans (Gayral 1998) : absence de découpage évident de l'espace des significations, niveau de détail potentiellement infini, perte des relations entre les items une fois isolés, ignorance

L'utilisation d'un lexique, donc d'un ensemble d'unités *données*, dans un traitement automatique, revient à disposer d'un ensemble de résultats d'interprétation[8], que l'on s'efforce d'ajuster au texte. Le préfabriqué peut rendre compte d'un certain nombre de choses, il est commode, rapide et économique, mais, tout aussi modulaire soit-il, il reste limité. Surtout, il n'a pas la créativité, la souplesse et la finesse d'une construction en contexte. Une approche plus juste serait d'observer comment procèdent les mécanismes interprétatifs, et si ces mécanismes peuvent inspirer des procédures de construction dynamique d'unités.

b) Le global détermine le local

L'apprentissage de la lecture passe classiquement par des étapes méthodiques : identification de chaque lettre de l'alphabet et d'un son associé, traduction sonore d'une juxtaposition de lettres (le « b.a.-ba »), enchaînement des syllabes d'un mot et reconnaissance du mot. Très vite, insidieusement, ces principes généraux s'avèrent en léger décalage avec la réalité. Le *c* français rend des sonorités manifestement différentes selon son contexte immédiat (*ci*, *cas*, *chou*...). Le déchiffrement haché des syllabes modifie les sonorités et perturbe la reconnaissance du mot : un ajustement se fait dans le liage des syllabes et l'intonation, qui donne cette fois-ci consistance au mot. Et ainsi de suite : la lecture ânonnante, mot par mot, peine à donner sens au texte qu'elle déroule.[9]

La lecture commence en fait dès avant la première lettre ou le premier mot. La première perception est une perception d'ensemble. Elle concerne notamment le genre du texte, compte-tenu de la pratique dans laquelle on se situe.

> Les formes de langue et les formes types d'énoncés, c'est-à-dire les genres du discours, s'introduisent dans notre expérience et dans notre conscience conjointement et sans que leur corrélation étroite soit rompue. Apprendre à parler c'est apprendre à structurer des énoncés (parce que nous parlons par énoncés et non par propositions isolées et, encore moins, bien entendu, par mots isolés). Les genres du discours organisent notre parole de la même façon que l'organisent les formes grammaticales (syntaxiques). Nous apprenons à mouler notre parole dans les formes du genre et,

---

de la capacité à interpréter les néologismes. L'observation des comportements des unités en contexte aggrave encore l'inadéquation de principe atomiste : certaines unités présentent simultanément plusieurs sens, et la plupart voient leur sens évoluer progressivement au fil du texte.
Et de fait, les prétentions d'un tel *répertoire* complet des significations oublient le sage et sain équilibre qui guide l'établissement des *dictionnaires* de langue. L'objectif du dictionnaire n'est pas de consigner tous les sens que pourrait prendre un mot, mais de noter ses pôles de signification et ses usages conventionnels. Il peut faire l'économie de l'explicitation extensive de régularités qui permettent de décliner certaines familles de sens : glissements usuels, focalisation sur un aspect, etc. C'est un guide à l'interprétation, donnant des points d'appui pour la construction d'un sens en contexte. (Martin 1994) éclaire ainsi le malentendu qui règne autour de l'usage des dictionnaires pour les traitements automatiques, et fait une mise au point méthodique par rapport aux accusations qui ont été portées à l'encontre des dictionnaires courants : incomplétude, redondance, découpage artificiel en significations (rompant la continuité sémantique). Il précise alors comment concevoir l'apport des dictionnaires, leur réaménagement et leur rôle dans l'analyse automatisée.

[8] Marc CAVAZZA, en prenant la voie de la constitution manuelle fine d'un lexique sémantique, destiné à un contexte applicatif bien défini, confirme : « *la description n'est que de l'interprétation figée*. » (Cavazza 1996, p.62). Que l'on entende bien cependant, que le caractère « figé » n'interdit pas l'existence de procédures d'actualisation, c'est-à-dire qu'un élément sémantique n'est effectivement reconnu comme tel que s'il remplit certaines conditions contextuelles :
« On peut [...] proposer une définition simplifiée de la notion de lecture dans la perspective des applications informatiques : une lecture se compose d'un ensemble de descriptions initiales (le lexique sémantique) et d'un ensemble d'interprétants (interprétants syntaxiques, normatifs, argumentatifs, etc., formulés comme des règles d'inférences sur le contenu des sémèmes). En tant qu'interprétation résultante, elle produit des structures actualisées utilisables pour la description, et dont la validité dépasse celle du texte étudié pour s'étendre à une classe de textes du même genre. la formalisation des lectures ne permet sans doute pas d'automatiser de nouvelles lectures, mais peut être susceptible de réutiliser partiellement des lectures, ce qui suffirait à justifier pleinement l'approche. » (Cavazza 1996, p.67)

[9] Quant aux algorithmes de reconnaissance optique de caractères, et plus spécialement ceux dédiés à l'écriture manuscrite, ils ont clairement mis en évidence la quasi impossibilité de déchiffrer une lettre sans considérer son voisinage sur la ligne d'écriture. Le détail de l'image de la page qui correspond à une lettre, et que l'on présente isolé et grossi, laisse perplexe le lecteur humain... et la machine.

> entendant la parole d'autrui, nous savons d'emblée, aux tout premiers mots, en pressentir le genre, en deviner le volume (la longueur approximative d'un tout discursif), la structure compositionnelle donnée, en prévoir la fin, autrement dit, dès le début, nous sommes sensibles au tout discursif qui, ensuite, dans le processus de la parole, dévidera ses différenciations. Si les genres du discours n'existaient pas et si nous n'en avions pas la maîtrise, et qu'il nous faille les créer pour la première fois dans le processus de la parole, qu'il nous faille construire chacun de nos énoncés, l'échange verbal serait quasiment impossible. (Bakhtine 1984, p. 285)

Pour un document écrit, le support matériel influe immédiatement sur la perception du texte. S'agit-il d'un parchemin ? Je peux être prédisposée à reconnaître du latin. Un journal de petites annonces ? Je m'attends à une avalanche d'abréviations, que d'ailleurs j'aurai peu à peu appris à décoder sans le secours du dictionnaire. Un roman du siècle dernier ? Je m'amuserai de trouver un personnage ayant le même patronyme que mon voisin, sans avoir l'idée de les confondre. Une épaisse compilation philosophique ? Par delà les termes savants qui me sont inconnus, le titre, qui m'a motivée en annonçant le sujet débattu, oriente à l'avance ma lecture comme un aimant. Si je pratique une lecture rapide, la disposition d'ensemble des mots sur la page influence le choix des points où s'arrête mon œil.

Comme dans un mouvement d'approche, la perception globale précède et conduit à l'entrée locale dans le texte. La lecture se termine souvent par le mouvement inverse, récapitulatif : en quittant l'attention à un point de la chaîne des mots, il reste une impression, une idée, une émotion, attribuée au texte ou à l'un de ses passages (Bommier 1994b). Si ensuite la lecture reprend, le rappel de cette impression, cette idée ou cette émotion rendent plus saillant tous les éléments du texte qui y concourent. La détermination du local par le global est dans le registre des attentes, des présomptions, des anticipations, des intuitions, de l'accomodation.

La mise en valeur de l'incidence du global sur le local ne conduit pas à exclure l'effet réciproque, l'incidence du local sur le global. Certes, si l'on n'identifie des mots, ou si l'on ne repère les contributions au fil du texte au retour d'une notion ou d'une sonorité, l'interprétation ne peut se construire. Il reste que la perception globale est irréductible à une série de perceptions locales ; elle seule est en mesure d'embrasser les propriétés *holistiques*[10] de la réalité à décrire, en l'occurrence du texte.

Voilà bouclé un cercle qui peut laisser perplexe : pour comprendre un texte, je dois y reconnaître des mots –mais pour y reconnaître des mots, je dois en avoir une perception globale –et pour en avoir une perception globale, il faut bien que je saisisse l'ensemble des mots, etc. C'est le fameux cercle herméneutique. En dépit des inquiétudes du logicien, il s'agit là d'un *cercle vertueux* (Rastier 1993a), puisqu'il est constructeur de sens. Son retour sur lui-même ne lui confère par le caractère statique et l'uniformité d'un radotage, elle autorise plutôt une *circul*ation, une interaction, une respiration continuelle entre un aperçu local et un contexte d'ensemble. Chacun des parcours sur le cercle est un apport, le retour se fait un peu plus loin, le cercle se déploie comme en spirale ouverte[11], ou gagne profondeur et hauteur, une nouvelle perspective, dans une figure d'hélice[12].

---

[10] « Généralement, les sciences cognitives distinguent deux types de processus cognitifs. D'un côté, les processus holistiques construisent des représentations d'une manière globale et synthétique, alors que de l'autre, les processus analytiques combinent un ensemble de traits explicites et indépendants d'une manière fine.
[...] les propriétés holistiques, parmi lesquelles se trouvent les propriétés configurales, [...] correspondent à des dimensions intégrables décrivant l'organisation globale du monde physique observé. Elles se réfèrent à des relations entre parties du stimulus. Ce sont par exemple la symétrie ou la périodicité d'un motif.
[...] La cognition est alors le fait de la conjonction interactive des activités de ces deux types de processus. Tout d'abord, les modules holistiques, plus rapides, élaborent des heuristiques guidant la recherche des modules analytiques. Ces derniers structurent alors en retour le comportement des processus holistiques. C'est ce type de boucle informationnelle que nous modélisons dans le système MICRO. » (Antoine 1994, §1.III.2.1 et §1.III.3.2)

[11] Fin de (Rastier 1993a) : « [Les cercles] ne se referment pas sur eux-mêmes, et se muent en spirales ouvertes, qui reviennent à leur point de départ, mais plus loin. En effet, l'analyse sémantique est un immense travail de paraphrase (un sème n'est d'ailleurs qu'une paraphrase intralinguistique standardisée), mais cette paraphrase n'a rien de tautologique. Comme toute définition, elle apporte une connaissance –car il n'y a pas deux expressions identiques, ni même deux synonymes exacts.
C'est le sens qui se déploie, indéfiniment sans doute, dans la description des lectures. Car chaque lecture accroît le livre. »

Les deux moments, perception locale *vs* perception globale, ne sont pas la réplique symétrique l'un de l'autre. Le local est fréquemment un point d'entrée, et le global domine en dernière instance. Le local est fréquemment un point d'entrée : un détail qui suscite la recherche d'un ensemble, ou – comme nous le verrons plus loin– un sème qui suggère la présence d'une isotopie. Le global domine en dernière instance : il n'y a confirmation (ou infirmation) que dans le renforcement mutuel (respectivement, la dispersion générale) au plan d'ensemble.

c) La compositionnalité en question

Le principe de compositionnalité est respecté quand une règle établit le passage des constituants au tout. Autrement dit, le tout est entièrement décrit par l'agencement, la combinaison de ses constituants.

En sémantique, ce principe se transpose par exemple comme suit : le sens d'une expression se déduit du sens de ses constituants et de la construction qui les rassemble. Ce principe est mis en défaut par la réalité des textes, au point d'être définitoire des mots composés et de leur justifier une entrée particulière dans le dictionnaire.

Déjà le fonctionnement diachronique de la langue explique des décalages. Ainsi, un groupe de mots se fige en expression, et devient une entité autonome. Cette expression trouve sa place dans d'autres contextes d'emploi que ses composants initiaux pris isolément. Les sens du groupe et des composants évoluent séparément et s'éloignent. Dans d'autres cas, la forme complexe se définit bien dans la continuité de ses constituants, mais semble dotée d'un surplus de sens. Ce peut être le cas de la dénomination d'une discipline, d'une méthode, etc., par exemple *l'analyse des correspondances*. La personne qui ignore l'existence du corps théorique établit par J.-P. Benzécri peut ne lire dans ce groupe nominal que ce que lui livre ses constituants, sans se douter que l'on puisse y lire davantage : on effectue une analyse, qui s'intéresse aux correspondances entre plusieurs entités. Le lecteur féru de statistiques ne contredira pas cette lecture « naïve », mais il voit d'emblée comment est conduite cette analyse, dans quelle courant elle s'inscrit, etc. Dans les textes scientifiques et techniques, des initiales majuscules quelquefois signalent les dénominations établies (on aurait : *l'Analyse des Correspondances*).

d) Faut-il renoncer au calcul ?

Une inquiétante convergence

La non compositionnalité à l'œuvre dans la langue et dans le texte ne fait que redire, sous l'angle opposé, la détermination du local par le global. Affirmer que le local ne détermine pas le global (sans exclure qu'il puisse apporter sa contribution), c'est bien constater la non compositionnalité. De même, le fait que les unités ne soient pas données, mais construites, sape dès son démarrage tout calcul compositionnel[13].

---

[12] Dans nombre de cas, le retour exact au point initial est illusoire : la perspective trompeuse est celle qui considère l'hélice dans l'orientation de son axe (cf. les dernières pages du tome 2 de (Bommier 1993), sur *La troisième dimension du schéma de Chandon & Pinson*). En physique l'entropie entraîne dans une hélice descendante ; l'interprétation ouvre une hélice montante.

[13] « pour la linguistique formelle, le sens d'une expression est composé du sens de ses sous-expressions [, ce que l'on appelle compositionnalité sémantique]. [...] Cependant le sens des sous-expressions n'est aucunement donné, [...] il est construit en fonction de contraintes globales exercées par le discours (en tant qu'il reflète une pratique sociale), le genre du texte, et la situation concrète de communication. Si bien que les unités les plus simples, les traits sémantiques, ne sont pas le point de départ d'un parcours interprétatif, mais pour ainsi dire son aboutissement. Leur simplicité ne doit au demeurant pas faire illusion : ils ne sont élémentaires que par décision de méthode, et parce que l'on n'a pas besoin d'aller plus loin que ne le font le texte et la langue décrits. En fait, un trait sémantique n'est pas moins complexe que les unités de rang supérieur dans la définition desquelles il entre. Il est simplement plus précis, en ceci qu'il résulte de leur analyse, et que le parcours du global au local n'est pas une simple décomposition, non plus que le parcours inverse n'est une composition. L'acte même de l'analyse ou de la synthèse perceptive ou descriptive modifie l'appréhension et la nature des unités initiales ou finales du parcours interprétatif. Les 'unités' ne sont pas déjà données et identiques à elles-mêmes mais résultent

Cette dernière formulation, l'affirmation de la non compositionnalité[14], interpelle cependant plus crûment l'informaticien. Les calculs qu'effectue l'ordinateur sont intrinsèquement compositionnels : les fonctions sont *fonction de* leurs arguments, depuis les opérations élémentaires (addition, concaténation,...) jusqu'au compositions complexes et particulières au programme. De même, l'algorithmique séquentielle ou parallèle fait-elle place à l'introduction d'un cercle herméneutique ? *Cercle* renvoie plutôt en écho bouclage et divergence...

Une compositionnalité désirable

Comme l'explique (Nazarenko 1998), le principe de compositionnalité fournit « une méthode de calcul précieuse » :

> La méthode de calcul induite par le principe de compositionnalité est par ailleurs incrémentale et monotone. On peut élaborer une structure complexe à partir des structures les plus élémentaires, en construisant pas à pas des structures de complexité croissante (incrémentalité). Les structures sémantiques sont construites une fois pour toutes : une sous-structure ne peut pas être remise en cause lors de son intégration dans une structure plus complexe (monotonie). Ces deux propriétés sont importantes parce qu'elles caractérisent un type d'opérations que l'on sait modéliser, pour lesquelles on dispose de modèles de calcul [...]. (Nazarenko, p. 4)

Une optique compositionnelle des traitements de la langue va de pair avec une attention importante portée au lexique (puisque ce sont les unités qui fourniront toute la matière des résultats). En outre, tout l'investissement de recherche consacré à la syntaxe ces dernières décennies semble ainsi justifié et valorisé : la syntaxe se prête tout naturellement à la description des modes de combinaison des unités lexicales[15]. On tient donc les deux données requises : les composantes, et la fonction de composition.

Des formalismes sont développés pour rendre compte d'opérations de composition toujours plus complexes. Très populaire, l'unification est une mise en relation avec apport d'informations. Elle permet de représenter, tout en restant dans un cadre de calcul compositionnel, des phénomènes d'influence entre constituants (Nazarenko 1998). La traque du non compositionnel est assimilée par certains à la progression de la formalisation du domaine linguistique, au point de considérer que « ce qui n'est pas compositionnel, c'est souvent ce qui est trop peu ou mal formalisé. C'est presque dire que « ce qui se formalise bien se 'compositionnalise' aisément » (Nazarenko 1996)...

Enfin, on trouverait satisfaisant d'expliquer ainsi, en se calquant sur le raisonnement calculatoire, les capacités interprétatives prodigieuses de l'homme, qui sait comprendre ce qu'il n'a jamais entendu. Tout procéderait d'une combinatoire économique : à partir d'un nombre limité (voire réduit) d'atomes à combiner, et d'un jeu de règles permettant de calculer le sens d'une combinaison, seraient produites un nombre indéfini de significations. Comme nous l'avons vu, l'explication est tentante, mais réductrice et démentie par la réalité linguistique.

Des calculs qui font place à la dimension globale

La lecture machinale reste en deçà de l'interprétation humaine. Mais il existe des voies pour introduire les propriétés mentionnées (*i.e.* détermination du local par le global, *etc*.). Les statistiques

---

de processus de discrétisation et de stabilisation toujours modifiables. » (Rastier, Cavazza, Abeillé 1994, §II.5, p. 37)

[14] La famille de *compositionnel* / *compositionnalité* renvoie de multiples échos : articulation entre un tout et ses parties (*composant*, *composante*), assemblage d'éléments autonomes (*composite*), opération (*composition*) et raisonnement analytique (*composé*).

[15] Mais évidemment, la syntaxe échoue à décrire le texte au-delà du palier de la phrase, et la compositionnalité qui reposait sur elle est démunie de la même manière face à la textualité.

Autres obstacles : rendre compte de la compréhension effective d'énoncés que la syntaxe refuse comme agrammaticales (Godart-Wendling 1998, p. 19), des « ambiguïtés syntaxiques » (quand la structure syntaxique n'est déterminable qu'en faisant appel à la sémantique globale du contexte),...

Il est maintenant de notoriété publique –l'expérience l'a démontré– que les traitements qui enchaînent séquentiellement analyse morphologique, analyse syntaxique, analyse sémantique, analyse pragmatique, ne sont pas en mesure de produire une analyse satisfaisante.

Sur tout ceci, voir aussi (Gayral 1998).

ou les réseaux de neurones multicouches, dont chacun sait les solides mises en œuvre informatiques, se situent déjà dans une perspective d'analyse globale des données. Ces types de traitements ont également l'avantage de la robustesse : en dépit d'éventuelles difficultés d'analyse locales, ils proposent un résultat[16].

A la crainte d'une impossibilité informatique, se substitue une autre manière d'envisager les traitements. La compositionnalité à base de primitives et de règles de syntaxe s'assouplit en une combinatoire, qui n'est plus prédéterminée, explicitée *a priori*[17]. L'enchaînement de procédures séquentielles et la monotonie (mathématique) du traitement, fait place à des modules en interaction, à des composantes qui communiquent et s'influencent mutuellement, et peuvent réajuster un résultat en cours de construction[18]. La préinterprétation, codée dans le lexique, s'estompe devant la description de mécanismes sémantiques d'activation, de propagation, d'inhibition, d'afférence,[19] qui traduisent les résonances en contexte, les interférences constructives et destructives.

Pour un corpus de textes, le traitement peut prévoir une pré-analyse locale, puis remodeler les résultats de cette analyse en fonction de la vue d'ensemble qu'il a à l'issue de cette première passe[20]. Les unités de la pré-analyse ne sont alors pas définitives ; elles n'ont pas non plus à être exclusives : la construction de nouvelles unités, construites à partir des premières mais non redondantes avec elles, peut être prévue. L'ordinateur s'arrête là où le véritable travail interprétatif commence. Ses résultats visent à mettre en valeur des points d'appui, des configurations, à partir desquels l'utilisateur élabore sa propre lecture. Ceci présage le rôle essentiel de l'interface.

[16] Des indicateurs de fiabilité et de représentativité guident l'interprétation, pour qu'elle puisse être lucide et critique. Il faut ajouter à cela « l'avertissement malicieux de L. Lebart [LEBART Louis (1975) - *Analyse des correspondances et méthodes dérivées*, Ecole d'été d'analyse numérique, C.E.A. - I.R.I.A. - E.D.F., §VI.1] : 'Les méthodes d'analyse factorielles [...] ont un assez grave inconvénient : elles fournissent toujours un résultat !' Ce que nous prenons la liberté de comprendre comme une confirmation que le résultat ne se définit pas par lui-même, mais en tant qu'aboutissement (ou étape) d'un cheminement expliqué. » (Bommier 1993, §I.V.a, p. 37)

[17] Dans le domaine de la recherche d'information, c'est le passage de l'interrogation par équation booléenne aux modélisations par espace vectoriel et calculs de similarité.

[18] Les limites ont été largement dénoncées, d'une conception qui définit morphologie, syntaxe, sémantique, et pragmatique, comme autant de domaines étanches, en forte complémentarité (par exemple, ce qui déborde de la sémantique échoit à la pragmatique), et intervenant dans cet ordre, irréversiblement. C'est pourtant encore le fil conducteur qui organise le découpage et l'ordre des chapitres des ouvrages classiques de synthèse sur le Traitement Automatique du Langage Naturel (Sabah 1989) (Fuchs & *al.* 1993).
Les nouvelles architectures logicielles modulaires ont fleuri ces dernières années, dans la modélisation de procédures « intelligentes ». Parmi elles, *tableau noir* et *systèmes multi-agents* explicitent et formalisent les passages de la détermination du local par le global ; les modèles *statistiques* et *neuronaux* offrent des descriptions moins analytiques, plus synthétiques.

[19] François RASTIER décrit ainsi les opérations d'actualisation, de construction, et de virtualisation d'un sème.

[20] Jean-Yves ANTOINE implémente ainsi, dans son système MICRO, une phase dite d'*amorçage* (sémantique), qui se dédouble en *amorçage isotopique* (plutôt paradigmatique) et *amorçage relationnel* (plutôt syntagmatique) (Antoine 1994, §6.II.2).

B. EXAMEN DE TRAITEMENTS LINGUISTIQUES AUTOMATIQUES EXISTANTS

1. Orientations d'utilisation des traitements courants

a) DECID a-t-il besoin d'un traitement linguistique ?

Une des quatre facettes du texte est qu'il est rédigé dans une langue. Sa matière est linguistique.

Les traitements effectués respectent donc d'autant mieux l'essence du texte si, par l'intermédiaires de méthodes issues de la lexicométrie, de modes de représentations de documents textuels, ou d'outils de Traitement Automatique du Langage Naturel (TALN), ils peuvent s'appuyer sur des propriétés linguistiques des données ou avoir accès à des informations linguistiques qui leurs sont utiles.

Cette étude sur la manière opportune de recourir aux outils linguistiques est loin d'être vaine ou spéculative. Dans la communauté de l'informatique linguistique, certains s'interrogent : « Tous ces outils d'analyse syntaxique c'est très beau, mais une fois qu'on a décortiqué ses phrases et qu'on a remplacé son texte par une collection d'arbre syntaxiques ou une série d'étiquettes, qu'en faire ? qu'est-ce que cela nous apporte ? » Des réponses sont données du côté de l'étude des phénomènes linguistiques et de la description des usages linguistiques (Habert, Nazarenko, Salem 1997), ce qui ne préjuge pas de l'utilité et de l'adéquation de ces ressources pour une application comme DECID.

D'autres constatent aussi : les systèmes les plus performants et les plus robustes ne sont pas ceux qui reposent sur l'appareillage théorique et linguistique le plus élaboré, mais ceux qui font un traitement brutal, rudimentaire, heuristique[21]. Et d'en prendre pour illustration les résultats des conférences / bancs d'essai MUC[22], des moteurs de recherche qui trouvent leur place sur Internet (Smeaton & *al*. 1995) (Church & Mercer 1993).

Ce n'est pas parce que le lecteur peut reconnaître des informations d'ordre linguistique que le système doive les utiliser. Le double écart que résume la phrase précédente est celui de la pertinence et du non nécessaire mimétisme. *Pertinence* : ce n'est pas parce qu'une information est valide qu'elle est utile. *Non nécessaire mimétisme* : l'objectif ne détermine pas nécessairement les moyens. Le traitement artificiel d'une machine n'est pas nécessairement meilleur en essayant d'imiter au plus près la vision que l'on a des mécanismes naturels. L'illustration classique est celle de l'avion, dont le vol n'est pas obtenu par le battement des ailes.

Reste donc à cerner ce que DECID peut prendre en compte avec profit, compte-tenu des corpus pour lesquels il est conçu, et de l'application qui consiste à mettre en relation des textes entre eux, quand ils partagent un pôle d'intérêt. Les solutions les plus sophistiquées ne sont pas nécessairement les plus opportunes. De premières pistes sont déjà indiquées par (Renouf 1993a) : face aux techniques de recherche d'informations, la linguistique appelle à reconnaître la diversité des unités et de leur rôle, la question non évidence de leur segmentation dans le fil du discours, la relativité des choix représentationnels du thesaurus, le passage non univoque des mots à la thématique du texte. L'examen est conduit ici par type d'outils : comment envisager l'utilisation de chacun dans le cadre de DECID.

Dans la mesure du possible, la mention de logiciels existants et sérieux est faite, car le but de ce travail n'est pas de réinventer la roue ! Recourir aux outils existants permet de bénéficier des années d'expérience –et parfois du génie de leurs concepteurs– qui ont été nécessaires pour les mettre au point. Il faudrait ajouter les serveurs, généralistes, qui donnent accès à toute une gamme de traitements : le Département SID à la DER d'EDF dispose d'un tel serveur, réalisant des analyses

[21] « In summary, one concludes with an apparently counterintuitive observation : the more simple-minded approaches to text processing and language understanding produce the fewest mistakes and exhibit the best performance. » (Salton 19888, p. 269)

[22] MUC : *Message Understanding Conference*.

morpho-syntaxiques ; en France, le projet SILFIDE[23] prévoit de donner un accès centralisé à toutes sortes de ressources linguistiques (corpus, outils, lexiques). Quant aux traitements les plus frustes, ils ne donnent pas lieu à la mention d'outils particuliers : l'algorithme est simple et se passe de ressources grammaticales et lexicales ; l'habitude est alors de les réécrire soi-même si besoin est.

b) Les n-grammes

La représentation d'un texte par des *n*-grammes consiste en l'ensemble des suites de *n* caractères présentes dans ce texte. Chaque lettre est un caractère, chaque espace entre deux lettres est un caractère, les ponctuations sont ignorées. Les espaces font donc partie des suites de *n* caractères représentatives ; cependant usuellement on ne considère pas les suites qui chevauchent un espace : les espaces, quand il y en a, sont au début ou/et à la fin des *n* caractères. Par conséquent, le début de ce paragraphe fournirait les *n*-grammmes suivants (pour *n* égal à 3) : *_La*, *La_*, *_re*, *rep*, *epr*, etc. (qui ne comprend pas le motif *a_r*).

Ce type de représentation n'a rien de linguistique, sinon qu'il se justifie à partir de la double articulation du langage, qui se reflète (imparfaitement[24]) dans l'écriture alphabétique. Ainsi, les espaces sont la délimitation des mots, unités que compose diversement la syntaxe. D'où la condition de ne pas relever les suites de caractères s'étendant de part et d'autre d'un espace : issues de deux mots, ce sont des suites hétérogènes. En revanche, l'espace en position initiale ou finale est significatif en ce qu'il marque le début ou la terminaison d'un mot. Après l'articulation en mots, vient l'articulation en lettres. Tout mot (français) s'écrit à partir d'un alphabet de 26 lettres (un peu plus si l'on ajoute les caractères diacritiques ; mais cela ne change pas le raisonnement). Un nombre extrêmement petit d'unités-lettre permet donc la représentation de tous les mots de la langue.

Seulement, ces unités-lettre sont par conséquent peu représentatives de chacun des mots. Savoir que mon texte comporte toutes les lettres de l'alphabet, et beaucoup de *e*, ne me renseigne guère sur les mots qui le composent et sur son contenu[25]. Les mots me renseigneraient davantage, mais leur nombre total est trop grand. En considérant des *n*-grammes, les unités sont plus informatives que les lettres, et leur nombre reste limité. Les utilisations les plus fréquentes se font pour *n* égal à 2 (digrammes), ce qui correspond à au plus $27 \times 26 \times 2 - 27$ unités différentes, *n* égal à 3 (trigrammes) soient au plus $27 \times 26 \times 27$ unités, ou *n* égal à 4 (quadrigrammes ou tétragrammes) soient au plus $27 \times 26 \times 26 \times 27$ unités.

Intermédiaires entre la lettre et le mot, à l'échelle de quelques lettres, les *n*-grammes se situent au niveau de la syllabe, et des éléments morphologiques qui composent les mots : préfixes, racines, suffixes, flexions. En effet, les mots qui ont le même préfixe, ou la même racine, ont les *n*-grammes correspondants en commun dans leur représentation. En ce sens, les *n*-grammes peuvent constituer une forme de réduction flexionnelle et dérivationnelle : les préfixes et suffixes qui donnent des formes différentes aux mots d'une même famille, ou les terminaisons de conjugaison, de genre, de nombre, qui séparent les variations d'une unité lexicale, empêchent de retrouver la racine commune au niveau des mots (non analysés), alors que les *n*-grammes y donnent accès.

L'approche par *n*-grammes a trouvé sa place dans les travaux autour du multilinguisme pour deux raisons. La première, c'est que la méthode est directement utilisable dans n'importe quelle langue alphabétique, en particulier il n'y a besoin d'aucun lexique, aucun dictionnaire, aucune grammaire qui seraient spécifiques à la langue. La seconde, c'est que pour des langues ayant des origines apparentées, certaines racines peuvent se retrouver d'une langue à l'autre : les *n*-grammes donnent alors un terrain de représentation commun intéressant pour les langues considérées.

---

[23] SILFIDE : *Serveur Interactif pour la Langue Française, son Identité, sa Diffusion et son Etude*. Ce projet est mené conjointement par le Centre de Recherche en Informatique de Nancy (CRIN-CNRS), l'Institut National de la Langue Française (INaLF-CNRS), le Laboratoire Parole et Langage (LPL-CNRS), le Groupe d'étude pour la traduction automatique (Geta), et le Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI-CNRS). Son adresse Internet est :
`http://www.loria.fr/Projet/Silfide/`

[24] Aussi tout le développement qui suit, qui néglige cet écart, ferait à bon droit bondir le linguiste.

[25] Sinon qu'il ne s'agit pas d'un certain roman de Pérec.

La grande faiblesse des *n*-grammes, c'est leur définition purement formelle (suite de *n* lettres), qui rend obscure leur interprétation. Un *n*-gramme isolé est rarement associable à un ensemble sémantiquement cohérent de mots. Un ensemble de *n*-grammes est encore moins décryptable ; et la combinatoire sous-jacente est en fait très lourde.

Cette difficulté majeure d'interprétation de la représentation, et partant de là des résultats des calculs, est une motivation majeure pour ne pas intégrer cette technique dans DECID. De plus, comme l'analyse en *n*-grammes commence par réduire le texte en confettis, elle ne se combine avec aucun autre traitement linguistique, et est donc exclusive à l'égard d'autres analyses, ou reste à l'écart. Cette fermeture n'est pas non plus souhaitable.

c) Le découpage ; les ponctuations

Le découpage se fonde sur l'articulation en mots[26]. Réalisé automatiquement, il s'en tient à la délimitation matériellement marquée par les espaces, qui n'est qu'une approximation de la délimitation des unités lexicales. Les tirets et les apostrophes sont souvent aussi comptés comme délimiteurs (*l'articulation* est scindé en article et nom, *prend-il* en verbe et pronom), au détriment de formes comme *aujourd'hui* ou *peut-être*.

Un dictionnaire peut recenser des cas de regroupement des mots graphiques : locutions, mots composés. La tâche d'explicitation de toutes les formes composées n'est en soi jamais achevée (cf. les dictionnaires, très approfondis mais toujours en chantier, du LADL à Paris), et les critères qui déterminent l'existence et la présence de la forme composée ne sont pas tranchés.

C'est dès un simple découpage que se pose la question de la prise en compte des signes de ponctuation. Ils doivent être détachés du mot précédent, le cas échéant, et font aussi office de séparateurs. Dans certains cas, la ponctuation fait partie du mot lui-même : abréviations (*etc.*, *M.* pour *Monsieur*), sigles (*E.D.F.*), chiffres (*12.345,67* –avec un rôle inversé des points et des virgules dans le monde anglophone). Son rôle de délimiteur d'unité lexicale et de contribution au rythme prosodique est alors inhibé.

Souvent négligées (et donc éliminées par le découpage), les ponctuations ont un rôle sémantique fort[27], rappelé par des études linguistiques récentes[28]. Leur présence habituelle dans tous les corpus, et leur fréquence élevée, fait des ponctuations un bon objet d'analyse par des mesures statistiques.

Les ponctuations sont elles-mêmes des unités sémantiques, qu'on peut appeler *ponctèmes* : elles peuvent avoir des affinités thématiques (Dupuy 1993), et contribuent par exemple à l'expression d'un sentiment (phrases hâchées, ton monotone ou éclats de voix, etc.) (Bonhomme & *al*. 1996, §V.I).

---

[26] Autant la notion intuitive et générale de mot est comprise de tous (et c'est ce qui nous permet dans ces lignes d'utiliser ce terme), autant la définition linguistique et précise du mot est problématique (Martinet 1986) (Tyvaert 1998).
Dans cette thèse, nous emploierons néanmoins le terme *mot*, soit comme un moyen d'éviter d'alourdir le discours (en renvoyant à la notion intuitive et générale), soit en lui associant une définition technique provisoire (par exemple : 'un mot est une suite de lettres délimitée par des séparateurs').

[27] Les ponctuations sont notamment recensées au titre des *interprétants*, contribuant à une analyse sémantique (Cavazza 1996, p. 74).

[28] (Bonhomme & *al*. 1996, § V.I.1) cite des numéros spéciaux de *Langue française* (La Ponctuation, n°45, 1980), *Traverses* (Le génie de la ponctuation, n°43, 1988), *Pratiques* (La ponctuation, n°70, 1991).
La communauté de l'informatique linguistique a fait rentrer cette question dans le champ de ses préoccupations. Le 28 juin 1996 s'est tenu un *International Workshop on Punctuation in Computational Linguistics* (dans le cadre de SIGPARSE / *34th Annual Meeting of the Association for Computational Linguistics*). L'appel (diffusé sur la liste électronique LN le 27 février) fait de la ponctuation un phénomène très général :
« Almost any structure-giving, or graphical, device in text could be described as punctuation - this means that punctuation falls into roughly three categories :
- within word : marks like hyphens and apostrophes ;
- between words : what we conventionally think of as punctuation, e.g. commas, full-stops, colons ;
- higher-level graphical punctuation : paragraphing, indentation, underlining, font changes etc... »
Egalement, un numéro de *Computers and the Humanities* (vol. 30, n°6, 1996) est consacré à la ponctuation : *Current Approaches to Punctuation in Computational Linguistics*, B. SAY & V. AKMAN (dir.).

Les ponctuations délimitent également des zones d'interaction sémantique de degré intermédiaire : elles régulent les propagations de traits sémantiques, et jouent donc à la fois le rôle de délimiteur et de connecteur. Elles structurent notamment les énumérations. Autre exemple, les guillemets. Ces délimiteurs se chargent de multiples significations[29]. Autour d'un mot, ce peut être la dénégation, l'emprunt à un autre registre de langue, l'emphase, etc. Marquant une citation plus longue, ils introduisent un changement de ton, de point de vue, etc. Leur nécessaire décryptage interprétatif traduit déjà la connivence qu'instaure le rédacteur avec son lecteur (Maingueneau 1991, pp.140 sq.).

d) Catégorisation : Etiquetage par les parties du discours (nature des mots)

Un outil qui découpe les unités lexicales et les identifie en tant que *nom*, *verbe*, etc, est appelé catégoriseur[30]. Souvent le catégoriseur ajoute quelques informations flexionnelles : par exemple, il distingue *verbe conjugué vs infinitif*.

L'usage d'un catégoriseur peut être envisagé de deux façons. Soit le catégoriseur est considéré comme un découpage amélioré : outre la décomposition de la chaîne textuelle en unités lexicales, il ajoute une information linguistique sur la nature des unités découpées. La question est alors celle de l'utilité de cette information pour le traitement envisagé. Soit le catégoriseur est une étape intermédiaire : il est le préalable requis à un extracteur de terminologie par exemple. La question de la pertinence de son utilisation est alors reportée. C'est donc le premier cas d'usage qui est considéré dans ce paragraphe, celui de l'utilisation directe des unités fournies par un catégoriseur pour construire la représentation d'un texte.

Supériorité sur le découpage pour la délimitation linéaire des unités

Du point de vue de la délimitation des unités, un catégoriseur est supérieur à un découpage sur caractères délimiteurs. Il est capable de reconnaître les locutions les plus courantes, celles qui constituent une unité grammaticale (*parce que*, *auprès de*, *à partir de*, *si bien que*) ou lexicale (*mettre au point*), et donc de regrouper leurs constituants comme une seule unité. Cela évite des confusions incongrues qui se produisent lorsque l'on retient comme caractérisation les noms ou adjectifs issus de *par rapport à*, *au fur et à mesure*, *en vertu de*, *de nouveau*, etc. Le traitement des tirets peut être ajusté, du cas où il est maintenu (*passe-passe*) au cas où il rattaché à un des composants (*-même*, *-là*). A l'inverse, il peut analyser des formes contractées en leurs composants élémentaires (*du* en *de* et *le*), et les rend ainsi comparables aux formes non contractées du même paradigme (*du*, réécrit *de le*, est rapproché de *de la*). Le catégoriseur donne donc un découpage fondé linguistiquement, et plus régulier et plus homogène que celui du découpeur : c'est un point fort pour motiver son utilisation dans DECID.

Le jeu d'étiquettes : un choix heuristique

Le jeu d'étiquettes varie d'un catégoriseur à l'autre. Il dépend à la fois de ce que l'on *veut* et de ce que l'on *peut* reconnaître.

Du fait de l'algorithme qui régit son fonctionnement, tel catégoriseur ne fait pas bien la part entre *participe passé* et *adjectif*, qui ont « en surface » le même comportement : on ne définit alors qu'une seule étiquette qui regroupe les deux possibilités. Quelques mots particuliers peuvent même nécessiter l'introduction d'une catégorie syncrétique spéciale, qui ne sert qu'à eux (par ex. *que*). A l'inverse, certaines distinctions sont essentielles au bon fonctionnement : par exemple, repérer les

[29] Il faut se méfier que les guillemets ne sont pas toujours délimiteurs : il arrive que, pour tout un passage qui est mis entre guillemets, les guillemets ouvrants soient rappelés en début de chaque ligne. Ils jouent alors plutôt le rôle d'un élément de présentation, de mise en page, en dehors de la succession des mots.

[30] L'anglicisme *taggeur*, littéralement « étiqueteur », désigne aussi ce type d'outils. Il s'étend plus généralement à toutes formes d'étiquetage, notamment avec des catégories sémantiques.

pronoms relatifs ou les *clitiques*[31] parmi les *pronoms*. L'algorithme dicte ainsi la formation de certaines catégories.

L'intérêt de l'étiquetage fournit repose aussi sur la finesse suffisante du jeu d'étiquettes : on ne veut pas que soient confondus un nom et un verbe par exemple. Cependant, d'une façon générale, plus le jeu d'étiquette est détaillé, plus la procédure d'étiquetage est lourde (nombre de règles, combinatoire des cas à parcourir), et plus sont grands l'indétermination (*i.e.* l'algorithme ne sait pas trancher entre plusieurs solutions d'étiquetage) et les risques d'erreur (*i.e.* l'algorithme tranche, mais retient une solution incorrecte). Une trop grande variété d'étiquettes suscite des difficultés interprétatives : deux personnes différentes ne s'accorderaient pas sur l'étiquetage correct d'un passage ; une même personne peut faire des choix d'étiquetages non cohérents, à des moments différents. Alors qu'un étiquetage trop grossier n'apporte pas d'information significative, la tentation d'un étiquetage trop fin se solde par une dispersion confuse des analyses.

Le jeu d'étiquettes est donc un compromis pour avoir un fonctionnement efficace, fiable, et produisant des catégories d'un niveau de détail adéquat pour être utiles. C'est en général l'aboutissement d'une recherche heuristique itérative, lors de laquelle le jeu d'étiquettes est peu à peu ajusté. Aussi la comparaison directe de deux catégoriseurs n'est-elle pas possible, du fait que leurs jeux d'étiquettes ne sont pas superposables : elle se fait par l'intermédiaire d'un jeu d'étiquettes de référence, constitué autant que possible comme un plus grand dénominateur commun, auquel rapporter les étiquettes de chaque système.

L'attribution des étiquettes

Indécision

La performance d'un catégoriseur se situe toujours comme un compromis entre *précision* (taux d'étiquettes correctes) et *décision* (taux d'étiquettes univoques) [32]. A partir d'un certain degré de finesse, la décision n'est jamais complète : l'expert humain lui-même ne peut trancher entre plusieurs étiquettes pour une même forme. L'indécision peut être perçue comme négative : il n'a pas été possible de trancher par manque d'information contextuelle, mais un complément d'information éliminerait certaines étiquettes ; les étiquettes sont en disjonction exclusive. L'indécision peut à l'inverse avoir une signification positive et traduire une égale admissibilité, voire une coexistence, des interprétations. L'indécision purement négative et l'indécision purement positive sont deux extrêmes entre lesquelles se positionnent les cas réels. En raison de ce reste d'indécision irréductible, il est discutable d'exiger qu'un catégoriseur propose toujours une étiquette et une seule pour chaque élément. Pour les cas de figure les plus fréquents, des catégories composites peuvent être prévues en complément aux catégories résolues : par exemple, un jeu d'étiquettes comporte « adjectif ou participe passé » et « adjectif ou participe présent », tout en permettant d'enregistrer plus précisément les valeurs « adjectif » (pour une forme non verbale, une construction épithète), « participe passé » (pour un verbe à un temps composé avec l'auxiliaire *avoir*), et « participe présent ».

Incertitude et inconnu

Une catégorie spéciale peut être prévue pour rendre compte des cas où l'étiqueteur ne trouve pas du tout de catégorie morpho-syntaxique à affecter à une unité donnée.

Il y a aussi le cas intermédiaire où le catégoriseur n'identifie pas clairement la catégorie d'une unité (en général parce qu'elle n'est pas recensée dans son lexique), mais peut néanmoins proposer une catégorie plus vraisemblable que les autres en fonction de certaines heuristiques. Les critères sont morphologiques (une majuscule initiale indiquant un nom propre, un suffixe ou une terminaison caractéristique) et syntagmatiques (ce qui est précédé de « n' » est un groupe verbal). Il est

[31] Les clitiques sont les pronoms personnels adjacents au verbe, par exemple les trois pronoms de *je le leur ai donné* ; *lui* et *moi* sont des pronoms personnels qui ne sont pas des pronoms clitiques.

[32] Les concepts de *précision* et de *décision* pour l'évaluation des catégoriseurs morpho-syntaxiques ont été mis au point et utilisés lors de la compétition GRACE (1995-1998). Un article de Martin RAJMAN et Patrick PAROUBEK est prévu sur ce sujet dans un numéro spécial de la revue *Langues*.

souhaitable que le catégoriseur puisse indiquer la catégorie qu'il trouve, tout en marquant le caractère moins fiable de l'étiquetage de cette unité par rapport aux autres.

Utilisation des informations sur la nature des mots

Conventionnellement, les classiques parties du discours (substantif, verbe, adverbe, etc.) sont distinguées. En effet, ces distinctions correspondent à la description qu'a forgé la linguistique pour rendre compte des constructions et enchaînements syntagmatiques. D'autre part, ce sont les distinctions que demandent la plupart des applications qui utilisent les résultats d'un catégoriseur. L'étude de l'utilisation possible par DECID de l'information apportée par les étiquettes d'un catégoriseur peut donc commencer en considérant chacune de ces grandes catégories.

Nom

Les procédés morphologiques et les constructions qui permettent de transformer une unité linguistique en une forme nominale sont extrêmement productifs. C'est une des raisons qui a favorisé l'usage conventionnel des noms comme forme normalisée pour exprimer un concept[33]. Le terme (d'une terminologie), le descripteur (d'un thesaurus), le mot-clé (d'un langage documentaire) adoptent une forme nominale. Un certain nombre de traitements et d'analyses utilisent cette dérivation vers le nom pour réduire les variantes d'expression[34]. L'erreur serait d'en déduire que seuls les noms ont une charge significative[35].

---

[33] Pour une analyse critique de la précellence du nom dans les terminologies et les ontologies, voir (Rastier 1995c, §I.A).

[34] Une application documentaire (Fluhr 1977, §III.4.2.1) recourt à une telle « normalisation vers le concept ».
La description structurale proposée par (Greimas 1966, §IX.2.d) rapporte tout son lexique à des substantifs, ce qui évite de disperser le mode d'expression dans deux procédés parallèles, à savoir l'affectation à une catégorie morpho-syntaxique (dite classe grammaticale) ou le choix d'un suffixe adéquat :
« On sait que les langues naturelles possèdent, en général, deux systèmes caractérisés de lexicalisation : le premier consiste à verser les sémèmes dans les classes grammaticales (verbes, adjectifs, etc.) ; le second procède par dérivation. Ainsi, tout sémème fonctionnel peut, en principe, être lexicalisé soit comme verbe : *résoudre*, *marcher*, *déménager*, etc., soit comme substantif déverbal : *solution*, *marche*, *déménagement*, etc. De même, tout sémème qualificatif peut se présenter soit comme adjectif : *long*, *certain*, *intransitif*, etc., soit comme substantif dérivé : *longueur*, *certitude*, *intransitivité*, etc. Cette redondance naturelle ne peut être qu'une source d'hésitation dans la pratique de la description. [...]
En face de telles ambiguïtés, il paraît plus économique d'éliminer l'un des deux procédés de dénomination, en excluant la lexicalisation par classes grammaticales, et d'adopter une procédure unique, qui ne conserve la motivation lexicale des classes de sémèmes que par le seul moyen de la dérivation suffixale. L'opération consiste :
1. A attribuer à tous les sémèmes la forme substantivale : comme il ne restera plus d'autres classes grammaticales auxquelles il pourra être opposé, le substantif, en tant que classe, se trouvera ainsi neutralisé ;
2. A lexicaliser les sémèmes par l'adjonction des seuls suffixes substantivaux appropriés : *-ement*, *-age*, *-tion*, zéro, etc., lorsqu'il s'agit de fonctions ; *-ité*, *-itude*, *-ance*, *-eur*, etc., pour lexicaliser les qualifications. Dans les cas où les moyens dérivatifs font défaut, les procédés périphrastiques du type *le fait de*... devront être employés.
La description systématique des classificateurs (ou des définissants) utilisés par la lexicographie, et qui sont des synonymes, ou des équivalents, au niveau des définitions, des suffixes employés au niveau de la dénomination, pourrait être, à ce stade, d'un grand secours. »

[35] C'est un peu le reproche que l'on pourrait adresser à certains travaux de construction automatique de thesaurus ou de graphes de concepts. Par exemple :
« Les mots appartenant à la catégorie des substantifs sont les mots considérés comme représentant des concepts » pose (Barakat-Barbieri 1992, p. 64), et de là il se cantonne aux noms et aux adjectifs qui les qualifient. Il envisage néanmoins également les adjectifs (et participes passés à valeur adjectivale), en tant qu'ils peuvent être substantivés, selon des transformations du genre : *molécule identifiée* → *identification de molécule*.
Ce n'est pas parce que les descripteurs finaux seront des groupes nominaux, qu'il (ne) faut glaner dans le corpus (que) des groupes nominaux : ou du moins, c'est une heuristique, pas une évidence ou une nécessité.
Les noms sont souvent associés de manière privilégiée à l'analyse de la *thématique* du texte : or ils ne sont pas les seuls à construire cette thématique (verbes, adjectifs...), et ils participent également à d'autres dimensions sémantiques du texte (par exemple la *dialogique*, avec l'étude des connotations apportées par tel ou tel choix lexical). Les études sociologiques de textes et les pratiques d'Analyse du Discours sont souvent très attentives à

La possibilité de rabattre les mots des diverses catégories morpho-syntaxiques sur la catégorie des noms, par substantivation ou construction périphrastique, contribue à l'hétérogénéité de la classe des noms et bannit les frontières trop étanches que l'on voudrait dresser entre les catégories. L'hétérogénéité des formes nominales se retrouve au niveau des constructions dans lesquelles elles entrent : par exemple, un verbe substantivé garde une valence riche ; ses arguments, transposés en compléments du nom, s'organisent selon des relations structurées et typées (cas sous-jacent). La perméabilité de la catégorie nominale montre que la différence entre le nom et les autres catégories n'est pas d'ordre conceptuel : ce sont plutôt des différences d'emploi, de point de vue.

La forme nominale pose ce qu'elle représente comme un objet, auquel peut s'appliquer un prédicat. Elle apparaît ainsi comme une marque de conceptualisation. Nommer est une manière d'affirmer l'existence d'une réalité[36] : d'où l'idée que les substantifs d'un texte rassemblent les notions mobilisées et pertinentes pour le domaine et le sujet concernés, et donc la place importante qui leur est toujours accordée dans les systèmes de capitalisation des connaissances, de recherche documentaire, etc.

Les noms propres et les sigles constituent généralement une classe à part. Ils se dotent de propriétés particulières à tous les plans d'analyse du catégoriseur : *morphologie* (majuscule initiale, points intercalés pour les sigles,...), *syntaxe* (constructions sans déterminant, sans adjectif,...). Au plan *sémantique*, les noms propres mentionnent des personnes, des organismes publics ou privés, des produits et des marques, des revues, des lieux, qui sont très saillants dans des lectures professionnelles, notamment pour toute l'activité de veille (veille technologique, veille stratégique)[37]. Or la plupart des entités ainsi désignées ont un espace ou / et une durée de renommée relativement faible. Leur renouvellement est constant ; leur reconnaissance, leur importance significative et leur rôle sont relatifs au domaine dans lequel on se place. Ces deux facteurs ont une conséquence *lexicale* : les noms propres échappent en grande partie aux dictionnaires, ainsi qu'aux terminologies de référence qui seraient trop statiques ou à vocation trop générale.

L'utilisation de l'information apportée par un catégoriseur pour l'identification des noms propres est donc intéressante, sous réserves que l'outil ait les moyens de repérer ces noms propres avec une régularité suffisante (analyse morphologique contextuelle, prise en compte de répertoires centraux pour le domaine concerné, etc.). Ce repérage est d'autant plus sensible que nombre de noms propres adoptent, à la majuscule initiale près[38], la forme de mots courants quelquefois éliminés comme peu significatifs (le magazine *Elle*, l'Ile de la *Réunion*,... et jusqu'à l'omniprésent *Bill Gates* !). Si l'on se replace dans le contexte d'un texte, le risque de confusions gênantes s'estompe, les domaines d'emploi du nom propre et du mot commun étant souvent disjoints.

Adjectif

Occultés par l'attention prépondérante accordée aux noms, ce sont quelquefois les adjectifs qui ont une incidence sémantique essentielle. Dans les textes techniques de la DER, on peut citer *vibro-acoustique* (« comportement vibro-acoustique », « énergie vibro-acoustique »,...), *tridimensionnel* (d'après une observation de Richard Quatrain), *flou* (« logique floue », « contrôleur flou »), *nucléaire* (« énergie nucléaire », « parc nucléaire », « sûreté nucléaire »). Sur un corpus technique de fiches d'incidents en centrales, (Lefèvre, Chellali 1993, §3.1.8) observent d'ailleurs, dans les effets du style télégraphique, l'effacement du nom au profit de certains adjectifs :

---

tous ces indices, et recueillent des unités de tous ordres : noms, mais aussi personne des verbes, modalités, etc. (Le Roux 1992) (Piat 1996)

[36] On peut cependant contester que le nom (pris isolément) puisse *désigner* une *référence* : la construction du rapport à une réalité mobilise incontestablement aussi les verbes (Tyvaert 1997), et d'une façon plus générale le contexte linguistique et extralinguistique.

[37] Concernant la détermination de la proximité entre des textes, les noms propres sont interprétables différemment d'autres unités : la présence d'un nom propre est un indice *précis*, utile pour spécifier le rapprochement, et son absence est généralement *non pénalisante* (pour ne pas s'en tenir à une situation particulière).

[38] Travailler en typographie riche, en particulier en gardant la distinction entre lettres majuscules et lettres minuscules (la « casse » des caractères), permet donc d'éviter un certain nombre de confusions artificielles.

changement du circuit hydraulique devient *changement de l'hydraulique*, voire *changement hydraulique*.

La catégorie adjectif des étiqueteurs concerne généralement les *adjectifs qualificatifs*, mais pas les *adjectifs déterminatifs* (possessifs, démonstratifs, etc.) qui sont reversés dans les déterminants. La catégorie adjectif ainsi définie ne comporte quasiment aucun mot grammatical. On oppose quelquefois, pour le français, les *adjectifs antéposés* et les *adjectifs postposés* : les premiers auraient une signification beaucoup plus vague et générale que les seconds[39]. Une distinction plus fine pourrait encore diviser la catégorie des adjectifs en *adjectifs qualificatifs* et *adjectifs relationnels* : les adjectifs relationnels sont ceux qui peuvent être réécrits comme un complément de nom (ex. *élection présidentielle* / *élection du président*)[40]. L'opposition qualificatif *vs* relationnel peut séparer les emplois d'un même adjectif (Habert, Nazarenko, Salem 1997, §I.3). Ayant une forme nominale équivalente, les adjectifs relationnels ont donc une affinité particulière avec la catégorie des noms dont ils semblent dérivés.

Les linguistes soulignent la parenté étroite entre adjectif et nom (Vendryes 1923, §3) (Ducrot, Todorov 1972, § *Parties du discours*). Pour une forme nominale et une forme adjectivale qui se correspondent, la différence de nature ne traduit qu'une différence de point de vue sur une même notion : le nom *cerne* une (des) entités, il détermine un objet, une catégorie, alors que l'adjectif renvoie à une propriété dont le domaine d'application *déborde* l'entité considérée.

Réserver un sort opposé aux noms et aux adjectifs pour une application comme DECID serait donc injustifié.

L'adjectif tient aussi du prédicat, comme expression d'un état. C'est ce qu'illustre le rôle effacé de la copule (le verbe être introduisant l'adjectif attribut), ou encore le glissement des participes entre emploi verbal et adjectival.

Verbe

Telle quelle, la classe des verbes est partagée entre lexique et mots outils. De ce qui relève plutôt de la grammaire à ce qui relève plutôt du lexique, on trouve : les auxiliaires, et les catégories marquées par la conjugaison (*personne*, *temps*, *mode*)[41] ; les semi-auxiliaires de mode (*pouvoir*, *vouloir*) ou d'aspect (*commencer à*, *venir de*, *être en train de*, *aller*), le verbe *être* en tant que copule (par exemple suivi d'un adjectif attribut) et des verbes supports génériques (*faire*) ; puis d'une manière générale les autres usages des verbes, à valeur prédicative.

Les catégoriseurs distinguent habituellement les verbes *être* et *avoir* (sans toujours préciser leur emploi (Benveniste 1966, §16) : auxiliaire, copule, prédicat à part entière exprimant l'existence ou la possession). Ils précisent encore si l'on a affaire à un participe présent, passé, à un infinitif, ou à un temps conjugué. En revanche, la notion de verbe d'état est difficile à cerner concrètement (Bronckart & *al.* 1985, §IV.A).

Le participe est souvent considéré comme une partie du discours différente du verbe, car il se décline plutôt qu'il ne se conjugue. Le participe se positionne de façon intermédiaire entre le verbe et l'adjectif. Il relève d'un emploi prédicatif, verbal, quand il dépend d'un auxiliaire, quand il prend la forme d'un gérondif, ou quand il est accompagné de compléments qui correspondent à la structure des arguments du verbe. Il peut aussi se comporter comme un simple adjectif, s'accordant avec le nom

[39] (Barakat-Barbieri 1992, p. 70) se rallie à cette conception, et l'utilise : les adjectifs antéposés ne sont pas retenus pour former les candidats descripteurs.

[40] Cette distinction manque à (Fluhr 1977), pour caractériser les adjectifs qui se prêtent à une normalisation vers la forme nominale.

[41] (Dupuy 1993) base son exploitation sémantique des temps (et modes) verbaux sur une articulation de trois dimensions, qui lui fournissent ensuite trois plans d'analyse :

	présent	p. composé	p. simple	p. antérieur	imparfait	pl.-q.-parf.	futur	conditionnel
attitude	commenté	commenté	raconté	raconté	raconté	raconté	commenté	raconté
perspective	temps zéro	rétrospect°	temps zéro	rétrospect°	temps zéro	rétrospect°	prospection	prospection
mise relief			1er plan	1er plan	arr. plan	arr. plan		

dont il dépend. Dans ce cas, il n'y a guère de raisons de le traiter différemment d'un adjectif qualificatif (au moins dans le cadre d'une application comme DECID).

Le propre du verbe est de se conjuguer, ce qui pourrait constituer un critère morphologique pour le caractériser. Ainsi, par nature, il rend compte d'une dynamique (notamment décrite par le concept linguistique d'*aspect*), ce qui est un facteur de contraste avec l'adjectif (Pottier 1974, §310). C'est aussi un support privilégié pour l'expression de modalités : une série de verbes au conditionnel dans un document scientifique peut traduire, de façon simple et régulière, les réserves de l'auteur à l'égard de ce qu'il présente.

L'infinitif est un premier degré de substantivation, qui conserve néanmoins l'essentiel du dynamisme verbal (*mourir d'aimer* vs *mourir d'amour*, *une nouvelle manière de vivre* vs *de vie*) (Pottier 1974, §312). L'infinitif s'approche effectivement du substantif lorsqu'il s'abstrait par effacement des compléments, alors qu'il garde une valeur verbale lorsqu'il résulte de certaines transformations syntaxiques comme l'enchâssement (proposition infinitive) ou la modalisation. (Bronckart & *al.* 1985, annexe B.1) propose une grille de critères pour séparer ces différentes valeurs de l'infinitif.

Adverbe

Outre les formes en *-ment* (obtenues par dérivation des adjectifs : *économiquement*, etc.), dotées d'une certaine charge lexicale grâce à leur racine (par ex. ici *économ-*), les adverbes transcrivent en français un certain nombre d'indications grammaticales importantes comme la négation (*ne...pas* et toutes ses variantes), le comparatif et le superlatif. On trouve aussi dans la catégorie des adverbes des interrogatifs (*pourquoi*, *comment*) qui (à la différence des pronoms interrogatifs) appellent une réponse potentiellement sous forme de proposition.

Il y a de fait un continuum entre des adverbes « qualificateurs » (à contenu lexical) et des adverbes « quantificateurs » (à valeur grammaticale), voire des emplois qualificateurs ou quantificateurs d'un même adverbe, que l'on pourrait illustrer par l'exemple suivant (Pottier 1974, §313) :

comiquement, *bizarrement* : « parler drôlement » ← *drôlement* → « drôlement triste » : *très*

Et les adverbes dérivés (en *-ment*) n'ont pas le monopole des contenus lexicaux : les contre-exemple, comme *hier* et *vite*, ne sont pas rares. La classe des adverbe est donc un cas exemplaire de mélange d'unités lexicales et grammaticales.

Les adverbes même les plus « grammaticaux » peuvent jouer avoir rôle sémantique déterminant. Dans un corpus technique (EDF), (Sta 1997, §6.2.2) note par exemple le contraste entre *très haute tension* et *haute tension*, ou entre *analyse non linéaire* et *analyse linéaire*, qui renvoient à des domaines d'étude tout à fait séparés dans la pratique.

Contrairement peut-être à ce que laisserait entendre son nom, l'adverbe ne détermine pas exclusivement des verbes[42]. Certains (ne) se construisent (que) avec des adjectifs (ex. *très* ; *beaucoup* a une distribution complémentaire, uniquement avec les verbes) (Martinet 1991, §4.45). On peut également voir un adverbe modifier un autre adverbe, ou une phrase entière (heureusement, jadis). En somme, l'adverbe se comporte comme indiquant une propriété d'une propriété, à la différence de l'adjectif ou du verbe qui indiquent une propriété d'une entité. C'est en quelque sorte des « attributifs du second ordre » (cf. James Harris), au deuxième degré, des méta-attributifs.

L'hétérogénéité de la catégorie des adverbes se reflète dans la diversité de manière de les considérer lors d'un étiquetage. Des choix sont motivés par la finesse souhaitée de la description, comme par des considérations heuristiques. Par exemple, le catégoriseur Papin-Maucourt (INaLF) s'applique à faire la part entre les valeurs positive et négative de *plus* (ce que l'oral rend par une différence de prononciation) ; en revanche, *non* ne fonctionne pas comme un adverbe de négation et n'est donc pas regroupé avec ceux-ci.

[42] Déjà, plus précisément, il porte dans ce cas sur le prédicat et non sur le verbe (dans *Il est tristement célèbre*, *tristement* se rapporte non à *est* mais à *célèbre*) (Martinet 1991, §4.45).

Pronom

Pour étudier ce que l'on peut faire de cette catégorie, il faut d'abord prendre conscience de son étendue. En suivant la grammaire traditionnelle, on a :

- les pronoms personnels, qui forment système en fonction du nombre, du genre, des personnes et des cas. A noter, le pronom personnel indéfini *on*, et le *il* neutre pour les formes impersonnelles (*il faut*, *il pleut*). Les pronoms sont dits *clitiques* quand ils se situent dans la zone précédant immédiatement le verbe ; les pronoms non clitiques occupent les même places que les groupes nominaux, et comportent des formes qui leur sont propres (ex. *moi*, *toi*, *eux*).
- le pronom réfléchi *se* et le pronom adverbial *y* occupent des positions analogues aux clitiques.
- on retrouve des familles de pronoms parallèles aux familles d'adjectifs déterminatifs : pronoms démonstratifs (*ce*, *c'*, *ceci*,...), possessifs (*(le) mien*,...), interrogatifs (*qui*, *lequel*,...), indéfinis (*rien*, *nul*, *tous*, *quelqu'un*, *personne*, *autrui*, *quiconque*, *certains*, *plusieurs*, *chacun*...). Les pronoms indéfinis ne se laissent pas décrire par un système de catégories grammaticales (nombre, personne, etc.) comme les précédents.
- les pronoms relatifs (*qui*, *que*, *qu'*, *dont*, *où*, *auquel*,...) ont la particularité d'être des connecteurs, du fait qu'ils introduisent une proposition subordonnée ; en ce sens, on peut les analyser comme équivalents à une conjonction, une préposition éventuelle (qui n'est autre qu'une marque de cas), et un pronom personnel (non clitique)[43].

Les catégoriseurs s'en tiennent généralement aux distinctions qui leur sont le plus utiles pour leur analyse (comportement distributionnel différent), à savoir l'opposition clitique vs non clitique, et la catégorie à part des pronoms relatifs (connecteurs).

L'importance d'un pronom est ambivalente. Soit l'on regarde son contenu intrinsèque, général et standard : savoir qu'un texte comporte le mot il ne m'apprend pas grand'chose sur son contenu (sauf contexte très particulier). Soit l'on vise, à travers le pronom, l'entité qu'il représente, la valeur qu'il prend en contexte. La première visée tend à rendre le pronom transparent, à l'effacer, à l'intégrer au mot dont il dépend. La seconde le renforce et lui confère une certaine autonomie, en le rapprochant des noms.

Parmi les arguments avancés en faveur de la résolution des références des pronoms, le plus fruste concerne les applications où la répartition et la fréquence des mots sert à construire une représentation du texte. On considère qu'un mot représente le concept auquel il renvoie. D'autre part, toutes choses égales par ailleurs, la fréquence d'un mot dans un texte donne une mesure de l'importance du concept dans ce texte. On juge donc crucial de faire un décompte des occurrences aussi juste que possible, et notamment de ne pas sous-évaluer un concept dont il constamment question du fait de sa reprise par des pronoms (c'est le phénomène d'anaphore). On sait en effet que les conventions de rédaction dans notre culture invitent à éviter la répétition d'une expression, au moins dans un voisinage immédiat, sauf effet de style particulier.[44]

La première remarque à faire est que l'usage de pronoms est un des principal mode de reprise, mais que l'on peut également faire des reprises elliptiques ou varier indéfiniment la désignation : les journalistes sont d'ailleurs coutumiers du fait, ils distillent ainsi peu à peu des indications sur leur

[43] Une telle décomposition est réalisée par le catégoriseur de l'équipe Cristal / GRESEC à Grenoble.

[44] C'est une thèse qui a été défendue par Christian Fluhr (concepteur du système de recherche documentaire SPIRIT) :
« Un autre élément qui perturbe la statistique est dû au fait que le style de rédaction, en français surtout, impose d'éviter les répétitions, ce qui amène à l'emploi de pronoms et de synonymes. La reconnaissance du référent des pronoms est donc utile si l'on veut avoir la fréquence exacte d'utilisation d'un concept dans un texte. » (Fluhr 1977, §III.3)
Sont ensuite identifiés quatre formes de reprise : emploi d'un (quasi)-synonyme, d'un terme plus général (un hyperonyme), reprise elliptique, et pronoms anaphoriques. Les modes de résolution proposés sont, pour les deux premiers cas, la consultation d'un thesaurus, pour le troisième des critères morphologiques utilisés avec « grandes précautions » (sans autres détails), et pour le dernier un algorithme récursif indiqué par l'auteur. L'algorithme considère les pronoms personnels, une fois écartés les pronoms impersonnels, *i.e.* sans référent (*personne*, *rien*, *il* dans *il pleut*). « Cet algorithme testé manuellement a donné sur un corpus d'environ 250 000 mots, entre 80 et 85 % de réussite suivant les sous-corpus testés. Ce résultat peut être suffisant dans la mesure où le traitement documentaire est statistique et donc a une certaine tolérance. » (Fluhr 1977, §III.4.2.4)

sujet. La logique serait donc finalement de décompter non seulement les reprises anaphoriques, mais aussi toutes les formes de coréférence. A ce point, deux objections se dressent. La première est d'ordre pratique : la résolution générale des coréférences est évidemment extrêmement complexe à mettre en œuvre dans un traitement automatique. La seconde objection est plus fondamentale : l'idée même que l'on puisse déterminer un ensemble d'expressions dont la signification est identique, au sens où elles désignent un même concept, ressort d'une conception référentielle et dénotationnelle de la sémantique. Or la langue n'est pas transparente au réel, elle a sa propre épaisseur. En l'occurrence, les diverses manières de cerner une même idée apportent chacune un effet de sens propre, qui n'est négligeable que par choix (réducteur). Et il y a un travail interprétatif, pour homologuer les différentes « versions » d'un même concept (et d'ailleurs ainsi ajuster le concept, qui n'est pas préexistant et figé). Comme dans toute interprétation, il y a des choix, des écarts plus ou moins grands, des variations de parcours, de sorte que l'affectation d'une expression à un concept n'est pas de l'ordre du « tout ou rien ». Enfin, sans reprendre ici ce que nous développons par ailleurs, le sens ne réside pas dans les mots mais se construit dans une interaction de tous les paliers (mot, phrase, texte), aussi l'identification de relations autonomes de type « 1 mot - 1 concept » n'est pas en mesure de rendre compte de la sémantique d'un texte.

La recherche d'un décompte exact des occurrences d'un concept a donc valu une première remarque, sur la problématique de la coréférence. La seconde remarque porte sur le principe même de décompte. Les chiffres sont utiles, mais il ne faut pas y croire avec trop d'orthodoxie, ou leur donner une précision qu'ils n'ont pas ! Un nombre d'occurrences donne une mesure, un reflet approximatif d'une réalité, dont l'interprétation se fait en termes d'*ordre de grandeur* et d'*équilibre général*. Si un mot contribue à une idée centrale du texte, peut-être n'est-il pas répété littéralement de phrase en phrase, en revanche quasiment chaque nouveau paragraphe lui redonne sa chance d'être formulé. Le principe de « non répétition » prévaut au niveau transphrastique (enchaînement local de deux phrases), et non à un niveau textuel. De plus, du point de vue interprétatif, le passage d'un paragraphe à l'autre est un seuil, la fin d'un paragraphe et le début du paragraphe suivant sont une courte zone de transition, pour passer en quelque sorte d'une unité de pensée à une autre : une reprise par anaphores et variantes n'a pas couramment à renvoyer à une formulation faite plusieurs paragraphes avant –c'est cognitivement une distance trop grande pour n'être pas exceptionnelle. Ceci permet de compter sur la reprise d'un mot pour refléter une certaine importance pour le texte, sans jamais croire tenir (ni même approcher) une (illusoire) mesure absolue de son rang au plan de la signification dans le texte.

Des résultats expérimentaux confirment cette analyse (Bonzi, Liddy 1989). Dans la cadre d'application de recherche d'informations sur des bases de documents en texte intégral, on a vérifié et constaté la non nécessité d'une résolution des anaphores. Les résultats des calculs de sélection de documents sont également satisfaisants, que les anaphores aient été résolues ou non. Même si les conditions de l'expérience sont par nature limitées (choix des variables linguistiques dans la représentation du texte, choix de la fonction de pondération, etc.), elles n'altèrent pas la portée du résultat : c'est l'équilibre général qui est ici en jeu, et qui est bien retrouvé dans les deux cas de figure. Le bon comportement général des divers systèmes qui procèdent à des calculs fréquentiels sur les textes sans analyse linguistique poussée (et donc sans résolution d'anaphores ou de coréférence) complète la démonstration.

(Pirkola & Järvelin 1996) considèrent l'incidence, effectivement plus problématique, de la résolution des ellipses et anaphores lorsqu'il est fait usage d'opérateurs de proximité. Pour une base d'articles de journaux, ils concluent que *seules* les reprises anaphoriques ou elliptiques de noms propres mériteraient d'être résolues. Plutôt que de se lancer dans des mécanismes complexes de résolution exacte, une notion de persistance sémantique d'un nom propre, sur un empan de quelques phrases ou de l'ordre du paragraphe, pourrait suffire à l'établissement de similarités entre textes (sans chercher à construire une représentation déterminée du contenu du texte).

La résolution des pronoms se conçoit si l'on veut procéder à l'analyse fine et systématique des relations locales, marquées par les dépendances syntaxiques. Par exemple, on veut recenser l'ensemble des sujets « réels » de tel verbe, et de part en part former des classes d'actants qui occupent les mêmes fonctions (dans la lignée de Greimas). Une telle recherche s'attire, à un degré moindre, les objections faites précédemment : continuum des significations et multiplicité des paliers

contribuant à la sémantique, construction globale de l'interprétation, dont la sensibilité ne doit pas être réglée au niveau d'une relation syntaxique dans une phrase particulière.

Déterminant

L'étiquette *déterminant* recouvre plusieurs natures grammaticales, qui ont un comportement syntaxique analogue (grosso modo : les déterminants précèdent les noms, un nom n'a qu'un seul déterminant), et qui se laissent énumérer comme des systèmes articulés notamment par le genre et le nombre. On a ainsi :

- les articles : définis, indéfinis, partitifs ; retrouver l'amalgame de l'article avec la préposition *à* ou *de* dans *au(x)*, *du*, *des* oblige le catégoriseur à faire des distinctions qu'il n'est en général pas capable de faire correctement, pour *du* (forme contractée ou partitif) et *des* (forme contractée ou indéfini).
- les adjectifs démonstratifs (*ce*,...), possessifs (*mon*,...), interrogatifs (*quel*,...), indéfinis quantitatifs (*aucun*, *chaque*, *plusieurs*, *tout*, *certain*, *quelques*,...). Ces derniers ne se laissent pas énumérer systématiquement de la même manière que les précédents.

Certains déterminants sont sensibles au fait que le mot suivant commence (phonétiquement) par une voyelle : élision (*l'*), ajout euphonique (*cet*).

Des linguistes expliquent le rôle du déterminant par son effet d'actualisation, nécessaire vis-à-vis des noms communs. Selon le déterminant choisi et le contexte, l'actualisation peut prendre de multiples valeurs, traduisant la généricité, l'unicité, le caractère fixé de la détermination, etc. Par exemple, dans le cadre de l'extraction de terminologie sur des corpus EDF, on est amené à distinguer trois valeurs pour l'article défini singulier : valeur unique (exprime l'unicité dans le domaine considéré), valeur anaphorique (reprise de ce qui vient d'être évoqué), et valeur spécifiante (ce que le nom désigne est unique suite aux précisions apportées par des compléments (relative, etc.)) (Bourigault, Gros 1994, §IV.3.1).

Le petit nombre de déterminants, et leur utilisation systématique imposée par la syntaxe, font que les déterminants (et surtout les articles) se caractérisent par leur très forte fréquence dans les études de statistique lexicale (Muller 1977, §26)[45].

Au plan textuel, et pour DECID, on est intéressé de repérer les phrases nominales ne commençant pas par un déterminant : ces deux indices (absence de verbe, absence de déterminant introductif), combinés à des indices de présentation (ouverture d'une nouvelle ligne, longueur relativement faible, absence ou utilisation particulière des ponctuations), aident à repérer titres et intertitres.

La valeur anaphorique (que peut prendre l'article défini) pourrait signaler une entité centrale et thématique. Une première raison, serait que l'entité en question (exprimée par le syntagme nominal défini) a été introduite et que la suite du passage « tourne autour » d'elle. Utiliser cette propriété s'annonce cependant complexe (il faut savoir quand on a affaire à une valeur anaphorique), peu caractérisant (grand nombre d'entités repérées à l'issue de l'analyse d'un texte), et insuffisamment justifié d'un point de vue sémantique (doute sur la pertinence du critère, si l'on examine les textes). Une autre raison concerne le cas où l'entité est directement présentée comme connue du lecteur. Elle fait donc partie des entités de base pour le genre du texte, « évidentes » et de définition implicite. On aurait là un indicateur de la présence significative de l'entité dans le contexte (domaine) du texte (Sågvall Hein 1989).

Les déterminants ne sont pas tous équivalents au regard de la description des syntagmes, ce qui rend nécessaire soit une restructuration du jeu de catégories, soit une finesse suffisante pour définir les catégories souhaitées comme un regroupement d'étiquettes. Par exemple, la possibilité de reconnaître la présence d'un *article défini* après la préposition *de* sert à établir une distinction fondamentale pour l'application LEXTER : elle sépare les structures d'*unité phraséologique* (dont la

[45] Une exception manifeste (et marginale ?) : les textes qui font appel à un « style télégraphique » des plus succincts. Le corpus de (Lefèvre, Chellali 1993), à savoir les résumés figurant sur les fiches consignant les événements dans les centrales, se caractérise par l'abondance des phrases nominales mais la conservation des déterminants qui articulent les termes entre eux.

décomposition est notée *T'* et *E'*)[46] des structures de *terme complexe* (notation *T* et *E*, sans prime). Ceci conduit à se donner trois classes de déterminants : *(i)* déterminant absent (non exprimé), *(ii)* articles définis (y compris ses occurrences dans les formes contractées), *(iii)* tous les autres déterminants (Bourigault, Gros 1993, §I.3.6.3, §III.1.1.1, et §IV.3.2).

Un cas particulier : les nombres cardinaux

En ce qui concerne leur comportement syntaxique, les nombres cardinaux peuvent être décrits comme des déterminants (*Trois points ont été abordés à la dernière réunion*) et des adjectifs (*Ces trois points sont :...*). Leur prise en compte appelle néanmoins quelques remarques particulières.

Il faut choisir de distinguer (ou non) les écritures en lettres et en chiffres[47]. Il y a certes équivalence de signification mathématique, mais ces deux formes ressortent d'usages différents : conventions qui privilégient l'écriture en lettres pour les petits nombres et l'écriture en chiffres ensuite ; écriture en lettres demandée dans certains contextes, administratifs ou solennels[48] ; écriture en chiffres pour des données issues de mesures et se prêtant au calcul. Cela peut même être une indication précieuse pour le traitement, pour cerner certains cas de figure : une écriture en chiffres peut être reconnue comme une numérotation de paragraphe, ou peut être suivie d'une abréviation désignant une unité physique (de longueur, de masse, etc.).

D'autre part, l'écriture des nombres cardinaux se laisse bien décrire par une grammaire formelle, si bien que un nombre peut être reconnu et identifié indépendamment de tout contexte.

Préposition

Les prépositions introduisent un complément (d'un substantif, d'un verbe, d'un adjectif, d'un adverbe). Elles indiquent donc une *dépendance*, à l'*intérieur d'une proposition* (entre les propositions la dépendance est marquée par d'autres unités : conjonctions, pronoms relatifs), et contribuent à marquer la *fonction*. Ces trois propriétés appellent un commentaire.

La relation de dépendance s'accompagne de l'asymétrie des constituants reliés par la préposition. C'est une différence majeure avec les conjonctions de coordinations, dont elles partagent le caractère invariable et la forme souvent très brève.

Les prépositions traduisent des liens internes, au palier du syntagme voire de la lexie (pomme de terre). Elles peuvent donc être « traversées » en restant dans une même zone de localité syntaxique.

Malgré l'éventail de prépositions disponibles, la préposition ne suffit en général pas à déterminer à elle seule la fonction (le *cas*) représentée. Certaines prépositions cumulent des interprétations extrêmement contrastées : *de* peut indiquer la provenance, l'appartenance, l'objet (accusatif : *le retrait du permis*),... *par* peut indiquer le moyen, l'agent,... La consultation d'un dictionnaire de langue poursuivrait cette démonstration.

Il est difficile d'établir un inventaire exhaustif des prépositions. On peut citer –en procédant à des regroupements intuitifs et qu'il serait intéressant d'ajuster expérimentalement– des prépositions participant (notamment) à la formation de termes composés (outre *de* et *à*, on a *en*, *par*, *avec*, *sans*,

[46] Une condition s'ajoute en fait à la présence de l'article défini pour reconnaître une unité phraséologique : l'expansion *E'* est une autre unité phraséologique ou un terme complexe, pas un simple nom.

[47] Les chiffres romains sont une écriture en chiffres, même s'ils usent des signes alphabétiques. Ils peuvent également être considérés à part, en raison de leurs usages propres : éléments de dates (mois, années, siècles) dont les ordres de grandeur sont connus ; numérotation (chapitres, volumes,...), généralement pour des divisions de haut niveau, avec une variante en minuscules pour des divisions plus fines.
D'autres systèmes de numérotation vaudraient d'être décrits (lettres alphabétiques, majuscules ou minuscules ; lettres grecques), ainsi que la syntaxe de leur combinaison (séparateurs tels que tiret, point, parenthèse fermante, encadrement par des parenthèses ou des crochets, espacements,...).

[48] On écrit en toutes lettres la somme d'un chèque, et souvent la date pour un diplôme, un acte de naissance, une plaque commémorative. Cela pourrait être hérité du souci d'éviter le risque que l'écriture (manuscrite) soit mal lue. L'écriture en lettres est adoptée car plus redondante et contextuelle : la difficulté à déchiffrer une lettre est estompée par le contexte des autres lettres, alors que chaque chiffre est indépendant. Cette attention souligne l'importance du nombre ainsi inscrit : manifestement, seule cette marque d'importance subsiste quand il s'agit d'écriture dactylographiée ou soigneusement gravée.

pour, *sous*, *sur*, *dans*[49]), des prépositions à valeur négative, utile pour analyser des formules de type requête documentaire (*sauf*, *hors*, *excepté*, *hormis*, et *sans*, déjà cité), des prépositions courantes dont certaines sont à distinguer de leur(s) homographe(s) (*contre*, *entre*, *jusque*, *vers*, *avant*, *après*, *devant*, *derrière*, *durant*, *depuis*, *dès*, *passé*, *chez*, *selon*, *envers*, *malgré*, *outre*, *parmi*). Les prépositions couvrent une large gamme de fréquences, depuis les fréquences les plus hautes (*de* vient habituellement largement en tête des inventaires lexicaux de corpus) ; l'« oubli » des prépositions plus rares n'aurait qu'une incidence limitée au plan quantitatif.

Conjonction

Le catégoriseur distingue quelquefois les coordonnants et les subordonnants. Il y a en effet plusieurs motivations pour cela :

- les principales conjonctions de coordination sont bien connues (*Mais où est donc Ornicar ?*)[50], alors que les conjonctions de subordination se déploient dans un grand nombre de locutions conjonctives ;
- les conjonctions de coordination relient deux éléments de même nature et de rôle identique, alors que les conjonctions de subordination marquent une relation asymétrique et une dépendance grammaticale ;
- les conjonctions de coordination unissent des éléments de toute « taille » (mot, syntagme, proposition), alors que les conjonctions de subordination articulent des propositions ;
- les conjonctions de coordination donnent lieu à de délicates questions de logique, de portée, de factorisation et de distributivité.

Les connecteurs sont finement étudiés pour leur valeur argumentative. Les significations associées à un connecteur s'avèrent multiples, à la fois fortes et très différentes, ce qui en fait un terrain linguistique captivant mais rend délicate toute détermination puis exploitation dans un système automatique.

De part leur rôle de connecteurs, les conjonctions ont une importance majeure dans la description des relations transphrastiques. Pour DECID, on s'intéresse davantage à la textualité globale (macrosémantique) plutôt qu'aux interactions transphrastiques, qui apparaissent comme une extension des analyses au palier de la phrase (mésosémantique).

En deçà de la phrase, et dans les cas où sa portée peut être évaluée par l'analyse automatique, la coordination peut signaler une parenté sémantique : alternatives issues d'un même paradigme (cf. le déploiement en énumération), éléments qui jouent un rôle comparable dans le domaine concerné[51]. En d'autres termes, la coordination contribue à créer une zone de propagation de traits sémantiques (Cavazza 1996, p.60). Une telle indication est exploitable par DECID.

Mot inconnu

Comme les unités grammaticales forment des paradigmes fermés (sauf peut-être les locutions conjonctives), elles sont toutes connues du système. Aussi un mot inconnu est très vraisemblablement un item lexical (Fluhr 1977, p. 99).

---

[49] (Bourigault, Gros 1993, §III.1.1) ajoute encore *contre* et *vers*.

[50] La formule mnémonique ne fait pas le tour des conjonctions de coordinations (il y manque par exemple soit, en tant que marquant une alternative).
La notion de coordination pourrait s'élargir à certains connecteurs –adverbes, locutions, éléments qui fonctionnent en système et se répondent–, que d'aucuns appellent *marqueurs d'intégration linéaire* (MIL), qui « coordonnent » les passages du texte : *tantôt*, *primo* / *secundo* / *tertio*, etc. Voir par exemple (Adam 1990, §II.1.2.1)

[51] Au plan syntaxique, « deux segments d'un énoncé sont coordonnés lorsqu'ils ont la même fonction (c'est le cas pour « le soir » et « avant le déjeuner » dans « Téléphonez-moi le soir ou avant le déjeuner »). Or on ne peut se passer de la coordination si l'on veut décrire certaines conjonctions comme le *et* et le *ou* du français, qui ne peuvent relier que des segments coordonnés : on ne peut pas dire, sans effet de style particulier, « Il travaille le soir et son examen », ni « il travaille le soir et à Paris ». » (Ducrot, Todorov 1972, § *Fonctions syntaxiques*).
C'est une observation qui, bien au-delà de la syntaxe, peut être rapportée à des processus interprétatifs : la coordination favorise la propagation de traits sémantiques (Cavazza 1996, p. 60).

Si l'analyse s'appuie sur un dictionnaire de langue général relativement complet, on peut augurer, selon le corpus, que l'on a affaire à un terme de spécialité, ou à un néologisme, ou à un nom propre,... ou encore à une coquille. Sauf dans le dernier cas, le mot inconnu est *a priori* important pour la caractérisation du texte.

Des indices morphologiques peuvent orienter les prédictions : une initiale majuscule, un certain suffixe typiquement nominal, etc. La syntaxe réduit aussi l'éventail des valeurs vraisemblables. Des considérations fréquentielles sont envisageables pour limiter l'impact des coquilles : un mot inconnu répété est peu vraisemblablement une coquille ; un hapax (une seule occurrence) n'est peut-être pas un terme central, si le texte est assez développé.

Autres...

Certains jeux d'étiquettes proposent de repérer des catégories marginales, ou/et n'étant pas une partie du discours au sens classique du terme.

- *Mot étranger* : il apparaît comme une incrustation dans la phrase de la langue considérée. S'il s'agit d'un passage entier, les règles (et les ressources : le dictionnaire éventuel, le jeu de catégories lui-même et l'interprétation qu'en en fait) utilisées pour une langue ne s'appliquent pas à une autre, et le catégoriseur n'est pas en mesure d'analyser deux langues de la même façon. Avoir cette étiquette est une manière de prévoir quelques (rares) mélanges de langues dans le corpus.
- *Interjection* : l'étiquette est sans doute davantage utile pour des corpus oraux, ou qui comportent des passages dialogués, des monologues, plutôt que dans des documents rédigés, d'information scientifique et technique. Les interjections ne forment pas une classe que l'on peut énumérer systématiquement (*oh*, *zut*,...). Elles restent en marge de la construction syntaxique de la phrase, dont elles ne doivent pas perturber l'analyse. L'intérêt serait de les considérer en interaction avec certaines ponctuations (le point d'exclamation, d'interrogation, les points de suspension), pour l'expression d'attitudes affectives. Etiqueter les interjections mais effacer les ponctuations apparaîtrait comme une incohérence au plan du traitement.
- *Onomatopée*, *abréviation* : ceci est une information sur la formation d'un mot, et incidemment sur sa forme. Par exemple, on peut s'attendre à de longues suites de consonnes, à des lettres plus que doublées, à l'utilisation du point comme marque de troncature. C'est supplémentaire (et non complémentaire) par rapport aux parties du discours traditionnelles. Les abréviations se comportent différemment les unes des autres, et seraient chacune à rapprocher de telle ou telle partie du discours : *cm* se comporte comme un *nom*, *vs* comme une préposition ou un coordonnant, *etc* comme une ponctuation ou un adverbe, *cf.* comme l'infinitif d'un verbe transitif,... De même pour les onomatopées : *atchoum* rejoint les interjections, *couiner* les verbes, et ainsi de suite. Par conséquent, ces étiquettes ne se positionnent pas comme des alternatives possibles au sein des parties du discours, elles peuvent éventuellement intervenir à titre d'information supplémentaire.
- *Préfixe* : un préfixe peut apparaître détaché et isolé pour diverses raisons : coordination « factorisante » (*du courant mono- ou triphasé*, *les super- et hypermarchés*), construction morpho-syntaxique (*l'amitié franco-italienne*), des ellipses ou abréviations qui peuvent se stabiliser (*auto*). Là encore, l'étiquette *préfixe* n'aurait pas à évincer les étiquettes des parties du discours : en fonction de son emploi, de sa place dans la construction syntaxique, le préfixe prend le rôle de telle ou telle partie du discours. Un autre choix consiste à décrire le préfixe comme une unité dont le comportement est de s'adjoindre à gauche d'une unité lexicale (nom, verbe, adjectif, certains adverbes) à l'aide d'un tiret : cela rendrait compte de *quasi-* et de *pseudo-*. Compte-tenu des réserves exprimées tout d'abord, la constitution et la pertinence d'une telle catégorie est à étudier.

Des catégories linguistiques aux catégories heuristiques

Le jeu de catégories associé à un catégoriseur a deux rôles, un rôle interne et un rôle externe. Son *rôle interne* oriente le choix des étiquettes sur des distinctions et des regroupements efficaces pour mener à bien l'analyse : catégories qui correspondent aux formes occupant une place dans des règles d'enchaînement ; catégories qui fournissent ensuite un contexte suffisamment précis pour aider à la détermination des catégories voisines. Le *rôle externe* du jeu de catégories est son élégance (on lui trouve une cohérence d'ensemble, on l'interprète bien), et son adéquation aux applications visées

(c'est traditionnellement le passage à un module d'analyse syntaxique)[52]. Ces deux rôles sont en interrelation complexe pour façonner le jeu de catégories, car les représentations qu'ils donnent sont interdépendantes.

D'où des confusions : un analyseur morphologique (qui donne toutes les catégories attribuables à une forme prise isolément) est-il un catégoriseur ? un étiqueteur qui s'affranchit des distinctions les plus élémentaires de la grammaire traditionnelle est-il un catégoriseur ? Pour illustrer ce dernier point, l'étiqueteur mis au point par l'équipe Cristal / GRESEC[53], à Grenoble, souligne le poids du rôle externe du jeu d'étiquettes, en dénonçant l'emprise du découpage institué par la grammaire traditionnelle et en créant des catégories originales, jugées appropriées à une application de recherche documentaire : notamment, les pronoms se répartissent en pronoms clitiques et pronoms toniques (qui se comportent comme un nom) ; *une seule catégorie confond le nom, l'adjectif, le pronom tonique et les participes* ; un pronom relatif n'est pas identifié comme tel, mais est analysé comme un motif à base de subordonnant, éventuellement de préposition, et de pronom tonique.[54]

On a donc pu envisager les catégoriseurs sous l'angle des catégories qu'ils mettent en œuvre, et qui sont nécessairement les plus précises qu'ils puissent retourner. Il est tout aussi acceptable –c'est même complémentaire– de partir du point de vue applicatif, pour explorer les types d'informations dont on voit l'utilité.

Plusieurs modes d'utilisation de la catégorisation se présentent :

- une définition des unités, alternative à celle du découpage : nouvelles délimitations (mots composés, locutions), dissociations syntagmatiques (décomposition des formes contractées comme *au*) et paradigmatiques (résolution d'homographies transcatégorielles) ;
- la fusion des éléments d'une catégorie en une seule unité (réduction par regroupement) ;

---

[52] Pour une analyse basée sur des statistiques, la grille de catégories morpho-syntaxiques doit trouver un équilibre quantitatif et qualitatif : elle doit être assez fine pour « rendre compte d'un maximum d'oppositions grammaticalement pertinentes », mais aussi le nombre d'étiquettes doit être « assez réduit pour que des corrélations soient possibles et significatives » (Dupuy 1993, p. 149).

[53] Voir par exemple Geneviève LALLICH-BOIDIN, Marc BERTIER.

[54] Le système ALCESTE s'appuie quant à lui sur un jeu de catégories extrêmement hétérogènes, puisqu'il mêle :
• *des catégories morpho-syntaxiques traditionnelles* (ou ce qu'il en reste après redistribution sur les autres catégories, plus sémantiques) : les noms (non classés dans une catégorie sémantique) ; les verbes (autres que les verbes modaux) ;
• *des sous-familles de catégories morpho-syntaxiques* : les verbes modaux ; les auxiliaires (*être* et *avoir*) ;
• *des regroupements de catégories morpho-syntaxiques* : les mots outils sans valeur particulière ; les pronoms et adjectifs indéfinis ou démonstratifs ; les adjectifs et adverbes ;
• *des catégories morpho-syntaxiques distinguées, reconstruites* : les nombres en lettres ; les interjections ;
• *des marqueurs* : modalisation assertive (essentiellement des adverbes contribuant à exprimer l'affirmation, la négation, le doute) ; espace, temps (à chaque fois, un choix d'adverbes et de prépositions) ; quantité (adverbes) ; organiseurs de l'argumentation (principalement des conjonctions et des pronoms relatifs) ; adverbes en *-ment* pouvant jouer un rôle dans l'argumentation ; personne (pronoms et adjectifs se déclinant suivant la première, la deuxième et la troisième personne) ;
• *des notions particulières* : les couleurs ; les noms de mois et de jours ; les numéros d'époques et mesures ; les lieux, pays, zones géographiques ; les prénoms et noms propres usuels ;
• *des mots répondants à un critère de forme* : les nombres en chiffres ; les mots transcrits en majuscules ;
• *des critères fréquentiels* : formes non reconnues et fréquentes
• *une catégorie artéfactuelle* : formes reconnues mais non codées car ambiguës.
Ce jeu de catégories est présenté comme une proposition, révisable, mais qui correspond à l'introduction progressive de la sémantique, appelée par l'interprétation des résultats :
« L'organisation des mots dans les résultats, leur plus grande accessibilité par l'interprète, nous a conduits petit à petit à considérer un marquage plus fin [que la séparation mots outils *vs* mots pleins] même si celui-ci doit être considéré comme exploratoire et peut du reste être changé par l'utilisateur. [...]
Ce marquage est plutôt de type grammatical mais s'oriente de plus en plus vers un marquage sémantique. [Les catégories mentionnées ici sont] les catégories provisoires actuellement retenues. » (Reinert, Piat 1995)
La richesse des informations dont on souhaite ainsi rendre compte est intéressante. Ce qui est le plus inquiétant, ce n'est pas le caractère non systématique, subjectif, ou apparemment *ad hoc* de certaines catégories, mais la présentation de ces informations comme un ensemble de valeurs interchangeables et exclusives : car chaque unité lexicale ne reçoit qu'une seule catégorie.

- le filtrage ou la sélection de certaines catégories (réduction par projection) ;
- la structuration des unités, en typant celles-ci en fonction de la catégorie affectée.

Dans ce dernier cas, la distinction de plusieurs types d'unités au niveau de la représentation du texte analysé n'a d'utilité que si elle est relayée par la suite du traitement. Par exemple, une proposition simple et judicieuse de Max Reinert est de pouvoir associer à chaque catégorie un statut, qui spécifie si les unités de la catégorie sont *analysées* (*i.e.* utilisées dans les calculs d'analyse du corpus), *illustratives* (*i.e.* n'intervenant qu'une fois la structure recherchée déterminée, pour aider à son interprétation), ou *éliminées* (ni prises en compte pour le calcul, ni prises en compte pour l'interprétation) (Reinert, Piat 1995).

Interprétations ontologiques

Une lecture des catégories morpho-syntaxiques est d'y voir le reflet d'une répartition fondamentale entre deux valeurs :

- la représentation d'une entité, d'un support, d'une substance, donc généralement attribuée aux *noms* (les substantifs). Il est coutume d'opposer les êtres (animés) aux choses (inanimés). Les marques du *nombre* et du *genre* s'expliquent par le fait que l'entité puisse être multiple et sexuée.
- la représentation d'un comportement, de ce qui est dit à propos de quelque chose (*i.e.* d'une entité), attribuée aux *verbes* et à ce qui a valeur de prédicat. Les valeurs sont cependant diverses : propriété, état, événement, action, processus qui se déroule. Mouvement ou état, il a un dynamisme interne car il s'inscrit dans une chronologie (*temps*) et dans la durée (*aspect*) ; point de vue, il s'ancre dans une situation d'énonciation (*personne*) et est diversement engagé (*mode*, *voix*).

Cette conception trouve des partisans convaincus au sein des sciences cognitives (l'une des interventions les plus fameuses en France est celle de (Langacker 1991)), alors que les linguistes la combattent, au moins dans sa forme la plus naïve et la plus intransigeante. Sans interdire que l'articulation entité / comportement puisse être pertinente au plan conceptuel, on dénonce sa projection directe dans les catégories morpho-syntaxiques de la langue.

> La mise en signes
>
> A - Sur le plan conceptuel, on peut hiérarchiser :
>
> Entité ← Comportement.
>
> L'entité peut avoir une existence autonome [...].
>
> Un comportement, par contre, suppose au moins une entité ; *rire*, *se disputer*, *comploter*, sont des lexèmes de comportement qui renvoient nécessairement à des supports, de même que (être) *bleu*, *gentil*, *carré*, ou également *regard*, *destruction*, ou *acceptabilité*.
>
> L'exception (apparente) serait le cas des impersonnels. Dans *il pleut*, le comportement renvoie à u support déictique (situationnel).
>
> B - Il n'y a pas de correspondance automatique entre entité et nom, comportement et verbe ou adjectif.
>
> 1) En latin, *maneo* exprimait un comportement, « rester ». Un dérivé, *mansio* 1 a d'abord signifié « le fait pour N de rester quelque part ». Puis il a désigné (*mansio* 2) « le quelque part où N reste », et enfin « demeure, édifice », la *maison*. On peut suivre ainsi historiquement le déplacement de la focalisation, d'un élément à un autre, dans un schème événementiel.
>
> 2) Dans l'ensemble des langues du monde, on note que la presque totalité des « objets matériels » (sous-ensemble des entités) sont exprimés spontanément par des « noms », quelles que soient les propriétés qui les opposent à une autre classe. Mais l'inverse est totalement exclu : un nom n'entraîne aucun savoir sur la distinction « entité / comportement » [...].
>
> (Pottier 1987, §VI.5)

Ne serait-ce que les multiples transferts d'une catégorie à l'autre manifestent la perméabilité des classes et leur impossibilité d'enfermer chacune ce qui serait intrinsèquement une chose ou une action.

> Le substantif ne désigne pas plus une « substance » qu'autre chose. Que l'on juge ou non qu'il désigne une substance, tout sémème peut être lexicalisé par un substantif, sans autres limites que celles des règles dérivationnelles de la langue considérée. (Rastier 1987, p. 140)

Plus fondamentalement, c'est la relativité de toute ontologie qui est réaffirmée : le mode n'est pas formé d'entités et d'actions, mais plutôt la reconnaissance d'une entité ou l'identification d'une action est une manière de percevoir et de décrire la réalité.

> Sur la différence entre verbe et nom, souvent débattue, les définitions proposées se ramènent en général à l'une des deux suivantes : le verbe indique un procès ; le nom, un objet ; ou encore : le verbe implique le temps, le nom ne l'implique pas. Nous ne sommes pas le premier à insister sur ce que ces définitions ont l'une et l'autre d'inacceptable pour un linguiste. Il faut montrer brièvement pourquoi.
>
> Une opposition entre « procès » et « objet » ne peut avoir en linguistique ni validité universelle, ni critère constant, ni même sens clair. La raison en est que des notions comme procès ou objet ne reproduisent pas des caractères objectifs de la réalité, mais résultent d'une expression déjà linguistique de la réalité, et cette expression ne peut être que particulière. Ce ne sont pas des propriétés intrinsèques de la nature que le langage enregistrerait, ce sont des catégories formées en certaines langues et qui ont été projetées sur la nature. La distinction entre procès et objet ne s'impose qu'à celui qui raisonne à partir des classifications de sa langue native et qu'il transpose en données universelles [...].
>
> (Benveniste 1966, §13)

> Il apparaît donc que, pour caractériser en propre, et sans considération de type linguistique, l'opposition du verbe et du nom, nous ne pouvons utiliser ni des notions telles que objet et procès, ni des catégories comme celle du temps, ni des différences morphologiques. Le critère existe cependant, il est d'ordre syntaxique. Il tient à la fonction du verbe dans l'énoncé.
>
> Nous définirons le verbe comme l'élément indispensable à la constitution d'un énoncé assertif fini.
>
> (Benveniste 1966, §13)

> La distinction du verbe et du nom, qui n'apparaît pas toujours dans un mot anglais ou chinois pris isolément, se révèle immédiatement lorsque ce mot est placé dans une phrase ; ce n'est pas une question de forme, c'est une question d'emploi. [...] S'il y a des langues où le nom et le verbe n'ont pas de forme distincte, toutes les langues s'accordent pour distinguer la phrase nominale et la phrase verbale. (Vendryes 1923, §3)

Pour s'en tenir aux catégories fournies par un catégoriseur, le fonctionnement des catégoriseurs illustre, par construction, une définition morphologique et syntagmatique des catégories. En effet, l'étiquetage recourt à des critères sur la forme d'un mot (reconnaissance dans un lexique, suffixe ou terminaison caractéristique,...) et sur l'enchaînement local des catégories (règles sur fenêtres de 5 mots, chaînes de Markov d'ordre 2 ou 3,...). Ces catégories morphosyntaxiques peuvent avoir des affinités sémantiques et des effets sémantiques, mais sans que l'on puisse conclure à un quelconque déterminisme[55].

Isoler les mots grammaticaux (syncatégorématiques)

La constitution de listes de mots vides, dans les systèmes faisant des calculs sur des documents en texte intégral, se base pour une part importante sur la distinction entre mots grammaticaux et mots du lexique.

Les mots grammaticaux correspondent aux *syncatégorématiques*, traditionnellement reconnus comme les unités linguistiques qui n'auraient pas de signification en elles-mêmes mais servent à articuler les autres unités. Ces mots grammaticaux ont trois propriétés intéressantes : *(i)* leur

[55] (Dupuy 1993) ouvre ainsi la partie de son étude des rythmes sémantiques où il se propose d'observer les structures itératives morphologiques :
« est-il raisonnable de parier que le lecteur va percevoir [...] les itérations morphologiques [...] ? [...] Dévoilées par le linguiste, les structures itératives morphologiques ne sont-elles pas en effet pur artefact ? [...] Jackobson [...] montre [...] qu'il faut bien distinguer la capacité à ressentir intuitivement un effet, à appréhender un sens, de l'aptitude à expliciter ce qui les produit ». (*ibid*., p. 146)
L'issue des expérimentations et des mesures est une conclusion sur le même ton, convaincu et nuancé :
« les structures itératives morphologiques se révèlent des éléments sémiotiquement pertinents [...]. Il reste cependant que, dans bien des cas, l'interprétation n'est proposée qu'à titre d'hypothèse : autant le passage du lexical au sémantique paraît aisé (le lexique dénote certaines significations), autant celui du morphologique au sémantique est incertain. [...] Force est de convenir que les rapports entre les itérations morphologiques et le sens du texte sont remarquablement non déterministes : une configuration morphologique ne produit pas nécessairement *un* effet sémantique [fixé] » (*ibid*., pp. 247-248).

implication constante dans la construction des phrases, quel que soit le sujet traité ; *(ii)* leur organisation en paradigmes fermés, regroupant des formes invariables. La première propriété se traduit par le fait que l'élimination de ces mots ne fait pas perdre d'information sur la thématique du texte, tout en réduisant considérablement le volume de données à retenir pour la description. La seconde ajoute la possibilité de les énumérer tous et sans difficulté (liste des pronoms personnels sujet, liste des conjonctions de coordination, etc.), et de les reconnaître par identification à un élément de la liste (sauf cas d'homographies).

Les mots du lexique –ou *catégorématiques*–, non entièrement énumérables, se laissent alors définir comme le reste, à savoir ce qui n'a pas été reconnu comme mots grammaticaux.

L'utilisation d'un catégoriseur présente l'avantage de définir les mots grammaticaux de la liste de mots vides de façon plus *synthétique* et *systématique*. En effet, il suffit de mentionner les catégories correspondantes (*prépositions*, *déterminants*, etc.), et l'on ne court plus le risque d'oublier une forme particulière. De plus, le catégoriseur résout les cas d'*homographie*, et donc par exemple d'éliminer comme mot vide la conjonction *car* tout en gardant comme unité descriptive le nom *car*.

Quelques considérations linguistiques nuancent cette présentation trop simple.

La ligne de partage qui séparerait items grammaticaux et items lexicaux ne suit pas le découpage qui distingue les différentes parties du discours. Certaines catégories, comme les conjonctions ou les prépositions, semblent bien n'être constituées que de mots grammaticaux. Pour autant, l'éventail des locutions met en question la possibilité d'énumérer entièrement tous les éléments de ces catégories. Quant aux catégories qui semblent les plus lexicales (verbes, noms), elles n'ont rien d'homogène (par exemple, auxiliaires, verbes modaux, gallicismes), et certains de leurs éléments se grammaticalisent (Pottier 1992, §IV.2).

L'unité grammaticale ou lexicale est généralement en deçà du mot : la marque du genre ou du nombre est un grammème qui se combine à une racine lexématique dans l'expression d'un nom. *Grammème* et *lexème* sont donc des concepts linguistiques plus justes que *mot grammatical* et *mot du lexique*, étant donné que l'unité grammaticale ou l'unité lexicale ne suit pas le découpage des mots. Si certaines parties du discours (prépositions, etc.) recouvrent des éléments grammaticaux qui se manifestent par des mots autonomes, d'autres éléments grammaticaux sont greffés et intégrés aux éléments lexicaux. Une même indication grammaticale a d'ailleurs une autonomie d'expression variable selon les langues, et ce qui est traduit ici par un mot sera représenté là par un procédé morphologique (Vendryes 1923, §2.III).

Enfin, les catégories grammaticales ont une contribution sémantique, dont l'intérêt pour une application de recherche documentaire et de confrontation texte à texte n'est pas nécessairement nul. L'expression des modalités fait par exemple largement appel aux grammèmes, et participe à la définition d'un point de vue sur les thèmes abordés ; les marqueurs temporels peuvent aider à situer la phase d'avancement d'une réalisation ; etc.

Un juste milieu peut être de prendre en compte des éléments grammaticaux dans le traitement, mais en leur accordant une place spécifique, compte tenu de leurs propriétés bien particulières (notamment leur fréquence élevée, qui domine celle des autres unités lexicales). Max Reinert (Reinert 1990) propose, dans les analyses effectuées par son système ALCESTE, d'en faire des éléments illustratifs, c'est à dire ne participant pas aux calculs de construction des univers thématiques sous-jacents au corpus, mais pouvant ensuite être rapportés à tel ou tel univers, et concourir ainsi à l'interprétation des résultats.

Le recours aux catégories identifiées par un catégoriseur est donc une aide pour identifier des éléments qui jouent un rôle grammatical, de façon régulière (tous les éléments d'une catégorie sont pris en compte), et en déjouant les homographies.

Discerner certains homographes

En s'appuyant sur l'enchaînement syntagmatique des unités qui précèdent et qui suivent, des unités lexicales qui ont la même forme graphique se trouvent répertoriées dans des catégories distinctes. Or quelquefois le lien des formes graphiques trahit l'appartenance à une même famille, et le simple découpage, qui confond ces formes, garde une information sémantique que le catégoriseur efface.

Ce qui est intéressant sémantiquement, ce sont les formes analogues correspondant à différentes dérivations d'une même racine : *une analyse* / *l'outil analyse* ; *une information valide se trouve...*/ *l'expérience valide l'hypothèse* ; *en informatique / la réalisation informatique*. Les exemples se multiplient pour l'anglais, du fait de sa morphologie moins variée (notamment pour la conjugaison des verbes, et les homographies *verbe - nom*). Cependant, l'analogie fortuite des formes de deux mots de la même famille constitue une information très partielle et irrégulière, et donc difficilement exploitable. L'utilisation explicite de ce type d'information est plutôt à envisager avec des outils qui repèrent *systématiquement* des liens dérivationnels. Inversement, des analogies de forme se produisent tout aussi bien entre des mots de sens très éloigné[56], et toutes les homographies ne se laissent pas décrire par une différence de catégorie grammaticale.

En revanche, pour les cas d'homonymie avec des mots grammaticaux, l'apport positif d'un catégoriseur est indiscutable. Il est tout à fait fondé de faire la distinction (*est*, verbe (ou auxiliaire) *être* ou point cardinal). L'homonymie avec des mots grammaticaux ne concerne pas une proportion importante de mots différents, mais est quantitativement très significative du fait des fréquences d'usage élevées des mots grammaticaux. L'usage d'un anti-dictionnaire ou d'une liste de mots vides devient alors beaucoup plus sûr et plus précis : on ne se prive plus du *car* (autobus ou navette), de l'*or* noir, du *la* musical autour duquel on se met au diapason, du *son* (sonorité), d'un résultat *nul*, etc. Le *pouvoir* et le *devoir* redeviennent des concepts à part entière, et ne se confondent pas avec l'expression d'une modalité (ex. : *pour pouvoir réaliser ceci*).

Trois points peuvent relativiser l'importance de ces distinctions.

D'une part, les homonymes d'une forme peuvent être quasi inexistants dans les genres de documents auxquels on s'intéresse. Les domaines concernés excluent naturellement les formes qui pourraient prêter à confusion : pour la mise en œuvre de DECID à EDF, on reste dans le registre des documents professionnels techniques, économiques, stratégiques ou administratifs, à l'exclusion des œuvres de fiction ou de documents culturels. A ce titre, la prise en considération des noms *la* ou *savons*, ou de la 2ème personne pour *tu* et *ton*, peut s'avérer purement spéculative. Le cas de *pas* (nom ou négation) mérite examen. A l'inverse, le cas de *as* peut tout simplement disparaître, en l'absence des dénominations superlatives, des jeux de cartes et de l'usage de la deuxième personne du singulier pour les genres considérés.[57]

D'autre part, il arrive que les distinctions apportées ne soient pas utilisées pour la suite du traitement. Par exemple, si le choix du traitement est d'éliminer les mots grammaticaux, peu importe d'avoir distingué le pronom et le déterminant pour les formes *l'*, *le*, *les*, *leur*, *ce*, de savoir séparer *si* adverbe (intensif) de *si* conjonction (condition), de décrire toutes les valeurs de *tout*, *comme*, *en*. Or c'est justement ce genre d'ambiguïté entre formes grammaticales qui domine lors de l'étiquetage morpho-syntaxique d'un corpus (Habert, Nazarenko, Salem 1997, §VIII.3.1). L'identification correcte de *des* comme article indéfini ou bien contracté *de* + *les*, sur laquelle achoppent clairement la plupart des catégoriseurs, n'est alors plus un handicap. Plus généralement, même des distinctions entre un

[56] (Fluhr 1977, p.147) cite :
« - des ambiguïtés verbe-substantif : *volant*, *marche*, *couvent*, *avions*
- ou verbe-adjectif : *content*
- ou substantif-adjectif : *commune*, *communs*, *fin*, *fine*
- ou substantif-adverbe : *bien*, *rien*, *mal* »

[57] De ce point de vue, il est instructif d'observer la constitution des listes de mots vides en ce qui concerne ces homographes.
- dans (Fluhr 1977), *or*, *car*, *la*, *ton*, *son* sont explicitement donnés à titre d'exemple d'homographies lexique - grammaire ; dans la liste de mots-vides donnée en annexe, et qui a été constituée pour être utilisée en l'absence d'analyse permettant de résoudre ces homographies, on trouve *la* et *son* (dont on peut penser que l'usage en tant que mot grammatical est largement dominant), mais pas *or* et *car* (dont on ne veut sans doute pas perdre les occurrences en tant que mot du lexique), ni *ton* (sans doute trop rare, aussi bien en tant que mot grammatical que en tant que mot lexical).
- dans (Chartron 1988), la liste de mots vides utilisée pour le corpus de documents EDF comporte *car* et *la*, mais pas *or*, *ton* et *son*. Là encore donc, des considérations de fréquence et d'importance respective des réalisations lexicales *vs* grammaticales départagent le sort fait aux cinq homographes.

mot grammatical et un adjectif (ex. *certain*) ou un nom (ex. *fait*), tous d'usage (trop) courant, ne sont pas cruciales pour la plupart des applications documentaires.

Enfin, si éliminations ou erreurs il y a, elles ne touchent que très peu de mots, et le reste du contexte peut suffir à caractériser le texte, le défaut étant comme interpolé.

Des indices sur les (non) relations syntagmatiques

Pour rendre compte des zones d'interaction forte possible entre les mots, Ghislaine Chartron se donne une « liste de mots qui sont sous-entendus dans le libellé des formes composées », et qui regroupe les articles définis et indéfinis, les articles définis contractés, et un choix de prépositions (*à*, *de*, *d'*, *en*, *par*, *pour*) (Chartron 1988, annexe C.1)[58]. Didier Bourigault, pour repérer des termes (nominaux) avec le logiciel LEXTER, adopte l'approche inverse, en s'attachant à repérer des catégories frontières (Bourigault, Gros 1993, §I.3.5). Chartron définit donc en quelque sorte une méta-catégorie d'unités « perméables » aux connections syntagmatiques formes, Bourigault une méta-catégorie d'unités « imperméables » (incluant par ex. : *ponctuation*, *verbe conjugué*, *conjonction de subordination*, *pronom*)[59].

Si le catégoriseur retourne des informations sur le genre, le nombre, la personne, un lien de dépendance possible entre deux unités lexicales peut être infirmé par l'absence d'accord entre les deux formes. En revanche, le seul fait que deux formes soient accordées au plan de ces catégories grammaticale ne suffit pas à indiquer un lien syntaxique direct entre ces deux formes. Les indications d'accord servent donc à restreindre la combinatoire des relations syntaxiques possibles plutôt qu'à suggérer des relations.

Les mots de certaines catégories se comportent syntaxiquement comme des compléments, des adjonctions à la structure centrale de la phrase. Les masquer peut aider à retrouver le squelette syntaxique, et en particulier certaines relations que ces compléments, en s'intercalant, rendaient plus difficiles à repérer par l'outil d'analyse.

> Il existe un certain nombre de valeurs grammaticales telles que si l'on supprime le mot correspondant, la phrase reste syntaxiquement correcte. On peut citer certaines catégories d'adverbes, d'adjectifs épithètes, etc. [...] [Un tel mot] peut être supprimé pour refaire une analyse locale qui traitera [ses deux voisins] comme mots consécutifs. (Fluhr 1977, §II.3.5.1.C)

Repérage de genres textuels et de parties

Cette piste semble encore peu utilisée. Pourtant, il a été montré que la morphosyntaxe utilisée varie selon les genres des textes (Biber 1988), et même selon les parties à l'intérieur d'un document (voir notre étude sur le corpus des ARD, en annexe).

---

[58] AlethIP/GN inclut davantage de prépositions dans ses patrons ; en particulier il sélectionne les suites *nom préposition nom*, où la préposition peut être non seulement *à*, *de*, *en*, *par*, *pour* mais aussi *avec*, *dans*, *entre*, *sans*, *sous*, *sur* (Gros & *al*. 1997). Ceci évite de trop fortes restrictions *a priori*, mais multiplie les difficultés de découpage et de rattachement, du fait des confusions possibles entre un complément du nom, une expansion d'un adjectif, et un complément circonstanciel dépendant du verbe.

[59] La simplicité de constitution et d'utilisation d'une telle liste n'est qu'apparente. Ainsi, chez Bourigault, la détermination des unités frontières ne se fait pas uniquement par des catégories, mais aussi par des « patrons » qui décrivent des enchaînements de catégories (par exemple : *préposition immédiatement suivie d'un adjectif possessif*). De plus, il y a des cas particuliers : les prépositions qui entrent dans une construction de sous-catégorisation sont préalablement repérées, rattachées au nom ou à l'adjectif dont elles introduisent le complément, et ne sont donc pas considérées isolément. Ainsi, pour un nom suivi d'une des prépositions *à*, *avec*, *contre*, *dans*, *par*, *pour*, *sous*, *sur*, *vers* et qui apparaît dans des contextes différents, la préposition ne constitue pas une frontière. Pour un adjectif qui peut entrer dans la construction *auxiliaire être + adjectif + préposition à*, si la préposition à contribue à indiquer une frontière, alors si elle suivait l'adjectif celui-ci prend également le statut de frontière.
D'autre part, certaines catégories morphosyntaxiques sont redivisées plus finement. Par exemple,
- les adverbes sont des frontières, sauf *très*, *non*, *tout*, *ultra* (devant un adjectif) ;
- à l'inverse, les adjectifs ne sont pas des frontières, sauf certains (antéposés) : *autre*, *bon*, *certain*, *deux*, *différent*, *divers*, *fort*, *même*, *nombreux*, *nouveau*, *principal*, *seul*, *tel*.
Les catégories morpho-syntaxiques seules sont trop grossières vis-à-vis du traitement (Gros & *al*. 1997).

Des heuristiques toutes simples, et utiles pour cerner la structuration interne d'un texte, peuvent être imaginées, en s'appuyant sur l'information de catégories morpho-syntaxiques. Ainsi, une phrase nominale isolée, se signalant par l'absence d'un verbe (conjugué), et éventuellement l'absence de déterminant (en particulier qui ne commence pas par un déterminant autre qu'un article, ou qu'un article défini), renforce la présomption que l'on a affaire, selon le contexte d'analyse, à un titre, ou à une mention synthétique dans une liste, ou peut-être qu'il s'agit d'une requête de type mots-clés, etc.

Outils existants

Le catégoriseur le plus connu est celui d'Eric Brill, pour l'anglais, dont la tactique s'apparente à celle de l'algorithme de Fluhr en France (Fluhr 1977)[60]. Il repose sur un apprentissage des successions des catégories sur des fenêtres de sept mots. Grâce à ce principe, il se prête à une transposition à divers corpus et à d'autres langues. Ainsi, il a été adapté au français à partir d'un corpus littéraire (essentiellement la première moitié de *Germinal* de Zola) : voir les travaux de Josette Lecomte à l'INaLF. A l'INaLF également, un très bon catégoriseur « maison » a été mis au point par Marc Papin (linguiste) et Jacques Maucourt (informaticien). Aux dires de Didier Bourigault (qui a expérimenté plusieurs catégoriseurs en amont de LEXTER), ce catégoriseur serait l'un des meilleurs actuellement. Les outils commercialisés (par exemple *AlethCat*, de la société GSI-Erli) peinent à faire valoir leur prix par un gain de qualité significatif[61].

Il faut savoir aussi que la plupart des analyseurs syntaxiques sont également en mesure de fournir un étiquetage morpho-syntaxique d'un corpus. Se construisant une représentation à l'échelle de la proposition, ils savent traiter des relations à plus longue distance que les catégoriseurs standards travaillant sur une fenêtre de quelques mots, ce qui leur permet d'atteindre les meilleures performances. Les catégoriseurs standards restent néanmoins intéressants : ils sont plus simples à développer, les règles sur contexte local sur lesquelles ils s'appuient se modifient plus facilement, et ils atteignent déjà des performances honorables (96 à 98 % d'étiquettes correctes).

De même, on trouvera des catégoriseurs parmi les outils de lemmatisation.

Un panorama général des catégoriseurs pour le français est fourni par l'initiative GRACE (*Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation*). La première session de GRACE a été lancée fin 1995 et concerne précisément les systèmes d'assignation de catégories grammaticales. Elle a réuni 13 participants[62], que l'on peut estimer représentatifs de l'état de l'art en la matière.

e) Lemmatisation

Définition et motivation

La lemmatisation consiste à neutraliser les variations flexionnelles d'une même unité lexicale. Tout mot est alors rapporté à son lemme, à savoir une forme canonique qui dépend de la catégorie grammaticale : un nom est mis dans sa forme singulier[63], un adjectif au masculin singulier, un verbe à l'infinitif.

Le traitement a ainsi accès à une information linguistique importante. Les régularités morphologiques sont un indicateur fort d'une sémantique commune. L'organisation des dictionnaires l'indique bien, puisque les entrées sont justement des lemmes (par exemple, l'entrée d'un adjectif

[60] Christian FLUHR est le concepteur du système documentaire SPIRIT (*Système Syntaxique et Probabiliste d'Indexation et de Recherche d'Information Textuelle*). Il a mis au point son catégoriseur en vue d'une telle utilisation –c'est d'ailleurs au traitement effectué par le catégoriseur que renvoient les qualificatifs *Syntaxique* et *Probabiliste* du nom du système.

[61] De fait, la société Erli a redéfini sa gamme d'outils autour d'AlethIP, et ne propose plus dans ses produits le catégoriseur seul.

[62] Les 13 participants finaux à l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français (1995-1998) sont les entreprises et laboratoires suivants : CITI, CLIPS, CNET, CRISTAL, GREYC, IBM, INaLF, Ingénia, ISSCO, LATL, LIA, LIMSI, Rank-Xerox.

[63] Certains noms se présentent sous les deux genres, masculin et féminin. Il s'agit quelquefois d'un même lemme (*boulanger*, *boulangère*), mais pas toujours (*guide*, *livre*, *manœuvre*, *mémoire*, *mode*, *physique*, *poste*, *tour*,...).

regroupe ses variations en genre et en nombre) : la définition du sens est donc la même pour toutes les flexions du lemme (dans le cas général). Pour illustration, le lecteur voit immédiatement que *un profil* ou *des profils* désigne la même réalité ; grâce à la lemmatisation, la machine peut aussi s'appuyer sur cela pour construire la représentation du texte.

L'effet quantitatif le plus évident est le regroupement de formes, différentes dans leur réalisation graphique en chaînes de caractères : cela va dans le sens d'une réduction du nombre d'unités, ce qui est *a priori* bénéfique pour la suite des calculs. Il s'y superpose toutefois un effet inverse, de moindre ampleur, d'augmentation du nombre d'unités. En effet, des homonymes, confondus dans leur apparence graphique, sont distingués par une lemmatisation morphosyntaxique. Par exemple, *avions*, qui constitue une unité d'un point de vue graphique, peut être associé à plusieurs lemmes : *avoir (verbe)*, *avion (nom)*. Des locutions (comme *bien que*) peuvent aussi être reconnues, complétant donc la description par de nouvelles unités.

Discussion : sémantique et usages

Cependant, contrairement à l'intuition, l'unité lexicographique n'est pas toujours une unité sémantique, tant s'en faut. Les formes différentes correspondent parfois à des usages tout à fait distincts, et les regrouper est une perte importante d'information sémantique. Cela ne se résume pas à quelques exceptions « voyantes » comme *lunette* vs *lunettes* ou *ciseau* vs *ciseaux*.[64]. On peut par exemple à la DER évoquer « le travail de bureau », et « les travaux devant le bâtiment T ». Dans l'étude d'un corpus de romans français, Evelyne Bourion (Bonhomme & *al*. 1996) a montré que « au pied de » et « aux pieds de » renvoient à des contextes et à des thématiques disjointes : la forme singulier introduit une indication de lieu par rapport à un objet (montagne, escalier, autel,...), la forme plurielle sert à décrire une scène d'imploration et s'applique à un personnage. L'excellent *Plaidoyer pour une non-lemmatisation* (Geoffroy, Lafon, Tournier, 1974) pourra encore compléter l'argumentation, chiffres à l'appui.

En fait, lemmatiser un texte, c'est lui appliquer les transformations édictées par un modèle morpho-syntaxique. Ceci se justifie pleinement si l'étude que l'on fait du texte est du même ordre : morpho-syntaxique, lexicale. Par exemple, étudier la proportion des substantifs, ou rechercher des contextes pour illustrer un article de dictionnaire. En revanche, rien ne dit que cette grille concorde avec les effets sémantiques davantage que les formes non-lemmatisées n'articulent les parcours interprétatifs (Tournier 1985).

Eviter la réduction éliminatrice et irréversible

D'autre part, la lemmatisation devrait se concevoir en termes d'analyse plutôt que de réduction. En effet, l'effacement des flexions est précédé de leur reconnaissance. La flexion traduit un certain nombre de valeurs de catégories grammaticales (genre, nombre, temps, mode, personne, etc.). Ces valeurs peuvent être elles-mêmes considérées comme des unités. Certaines analyses du discours font grand cas des usages des personnes (emploi de la première personne, emploi de la deuxième personne, etc.). Pour l'analyse des descriptifs d'activité dans DECID, la connaissance des temps et modes dominants dans un passage renseigne sur la phase décrite (résultat acquis, propositions prospectives, etc.) La lemmatisation serait alors non pas une pure réduction par élimination des flexions, mais la dissociation d'unités lexicales et grammaticales (ces dernières étant issues des marques amalgamées dans la flexion).

La lemmatisation n'est donc ni toute bonne, ni toute mauvaise, pour un traitement comme celui de DECID. La voie médiane est de disposer de l'information, sans l'imposer. Trancher en interdisant au système de percevoir des relations morphologiques fondamentales, ou en gommant des différences très significatives au plan sémantique, ne peut pas être satisfaisant. Si en revanche le système dispose de l'information qu'il y a une relation sémantique vraisemblable entre plusieurs unités d'analyse, il peut évaluer si cette information se révèle utile pour sa description, si elle

[64] On se voit néanmoins quelquefois demander « un ciseau pour découper ce papier » : la mise en exergue de l'opposition *ciseau* vs *ciseaux* procède d'une approche *normative* de la langue, mais peut-être pas tout à fait *descriptive* des usages.

« fonctionne ». Un rapprochement incongru de deux unités a toutes les chances de rester une hypothèse infructueuse et de s'éteindre d'elle-même.

Tous les lemmatiseurs ne permettent pas ce suspens de l'analyse : souvent, les résultats sont un ensemble de lemmes, sans mise en relation explicite avec la forme fléchie trouvée dans le texte. Pour DECID, on souhaiterait avoir les propositions de lemmes en lien avec les unités du texte.

Outils

La société *Erli* a remplacé *AlethCat* (Herviou 1994) par *AlethIP* (Herviou-Picard 1996), plus général (il fournit au choix lemmatisation, extraction de syntagmes ou indexation contrôlée), et qui opère avec une grammaire et un moteur entièrement nouveau. La société *Cora* commercialise le module *Lemma* (utilisé dans l'application d'indexation automatique *Darwin*). Là aussi, un paramétrage permet d'aller un peu plus loin en ajoutant pour chaque verbe les noms de la même famille : si l'analyseur rencontre le verbe *juger* employé dans une phrase, alors il renvoie non seulement l'infinitif *juger* mais aussi les noms *juge* et *jugement*. Le logiciel *INTEX* (conçu par Max SILBERZTEIN, au LADL, université de Paris 7) est fort des très riches dictionnaires sur lesquels il s'appuie, pour la reconnaissance des formes fléchies et des formes composées (expressions figées). Il offre également un jeu d'automates pour repérer certaines formes particulières, comme les expressions de date.

f) Réduction flexionnelle et dérivationnelle : retour à la racine

Principe et motivation

Chaque mot est dissocié en ses parties lexicales (racine, affixes) et ses parties grammaticales (déclinaison, conjugaison).

En général, on choisit ensuite de négliger les composantes ajoutées à la racine, afin de regrouper les mots de la même famille (au moins apparemment). L'idée est de retrouver la notion ou le concept sous-jacent, indépendamment des variantes induites par les différents contextes syntaxiques. Par exemple, on pourrait regrouper ainsi : *expérience*, *expérimental* (et *expérimentaux*), *expérimenter*, *expérimentalement*, *expérimentation*, avec peut-être aussi *expert*, *expertise*, etc. La réduction est également avantageuse d'un point de vue algorithmique, pour les traitements effectués : diminution du volume des données à gérer, ordre de grandeur mieux adapté aux techniques courantes d'analyse des données, renforcement des caractéristiques associables à chaque unité (sinon dispersées et affaiblies), récupération d'unités qui, isolément, sont trop rares pour être significatives (assimilation à des écarts marginaux et négligeables). Enfin, ce regroupement apparaît comme une aide dans la lecture de résultats, dans la mesure où il évite une dispersion d'unités alors qu'elles sont perçues en rapport les unes avec les autres (Le Roux 1995).

Une simple procédure manipulant les chaînes de caractères peut décrire le lien entre *nocif* et *nocivité*, mais ne saura pas reconnaître des transformations comme *nuire* / *nocif*. Ces cas peuvent être enregistrés dans des dictionnaires (jamais exhaustifs), qui couvrent *a minima* les cas les plus fréquemment rencontrés, notamment dans les formes conjuguées des verbes *être* et *aller*.

Discussion linguistique : la question des relations entre forme et sens

Si le procédé est calculatoire (extraction de préfixes, de suffixes, de terminaisons et de racines), il appuie une conception compositionnelle de la morphologie (en isolant des constituants). Au plan sémantique, cela rejoint la question de la motivation, à savoir de la relation entre la forme, l'expression qui représente le mot (appelée *signifiant*), et le sens qu'on lui associe (appelé *signifié*).

1. *L'absence de motivation*

Un francophone n'a pas le sentiment que *érable* ou *moisir* puissent se rattacher à d'autres lexèmes de la langue, et il n'en voit pas de structure analysable. De même la séquence graphique *-eau-* de *eau*, *seau*, *beau*, *veau*, *sceau* ne joue aucun rôle sémantique.

Par contre, nous remarquons que pratiquement tous les verbes français en *-oir* expriment des modalités (*savoir*, *pouvoir*, *devoir*, *valoir*, *vouloir*, *voir*, *percevoir*... [...]), et le phénomène a ses racines en latin (verbes en *-ere*).

> Le besoin de motivation est toujours puissant, et il s'exprime à travers les « étymologies populaires » et les faux rapprochements (du type *ouvrable*, de la série *ouvrier*, *œuvrer*, relié à *ouvrir*).
>
> 2. *L'isomorphisme signifiant / signifié*
>
> On a souvent remarqué qu'il y avait une corrélation entre la « marque » sémantique (un plus de sèmes) et une augmentation physique du signifiant : *chien* / *chienne*, [...] *pensons* / *pensions*, *venir* / *ne pas venir*, *partir* / *repartir*. On ne peut généraliser, mais il s'agit d'un exemple de la naturalité du comportement sémiologique.
>
> 3. *La motivation interne*
>
> Il est possible que des éléments du mot soient identifiables. *Quinze* rappelle difficilement *cinq* (lat. *quindecim*, *quinque*), alors que *dix-sept* est explicite. *Carnivore* dit bien qu'il s'agit de manger de la chair, mais *végétarien* ne suggère que le végétal sans autre précision.
>
> Lorsqu'une langue évite les emprunts, elle explicite la néologie avec ses propres éléments. [...] Le mot *automobile* était déjà bien motivé, mais on l'a vite oublié. [...]
>
> 4. *Noyau sémique et signifiant*
>
> La racine des langues sémitiques offre un bon exemple de la stabilité d'un noyau sémique (constante) à travers des séries dérivationnelles. Comparons ces mots français, reliés directement ou indirectement à l'arabe [qui présentent tous le motif] |S-L-M| : *Salomon*, *salamalec*, *islam*, *musulman*. [...]
>
> On remarque la variété des signifiants du français, comme dans la série suivante à travers laquelle on reconnaît sans peine un noyau sémique [*i.e.* une unité de sens qu'ils ont en commun] : *œil*, *voir*, *lunettes*, *télescope*, *regarder*, *miroir*, etc. [...]
>
> (Pottier 1992, §IV.6)

Si l'on fait appel à une réduction dérivationnelle, il faut donc être conscient des phénomènes de « silence » et de « bruit » associés. *Silence* : on (ne) décrit (qu')une partie des rapprochements possibles : l'opération n'est pas régulière et systématique, elle ajoute certaines relations mais pas d'autres. *Bruit* : les relations repérées ne sont pas toujours justes. Deux cas sont intéressants à noter :

- certaines ne sont pas justifiées étymologiquement, mais pour autant la proximité (notamment phonétique) des expressions induit une parenté de sens dans l'esprit du lecteur (ce sont notamment les « étymologies populaires »). Pour DECID, comme il s'agit de travailler au plan du sens et de l'interprétation, cette forme d'erreur (linguistique) n'est pas gênante, au contraire[65].
- certaines relations sont fondées étymologiquement, mais chacun des mots, de par son autonomie et ses usages propres, a évolué indépendamment, si bien qu'un lecteur ne perçoit plus spontanément de rapport sémantique entre eux (cf. la non compositionnalité de la langue). Ces rapprochements cocasses s'évitent en constatant les différences de contexte d'emploi.

Autrement dit, la mise en relation de formes qui dérivent d'une même racine apparente rend compte d'un lien que la morphologie rend perceptible, mais que quelquefois des usages très éloignés masque. La morphologie n'épuise pas, loin de là, les relations de proximité sémantique que l'on peut établir entre les mots. Ainsi, des affinités phonétiques peuvent, sans que ce soit systématique, se doubler d'une effet de sens (par confusion ou contagion). Les modes d'abréviation des mots mettent en rapport un mot et une forme contractée, mais ne sont pas prises en considération dans les transformations morphologiques classiques. Enfin, des séries de (para)synonymes traduisent l'affinité sémantique forte de mots extrêmement différents quant à leur forme. Quand elle effectue des rapprochements, la procédure de réduction dérivationnelle explicite des relations très pertinentes dans certains cas, accentue et redouble des effets de polysémie et d'homonymie dans d'autres. Pour être utilisée, elle ne doit donc pas être prise au pied de la lettre : les liens morphologiques trouvés sont particulièrement intéressants à considérer puisqu'ils sélectionnent des associations linguistiquement motivées ; ils sont donnés à titre indicatifs et peuvent être remis en cause par une exploration des contextes.

Variations dérivationnelles et contexte(s)

Les variations dérivationnelles rapprochent des formes de catégories morphosyntaxiques différentes, entrant dans des constructions syntaxiques différentes. Les contextes proches, à savoir ceux organisés par la syntaxe (expressions composées, groupes qui reçoivent une fonction

[65] (Hérault 1981) prend le parti inverse, de toujours respecter l'étymologie dans les regroupements à partir des racines.

grammaticale, phrase), peuvent donc être notablement différents (surtout si ces contextes sont considérés sans réduction dérivationnelle). En revanche, les contextes plus larges, extra-syntaxiques (paragraphe, texte), caractérisent le domaine d'emploi indépendamment des contraintes syntaxiques impliquées par la catégorie.

Certains effets contextuels des transformations par dérivation échappent à une description au palier du mot. Le passage d'un verbe à un nom s'accompagne d'une transposition de l'adverbe en adjectif qui ne conserve pas nécessairement la racine, il arrive qu'une autre forme supplée : *éliminer vite un problème → élimination rapide du problème*. Cette transformation peut dépendre du contexte, comme pour : *consommer beaucoup d'électricité → un gros consommateur d'électricité*, mais *confier beaucoup → une grande confiance*.

Techniques de décomposition, constitution et reconnaissance des racines

La réduction dérivationnelle n'est pas un simple réaménagement d'une réduction flexionnelle

La réduction dérivationnelle est d'un ordre de difficulté supérieur à celui de la réduction flexionnelle, car l'inventaire des flexions est grammatical et fini, alors que celui des préfixes ou suffixes est lexical et ouvert.

D'autre part, les variations dérivationnelles sont encore moins régulières et systématiques, ce qui joue pour la mise au point de règles dans les traitements automatiques :

> Si la prise en compte des flexions est relativement aisée dans un analyseur morphologique, en revanche celle des dérivations l'est beaucoup moins : les mots dérivés constituent en effet, comme d'ailleurs les mots composés, un inventaire ouvert, en constante évolution ; de plus, leur analyse pose des problèmes théoriques difficiles. Certes il existe des dérivations extrêmement productives en français, comme la dérivation d'un adverbe à partir d'un adjectif suivi du suffixe *-ment* ou la préfixation d'un verbe par la forme *re-* : la représentation de ces découpages fournit un gain appréciable pour la description syntaxique et sémantique des unités dérivées (même si un préfixe comme *re-* est lui-même polysémique, son apport sémantique à un verbe donné obéit à des régularités qui peuvent être explicitées). Mais là encore, il est difficile de savoir jusqu'où systématiser ces mécanismes. Les nombreux « trous » et irrégularités apparentes du système de dérivation peuvent faire douter de l'efficacité d'une prise en compte des faits de dérivation dans un analyseur morphologique. Pour n'en citer que quelques exemples : pourquoi existe-t-il les deux dérivés *décentrage* et *décentrement* dans le cas du terme préfixé, mais seulement *centrage* et pas **centrement* dans le cas du terme non-préfixé ? si *carpette* (au sens de « tapis ») est à considérer comme une forme complexe, faut-il dériver celle-ci d'une base **carp-* inexistante (aucun rapport étymologique ou sémantique avec *carpe* « poisson ») ? pourquoi *truchement* n'est-il pas décomposable en *truch-ement*, comme *err-ement* ? pourquoi un *anti-clérical* est-il un mot dérivé signifiant « opposé au clergé », alors que l'*antimoine* ne signifie pas « opposé au moine » et ne doit pas plus être décomposé que l'*antilope* ? [...]
>
> (Fuchs & *al*. 1993, §3.1.2)

Dans l'optique de l'analyse d'un corpus, une manière efficace d'écarter la plupart des cas artificiels (incorrects) est de s'en tenir aux formes que le corpus atteste. De toutes façons, si l'on vise une *réduction* dérivationnelle, alors une dérivation supposée n'est effectivement utile que si elle permet un rapprochement avec au moins une autre unité du corpus. Les réductions ainsi opérées sont relatives : elles ne se superposent pas exactement avec ce qu'un dictionnaire de langue le plus complet pourrait prévoir.

> La réduction d'un mot n'est effectuée que dans la mesure où elle permet un regroupement et donc dépend de la distribution du vocabulaire dans un corpus donné. D'autre part, cette réduction est applicable à des mots de la langue parlée qui ne sont pas forcément répertoriés dans un dictionnaire, réduction utile lors de l'analyse d'entretiens ou de récits d'enfants, par exemple. (Reinert 1990, §2.2)

Recherche automatique de racines, en se basant sur un corpus

Les techniques de recherche de racines à partir de l'exploration d'un corpus (Fluhr 1997, p. 152)[66] s'appuient sur les propriétés suivantes : *(i)* les terminaisons possibles sont connues (le système des flexions est fermé : conjugaisons, déclinaisons, marques du genre et du nombre...) ; *(ii)* les suffixes et préfixes sont des éléments utilisés pour la construction de diverses unités (on relève chacun dans plusieurs formes différentes si le corpus n'est pas trop petit) ; *(iii)* de même, la racine d'un mot se retrouve dans d'autres mots (de la même famille) ; *(iv)* la recherche la plus efficace commence par examiner, pour chaque mot initiant la recherche d'une racine, la racine possible la plus longue ; *(v)* le tri préalable par ordre alphabétique des mots à considérer regroupe les formes et les présente dans un ordre optimal (surtout si l'on ne cherche pas à détacher les préfixes).

La racine peut regrouper plusieurs racines au sens strict (notamment pour les verbes *être* ou *aller*). Même dans le cas général, une racine ne s'identifie pas à une (sous-)chaîne de caractères, elle connaît des variantes d'écriture souvent pour des raisons phonétiques (redoublement, ajout, transformation ou élision d'une lettre par exemple).

Le plus efficace consiste à allier un traitement par dictionnaire aux procédures d'exploration de corpus. Ainsi, les formes les plus courantes (et particulièrement celles qui sont irrégulières) sont reconnues et transformées sur la base de leur racine et de leur terminaison ; d'autres formes ne sont reconnues que sur la base de leur racine, suffisamment déterminante ; et pour les formes qui ne sont pas connues du dictionnaire, le traitement peut au moins tenter d'identifier les différentes terminaisons. C'est ce genre de technique qui est disponible[67] dans le logiciel ALCESTE (Reinert 1990) (Reinert, Piat 1995).

De la décomposition morphologique comme une combinaison d'éléments fondamentaux, pour une description partielle des textes

L'approche de Daniel Hérault (Hérault 1981), bien que discutable à certains égards, mérite d'être considérée. Sa conviction de départ est la suivante :

> il y a, pour le moins, dans chaque dictionnaire deux secteurs principaux et que l'on peut considérer comme disjoints : l'un de type « sémantique fermée », qui correspond au noyau prédicatif de la langue ; l'autre, du type « sémantique ouvert », qui contient, pour l'essentiel, les racines fortement nominales, associées principalement aux objets et à tout ce qui n'est pas une action ou un processus. (p. 86)
>
> [...] les éléments du spectre sémantique (racines, préverbes, certains infixes et suffixes), dès lors qu'ils correspondent à une action importante, impliquent presque toujours une construction syntaxique bien déterminée, apparaissent avec une grande stabilité, de même que les schémas effectivement utilisés du système dérivationnel (c'est-à-dire la « combinatoire » de ces éléments). (p. 84)

Par l'examen d'un corpus de quatre textes (38 000 mots, 140 pages), il établit une table de 656 *racines* (870 en comptant les variantes qui s'ajoutent à certaines racines), qui se combinent chacune avec quelques modificateurs principaux pour former 1124 *bases*. 65 racines (plus 18 variantes ou racines secondaires) sont sélectionnées sur un critère de stabilité et de productivité : « toutes variantes confondues, ces racines s'associent à trois préverbes ou plus, la négation *in-* et ses variantes étant exclues ». Il est intéressant de noter que ce critère rejoint un principe moteur que l'on retrouve dans les autres tactiques de recherche de racines. Ces 83 racines appartiennent au *noyau prédicatif* du français, dont la taille effective est estimée à 200 éléments.

---

[66] (Chartron 1988, §IV.2) reprend le même genre de méthode, mais s'en tient à une réduction flexionnelle (lemmatisation), avec des « listes réduites de suffixes visant uniquement le regroupement des formes verbales et le regroupement des formes différentes en genre et en nombre. » Elle ne veut pas amalgamer adjectifs, substantifs et formes verbales (qui « ne contienent pas le même degré d'information »), notamment dans le cas « où le substantif est un mot retenu pour le lexique » (*ibid*., p. 48).

[67] La réduction dérivationnelle n'est plus une étape nécessaire du traitement effectué par ALCESTE. L'analyse se décline maintenant en quatre possibilités (Reinert, Piat 1995) : *(i)* aucune réduction, *(ii)* étiquetage par des catégories, *(iii)* réduction « sûre », c'est-à-dire uniquement des mots décrits par les dictionnaires de racines et de suffixes, *(iv)* réduction dérivationnelle complète (dictionnaire et heuristique d'exploration du corpus). C'est un symptôme que la réduction dérivationnelle, aussi bonne soit-elle, n'est pas toujours souhaitable dans les traitements.

Tout mot à valeur prédicative se laisse alors décomposer en une racine et un ou plusieurs modificateurs principaux. Le *spectre sémantique* d'un texte est l'ensemble des racines (avec leurs modificateurs principaux) qui y figurent.

Ces mots de « nature prédicative », qui font l'objet d'une décomposition, (p. 99) sont des verbes, mais aussi des noms (construction), des adjectifs (intraitable), des adverbes (additivement). On établit par ailleurs d'autres listes, complémentaires pour mettre en œuvre l'analyse des textes :

- les *mots constants* : prépositions, conjonctions, déterminants,..., et l'auxiliaire *être* ;
- les *mots exceptionnels* : noms propres, de personnes ou d'entités géographiques, et leurs dérivés tels que *booléen*, *euclidien*, *wronskien* ; mots spécifiques au domaine : notations abrégées (*p.g.c.d.*, *resp*., *lim sup*) ou termes savants (*phlogistique*) ;
- les *mots spéciaux* : leur motivation semble être de décrire les titres de la forme : <mot spécial> *de* <syntagme> (p. 111). Hérault les présente comme énumérables[68]. Selon leur caractère prédicatif ou non, ils sont du type *classificateurs* (tels que *méthodes*, *type*, *notion*, *concept*, *théorie*, *élément*, *forme*, *système*) ou du type *foncteurs* (prédicatifs : *construction*, *approximation*, etc.).

Enfin, –troisième famille d'éléments repérés–, des syntagmes nominaux complexes représentent les objets du texte : leur identification est prévue en s'appuyant sur les marques typographiques (toutes les formes de titres, et de mises en relief dans le texte), et une validation globale, à l'échelle du texte[69].

Parmi les choix généraux de Hérault, on peut retenir ceux-ci, comme les plus décisifs et les plus originaux :

- *au plan de la morphologie et de la description du lexique* : *(i)* une distinction fondamentale entre des racines nominales et des racines prédicatives : seuls les mots à racine prédicative font l'objet d'une décomposition, l'analyse des dérivations est donc partielle ; *(ii)* la part belle faite à certains préfixes et le peu de cas fait des suffixes : les modificateurs principaux sont des « *préverbes* (modification sémantique profonde) [que l'on s'efforce de distinguer] de simples *préfixes* (constance du sens du préfixe ; fréquente existence du mot non-préfixé) » (p. 47) ; on ne se préoccupe pas « de la partie suffixale, ni même de la partie préfixale précédant le groupe préverbal : cette attitude a été totalement délibérée, et se justifie (*a posteriori*...) par le fait que ces deux segments [semblent] apporter une information beaucoup moins essentielle [...]. Bien entendu, la partie suffixale étant, de ce point de vue, très hétérogène, certains de ses éléments, dans certains cas, seront pris en considération, et même étudiés avec minutie » (p. 21).
- *au plan de la portée de l'étude des mécanismes de dérivation* : *(i)* l'optique d'emblée multilingue : à travers l'inventaire des racines et des modes de construction associés, on cherche à rejoindre la souche indo-européenne et à produire une description unifiée de plusieurs langues ; l'isomorphisme trouvé entre les modules prédicatifs associés à différentes langues est souligné (p. 101) ; on se place en mesure de considérer les divers systèmes dérivationnels « imbriqués » dans une même langue (par exemple les systèmes slave, germanique et roman) (p. 101) ; *(ii)* la restriction de l'étude au champ des textes scientifiques, tant il est « difficile de tirer des conclusions assez générales à

[68] Les classificateurs « appartiennent à une liste pratiquement close, qui, à peu de choses près, se retrouve identiquement dans toutes les langues que nous avons manipulées (langues indo-européennes, pour l'essentiel) » (p. 97) ; « les listes utiles de classificateurs et de foncteurs ne sont pas considérables et ne semblent pas, globalement, dépasser une centaine d'unités. » (p. 111)
Le doute reste permis quant à ces affirmations, d'autant que l'ouvrage de Hérault ne montre pas une telle liste. Une proposition de liste aurait constitué, pour Hérault, un argument en faveur de la possibilité de définir l'ensemble postulé, et, pour une étude critique, aurait pu servir de base à la discussion sur la validité (théorique) et la pertinence (pratique) des notions de *classificateurs* et *foncteurs*.

[69] « On voit tout de suite que, quelle que soit l'habileté stratégique des procédés utilisés, les résultats obtenus ont une probabilité non nulle d'être inexacts. Il est donc nécessaire de vérifier, à chaque instant, la *cohérence sémantique* des « objets » reconnus. C'est cette vérification qui forme l'essentiel de l'*analyse hypersémantique*, le préfixe *hyper-* étant là pour rappeler que l'analyse porte sur le texte entier, ou tout au moins sur de ses larges fractions et que, en tout état de cause, jamais la phrase n'est prise comme unité de raisonnement. » (p. 96)

partir d'un discours littéraire, qui semble soumis à des fluctuations incompatibles avec toute stabilité raisonnable » (p. 72)[70].

- *au plan de l'utilisation de l'analyse des dérivations dans une application informatique* : le spectre sémantique est intéressant non pour sa capacité à caractériser les textes, mais au contraire pour sa très grande régularité de manifestation[71], qui autorise une description statistique et prospective (mise en évidence d'une loi décrivant le comportement d'apparition des racines au fil d'un texte, en l'occurrence un processus probabiliste de Poisson)[72] ; cette propriété, « très importante d'un point de vue purement théorique, [apparaît] d'un intérêt presque nul du point de vue pratique : la compréhension [...] ne peut [...] être fondée sur l'extraction de tel ou tel élément par des méthodes probabilistes ; tout ce qui est utile, mais rien de plus, doit être obtenu » (p. 84). Hérault propose toutefois d'utiliser le spectre sémantique pour déterminer l'architecture du discours (p. 109) : indications sur les *verbes modalisés* (p. 95), repérage de formules (pour introduire, définir, nommer, marquer une déduction, une remarque, une conclusion, un renvoi, une notation, etc., chacune de ces formules étant « fondée sur un nombre très restreint d'éléments prédicatifs ») (p. 107).

Le (gros) point faible des propositions de Hérault tient à l'écart entre le système (trop ?) ambitieux qui doit leur donner leur valeur en les mettant en œuvre, et les doutes que l'on peut avoir quant à la faisabilité d'un tel système. La plupart des modules ne sont qu'esquissés –Hérault admet qu'il reste un travail *considérable* à faire–, et l'on entrevoit des difficultés de tous ordres dès qu'il s'agirait de les préciser et de les réaliser : existence confirmée d'un noyau prédicatif à l'échelle d'un corpus plus étendu, établissement des diverses listes fermées, exploitation directe des marques typographiques, etc. Cependant, la discussion de cette proposition originale, en connaissance des quelques résultats avancés par Hérault, a le mérite de renouveler le regard sur des acquis par trop évidents.

Rôle des affixes : variations à réduire, ou unités pertinentes ?

La *réduction* dérivationnelle a coutume d'éliminer les suffixes pour regrouper les unités lexicales ayant une même racine. Une analyse dérivationnelle peut aussi ajouter certains préfixes et suffixes comme unités significatives pour représenter le texte. En effet, les préfixes et suffixes n'ont pas qu'un simple rôle de « convertisseur de catégorie morpho-syntaxique » (par exemple passage de l'adjectif à l'adverbe avec le suffixe *-ment*, ou du verbe à l'adjectif avec le suffixe *-able*). Ils se chargent parfois d'une signification autonome et pourraient être retenus comme des unités participant à la description du texte.

> [...] certains suffixes sont très importants pour la sémantique et peuvent même être assimilés à des concepts. Dans ce cas il convient de conserver l'information apportée par le suffixe. A titre d'exemple, on peut citer en chimie des suffixes comme *ane* ou *ène* qui précisent que les produits sont respectivement des alcanes ou des alcènes (ex. : *propane*, *propène*). De même, en médecine on utilise le suffixe *ite* pour désigner une inflammation (*appendicite*) ou encore *ectomie* pour désigner une ablation (*appendicectomie*). (Fluhr 1977, §III.4.2.2)

> Pour réduire la taille des résumés, compte-tenu des contraintes qui leur sont imposées, les rédacteurs ont été amenés à employer souvent des néologismes obtenus à partir de mots courants par des opérations morphologiques particulières. Les textes sont ainsi truffés de termes techniques spécifiques et de néologismes. Exemples : *vannes fuitardes* (suffixation), *désolidarisation* (préfixation), *inétanchéité* (préfixation). [...]

[70] Le discours non-littéraire a pour lui « la considérable importance prise par la science et les techniques, [...] la masse gigantesque des textes qui leur sont consacrés », et « la notion d'ambiguïté [est] pratiquement inconnue dans ce discours, dès lors que le lecteur [a] une connaissance suffisante du domaine abordé » (p. 73).

[71] Le spectre sémantique « possède de remarquables propriétés de *stabilité*, d'*invariance* vis-à-vis des textes proprement dits, de leur déroulement, de leur composition, de leur auteur, du domaine auquel ils se rattachent et même, à un degré moindre, de leur langue. » (p. 21)

[72] Hérault dénonce l'application des statistiques aux mots issus d'un corpus de textes : « un texte n'est pas, ne peut être assimilé à un échantillon, *puisque les mots qui le forment ne sont pas indépendants les uns des autres*. » (p. 72)

> On peut exprimer le sens de ces mots avec des phrases entières portant sur la racine du mot et sur la signification de son préfixe et / ou de son suffixe. [...]
>
> [Cependant,] classer sémantiquement des noms simples et composés en s'appuyant sur leur morphologie est inefficace ; les préfixes fournissent de bons renseignements sur la sémantique des mots, ce qui n'est pas le cas pour les suffixes.
>
> La présence des suffixes *ation*, *ement*, *age*, *ateur*, etc. ajoutés à une racine verbale (*fixation*, *chargement*, *regénérateur*, *réglage*) ne nous permet pas de conclure sur la sémantique de ces noms, il est impossible de faire la différence entre une notion et un objet ; le nom *fixation* peut signifier deux concepts différents : l'objet *fixation* ou l'action de *fixation*.
>
> Les préfixes ajoutent en revanche une sémantique à celle de la racine à laquelle ils sont attachés : *inétanchéité = non-étanchéité*, *déréglage = il n'y a plus de réglage*, *incompétence = non-compétence*.
>
> (Lefèvre, Chellali 1993, §3.1.2 et §3.5.2).

En fait de signification, ce n'est quelquefois pas tant le contenu « objectif » (dénotation) qu'il est intéressant de retenir, mais l'effet interprétatif (connotation) : suffixes savants, péjoratifs, etc. (Maingueneau 1991, pp. 36-37).

Les descriptifs d'activités utilisés pour construire les profils dans DECID comportent notamment de nombreux mots formés par l'intégration d'un adjectif sous forme de préfixe terminé en *-o*, par exemple : *élastoplasticité*, *thermoélasticité*, *thermomécanique*, etc. Comme cette forme de composition est encore très perceptible, le rédacteur écrit quelquefois le mot tantôt avec tiret, tantôt sans : *thermohydraulique*, *thermo-hydraulique*. Ce n'est plus le cas pour *thermodynamique* ou *aérodynamique*, plus figés et complètement « entrés en langue ». A l'inverse, certaines sont très peu intégrées : *thermo-métallurgie*.

Un cas important : la nominalisation

La morphologie française permet dans un très grand nombre de cas de trouver pour un verbe, un adjectif ou un adverbe, une réécriture basée sur un substantif. D'où l'usage de la nominalisation comme normalisation, en vue d'homologuer différentes expressions d'un même concept, dans une recherche documentaire.

> Nous admettrons que les verbes les plus informationnels peuvent se décomposer en un verbe « outil » qui supporte le temps et la modalité d'action et qui a pour objet un substantif (ou groupe substantif) qui représente le concept.
>
> ex. : *travailler → faire un travail*
>
> [...]
>
> Une partie des adjectifs peuvent être transformés de manière simple en complément de nom :
>
> ex. : *syndrome hémorragique → syndrome de l'hémorragie*
>
> Pour certains adjectifs, cette transformation ne se fait pas facilement ou change le sens. Dans ce cas on pourrait, du point de vue du concept, considérer comme un tout le groupe substantif-adjectif ou bien, si ce dernier a un pouvoir informationnel faible, l'éliminer.
>
> Ex. : centrale nucléaire
>
> [...]
>
> Les adverbes : [...]
>
> ex. : *manuellement → avec les mains*
>
> (Fluhr 1977, pp. 150-151)

On peut souligner l'équivalence qu'instaure ce passage d'une catégorie à la catégorie nominale : on obtient alors un mode de réduction des données, en rapportant toutes les formes à la forme nominale (ou à la racine).

Mais inversement, constater ce passage invite aussi à prêter attention à la spécificité de l'expression nominale. Il est instructif d'observer les nuances sémantiques apportées par le changement de catégorie. La transposition d'un adjectif en substantif ajoute une certaine emphase, et souligne l'autonomie d'une propriété qui est à présent saisie comme un concept en soi, une entité définie (Pottier 1974, §309). La transposition d'un verbe en substantif autorise une formulation générique, car la syntaxe dispense alors de l'expression des arguments du verbe (notamment le complément d'objet direct : *découvrir qqch. → faire une découverte*, *acheter des livres → faire des achats*, *étudier la chimie → faire des études (de chimie)*) (Pottier 1974, §311). Le style en devient plus abstrait et plus dense (un nom peut être une évocation condensée d'une phrase entière).

Cette marque d'abstraction, de conceptualisation, expliquerait l'abondance des substantifs et des formes nominales dans les textes scientifiques et théoriques. Les expérimentations qui contrastent les genres textuels sont cependant partagées quant à la valeur vraiment caractérisante des nominalisations. Pour (Bronkart 1985, §IV.A), « les nominalisations [...] [se révèlent] trop peu discriminatives (occurrences trop rares, même dans les textes théoriques) ». (Biber 1988) teste lui aussi la spécificité de trois types de formes nominales : les nominalisations (*i.e.* « all words ending in *-tion*, *-ment*, *-ness*, *-ity* plus plural forms »), les gérondifs (*i.e.* « all participle forms serving nominal functions – these are edited by hand »), et les autres noms répertoriés dans le dictionnaire. L'analyse ne fait ressortir aucune de ces formes nominales comme caractéristique d'une dimension sous-jacente aux registres textuels, sauf la nominalisation, qui est néanmoins la caractéristique positive *la plus faible* de la *troisième* dimension (sur sept retenues comme significatives ; les dimensions sont rangées par influence décroissante). L'interprétation donnée de cette troisième dimension est l'opposition entre d'un côté la construction d'un univers de référence interne au texte (où les référents sont définis dans le détail à l'aide de relatives), et de l'autre le discours ancré dans une situation extérieure à laquelle il renvoie par des déictiques généraux et des indications de temps et de lieu (adverbes). C'est une autre dimension qui est étiquetée comme l'opposition entre information abstraite ou non abstraite : il s'agit de la cinquième dimension, caractérisée par la forte présence de la voix passive et des connecteurs (conjonctions, locutions conjonctives).

Le rôle pressenti de la nominalisation en ressort accrédité : parmi les groupes de textes étudiés par Biber, les trois qui se rapprochent le plus des écrits qui circulent dans le centre de recherche d'une entreprise (les lettres professionnelles, la prose académique et les documents officiels) se trouvent bien caractérisés par ces deux dimensions. (Assadi 1998, §I.5.2, p. 67) fait un constat analogue sur une documentation technique, un guide de référence. Chellali (Lefèvre, Chellali 1993) étudie un corpus professionnel encore différent, relevant plutôt des fiches techniques. Il s'agit de résumés d'incidents en centrale, qui se présentent comme de brefs commentaires (limités à 450 caractères) rédigés en style télégraphique. La nominalisation est là encore soulignée comme une des principales particularités du lexique.

Dans la langue des écrits de l'entreprise, la nominalisation pourrait être promue par divers facteurs, qui ne s'apparentent pas uniquement à des motifs scientifiques et techniques. Le vocabulaire abstrait et général, masquant les oppositions qui apparaîtraient dans le détail, est consensuel. Il confère un certain prestige, car donne l'impression de dominer le sujet traité. De fait, connaître un terme générique, serait avoir fait le tour de tous les termes spécifiques qu'il englobe. Le style administratif, requis pour un certain nombre de documents et circulaires, exige l'abstraction et certaines longueurs inévitables qui s'expliquent par la nécessité de s'entourer de précautions. Enfin, dans notre civilisation, le langage abstrait est certainement valorisé. Dans les pays du Maghreb, par exemple, le langage valorisé est celui qui contient de nombreuses citations du Coran, et dans les communautés africaines, c'est la multiplication des proverbes qui assure la qualité de langage. (de Almeida, Bellamy, Kassai, pp. 78, 90, 97)

Racine prédicative, racine nominale

Plutôt que de rapporter les mots d'une même famille à une forme conventionnelle (un substantif), on aurait intérêt à s'en tenir à la racine « naturelle », brute. Elle serait à rebours : la préfixation ou la suffixation est perçue comme une adjonction ; l'effacement des préfixes et suffixes suit le chemin inverse pour trouver « l'origine » dont les formes sont « dérivées ». Si la racine se trouve avoir la forme d'un nom, il ne faut pas en conclure à sa valeur substantivale, elle peut très bien être prédicative. Par exemple, *calcul*, *calculateur*, *calculatrice*, *calculer*, *calculette* se ramènent à *calcul-*, mais un *calcul* se définit comme l'*action de calculer*.

g) Segments répétés

L'algorithme a été mis au point au laboratoire de Lexicologie de l'ENS de Saint-Cloud (Lafon, Salem 1983). Il consiste à balayer le corpus à la recherche de séquences de 2, 3,... *n* mots ou plus qui sont employées à plusieurs reprises.

Un des points délicats est de bien faire la part entre les séquences significatives, et celles qui ne se définissent que comme des parties de celles-ci. Inclusions, chevauchements, multiplient et brouillent les inventaires de segments répétés, si l'on ne se munit pas d'indicateurs pour trancher entre les séquences sources et leurs « échos » (dans les sous-segments).

Il n'y a pas d'*a priori* restrictif sur la nature des composants des séquences. Les expérimentations montrent que l'on trouve essentiellement des composés nominaux, mais aussi des locutions (Lafon, Salem 1983). La recherche de segments répétés est un outil efficace pour mettre en valeur les phraséologies qui ponctuent un corpus. Ce que l'on retire donc n'est pas spécialement de nature thématique (unités qui concernent le sujet abordé), mais comporte un certain nombre d'éléments de toile de fond du genre, les formulations qui l'articulent, qui le rythment, –pour peu que le corpus soit relativement homogène du point de vue du genre. Les résultats sont probants pour l'analyse de textes politiques (tracts, discours) (Lafon, Salem 1983), mais sont tout aussi intéressants dans d'autres domaines comme l'analyse de réponses libres dans des enquêtes d'opinion (Lebart, Salem 1988).

Il n'y a pas non plus d'*a priori* sur la longueur des séquences, et donc sur le nombre de termes qui sont ainsi associés : la méthode se dégage du primat artificiel souvent conféré aux paires. Pour le calcul, on doit néanmoins préciser une longueur maximale au-delà de laquelle on arrête les recherches. Une longueur à l'échelle de la phrase fournit une borne naturelle : la répétition éventuelle, mot pour mot, de passages entiers, tient davantage du copier / coller, que de la reprise d'une unité sémantique. L'intérêt de les repérer relèverait du domaine de la génétique des textes, ou d'un contrôle de non réutilisation de textes précédents ou voisins. Cela doit être tenu à l'écart du champ d'investigation de DECID, pour que l'application ne soit pas perçue négativement comme un instrument de contrôle rédactionnel, mais positivement, comme mise en relation de textes et de personnes, indépendamment de contraintes organisationnelles.

En « soudant » des séquences qui étaient sinon analysées comme plusieurs unités, les segments répétés concourent à l'homogénéité et à la régularité de la description. On notera qu'ils procèdent dans ce sens à certains rapprochements analogues à ceux opérés par un catégoriseur morpho-syntaxique, mais par le chemin inverse : ainsi, alors que le catégoriseur décompose *du* en *de* +*le*, *au* en *à* + *le* (dans l'optique de faire ressortir le parallèle avec les occurrences *de la*, *à la*), les segments répétés groupent *à la*, *de la*, reflétant la forte association réalisée linguistiquement par les formes amalgamées des équivalents masculins.

Les segments répétés sont implémentés dans le logiciel SPAD-T d'analyse statistique de données textuelles. Une proposition d'algorithme est donnée dans (Bommier 1993).

h) Extraction de groupes nominaux et acquisition de terminologie

L'utilisation de groupes nominaux, en plus des formes simples (nom seul, adjectif seul, etc.), demande de trouver un point d'équilibre. L'avantage de disposer d'unités pour représenter les formes composées est d'abord d'avoir des unités plus *justes*. Pour les noms composés, les syntagmes en voie de figement, le sens du composé ne se déduit guère du sens des composants, ils sont donc mal représentés par les unités simples que sont leurs composants. Les groupes nominaux apportent aussi des unités plus *précises*. Ainsi, la terminologie spécifique, propre à la description d'un domaine, comporte une proportion importante de formes composées. On obtient également un gain sémantique, et une réduction des effets de polysémie, pour des termes simples très généraux, comme dans le cas de *base de données* par rapport à *base* et à *données*. En revanche, l'inconvénient d'ajouter à la description les groupes nominaux est de *multiplier* le nombre des unités, ce qui alourdit le système et peut *disperser* les représentations obtenues (le niveau de détail est trop grand pour trouver des points communs).

Il faut savoir aussi que toute une série de constructions linguistiques compliquent le processus de repérage des groupes nominaux :

- l'ellipse (reprise partielle, généralement uniquement de la tête du syntagme - ex. : « WWW, le World Wide Web... Le Web... ») peut fausser un décompte de fréquence, les occurrences de la forme complète n'étant pas rapprochées de celles de la forme allégée.

- la coordination de deux termes, ayant une tête commune, peut conduire à une mise en facteur de la tête : cela gêne ou / et complique le repérage automatique des deux termes. (Bourigault, Gros 1994a) montre que la présence d'un coordonnant dans un groupe nominal maximal ouvre trois alternatives d'analyse : soit conserver le groupe (*alarme sonore et lumineuse*), soit couper le groupe (les candidats termes sont de part et d'autre du coordonnant), soit distribuer (*circuit de soufflage et de reprise* donne *circuit de soufflage* et *circuit de reprise*). Dans la pratique il reste des cas ambigus, même après examen de la structure morpho-syntaxique et apprentissage endogène (repérage d'autres occurrences, non problématiques, dans le corpus).
- dès que le groupe comporte plus de deux composants de type nom ou adjectif, la détermination des rattachements, quand elle n'est pas résolue à l'aide d'un dictionnaire ou d'un apprentissage endogène, peut rester ouverte[73]. La combinatoire des découpages possibles encombre alors rapidement la description. L'autre extrême est de laisser le groupe inanalysé et de perdre ainsi les rapprochements possibles avec ses composants intermédiaires réels.
  Par exemple, si l'on repère le groupe nominal *ouverture de la vanne d'isolement du puisard d'enceinte*, on peut aussi vouloir y lire *ouverture de la vanne d'isolement*, *vanne d'isolement, puisard d'enceinte*, *ouverture de la vanne d'isolement d'enceinte*, *vanne du puisard, vanne*, *ouverture*, *puisard d'enceinte*, *ouverture de vanne*, etc. Si l'on ne fait pas entrer dans le décompte les prépositions et articles (qui font partie de l'ensemble des unités indépendamment de cette occurrence), le nombre d'unités passe ici au moins du simple au double, par rapport aux unités simples seules.

Un extracteur (grammatical) de groupes nominaux ne convient pas à notre usage, car il noierait la description avec une foule de groupements inefficaces, parce que apparus dans un ou quelques contextes, mais non stables, et sans valeur sémantique particulière. Les groupes nominaux qui nous intéressent correspondent plutôt aux termes, pour leur stabilité d'emploi et leur charge sémantique propre. Cependant, les extracteurs terminologiques les plus performants fournissent de longues listes de candidats-termes, dans lesquelles la proportion de termes effectifs est très faible. Le travail de dépouillement n'est pas automatisable, départager les termes mobilise un expert du domaine. Tout au plus peut-on concevoir des interfaces facilitant le travail : tableaux avec différents critères de tris, représentation graphique, présentation en premier des meilleurs candidats[74]. Ces interfaces interactives sont adaptées pour un travail de terminologie, mais ne conviennent pas à DECID qui requiert un traitement tout automatique.

Ce ne sont donc pas tant les *unités* fournies par un extracteur terminologique qui peuvent nous être utiles, que *l'information* qu'il y a une *relation* forte entre telle et telle unité, prédisposant celles-ci à former une unité syntagmatique plus large. L'intérêt d'un extracteur terminologique pour DECID est donc son travail sélectif sur les relations syntagmatiques pertinentes pour former des unités composées. Au lieu de se baser sur un critère de proximité linéaire (mot qui précède, mot qui suit), l'extracteur terminologique oriente tout de suite la recherche sur des combinaisons linguistiquement valides. De plus, LEXTER par exemple précise entièrement la structure en relations binaires de dépendance *Tête / Expansion* : cela permet d'éviter le gonflement artificiel de la combinatoire de décomposition des unités longues, les rattachements étant résolus.

Assadi utilise ainsi LEXTER en amont d'une procédure classification automatique : le fait de ne pas se baser sur une simple cooccurrence de voisinage, mais d'avoir dès le départ des associations linguistiquement validées et identifiées, est un des atouts majeurs de son traitement. (Assadi 1996)

---

[73] Lorsque LEXTER ne sait pas trancher (cela ne concerne que quelques pourcents des candidats termes extraits), il opère le rattachement à gauche, qui est largement dominant en français. Cette heuristique n'est effectivement satisfaisante que pour les cas difficiles résiduels, car cette structure n'est pas toujours vérifiée.
ex. : [[*mise en* [*arrêt à froid*]] *de la pompe primaire*]
Si l'on appliquait uniquement l'heuristique de rattachement à gauche, on aurait obtenu :
ex. : *[[[*mise en arrêt*] *à froid*] *de la pompe primaire*]

[74] Richard QUATRAIN (EDF/DER, Département SID) utilise un réseau de neurones. Ce réseau apprend à reconnaître les formes de termes les plus souvent validés. Pour un ensemble de candidats-termes donné, il les présente pour validation dans l'ordre du plus probablement validé au plus probablement éliminé, tout en continuant à adapter ses critères au fur et à mesure des jugements de validation.

La dominance *syntaxique* de la Tête sur l'Expansion ne préjuge pas de son importance *sémantique* supérieure : utiliser l'information de structuration des termes pour filtrer certains composants n'est satisfaisant ni théoriquement, ni heuristiquement (Pohlmann, Kraaij 1997).

A défaut d'utiliser directement un extracteur terminologique, DECID pourrait consulter une liste de termes validés, dont il tirerait des indications de relation pour les occurrences rencontrées dans le corpus. Cette solution est moins satisfaisante, car alors la liste de référence est statique, et rien n'assure son adéquation au corpus.

Outils et comparatifs

LEXTER a été développé par Didier Bourigault dans le cadre d'un contrat avec EDF. Egalement présent à EDF, GCE (*GRAAL Corpus Exploration*) réalise une extraction de groupes nominaux, avec la même visée terminologique. Parmi les comparaisons qui ont été faites entre les deux outils, on peut noter (Ferrier 1995) :

- les deux outils donnent à peu près le même nombre de termes en sortie, mais en ont seulement 50% en commun ;
- leurs grammaires sont différentes (et même leur approche générale : Lexter procède par érosion, il repère d'abord ce qui n'est pas un terme et forme donc frontière ; GCE procède par dotation, il dispose d'un ensemble de modèles d'enchaînement de catégories morphosyntaxiques, des patrons, qui lui servent à reconnaître des expressions qui ont l'allure d'un terme) ;
- GCE fonctionne de façon récursive à partir de ses patrons de base, pas Lexter ;
- Lexter abandonne le groupe nominal quand il n'arrive pas à résoudre une ambiguïté de rattachement, GCE choisit (arbitrairement) un des rattachements.

Une autre comparaison méthodique a été réalisée entre LEXTER et le successeur de GCE, AlethIP, paramétré pour obtenir des groupes nominaux en sortie (Gros & *al*. 1997). Les deux systèmes diffèrent de la même manière que précédemment, quant à leur approche et à leur visée. Du fait de la procédure d'apprentissage endogène[75], la qualité des résultats de LEXTER dépend de l'homogénéité et la représentativité du corpus (qui doit donc avoir un volume suffisant). Quant à AlethIP/GN, il est moins sélectif en ce qu'il génère des syntagmes plus longs, avec des expansions circonstancielles : il ne fournit pas des candidats-termes, devant représenter des concepts (et donc présentant un caractère de dénomination stable et générale), mais purement et simplement des groupes nominaux.

Toujours en regard de LEXTER, le progiciel TERMINO, de dépouillement terminologique, a fait l'objet d'une évaluation dans le contexte EDF, pour l'analyse des résultats d'une enquête sociologiques (Amar, Le Roux 1992). Les résultats de l'analyse de LEXTER et de TERMINO étant des candidats termes, un effort particulier est consenti au niveau de l'interface pour le dépouillement et la structuration de la terminologie extraite. Ces procédures semi-automatiques, bien qu'elles soient un aspect majeur de ces applications d'assistance à la constitution de terminologies, ne concernent pas DECID, qui lui est conçu pour un traitement tout automatique.

Bien d'autres logiciels seraient également à mentionner dans ce domaine, et même plus spécifiquement pour des applications à de la documentation technique. Par exemple, le système ANA (*Apprentissage Naturel Automatique*) a été conçu dans le cadre du projet *Retour d'EXpérience* (REX) développé au *Commissariat à l'Energie Atomique*, centre de Cadarache, et appliqué à un corpus concernant Super-Phénix. Il opère un traitement très robuste, initié avec un minimum de données terminologiques et linguistiques (Enguehard 1993).

i) Désambiguïsation

La question de l'ambiguïté est artificielle : engendrée par la combinatoire des alternatives avec lesquelles jongle le système informatique, elle se cantonne, dans la vie quotidienne de tout un chacun, à des usages très particuliers (jeux de mots, dénonciation de contrats,...). L'incidence du genre (qui gouverne les pratiques de lecture) et, dans le texte, du contexte et des résonances entre les

[75] L'annexe 3 de (Gros & *al*. 1997) résume très clairement, en une page et demie, la définition et l'utilisation de l'apprentissage endogène dans LEXTER.

mots, sont de premières clés pour expliquer l'échec des systèmes sans vision globale, opérant au ras des mots.

Le problème de l'ambiguïté sémantique se pose pour les traitements automatiques dès lors qu'il s'agit d'attribuer des unités à valeur sémantique ou conceptuelle à un texte. En un point du texte, le système trouve par le calcul plusieurs solutions équivalentes. Les deux écueils (pas toujours évités) sont : tout garder et finalement tout retourner à l'utilisateur ; trancher arbitrairement et immédiatement. La première stratégie n'est pas viable en dehors des expérimentations sur quelques phrases tests, dépouillées par un chercheur motivé. Le volume des résultats croît exponentiellement avec le volume des données à traiter, et leur analyse requiert de rentrer dans la logique du système. L'autre approche résout radicalement le problème du volume, mais c'est au prix d'une perte d'information précipitée et irrémédiable, abîmant sensiblement la qualité des résultats ultérieurs. (Sanderson 1994) montre, dans un contexte de recherche documentaire, qu'une mauvaise désambiguïsation détériore bien plus la qualité des résultats qu'une non-désambiguïsation (le contexte des autres mots clé orientant vers l'interprétation souhaitable dans ce second cas).

C'est l'introduction d'une dimension globale, contextuelle, qui permet de se sortir de cette mauvaise passe. Plusieurs tactiques vont dans ce sens :

- cerner le domaine de l'application et ajuster les ressources lexicales et sémantiques en conséquence : dans un domaine donné, la polysémie est quasi inexistante.[76]
- sélectionner une des propositions, celle la plus vraisemblable en fonction de probabilités calculées sur un corpus représentatif, ou d'un apprentissage. La prise en compte d'éléments de contexte locaux affine les prédictions mais complexifie le calcul : on s'arrête à une solution de compromis. Si un bon équilibre est trouvé, les erreurs résiduelles ne coûtent pas trop : la redondance naturelle de la langue pallie les manques ponctuels et estompe les écarts isolés.
- garder l'ensemble des solutions et le retraiter globalement avant de le retourner à l'utilisateur. Pour que cela puisse être pris en compte par les traitements ultérieurs, il semble utile de conserver l'information comme quoi tel et tel terme sont en alternative. Toutefois, une telle représentation est plus complexe. En général, on préfère dans un premier temps accumuler indistinctement l'ensemble des propositions et dans un second temps exploiter globalement les convergences pour faire ressortir les propositions qui se renforcent et éluder les propositions isolées, non cohérentes avec le reste. La caractérisation délayée, confuse et dissonante obtenue après le traitement local n'est pas lisible pour l'utilisateur, mais est exploitable par le calcul.

On retient que la prise en compte de l'incidence du global sur le local peut donc intervenir *a priori*, pour guider l'élagage des alternatives en un point donné, et *a posteriori*, pour dégager les lignes de force des informations collectées au fil du texte. Et dans tous les cas il a fallu sortir d'un traitement purement linéaire et immédiat.

j) Analyse syntaxique

Des catégories morpho-syntaxiques à l'analyse syntaxique

Avec un catégoriseur morpho-syntaxique, le résultat cherché est local et au niveau des mots ; il est au niveau des groupes de mots et de leur structuration pour un analyseur syntaxique. Alors que les catégoriseurs déterminent les unités lexicales et leur nature morphosyntaxique, les analyseurs syntaxiques apportent de surcroît une information sur les regroupements et dépendances entre ces unités, et sur les fonctions recensées par la grammaire. Une même séquence formelle –un enchaînement donné de catégories morphosyntaxiques– peut en effet recouvrir des relations variables,

[76] Le contexte est désambiguïsant, qu'il s'agisse de celui du texte, du genre textuel (lié à une pratique, à un domaine, à des conventions stylistiques et thématiques), ou des quelques autres mots clés de la requête :
« in the totality of text, a word tends to have one predominant sense and one or more much rarer ones. »
« A restrictied textual domain, such as finance or medicine, is likely to feature a word in one sens only. »
« if several keywords are submitted simultneously, they will of course serve to disambiguate each other »
(Renouf 1993a, pp. 180-181)
En particulier, une métaphore, conventionnelle ou originale, n'a pas à être prise au pied de la lettre et considérée indépendamment de ce qu'elle qualifie.

et la syntaxe contribue à déterminer les structurations valides à l'échelle de la phrase[77]. En somme, l'analyse syntaxique vise à déterminer quatre types d'information interdépendants :

- un *découpage* et des *regroupements*, par la formation de constituants immédiats élémentaires, et le rassemblement des composants discontinus (ce peut être le cas pour des négations, des verbes à des temps composés, des syntagmes tranchés par une incise, des constructions marquant l'emphase, cf. (Pottier 1974, §321)) ;
- la *nature* des unités ainsi délimitées : pour les unités initiales, c'est l'*étiquetage* par une catégorie morphosyntaxique ;
- des *rattachements*, symétriques (coordinations) ou orientés (dépendances), qui se traduit généralement par une structure emboîtée des constituants immédiats ; la linéarité du texte (chaque mot prend place après un autre et avant un suivant) « applatit » l'arbre syntaxique, non sans le déformer –et non sans l'*in*former : une inversion par rapport à un ordre de succession habituel devient par exemple un procédé de mise en relief (Pottier 1974, §320).
- la *fonction* (éventuellement le cas) de chaque relation de structuration.

Si l'on s'en tient aux deux premiers types d'information et aux unités les plus petites, on retrouve l'information apportée par un catégoriseur. En ce sens, un analyseur syntaxique peut être utilisé comme catégoriseur (« qui peut le plus peut le moins »), pour peu que l'on puisse bien séparer ce niveau d'information des niveaux suivants. Un analyseur syntaxique est normalement plus performant qu'un catégoriseur simple pour la tâche d'étiquetage : en effet, l'analyseur syntaxique mobilise davantage de connaissances pour effectuer son analyse, et en particulier s'appuie sur la représentation de l'ensemble du contexte syntaxique qu'il construit. Il reste toujours des cas de polycatégorie, où ce sont des considérations sémantiques ou / et prosodiques qui permettent de choisir une interprétation plutôt qu'une autre (Pottier 1974, §319)[78].

Propriétés et utilisations des structures syntaxiques : équivalences et réductions

D'une idée à ses expressions

Inversement, une même idée peut être mise en discours et linéarisée de multiples manières (cf. les paraphrases). De fait, de l'idée à l'expression linguistique interviennent plusieurs points de choix (Pottier 1987, §IX.5), dont : *(i)* le choix des lexèmes (racines lexicales utilisées, qui fixent déjà certaines relations) ; *(ii)* le choix prédicatif (choix d'une base de vision, d'un point de départ) ; *(iii)* la hiérarchie phrastique (choix de la dominante sémantique, qui entraînera la dominante syntaxique).

Le choix prédicatif est le choix de la base du parcours diathétique, et se traduit par le choix d'une voix (ex. : *Untel a écrit ce programme* vs *Ce programme a été écrit par Untel*).

Du point de choix sur la hiérarchie phrastique dépendent la mise en valeur de certains éléments par *topicalisation* (ex. : *Sa pertinence, l'application la tient de l'étude approfondie des pratiques des usagers*.) ou par *focalisation* (construction en *c'est...qui...*, par ex. : *ce sont des résultats concrets que veulent les directeurs*). L'expression peut encore être rendue *impersonnelle* (construction autour de il y a, ça / cela fait que ; passif impersonnel : il a été décidé ceci par la Direction), et certains actants peuvent même être *effacés* (il est connu que le passif permet d'effacer l'agent ; l'infinitif ou la substantivation est un moyen de généraliser en rendant facultative l'expression de l'objet (accusatif) : *j'achète des livres* vs *acheter (des livres)* / *mes achats (de livres) me condui(sen)t à faire mes comptes régulièrement*). Si l'on considère plusieurs propositions, différents agencements sont possibles, par *juxtaposition*, *coordination* ou *subordination*.

A cela s'ajoutent encore le choix d'un temps (vision prospective ou rétrospective) et d'un aspect (accomplissement ou action en cours), l'expression de modalités (*vouloir*, *pouvoir*, *devoir*,...),

[77] ex. : *Les (usagers des Halles) autorisés* vs *Les usagers des (Halles centrales)*
Le chat de la gamine qui étudie l'anglais (qui = la gamine) vs *qui miaule sans arrêt* (qui = le chat)
Dans le premier exemple, le rattachement pouvait être déterminé par l'analyse syntaxique grâce aux marques d'accord (Pottier 1974, §314).

[78] (Pottier 1974, §319) donne l'exemple *Un travail reposant sur des techniques nouvelles* : *reposant* est-il un adjectif, qualifiant *travail*, ou une participe présent, auquel rattacher « sur des techniques nouvelles » ?

la détermination des entités concernées et son degré de précision (ajout de qualificatifs, emploi de déictiques (*ici*, *ce*), etc.) (Pottier 1992, p. 225).

Réduction des variantes paraphrastiques : réécriture des phrases sous forme de propositions élémentaires normalisées

Pour retrouver une unité de signification par delà la dispersion des réalisations, chaque phrase est rapportée à une forme normalisée. Par exemple, elle est décomposée et traduite par un ensemble de propositions élémentaires, correspondant chacune à la structure d'un prédicat. On isole ainsi chaque action, et chaque qualification.

Cette approche est commune à un certains nombre de travaux concernant des méthodes d'analyse des textes. C'est un aspect commun à l'analyse propositionnelle (cf. les travaux de Rodolphe Ghiglione) (Bardin 1977, §4.IV), aux grammaires transformationnelles de Harris reprises, en France, par l'analyse du discours (cf. Pêcheux) (Maingueneau 1991, §3), à la réduction des équivalences syntaxiques en sémantique structurale (cf. Greimas) (Greimas 1966, §IX.3.b).

C'est centrer (à grands frais) la description sur les relations fonctionnelles. L'opération de réécriture, complexe, n'est à notre connaissance jamais entièrement automatisée[79] (c'est une préparation des données pour un module automatisé qui s'applique ensuite). Pour ne pas générer un volume de données fastidieux, on se donne une ou plusieurs entités à partir desquels s'organise l'analyse : ne sont retenues que les propositions élémentaires faisant intervenir la ou les entités choisies. Mais comment déterminer validement la ou les entités fondatrices (Maingueneau 1991) ? La transformation en propositions élémentaires passe aussi par l'explicitation de l'antécédent des pronoms (anaphores et cataphores), voire de l'identification d'une même réalité derrière plusieurs désignations (coréférence), ce qui, on l'a vu, peut forcer l'interprétation. L'ancrage dans les textes ne peut alors être revendiqué comme une garantie d'objectivité : l'analyse est une lecture, sélective, des textes.

Sélection selon des constructions significatives

A l'inverse, certains choix de construction peuvent être considérés comme significatifs, et participer à la construction de la représentation du texte.

Le logiciel d'Analyse du Discours DEREDEC prend soin de repérer, parmi les noms, verbes et adjectifs, ceux qui sont thématisés[80] et ceux qui sont déterminés[81]. Les premiers (mots thématisés) sont ceux qui font l'objet d'une mise en valeur explicite. Quant au repérage des seconds (mots déterminés), cela permet de mettre en œuvre l'hypothèse interprétative suivante : ce qui est peu souvent déterminé fait figure d'évidence, relève du consensus, dans le corpus considéré, et donc pour la communauté (de lecteurs, d'auditeurs) qu'il représente. A l'inverse, ce qui est souvent déterminé est ce qui vaut d'être expliqué : jargon, emplois propres à un domaine spécifique (Maingueneau 1991, pp. 67-69).

Relations syntaxiques et relations sémantiques

La syntaxe donne à certains mots ou groupes de mots le statut de facultatif : ils peuvent être effacés sans altérer la grammaticalité de la phrase (déterminations, compléments).Il est bien évident que ceci n'est pas transposable au plan sémantique.

Les fonctions grammaticales ne reflètent pas nécessairement le rôle profond des actants (cf. la distinction sujet apparent / sujet réel), et peuvent correspondre à une multiplicité de cas (au sens linguistique du terme : accusatif, locatif, etc.) [82]. En outre, le système de cas n'est pas une donnée, il

[79] Le système ALARIC (d'analyse des résumés d'incidents en centrale nucléaire) semble pourtant mettre en œuvre quelques transformations syntaxiques (Lefèvre, Chellali 1993, §3.3.3).

[80] La thématisation marque l'emphase et l'identification, à travers des tournures du type : « C'est X que P », « ce que P c'est X », « X c'est ce que P », « X, P ». Dans toutes ces constructions, X est thématisé, (P désigne une proposition).

[81] Les mots déterminés sont ceux qui reçoivent un complément : adjectif qualificatif, complément de nom, relative pour un nom, adverbe pour un adjectif, etc.

[82] Nous allons prendre pour illustration les exemples de (Pottier 1974, §315-316).

est à construire (ou à choisir) en fonction de sa pertinence linguistique et de son adéquation à l'application.

Les relations syntaxiques, en structurant la phrase, créent des voisinages privilégiés. Du point de vue de l'utilisation de ce fait dans une application de recherche documentaire, (Rajman 1995, §III.5) fait une proposition intéressante, mais à laquelle il donne une forme un peu trop radicale : focaliser le contexte (sémantique) sur celui défini par des relations syntaxiques.

Concrètement, pour l'énoncé :

Le chat de la voisine miaule sur la palissade blanche

au lieu de retenir les cooccurrences de toutes les paires entre toutes les unités lexicales, dont :

(voisine, miauler), *(palissade, miauler)*, *(chat, blanc)*, etc.

on se limiterait par exemple aux paires reliées par un lien syntaxique direct, soit simplement :

(chat, voisine), *(chat, miauler)*, *(miauler, palissade)*, *(palissade, blanche)*.

Aussi fine que soit la spécification des liens syntaxiques à retenir (en faisant la distinction entre plusieurs types de liens, comme le propose Martin Rajman pour éviter l'association « pathologique » *(miauler, palissade)*), il nous semble dangereusement réducteur de limiter les interactions sémantiques à la syntaxe.

Déjà, la mise au point d'un tel filtre syntaxique promet de n'aboutir jamais qu'à un compromis imparfait, ne serait-ce que du fait des ambiguïtés syntaxiques et de la « polysémie » des relations syntaxiques.

Mais aussi, en l'occurrence, une paire comme *(miaule, palissade)* ou *(miaule, voisine)* peut s'interpréter autrement que dans un schéma *sujet-verbe* ; *miaule* apporte le trait /*chat*/, et les paires en cause traduisent une relation entre le (un) chat et la palissade, entre le chat et la voisine, ce contre quoi il n'y a rien à redire.

Plus généralement, le sens transgresse allègrement les canaux syntaxiques, et des termes se font écho par delà la construction grammaticale, et par delà l'horizon syntaxique qu'est la phrase. Une figure de style comme l'hypallage (qui échange des groupements et des rôles syntaxiques) joue ouvertement de cette liberté.

Sans dénier que des constructions syntaxiques favorisent l'interaction sémantique (ce qui valide l'idée de s'appuyer sur l'information syntaxique pour repérer des voisinages privilégiés), il faut conjointement admettre que, pour cette étape de délimitation de contextes comme dans les autres traitements, la syntaxe ne détermine pas la sémantique, et donc que la syntaxe ne doit pas être l'unique et ultime source d'informations pour un traitement à visée sémantique.

Construction automatique de paradigmes

L'identification de la nature de la relation (syntaxique) entre deux unités permet de reconstituer des paradigmes, rassemblant toutes les unités qui ont la même relation avec une unité

Considérons le syntagme *la vente du collège*. Du point de vue des fonctions, *du collège* est un complément du nom déterminant *vente* ; l'information sur les cas est plus précise, car elle doit préciser s'il s'agit d'un accusatif (on vend le collège) ou d'un locatif (la vente a lieu au collège).

Les constructions avec la préposition de fourmillent de ce genre d'ambiguïtés (même schéma d'enchaînement des catégories morpho-syntaxiques et différentes interprétations possibles) :

une jarre de terre cuite (matière) vs *une jarre de vin (contenu)* vs *une jarre de qualité (appréciation)*...

Cette construction n'est pas la seule du genre (ex. *la machine à laver* (instrument) vs *le chocolat à croquer* (accusatif)).

L'identification de la relation joue parfois sur la détermination du rattachement :

ex. : *Pierre a trouvé un livre sur la table* (locatif spatial, rattachement au verbe) vs *sur le Japon* (locatif notionel, rattachement à *livre*)

recevoir une photo de Rome : selon l'interprétation, *de Rome* détermine *photo* (la photo représente Rome) ou complète *recevoir* (la photo vient de Rome).

Les difficultés que posent ces cas de figure à un traitement automatique de la syntaxe se passent de commentaires. On voit également, une fois de plus, que dès que l'analyse linguistique côtoie l'interprétation elle perd de son caractère déterministe, et que l'on ne peut exiger, sans réduction arbitraire, d'avoir toujours une et une seule représentation syntaxique.

donnée (Grefenstette 1997). De fait, ce sont, pour le corpus considéré, les unités attestées vérifiant un test de commutation.

Si par exemple on considère les unités qui sont en relation avec un prédicat, et que l'analyse permette d'identifier le cas sous-jacent à chaque relation (par exemple *agent*, ou *objet*, ou *bénéficiaire*), chaque paradigme décrit la diversité des acteurs occupant un même rôle pour le prédicat choisi (point de vue synchronique), ou l'évolution de ces acteurs (point de vue diachronique) (Maingueneau 1991, p. 88 sq.).

Les paradigmes s'affinent si l'on retient les regroupements stables, c'est-à-dire qui se retrouvent dans la même relation à *plusieurs* unités. La description s'enrichit encore lorsque l'on utilise ces paradigmes comme unités, à partir desquelles construire de nouveaux paradigmes. La démarche en devient itérative : regroupement d'unités à partir d'une relation à des unités pôles, ajustement par stabilisation (noyaux des regroupements qui se retrouvent dans les contextes de différentes unités pôles), définition de nouvelles unités (paradigmes), utilisation de ces nouvelles unités comme unités pôles.

Les choix opérés par les formalismes syntaxiques

La structuration en arbre binaire est un choix descriptif

Les formalismes classiques (marqués par le rayonnement de Chomsky) décrivant la syntaxe pour les traitements automatiques décomposent la phrase sous forme d'un arbre binaire orienté. Autrement dit, tout élément est dominé par un et un seul autre élément. Or on trouve bien sûr des cas où il est difficile de hiérarchiser plusieurs rattachements, qui ne sont pas perçus comme intervenant à des niveaux différents, des cas où plusieurs possibilités de rattachement coexistent, etc. (Fuchs 1993, §4.1.3) fait un rapide inventaire des principales difficultés rencontrées par l'analyse hiérarchique.

Analyse partielle

On distingue l'analyse syntaxique partielle d'une analyse syntaxique complète. L'analyse partielle identifie les rattachements syntagmatiques locaux (par exemple des *constituants immédiats*) mais ne reconstruit pas nécessairement l'arbre syntaxique complet couvrant la proposition grammaticale, voire les relations entre propositions. Plus robuste (car poussant l'analyse moins loin), elle fournit des indications suffisantes pour l'identification de termes composés ou de formes figées. De plus, au delà de ce premier degré d'analyse, la structuration des relations entre les termes n'a rien d'évident[83], si bien que les spécialistes eux-mêmes ne s'accordent pas sur la préséance de tel ou tel rattachement, ou sur le découpage de tel constituant (Habert, Nazarenko, Salem 1997, §II.1).

Quelle analyse syntaxique envisageable pour DECID ?

Dans notre perspective, l'analyse syntaxique donne des informations de portée et de liens. Un élément est en relation privilégiée avec certains autres, et non pas uniformément avec tous ses « voisins » dans une « fenêtre » de longueur fixée. Cette description est plus riche et moins dispendieuse que les rapports combinatoires induits à partir des positions. Elle rend compte de frontières et de seuils, d'unités non connexes, de rapports de dominance.

L'analyse syntaxique permettrait également de considérer non pas chaque unité syntagmatique pour elle-même, mais en fonction de la valence syntaxique qu'elle présente, et qui est un facteur sémantique (Martin 1994, §II.A.1.b, p. 94). En particulier, la construction d'un verbe pourrait faire partie de la définition de l'unité associée : des différences importantes de construction (transitive vs intransitive, pronominale ou non) se traduiraient par des unités distinctes.

Compte-tenu des intérêts que présente une analyse partielle, l'horizon de la proposition voire de la phrase pourrait nous suffire, sans entrer dans le détail de sa structuration interne, pour rendre compte des relations syntaxiques potentielles moins « proches » que les premiers regroupements en syntagmes.

---

[83] Très vite, des considérations morpho-syntaxiques seules sont insuffisantes, cf. par exemple (Poirier, Mathet, Enjalbert 1998).

Dans un premier temps, l'identification de la nature des relations, et l'enregistrement du mode de combinaison syntaxique des unités, n'est pas d'une importance déterminante. En effet, le contexte créé par les unités dessine déjà les relations qu'elles peuvent avoir. D'un point de vue thématique, la structuration syntaxique est presque redondante, les contraintes lexicales et sémantiques, très fortes, sélectionnent déjà un petit nombre de rapports possibles entre les unités, leur combinatoire est restreinte. Ajoutons, dans le cadre de la diffusion ciblée, que les unités identiques à des variantes de liens près semblent de toutes façons intéressantes comme facteur de rapprochement entre deux textes.

> On ne tient pas compte pour la recherche des prédicats à la fois dans les textes et dans les questions. Les erreurs sont toutefois limitées du fait que les systèmes de concepts-sujets ne peuvent être agencés n'importe comment dans n'importe quels types de prédicats. (Fluhr 1977, p. 134)

C'est ainsi que Jean-Yves Antoine peut décrire le rapport entre syntaxe et sémantique comme une relation de *vicariance*, qui est une manière plus technique et précise d'expliciter leur redondance partielle et leur interaction (Antoine 1994, §1.III.3.1). Pour une application qui vise à rendre compte d'une lecture et d'une signification globale d'un document, la sémantique prime sur la syntaxe :

> Johnson-Laird a pu montrer que, sauf consigne spécifique, les individus ne mémorisent que le sens des énoncés et oublient leur structure syntaxique [Johnson-Laird 1977][84] (Antoine 1994, §4.III.1.2)

Les modélisations syntaxiques se montrent moins robustes[85] (Jean-Yves Antoine travaille sur la « parole spontanée »). Ceci n'empêche pas de les utiliser, mais en seconde instance, pour réduire la combinatoire des solutions sémantiques. C'est à cette approche que nous nous rallions.

> De par la mise en avant de la sémantique, [la stratégie coopérative] réoriente l'analyse linguistique vers sa vraie finalité, la compréhension, au détriment des critères de bonne formation des énoncés. Elle autorise le traitement d'une grande diversité de constructions orales, tout en limitant les risques d'explosion combinatoire grâce à l'intervention de la syntaxe qui n'est pas négligée. Ainsi, les deux analyses se voient accorder un rôle important, même si, en dernier recours, la stratégie coopérative donne la priorité à l'agent sémantique : entre deux hypothèses concurrentes, on préférera celle qui satisfait le critère de cohérence sémantique à celle qui est grammaticalement correcte. (Antoine 1994, §7.III)

Outils

Le logiciel Sylex (conçu par Patrick Constant et commercialisé par la société Ingénia) indique l'*unité syntaxique de rattachement* de chaque unité (catégorisée et lemmatisée), et permet notamment ainsi la description du groupes (syntagmes) discontinus. Il faut souligner aussi que, pour chaque forme analysée, le logiciel indique précisément à quel *segment du texte initial* elle correspond : l'information sur la graphie source et la localisation est conservée. C'est une information précieuse pour DECID, et qui, contrairement aux apparences, n'est pas évidente à donner, compte tenu de la délinéarisation et des reformulations opérées par l'analyse (nombre d'analyseurs ne savent pas la fournir).

k) Une procédure complexe particulière : l'identification de négations

Repérage

La détection de la présence d'une négation (sur le plan sémantique, qui nous intéresse ici) est rendue délicate par la variété indéfinie des manifestations qu'elle peut prendre :

- *ne...pas*, *ne...plus*, *ne...guère*,...
- *pas de*, *ne*, *ni*,...
- *rien*, *nul*, *personne*, *aucun*,...
- *hormis*, *sauf*, *sans*, *à l'exclusion de*, *à part*,...
- *interdiction de*,...

[84] JOHNSON-LAIRD P.N. (1977) - « Psycholinguistics without linguistics », in N.S. SUTHERLAND, éd., *Tutorial essays in psychology*, vol. 1, Lawrence Erlbaum, Hillsdale, NJ.

[85] Mais, fait classique, la fusion (sous forme de coopération) des deux approches peut se montrer supérieure, puisque « il n'y a pas de corrélation entre les variations de la robustesse des deux analyses [syntaxique et sémantique], qui reposent sur des critères différents. » (Antoine 1994, §8.IV.1.3)

- préfixes privatifs : *in-*, *a-*,...

De plus, on voudrait pouvoir trancher entre *(i)* ce qui est mentionné pour être rejeté hors du champ du texte, et *(ii)* ce qui est une formulation négative d'une notion qui pourrait tout aussi bien être exprimée par une forme positive.

Portée

Qu'est-ce qui est nié, rejeté ? La question de la détermination de ce à quoi s'applique la négation, l'influence et la difficulté de formaliser une telle détermination dans le cas des langues naturelles, sont bien connues des adeptes de la logique.

La place occupée dans une construction (par exemple, le nom qui suit un *hormis*) peut résoudre un certain nombre de cas. Et peut-être ne nous est-il pas nécessaire d'en savoir beaucoup plus : voyons ci-après en quoi le repérage des négations peut nous être utile.

Exploitation

Nous partageons l'avis de Christian Fluhr sur la nécessité de distinguer la négation dans les requêtes documentaires (de l'ordre de quelques mots ou d'une phrase) et celle dans les textes. (Note : Dans le cas étudié par Fluhr, les requêtes ne sont jamais textuelles, les textes sont uniquement du côté des documents recherchés.)

> Le problème de la négation se pose différemment au niveau du texte et au niveau de la question.
>
> Dans une première approximation, on peut ignorer la négation dans les documents. Le fait qu'une action soit niée [...] est un jugement porté sur l'action, et le plus souvent fera partie de ce que le questionneur veut savoir sur l'action.
>
> ex. : *peut-on anesthésier suivant cette méthode ?*
>
> réponses dans le texte : - « *on a dans ce cas pratiqué une anesthésie* »
>
> ou encore : - « *on n'a pas anesthésié* »
>
> Les cas où l'interrogation porte uniquement sur les cas niés sont assez rares et demandent une analyse plus fine de la question. La négation dans la langue n'est pas aussi stricte que dans son sens booléen, elle est assortie de nuances qu'il est souvent délicat de manipuler. [...]
>
> Au niveau de la question la négation est souvent d'une autre nature, elle désigne généralement des concepts qui ne doivent pas être abordés dans le document.
>
> exemple de question : « *Ethnies d'Amérique du Sud sauf du Brésil* »
>
> Il faut dans ce cas interdire les documents contenant le mot « Brésil » ainsi que ses termes spécifiques et certains termes associés.
>
> (Fluhr 1977, p. 164)

La problématique de la négation dans un contexte documentaire est bien exposée, mais nous ne partageons pas toutes les conclusions tirées.

Cas des requêtes dites en langage naturel

Il s'agit de l'expression d'un thème que l'on soumet à un système de recherches d'informations. Cela peut se présenter pour DECID, en général lorsque l'utilisateur ne soumet pas un document par copier/coller mais frappe au clavier le sujet qui l'intéresse.

Entrant dans une pratique bien particulière, ces requêtes seraient à considérer comme un genre. Parmi les régularités à décrire, on pourrait envisager un inventaire relativement complet des formules de négation qui expriment un aspect que l'on veut écarter de la recherche. Cet inventaire comporterait vraisemblablement des tournures rappelant les expressions courantes des opérateurs des équations de recherche (tels que `SAUF`, `NON`).

La prise en compte de ces négations doit se faire avec grandes précautions : les professionnels de la documentation, qui interrogent fréquemment les bases documentaires, mettent en garde contre l'abus des opérateurs négatifs, dont la contribution au silence documentaire, à savoir le rejet par le système de documents pertinents, est connue. En outre, l'exclusion au niveau de la requête fait écho à l'absence au niveau de termes d'index (descripteurs ou mots-clés), non pas au niveau des textes. En déduire que la mention exclue ne doit pas figurer une seule fois dans le texte intégral d'un document est une extrapolation pour le moins hasardeuse.

Cas du texte d'un document

La négation d'une notion n'est pas équivalente à son absence ni à sa non pertinence. En effet, elle peut entrer dans une discussion, qui en tant que telle est une partie significative du contenu du document. Le fait que le rédacteur qui éprouve le besoin de mentionner une notion, même si c'est pour la rejeter, révèle que la notion en question a quelque affinité avec le thème du document. De plus, la négation n'a que rarement la valeur d'une exclusion au sens d'un opérateur logique ; elle est souvent modalisée, et sert à l'expression de contrastes, de mouvements de rectification, de concession, de refus, de commentaires, d'évaluation, de comparaison, de mise en relief... (Salazar Orvig 1988).Ce décalage entre le marquage morpho-syntaxique de la négation et l'exclusion d'une signification est une observation relevant de la linguistique générale.

> Pour faire sentir au lecteur le contraire d'une impression donnée, il ne suffit pas d'accoler une négation aux mots qui la traduisent. Car on ne supprime pas ainsi l'impression qu'on veut éviter ; on évoque l'image en croyant la bannir. Voulant décrire un jardin appesanti sous le soleil d'été, à midi, un poète contemporain dit :
>
> Et d'entre les rameaux que ne meut nul essor
> d'ailes et que pas une brise ne balance,
> dardent de grands rayons comme des glaives d'or.
>
> Ces vers sont bien faits pour donner l'impression du battement des ailes d'un oiseau ou du balancement de la brise, et l'emploi de la négation n'écarte pas cette impression de l'esprit du lecteur. En un seul vers, de Heredia dit plus juste :
>
> Tout dort sous les grands bois accablés de soleil
>
> Le morphème grammatical ne se confond pas avec ce qu'on pourrait appeler le morphème d'expression.
>
> (Vendryes 1923, §2.III)

Un indicateur plus décisif que la négation serait plutôt la place accordée à la notion dans le document. Si elle fait l'objet de multiples expressions négatives, mais pour autant est constamment évoquée dans le texte, alors on lui accordera de l'importance dans la représentation du texte. Si en revanche elle ne fait l'objet que d'une mention passagère et discrète de rejet, il faut éviter qu'elle devienne à elle seule un facteur de rapprochement. L'évaluation de cette place pourrait recourir à des grandeurs comme la fréquence et l'étendue de la zone où la notion apparaît.

Le raisonnement que nous venons de faire pourrait être conforté par ce que nous observons des travaux de Salton et de ses successeurs. Ils choisissent des fonctions de pondération croissantes vis-à-vis de la fréquence d'apparition du terme dans le document. Ils ne se préoccupent guère d'analyser les négations des les textes qu'ils traitent. Et les résultats obtenus n'ont rien de catastrophique.

l) L'indexation documentaire automatique, comme extraction de descripteurs puis filtrage

A EDF, l'application d'indexation automatique a été développée dans la perspective d'une indexation automatique du fonds documentaire : tout document enregistré dans la base interne EDF est ainsi automatiquement pourvu d'une liste de mots-clés, descripteurs issus du thesaurus, grâce auxquels peuvent ensuite s'effectuer des recherches.

L'outil qui effectue l'analyse linguistique est élaboré, il combine plusieurs des traitements évoqués précédemment : catégorisation, lemmatisation, repérage de groupes nominaux, transformations dérivationnelles (substantivation), analyse syntaxique (Monteil, Penot 1990) (Herviou 1992a) (Herviou 1992b).

La liste de mots-clés d'un document comporte raisonnablement au plus une quinzaine d'éléments. Or l'analyse automatique du texte (titre, résumé) fournit un nombre de descripteurs de l'ordre du nombre de noms différents présents, d'autant plus que le texte est long. Des critères de filtrage ont été mis en place pour sélectionner les descripteurs qui auraient été les plus vraisemblablement choisis par une indexation manuelle.

C'est l'outil d'indexation automatique qui a d'abord été utilisé pour la caractérisations des textes dans DECID. Il s'est avéré que la perspective de la diffusion ciblée est différente de celle de

l'indexation automatique. En effet, une de ses forces est de se baser sur l'ensemble du texte. On ne veut pas éliminer d'emblée certaines facettes du texte : les thèmes secondaires, le contexte, certains détails, doivent pouvoir être moteurs dans les rapprochements. D'autre part, l'indexation gomme la disposition des concepts dans le texte. Or de la localisation peuvent être déduites des informations de proximité, de groupement, de zone d'influence (locale *vs* globale) précieuses pour la caractérisation.

D'autres raisons se sont ajoutées à celles-ci pour conduire à l'abandon de l'indexation automatique dans DECID : cela fait l'objet de la partie suivante, qui s'appuie également sur la note (Bommier, Lemesle 1996). C'est une manière concrète de discuter à nouveau de l'utilisation des techniques de traitement automatique du langage naturel, car l'outil d'indexation incorpore la plupart des traitements classiques existants (analyse morpho-syntaxique, nominalisation, etc.)[86].

2. Un cas concret : raisons de l'abandon momentané de l'indexation automatique

a) Rappel : principes de l'indexation automatique

L'outil d'indexation automatique utilisé à EDF[87] parcourt le texte à indexer phrase par phrase. Il réalise une analyse morpho-syntaxique robuste. Le but de cette analyse est de reconnaître des unités lexicales (noms et groupes nominaux) qui correspondent aux termes du Thesaurus EDF. Ces termes sont alors retenus comme descripteurs pour le texte soumis. Une fois l'ensemble des descripteurs trouvés pour le texte, un filtre ajuste leur nombre à un ordre de grandeur raisonnable fixé (par exemple, on souhaite avoir entre 5 et 15 descripteurs).

Comme, par construction, les descripteurs qui servent à indexer le texte font partie du thesaurus, l'indexation réalisée est une *indexation contrôlée*. Par opposition à une *indexation libre*, les termes d'index d'une indexation contrôlée sont tous (pré)définis et organisés par un référentiel terminologique (ici le Thesaurus EDF).

L'analyse morphologique permet de retrouver une même unité lexicale derrière ses variantes de forme (*singulier / pluriel*, mais aussi *masculin / féminin* en particulier pour les adjectifs). Mais l'analyse linguistique va ici plus loin, car ce ne sont pas seulement les syntagmes nominaux du texte qui peuvent être traduits sous forme de descripteurs. Des nominalisations possibles des verbes servent également à retrouver des termes d'index : par exemple, *Gérer l'énergie* est transformé en *gestion de l'énergie*, ce qui permet ensuite de reconnaître le descripteur correspondant.

Ce que l'on cherche à relever dans le texte (et à travers ses transformations linguistiques), ce ne sont pas les descripteurs eux-mêmes, mais des mots et expressions qui leurs sont associés. En effet, les descripteurs du thesaurus sont des unités *conceptuelles*, et la forme d'un descripteur est une étiquette choisie pour désigner commodément le concept. Le thesaurus est mis en correspondance avec un dictionnaire, qui lui décrit des unités *linguistiques* (avec leur morphologie, leur syntaxe, etc.). Tous les mots et expressions consignés dans le dictionnaire, et qui expriment le concept représenté par un descripteur, sont reliés à ce descripteur. Le référentiel conceptuel (thesaurus) se double donc d'un référentiel linguistique (dictionnaire). Le dictionnaire donc joue un rôle central, tant pour mener à bien l'analyse grammaticale que pour servir de relais à l'identification des concepts.

Une telle mécanique est très complexe à mettre en œuvre, et rencontre ses limites dans la pratique. De plus, l'approche elle-même n'est pas entièrement satisfaisante dans l'optique de la diffusion ciblée. Les critiques exprimées ci-après relèvent de ces deux plans, théorique et pratique.

[86] Les observations consignées ici ne correspondent sans doute plus tout à fait à l'état de l'art actuel atteint dans chacune des procédures de traitement. Ce rapport a cependant l'intérêt de rendre sensibles les difficultés réelles que l'on rencontre avec ces types de traitements.

[87] Il s'agit de l'outil Aleth de la société Erli, dont on trouvera une présentation générale dans (Monteil, Penot 1990) et une description technique détaillée dans (Herviou 1992a) (Herviou 1992b).

b) Doublons, sources d'irrégularités

Attribution redondante de descripteurs

Liens entre sigles et formes développées

Quant elles sont systématiques, et effectuées à la fois sur les cibles et les projectiles, les reformulations de sigles doublent artificiellement les descripteurs sans apporter d'information utile aux rapprochements.

On trouve par exemple systématiquement conjointement : EAU et H2O, AMMONIAC et NH3, CONTROLE NON DESTRUCTIF et CND.

Elles pourraient cependant intervenir pour la lecture et l'interprétation des résultats, mais alors il faut se donner les moyens de distinguer ces descripteurs ajoutés et « passifs » pour le calcul de la similarité.) De plus, que cette transformation soit à sens unique (du sigle vers son développé) introduit des distorsions entre les documents, suivant leur style de rédaction (usage libre ou limité des sigles). On relève aussi des cas où le sigle n'est pas conservé (cas de MMS, traduit par MANUFACTURING MESSAGE SPECIFICATION) : la procédure d'indexation n'apparaît donc pas régulière sur ce point.

Mauvaise coordination des référentiels terminologiques

Au moment où la diffusion ciblée expérimentait l'indexation automatique, deux référentiels terminologiques étaient utilisés conjointement. Le principal est le Thesaurus EDF : 15 à 20 000 descripteurs, structurés en 350 champs sémantiques et en une quinzaine de thèmes. Il est complété par la Terminologie DER, liste complémentaire de termes plus récents et plus spécifiques. Mais l'indexation fournie se voit artificiellement gonflée par des désignations différentes d'une même notion repérée dans le texte.

Quelques exemples :

- redoublement simple : REP (thesaurus) / REACTEUR REP (thesaurus) / REP (terminologie DER) / REACTEUR REP (terminologie DER)
- variations flexionnelles : HP / HAUT PRESSION / HAUTE PRESSION
- variantes orthographiques : METHODE DE MONTE-CARLO / METHODE DE MONTECARLO
- forme morpho-syntaxique des descripteurs : GESTION DE L'ENERGIE / GESTION D ENERGIE
- redoublement en fait divergent du point de vue des concepts sous-jacents : ENQUETE (rattaché au champ sémantique PROCEDURE JUDICIAIRE) / ENQUETE (STATISTIQUE)
- combinaison de différents facteurs de variation : ASSURANCE DE LA QUALITE / NORME DE QUALITE / AQ / ASSURANCE QUALITE.

Il est facile d'épingler quelques cas de la sorte ; le travail pour les résoudre est long, austère et délicat. Il a été entrepris depuis, et à présent, la terminologie DER est intégrée au Thesaurus et n'existe plus en tant que telle.

Décompte d'une forme

On demande à l'indexation d'indiquer le nombre d'occurrences de chaque notion dans le texte. Or il arrive que le nombre d'occurrences donné pour une forme dans un texte soit inexplicablement élevé. Les données sont alors bruitées pour la suite des calculs (en particulier pour le calcul des pondérations), puisque le chiffre donné ne correspond pas au sens qu'on lui prête.

Il va sans dire que ce n'est pas le logiciel qui fait une erreur d'addition, mais que ce qui est perçu comme une erreur tient à la non correspondance entre le résultat fourni et le résultat attendu : on ne connaît pas suffisamment le traitement appliqué si bien qu'on ne contrôle pas son sens, et le sens utilisé est erroné.

c) Réductions muettes et irréversibles

Une réduction morphologique (ou lemmatisation) est opérée systématiquement.

Observons la pratique des bibliothécaires et de leur système canonique de fichiers : les clefs du fichier *auteurs* fonctionnent comme des pointeurs ; les entrées par titre opèrent sur le principe d'un matching exact ; l'accès matière correspond à un matching approximatif, modulo lemmatisation, réduction dérivationnelle (c'est-à-dire mots de la même famille), synonymies, indexation conceptuelle par thesaurus. Ces trois types d'identificateurs sont utiles et complémentaires pour l'identification des documents.

L'indexation opère aussi des réductions dérivationnelles, qui permettent le rapprochement des mots d'une même famille. Bien sûr, il est inévitable de relever des cas où une telle réduction aurait été bienvenue, mais n'était pas prévue dans le dictionnaire (stéréoscopique non mis en rapport avec le descripteur STEREOSCOPIE par exemple). Le problème de non exhaustivité est inhérent à l'usage d'un dictionnaire.

d) Termes composés : des choix compromettants

Règles d'identification

Au palier syntagmatique, l'indexation peut opérer certains rattachements hâtifs malheureux : elle croit par exemple reconnaître ANALYSE DE CONTENU dans « analyse des orientations et du contenu technique du prototype ». LEXTER (Bourigault, Gros, 1994a) considère d'ailleurs comme une différence essentielle la construction en *de vs* celle en *du* (= *de* + *le*).

Dans le cas de termes à plus de deux composants, pour ne pas multiplier les hypothèses de rattachement, l'analyseur ne garde *arbitrairement* que les rattachements à gauche (*i.e.* pour un complexe ABC, c'est uniquement la sous-partie AB qui est examinée). Quand on dispose d'un texte (et a fortiori d'un corpus), laisser la possibilité d'un apprentissage endogène (Bourigault, Gros, 1994a) semblerait préférable.

Utilisation en tant que critère de fiabilité pour un filtrage

L'indexation automatique s'est appuyée sur la forme composée de certains descripteurs effectuer un filtrage. L'hypothèse faite est la suivante : un terme composé est en général, de par sa nature, plus précis et moins sujet à d'éventuelles ambiguïtés qu'un terme simple. C'est donc un indicateur de fiabilité : un filtre retient en priorité les polytermes comme de bons candidats mots-clefs.

Or il faudrait au moins distinguer les polytermes *à la source* et les polytermes *au but*. Nous dirons d'un terme qu'il est polyterme à la source s'il représente plusieurs mots (« mots pleins ») du texte initial ; il est polyterme au but si le descripteur fourni est un polyterme. L'identification des polytermes au but est la plus simple, mais ce sont les polytermes à la source qui correspondraient le mieux à un critère de fiabilité, puisqu'ils sont davantage ancrés dans le texte à caractériser.

Par exemple, les mots *distribution* (générant l'index CENTRE DE DISTRIBUTION EDF-GDF) ou *symbolique* (indexé MOUVEMENT ARTISTIQUE ET LITTERAIRE) n'ont manifestement rien de très fiable. Le cas est différent pour *théorie de l'information* (indexé THEORIE DE L INFORMATION), qui à la différence des précédents est un polyterme non seulement au but mais aussi à la source.

Recouvrements et décomptes : brouillage

Une des difficultés introduites par l'utilisation des polytermes apparaît au niveau des décomptes de nombre d'occurrences. Si l'on retient comme index le polyterme seul, sans les index qui pourraient être dérivés de ses composants, on peut perdre une information importante, par exemple on se prive d'un rapprochement avec une variante elliptique (ex. *World Wide Web* non mis en relation avec *Web*). Si l'on choisit de garder aussi des composants, d'une part il faut identifier les formes valides (la combinatoire des composants n'est pas libre), d'autre part les composants doivent être recensés comme étant issus du même segment de texte.

D'après ses résultats, l'outil décrit ici fait le choix de retourner tout ce qu'il reconnaît : dans *bases de données relationnelles*, outre BASE DE DONNEES RELATIONNELLE, il semble extraire BASE DE DONNEES, DONNEE.

Deux choses sont dommageables pour notre application. Premièrement, on n'a aucune indication sur les descripteurs qui sont dérivés des mêmes occurrences, ce qui engendre un phénomène de redondance incontrôlée, comme pour le syntagme *bases de données relationnelles* qui se voit traduit par trois descripteurs indépendants. Deuxièmement, quand un composant peut être associé à un descripteur, ce descripteur est irrégulièrement extrait, ce qui rend peu fiable le décompte des composants (par exemple DONNEE n'est pas toujours retrouvé et comptabilisé dans *base(s) de données*).

e) Entre le jeu des ambiguïtés et le risque des contresens

La force du thésaurus est de se situer non pas au niveau des mots de la langue mais des concepts d'un domaine : cela fait aussi sa faiblesse pour son utilisation par l'indexation automatique expérimentée ici.

Le contexte perdu

En effet, face à une forme linguistique trouvée dans le texte, par exemple *plan*, l'application ajoute trois descripteurs indépendamment les uns des autres : `plan (dessin)`, `plan (objectif)`, `plan (surface)`. Cela multiplie insidieusement la présence du terme (toute occurrence d'un des sens étant traduite par l'occurrence de tous les homonymes) alors que justement son sens prête à confusion.

De plus, le recours au contexte est difficile. L'indexation n'indique pas quels sont les mots qui, localement, ont permis de trouver tel ou tel descripteur. La seule notion de voisinage que l'on ait entre les descripteurs est celle correspondant à l'organisation hiérarchique du thesaurus, essentiellement l'appartenance à un même champ sémantique ou non. En faisant l'hypothèse que le texte est centré sur une thématique, un filtrage est possible qui retient les descripteurs concentrés dans un même champ et rejette les descripteurs isolés dans des champs lointains. Cependant, on est alors extrêmement dépendant de la finesse et de la régularité de la structuration du thesaurus, et de sa bonne représentation du domaine du texte. Des études montrent que le premier point n'est pas acquis (Sta 1992) (Sta 1997, §3.6.4) ; quant au second, c'est une difficulté patente des référentiels *a priori*.

Une sémantique préétablie

Le risque inverse à la prolifération des index trouvés, c'est l'indigence des descripteurs proposés : un ou plusieurs concepts sont proposés, mais aucun ne convient. L'indexation s'écarte alors du texte des façons les plus inattendues, égarant l'interprétation et laissant l'utilisateur perplexe. Par exemple : une *méthode bayésienne de calcul* se transmue en `CALCUL (ORGANIQUE)`, la *distribution numérique* associée à une loi de probabilité introduit la `DIRECTION DE LA DISTRIBUTION EDF GDF`. Ce genre de coq-à-l'âne épinglent froidement les moindres décalages entre le thesaurus (terminologie utilisée pour l'indexation contrôlée) et le corpus traité.

f) Une terminologie en décalage

Complétude

Du fait de la perpétuelle évolution de la langue à travers ses usages, et de la perpétuelle évolution des pratiques et des champs d'étude, du fait aussi de l'impossibilité de définir algorithmiquement la frontière posée par un niveau d'analyse, les terminologies sont vouées à l'incomplétude. Les efforts pour l'entretien du Thesaurus EDF devraient maintenir une incomplétude seulement marginale.

Cependant, « le système manque de vocabulaire » (pour reprendre les propos d'un utilisateur) précisément sur la frange des nouveautés (pas encore intégrées), des dénominations précises de produits ou d'intervenants majeurs (leur évolution rapide n'entre pas dans la logique d'un vocabulaire contrôlé). Et ce sont là justement parmi les éléments les plus significatifs dans la description des projets de recherche.

Représentativité

La terminologie est représentative si elle est suffisante pour décrire sans lacune majeure ni contre-sens le texte considéré (sujet, jargon...) au niveau d'analyse correspondant. Dans notre cas, le thesaurus est assez représentatif des ARD, l'est moins pour des CV.

Reste que l'esprit encyclopédique du Thesaurus EDF, qui traite aussi bien des domaines techniques de l'électricité que d'anatomie ou de philosophie, peut systématiser des interprétations majoritairement erronées dans le contexte de la DER. D'où l'intérêt toujours présent à EDF pour des terminologies spécifiques et sur mesure.

Niveau de détail

Le thesaurus se déploie en partant des termes génériques, des branches ; il ne rentre pas dans le détail des termes les plus pointus, il privilégie plutôt les dénominations synthétiques. Passer à un plus grand niveau de détail rend plus exigeant le maintient de l'équilibre général du thesaurus, et met peut-être à l'épreuve l'organisation hiérarchique des descripteurs.

Cependant, rester à un trop grand niveau de généralité ne permet plus de singulariser chaque profil. L'analyse d'un article intitulé « Des outils pour la gestion des carrières et des compétences », qui a fourni pour le calcul des rapprochements des descripteurs du type PROGICIEL, OUTIL, INFORMATIQUE, a manqué la caractéristique saillante du thème, et a de ce fait généré des résultats sans valeur.

Expressivité

Le descripteur est choisi pour désigner un concept : de ce fait, il est normal qu'il puisse recourir à une formulation complètement différente des mots utilisés dans le texte. Or le terme choisi n'est pas toujours un bon représentant de la notion qu'il est censé traduire : notamment, certaines dénominations tombent en désuétude avec l'évolution des techniques (ORDINATEUR UNIVERSEL).

La déconnexion du descripteur au texte rend problématique l'interprétation, particulièrement quand il y a un contresens manifeste : on s'évertue quelquefois en vain à rechercher dans le texte ce qui a pu motiver l'attribution du descripteur, et l'interprétation du rapprochement tourne court.

g) Eloignement au texte

Dérive par synonymie hors contexte

Les liens sémantiques entre termes isolés conduisent parfois à des reformulations pour le moins déconcertantes. En théorie, ce n'est que modérément gênant pour le calcul des similarités (si on a la même procédure d'indexation pour le profil et le document, il ne peut y avoir que création de bruit, ceci dans le cas où le contexte formé par les autres termes communs est insuffisant ; il n'y a pas d'augmentation du silence). En revanche, l'utilisateur voit sa tâche compliquée pour évaluer la pertinence des rapprochements. Par exemple, on observe que *compte-rendu* devient CHROME (les étapes de transition étant l'abréviation CR puis la notation Cr). D'où une curieuse répartition du descripteur CHROME sur les Actions de la DER, puisqu'il caractérise à la fois le Groupe *Ingénierie des Systèmes d'Information*, qui compte parmi les documents qu'il traite les compte-rendus d'Actions, et le Département *Etude des Matériaux* ainsi que son Groupe *Métallurgie*. De même, le Groupe *Environnement social et socio-économique* se trouve, par le biais de ses *vagues d'enquête*, caractérisé par le même descripteur HOULE que l'hydro-électricité maritime[88].

[88] Le même genre de dérives ont été constatées au CNET, dans l'utilisation d'une propagation sémantique pour un moteur de recherche sur Internet :
sport de combat → *boxe* ← *mail boxes*,
bande dessinée → *BD* ← *boulevard*,
fax et email → *émail* ← *émaux*, etc.
Ces transformations, automatiques, sont incompréhensibles pour l'utilisateur. La conclusion est qu'un système interactif d'aide à la reformulation est bien préférable. Ou qu'il faudrait pouvoir / savoir prendre en compte le contexte...

Repérage

Suite aux divers traitements linguistiques (lemmatisation, nominalisation,...) et lexicaux (liens de synonymie), il est parfois difficile de savoir d'où vient un terme, c'est-à-dire à partir de quel(s) mot(s), groupe(s) de mots, voire même seulement de quelle(s) phrase(s) il a été généré. Or cela est important à plusieurs titres :

Aide à l'interprétation

- *justification* du terme par sa localisation : ceci est utile pour analyser l'origine des termes aberrants et pouvoir éventuellement trouver une manière propre d'y remédier.
- assistance au *balayage* : une assistance à l'interprétation des rapprochements proposés consiste à présenter à l'usager les passages du document qui ont le plus contribué au rapprochement (fonctionnalité intégré dans SPIRIT et qui rencontre la satisfaction des utilisateurs).

Apport d'informations pour le calcul

- utilisation des *zones de localité* : le contexte du document ne suffit pas toujours pour décrire les relations entre les mots. Le voisinage dans les mêmes paragraphes ou dans les mêmes phrases est une information utile à exploiter.
- étude de la *distribution* d'une notion : pour un document relativement long, il peut être intéressant de voir si un terme est en fait issu d'une sous-partie particulière du document ou s'il court tout au long de celui-ci. Un terme peut alors se voir reconnaître une portée plus ou moins locale ou globale.
- interprétation dans le cadre d'une *partie* pour un genre connu : par exemple dans certains documents très structurés, il est très informatif de savoir à quelle partie rapporter tel ou tel terme. La structure peut être plus finement prise en compte si les descripteurs y sont rapportés.

Une manière de contourner cela est de soumettre à l'indexation automatique non pas le texte tels quels, mais par exemple les phrases une à une.

h) Déconvenues par rapport au bilinguisme

Le bilinguisme (français-anglais) du thesaurus EDF apparaît comme une des motivations majeures de son utilisation pour la caractérisation des destinataires et des documents pour la diffusion ciblée. Le principe est d'être en mesure d'obtenir une caractérisation en descripteurs français d'un texte anglais : on réalise une indexation sur la langue anglaise, et on utilise le fait que tout descripteur anglais a un et un seul correspondant français pour passer d'une langue à l'autre.

Comme l'ont illustré les expériences de traduction automatique, traduire mot à mot est loin d'être satisfaisant. Disposer d'une terminologie bilingue est cependant un peu différent : en effet, la conversion d'une langue à l'autre ne se fait plus d'une forme linguistique à l'autre, mais d'un concept à l'autre, ce qui devrait limiter les dérives. Néanmoins, nous voudrions émettre plusieurs réserves.

Constituer et surtout tenir à jour un thesaurus bilingue est extrêmement délicat et coûteux, dans la mesure où chaque traduction nécessiterait la validation d'un expert du domaine (le vocabulaire technique n'est bien connu que des spécialistes, qui connaissent les formulations en vigueur et leurs emplois corrects).

De plus, même si l'on ne manipule que des termes (devant donc désigner de façon univoque les concepts d'un domaine), il n'est pas du tout évident que les termes que l'on a mis en correspondance recouvrent exactement la même réalité. Chaque langue véhicule une ontologie, une vision du monde, chaque langue a sa façon (et ses moyens) de découper et catégoriser l'environnement perçu. Ces décalages sont sans doute limités dans le vocabulaire scientifique et technique (qui nous intéresse ici), cependant il faut avoir conscience de ce phénomène, voire évaluer l'écart pour pouvoir l'estimer négligeable.

D'autre part enfin, au niveau du document dans son ensemble, les textes ne se prêtent pas nécessairement à une transposition directe d'une langue dans une autre : les formes et les habitudes de rédaction peuvent varier d'une langue à une autre.

Ces différents facteurs, se cumulant aux décalages déjà observés pour l'indexation monolingue, contribuent à expliquer les mauvais résultats obtenus par l'utilisation de l'indexation bilingue dans DECID. D'autres pistes se présenteraient mieux. L'une consiste à construire les profils

à partir de textes originaux dans les différentes langues que l'on veut traiter. Une autre consiste à s'en tenir au niveau de la langue (et non de raisonner sur le plan des concepts) et à utiliser un dictionnaire bilingue. Un mot donné peut avoir de multiples traductions possibles ; cependant, la participation à un même contexte sélectionne les traductions pertinentes et rend les traductions farfelues inopérantes. En effet, le texte dans la langue originale, avec lequel le rapprochement est calculé, fournit un contexte valide et attesté pour trouver les mots qui participent à une thématique cohérente. Cette stratégie a été expérimentée avec succès pour construire un SPIRIT multilingue[89]. La transposition à DECID n'est cependant pas immédiate dans la mesure où les requêtes usuelles de SPIRIT sont beaucoup plus courtes (mots-clés, phrase) que celles de DECID (texte).

[89] « As long as the amiguous word is surrounded by other words, the system will identify as most relevant document the one in which the ambiguous word is surrounded by a number of other words also appearing in the query. Experience shows that in such cases, the text considered as most relevant features the ambiguous word in the correct semantic interpretation. » (EMIR 1994, p. 7)

C. DEUX ÉTAPES MÉDIATRICES : CONSTRUCTION, ÉLECTION

1. De la nécessité de renoncer à une extraction directe des unités pour caractériser le texte

a) Non superposition du plan de l'expression et de celui du contenu

Le principe qui consiste à extraire du texte certaines formes (pour le découpage, les chaînes de caractères), et à les considérer comme des unités sémantiques (au sens où elles sont une indication d'un sens), s'avère intrinsèquement insatisfaisant.

Les linguistes déjà nous avertissent : les relations entre signifiant et signifié, entre graphie et unité lexicale, ne se laissent pas décrire en correspondances biunivoques et sont autrement plus complexes. Cet aspect fondamental des langues humaines a motivé la définition de langages formels, qui ont la propriété d'univocité voulue. Ce qui se gagne alors en simplicité de traitement se paye sur le plan de la richesse expressive : les langues seules ont une finesse et une dynamique à la mesure des besoins de communication et des désirs d'expression de l'homme.

b) Le cas du (meilleur) découpage

Dans la pratique, retenir les mots (graphiques) comme les unités de la représentation du texte engendre de multiples confusions et dislocations. Les confusions sont l'amalgame d'unités qui, pour avoir une description cohérente, devraient être identifiées séparément. Les dislocations sont à l'inverse l'instauration en unités (indépendantes) de composants d'une seule unité effective.

Une première série de confusions et de dislocations concerne la segmentation du texte en mots. Les formes contractées (par exemple, *du* qui s'analyse en *de* + *le*), les sigles (*TALN* pour *Traitement Automatique des Langues Naturelles*), condensent sur une seule unité l'équivalent de ce qui est traduit, dans le même corpus, par plusieurs unités. Le cas converse, celui des dislocations, se présente aussi rapidement : c'est la non reconnaissance des formes composées, en particulier des locutions (*parce que*, *a priori*).

Même pour des unités lexicales correctement délimitées, les confusions et dislocations persistent. L'homonymie, et à un moindre degré la polysémie, sont ignorées, et les unités concernées groupent de façon forcée des éléments de sens indépendants les uns des autres. Les variantes orthographiques, flexionnelles (singulier/pluriel par exemple), dérivationnelles (mots de la même famille) ou les synonymies sont traduites par des unités qui n'ont pas plus de rapport entre elles qu'avec n'importe quelle autre unité.

Cette présentation simplifiée suppose évident la reconnaissance du « bon » découpage, ou l'énumération des significations d'un mot. Or les unités ne sont pas des données absolues, elles procèdent de choix interprétatifs : Charles Muller (Muller 1977, §1 à §5) montre, en détaillant la liste des cas à traiter et en l'illustrant de nombreux exemples, la nécessité d'expliciter une *norme lexicologique* pour une telle tâche de repérage d'unités au fil du texte. Reste que ceci n'est qu'une grille d'analyse (morphologique, grammaticale, etc.), qui ne peut se poser en norme sémantique, et dont la pertinence n'est justifiable que par rapport à des objectifs linguistiques particuliers (Tournier 1985).

Avec ces procédures, ce que nous obtenons dans le meilleur des cas reste des unités lexicales. Or la sémantique ne se cantonne pas au palier lexical. Le mot a été débouté de son statut d'unité significative minimale : la linguistique comparative le dissocie en unités significatives plus élémentaires (Ducrot, Todorov 1972, § *Unités significatives*). Inversement, par delà le mot, un rapport paraphrastique, d'équivalence sémantique, peut être admis entre deux énoncés, sans qu'il y ait d'équivalence mot à mot. Plus généralement, les opérations d'expansion ou de condensation forment les passerelles sémantiques entre mot et phrase (la définition d'un mot dans le dictionnaire),entre paragraphe et texte (le résumé), etc.

c) Les mots du texte comme balisage (le sens est entre*) plutôt que comme code (le sens est* dans*)*

Les unités qu'une analyse linéaire extrait du texte ne sont pas adéquates pour être la représentation du contenu du texte. La langue n'enferme pas le sens, ce n'est pas un code qui capture l'information : au contraire, elle le laisse passer, s'infiltrer, se glisser dans la chaîne des signes. La vision coercitive doit céder le pas à une vision dynamique... C'est *entre* les mots que circule le sens. Ou encore, chacun a en mémoire ces visualisations scientifiques de phénomènes d'interférences[90]. Deux points-sources, activés, émettent des ondes ; aux croisements de celles-ci apparaît un motif régulier. Des mots d'un même contexte, diversement activés par la lecture, donnent à percevoir tel et tel sens possible.

Le traitement automatique, pour obtenir matière pour sa description, nous assigne comme point de départ une traduction du texte en unités d'analyse. La solution est donc là : prendre appui sur cette première récolte d'unités pour construire, en quittant la myopie du traitement déroulant les lignes, les unités qui font sens pour la représentation. La différence fondamentale introduite entre les unités d'analyse (élémentaires) et les unités descriptives (construites), ouvre un degré de liberté nouveau. Les unités descriptives s'affranchissent du confinement au niveau lexical, elles peuvent traduire des unités significatives *infra- et supra*-lexicales.

2. Statut des unités élémentaires

a) Présentation

Si l'on ne procède pas à son analyse, le texte fait bloc. La première intervention pour le scruter, le décomposer, fournit ce que nous appelons les *unités élémentaires*. Ces unités élémentaires sont le matériau à partir duquel construire les *unités descriptives*.

Le nom d'*unités élémentaires* est choisi pour souligner leurs propriétés principales. Elles font figure de ce qu'il y a de plus *simple* : le type le plus simple d'unité descriptive consiste en une unité élémentaire. Les unités élémentaires sont *atomiques* vis-à-vis de l'ensemble du traitement, c'est-à-dire qu'elles sont elles-mêmes indécomposables. Elles sont les briques, les *éléments primaires* dans lesquels piochent les constructions ultérieures.

b) Le seuil de discernement

Un niveau d'analyse trop grossier pénalise, en cascade, la qualité de la représentation obtenue pour le texte. Il occulte des articulations linguistiques signifiantes, qui auraient pu être directement reprises par les unités descriptives. N'offrant qu'un nombre insuffisant d'éléments sans nuances, l'analyse trop grossière conduit à forger les unités descriptives par des croisements complexes d'unités élémentaires, entrant dans un jeu combinatoire qui alourdit le traitement et est limité dans son pouvoir expressif. En effet, les opérations de composition pour former les unités descriptives ne sont pas de même nature que les découpages linguistiques : au mieux, on arrive à une approximation. C'est acceptable quand les besoins d'ajustement restent marginaux : grâce donc à ce jeu possible de croisements il y a une petite tolérance, essentielle pour la robustesse du système.

Un niveau d'analyse trop fin a surtout l'inconvénient de faire grossir inutilement le traitement, en temps de traitement et en ressources de mémoire. Or justement l'ensemble du traitement est déjà très complexe et affleure les limites des machines actuelles pour les plus gros corpus. Cela oblige donc à être économe, à commencer par une vigilance particulière sur ce point.

Le niveau de découpage en « tokens minimaux », dans l'évaluation GRACE, correspond bien à l'optique DECID[91]. L'idée est qu'il est plutôt plus simple de « recoller » ce qui a été abusivement morcelé, que de rendre compte de la nature composite d'un élément mal dégrossi.

---

[90] Ou à défaut, a jeté quelques cailloux dans l'eau lisse d'un bassin.

[91] Dans le contexte de l'évaluation des étiqueteurs morpho-syntaxiques pour le français GRACE, un token est une unité à laquelle il est demandé d'affecter une catégorie morpho-syntaxique. Concrètement, les tokens sont les

Le « bon » choix des unités élémentaires est gratifié par une construction plus directe des unités descriptives. Peuvent être utiles ici des connaissances propres à la langue, notamment sur sa morphologie.

3. Statut des unités descriptives

a) Présentation

Les unités descriptives sont les unités significatives utilisables pour la description des documents. Elles sont construites à partir des unités élémentaires, et relativement à un ou plusieurs points de vue : une description n'est jamais neutre.

Leur ensemble forme le vocabulaire utilisable pour la description des textes. C'est le dictionnaire de référence, qui cerne le champ de perception du système, son ontologie. Nous choisissons d'appeler *univers* associé à un traitement l'ensemble des unités descriptives disponibles lors de ce traitement. Ces unités peuvent préexister au traitement ou / et être construites dans une première phase du traitement.[92]

Si un texte présente des unités élémentaires qui ne sont intégrées dans aucune unité descriptive, alors ces unités élémentaires ne seront pas exploitables pour la description et la caractérisation du texte. Tout texte peut être soumis au système, mais moins il correspond à l'univers précisé, plus sa description est pauvre.

b) Construction automatique fondée sur un corpus représentatif

Le premier mode de construction des unités descriptives est de les calculer en fonction d'un corpus. Cette approche est centrale, elle apporte une contribution massive et déterminante à l'univers qu'elle fonde, et elle garantit aussi des qualités importantes à ses unités descriptives.

Les unités ainsi construites sont *attestées* : elles ne correspondent pas à une vision normative *a priori*, mais résultent d'une observation des faits. Elles sont ancrées dans la réalité.

Autre qualité de ces unités descriptives, elles sont forgées à partir d'une vue *en contexte* des unités élémentaires, et d'un traitement *global*. Alors que les unités élémentaires sont par nécessité définies localement, c'est avec ces unités descriptives que se fait le premier pas d'une détermination du local par le global.

Enfin, les unités sont globalement utiles et *efficaces*, relativement aux textes qui constituent le corpus. On sait toute la difficulté qu'il y a à valider des unités terminologiques par exemple : on ne peut se passer du jugement d'un expert, voire d'une commission d'experts ; et la définition même de

---

séquences de lettres, les séquences de chiffres, et tout caractère restant pris isolément (notamment la ponctuation).
Le format choisi pour l'expression des résultats de l'étiquetage ménage cependant la possibilité de décomposer un token pour lui assigner plusieurs étiquettes (ex. *du* → *préposition* + *nom*), et de regrouper plusieurs tokens sous une même étiquette (ex. *parce que* → *conjonction*). L'idée fondamentale est de minimiser le recours à ces réécritures complexes en ayant les tokens du bon niveau dès le départ. D'autre part, il vaut mieux un découpage un peu trop fin (il est simple de regrouper des tokens successifs) plutôt que pas assez (affecter une suite d'étiquettes à un token est peu précis : cela ne donne pas l'information sur la manière dont le token se décompose et est mis en correspondance avec chaque étiquette). Le bon niveau est donc un découpage « minimal ».
(cf. Journée ATALA organisée par Patrick PAROUBEK et Martin RAJMAN, « Le marquage morpho-syntaxique : résultats de l'évaluation GRACE et perspectives », Paris, 24 octobre 1998.)

[92] Cette conception des unités descriptives rejoint la position de (Nazarenko 1996) quant à la mise en œuvre d'une description lexicale sémantique :
- non pas signification absolue, mais perspective différentielle,
- non pas recensement de tous les sens possibles, mais description dans un discours donné,
- non pas définition a priori, mais construction endogène (navigation dans le corpus, expertise),
- non pas automatisation totale du traitement sémantique, mais guidage pour le choix et l'interprétation des traits sémantiques.
Nous ne la suivons plus dans l'organisation hiérarchique de concepts et l'utilisation des traits comme étiquetage de mots (le sens est atomisé, et ne vient pas du contexte).

ce qu'est un terme est toujours en débat. L'unité descriptive s'en tient, elle, à une critère d'efficacité au plan du traitement. Elle est validée pour autant qu'elle se montre utile, pertinente et économe pour la description. Elle n'a donc pas une justification intrinsèque, mais relative à un univers[93].

Le corpus apporte une intertextualité « dynamique » : en faisant varier le corpus, on fait varier l'environnement intertextuel dans lequel s'inscrit la lecture du texte. L'automatisme permet de multiplier ainsi, au gré des perspectives souhaitées, les représentations.

Le local propose...

Lors de l'analyse en unités élémentaires, certaines informations de relation entre les unités élémentaires peuvent être décelées et enregistrées. Il ne s'agit en fait que d'hypothèses ou de propositions de relation : signalées par l'analyse locale, c'est en faisant une synthèse de leur présence et de leur comportement global sur l'ensemble du corpus que ces relations sont validées en se traduisant en unités descriptives. L'analyse ne relève donc que les relations potentielles dont la nature est intéressante pour construire des unités descriptives.

Un découpage peut s'appuyer sur quelques critères de forme sur les sous-chaînes qu'il détache. Il peut examiner la présence de certains caractères, la casse. L'algorithme indique par exemple les chiffres et leur ordre de grandeur, les mots commençant avec une majuscule, ceux tout en majuscules, ceux qui sont une alternance de majuscules et de points. L'idée est d'en déduire ensuite des équivalences modulo la majuscule de début de phrase, des dénominations complexes sous forme de sigle ou de syntagme développé, la reconnaissance de données factuelles particulières (dates, prix), etc.

L'analyse d'un catégoriseur / lemmatiseur apporte des informations morphologiques plus riches : formes qui sont à rapporter à un même lemme (voire à une même racine ?) ; statut de verbe auxiliaire ; occurrence de tel temps, de tel mode. Cela suggère des classes d'expressions correspondant à une même notion, des équivalences entre mots-outils qui ont un statut analogue, des mots marqués par une même tonalité (ton assuré *vs* réservé, etc.).

Pour peu que l'analyse recoure à une description syntaxique, elle est en mesure d'indiquer les liens de dépendance, accord et rection. Ce qui peut en être tiré : des formes complexes qui correspondent à des désignations précises, des entités qui interagissent et font partie d'un même scénario.

Si l'analyse comporte un niveau sémantique, on aura par exemple des indications de synonymie ou de voisinage de sens., ou des catégories générales (action qui dure, être vivant). Là encore on en tire des suggestions de regroupements notionnels synthétiques.

Il y a aussi des informations de relation possibles en dehors de la phase d'analyse. Différentes zones de localité définissent des voisinages, des rayons d'(inter)action, au sein desquels des unités élémentaires entre en relation. On calcule par exemple des regroupements significatifs de termes qui apparaissent ensemble dans certains paragraphes, et quasiment jamais ailleurs. A des zones de localité de tailles d'ordre différent répondent des relations de natures *a priori* différentes.

...le global dispose

En fonction de la nature de chaque relation, des critères globaux sont établis pour adopter ou non la relation. L'examen de chaque récolte de relations n'intervient qu'une fois tout le corpus parcouru, afin d'avoir une vue d'ensemble. Il y a trois degrés de tests.

La première chose à faire est de compléter, et sinon d'éliminer, les relations partielles : une relation forme avec majuscule en début de phrase / forme en minuscules qui ne réalise sur tout le corpus que la forme avec majuscule ; ou encore, relation d'un terme avec un synonyme « fantôme », qui dans les faits (ceux du corpus !) n'apparaît jamais. Cette première étape vise à ne conserver que les relations bien définies.

[93] Le « dictionnaire » d'unités descriptives doit son existence aux textes et ne détient aucune primauté : « Les types sont des reconstructions transitoires, selon les objectifs de la pratique en cours, et ne jouissent d'aucune prééminence ontologique sur les occurrences. » François RASTIER, *Thématique et topique*, conférence à l'Université de Winnipeg, 7 octobre 1998.

Le deuxième stade sert à préciser le statut de la relation : en fonction de sa manifestation dans le corpus, la relation indiquée par l'analyse correspond à tel ou tel type d'unité descriptive. Par exemple, on mesure à partir de toutes les occurrences dans le corpus le degré de figement d'une expression composée : s'il est maximal, la relation servira à définir un certain type d'unité descriptive ; s'il est modéré mais significatif, c'est un autre type d'unité descriptive qui pourra s'appuyer dessus, autorisant l'interprétation des variantes. Le deuxième degré de test sélectionne donc les relations admissibles compte-tenu des objectifs de description, et les associe au type d'unité descriptive correspondant.

A ce point, on s'est assuré de la viabilité des relations : le dernier test ne garde que celles qui sont efficaces, qui ont un bon rapport rendement / coût. Autrement dit, la description est ouvertement opportuniste, avec d'autant moins de scrupules qu'elle est destinée à des calculs internes. Par exemple, on renonce à tel terme composé long plus précis mais d'usage nettement minoritaire, au profit d'un de ses composants ; l'autre composant, l'abandonné, peut toujours exister et être utilisé pour la description, mais le lien particulier qu'il avait en tant que composant n'est plus reconnu. Cette logique économique se doit d'être clémente : elle régule le développement de la description, mais elle ne peut faire disparaître un élément descriptif peu rentable mais seul à couvrir certaines zones du corpus.

c) Un apport particulier : unités distinguées

Certaines unités sont définies pour les besoins du traitement, dans lequel elles ont un rôle particulier. Elles sont alors consignées dans un fichier (consulté par le programme) et portent chacune un nom, qui aide à se rappeler leur rôle et sert à les désigner au moment où on en a besoin.

Par exemple, on peut intégrer la notion de mots-outils (mots vides) au moyen d'une seule unité descriptive, qui traduit la liste de ces mots. Ensuite, le sort des mots-outils se résume à ce que le traitement fait de cette unité. Ainsi, supprimer les mots outils pour ne pas encombrer la description revient à filtrer cette unité. Mais déjà, le seul fait de saisir l'ensemble des mots-outils sous une seule unité allège notablement la description.

Les ponctuations ont une bonne raison d'intervenir dans le traitement, pour participer à la définition des petites zones de localité, celles de l'ordre du syntagme et de la phrase. Des unités descriptives les représentent au niveau d'abstraction qui sert au traitement. Pour DECID on reconnaît les ponctuations semi-fortes (point-virgule et deux-points), les ponctuations fortes (les autres points), les ponctuations faibles (virgule), les frontières ouvrantes et fermantes (parenthèses, guillemets)[94].

d) Introduire les genres

L'examen approfondi d'un genre, basé sur un large ensemble de textes représentatif, permet d'expliciter des régularités qui le caractérisent. Ces régularités jouent un rôle essentiel dans la représentation que l'on se fait d'un texte du genre, car elles induisent une appréhension globale, des attentes avant et au cours de la lecture, elles établissent des implicites.

Prenons le genre des descriptifs d'activité de la DER : pour peu qu'il en ait déjà eu quelques-uns entre les mains, le lecteur *sait* qu'un texte d'Action s'organise en parties, comprenant généralement une description du contexte, le but principal visé, le détail des étapes planifiées pour l'année.

Il repère les intertitres qui marquent ce plan, car ils ont une formulation à peu près conventionnelle, même en l'absence d'une mise en page qui les fasse ressortir : *But de l'action*, *Objectifs et principales étapes pour l'année N*, font effet de frontières qui cernent des informations bien définies. *But de l'action*, entre deux paragraphes, est donc saisi comme une unité par la lecture, au même titre d'ailleurs que ses variantes comme *But* ou *Objectif de l'action*.

[94] Les choix faits dans le système ALCESTE sont notablement différents (Reinert, Piat 1995). Il n'y a pas de notion de frontières ouvrantes et fermantes, celles-ci sont redistribuées sur les séparateurs (guillemets) et les ponctuations faibles (parenthèses). Les points de suspension font partie des ponctuations faibles. La virgule est une ponctuation semi-forte, au même titre que les deux points. Le point virgule est hissé au niveau des ponctuations fortes, rejoignant le point, le point d'interrogation, le point d'exclamation, et la barre oblique.

La lecture occupée de trouver le contenu de l'Action, passe sur la phraséologie commune et sur les termes de construction qui servent de liant, la manière dont le travail de rédaction enchâsse dans des phrases, des paragraphes, les termes et expressions qui renvoient aux idées-clés du texte. Chaque partie a sa toile de fond qui la rappelle : *Cette action vise à* est une manière d'ouvrir la partie *But*.

Des formules conventionnelles servent pour certaines réponses à des renseignements attendus. Le texte d'Action est censé m'indiquer ce qui vient d'être fait l'an passé : la mention *action nouvelle*, en lieu et place du développement de la partie *Etat d'avancement*, précise le statut chronologique de l'action.

En généralisant ces observations, la description d'un genre peut déterminer et s'appuyer sur des unités descriptives traduisant les régularités attendues, que ce soit dans la forme des intertitres structurant le texte, dans le style rédactionnel adopté, dans la mention d'un cas pris dans une liste de possibilités prévues.

On retrouve bien tout ceci pour la description des CV : *langues*, *langues vivantes*, etc. est une unité descriptive qui annonce une partie ; *anglais*, *allemand*, *espagnol*, *chinois* etc. font partie des réalisations possibles d'un renseignement attendu, idem pour les systèmes d'alternatives *lu* / *écrit* / *parlé* ou *courant* / *(niveau) moyen* /...

L'étude ici a surtout été faite pour des documents avec une structure marquée référant à un plan type, mais que l'on songe à d'autres genres très différents et l'on trouve encore des unités descriptives particulières : le premier élément du conte n'est-il pas le canonique *Il était une fois* ? et le dernier, le non moins rituel *Ils se marièrent* / *Ils furent heureux*, et *eurent beaucoup d'enfants* ?...

Ces unités sont autant d'informations que l'on peut apporter au système pour l'aider dans sa construction de représentations des textes. Puisque ces unités lui sont données, le traitement peut les utiliser avec des règles spéciales prévues pour un genre (qui par exemple distinguent des parties) : ce sont des unités distinguées supplémentaires (cf. § précédent). Si cette description initiale est bonne, les unités s'intègrent de façon cohérente dans des unités plus complexes calculées par le système. Encore une fois, ce sont la réalité des textes concrets et le contexte global du corpus qui façonnent la représentation : les unités pertinentes sont promues et utilisées, les autres sont négligées.

Les unités d'un genre ne doivent bien sûr être prises en compte que pour un texte de ce genre, elles peuvent même être associées à une partie définie pour le genre. C'est donc un paramètre du traitement. Les genres à décrire sont bien sûr fonction de l'application envisagée : pour DECID, cela commence par les textes d'Action, puis peut se poursuivre par les Notes internes, les CV, etc. en fonction des utilisations principales.

La description fine d'un genre demande un travail conséquent, puisqu'il faut étudier un bon nombre de textes. Elle peut être assistée par des procédures automatiques. Des critères statistiques permettent de retenir les éléments qui contrastent des genres : on recherche les éléments spécifiques à la partie d'un corpus correspondant à un genre (indices de spécificité : écart réduit, etc.), ou on extrait ce qui oppose les différents genres les uns aux autres (critère discriminant), ou on fait ressortir les éléments « fond de sauce » communs et répandus sur un corpus « mono-genre » (mesure de dispersion, de distribution uniforme). Avec une analyse factorielle, Biber découvre un système d'axes dans lequel chaque genre prend place, et qui associe un certain nombre de propriétés linguistiques à un genre en fonction de sa position.

4. De l'univers descriptif au texte : l'exigence du sur mesures

a) Le pouvoir décisif du texte

Le texte, en tant qu'il constitue une unité autonome, imprime son individualité à la fois en ce qui concerne son contenu et vis-à-vis des autres textes environnants. Il organise ses constituants : il relativise chacun par rapport à l'ensemble, et donne un contexte qui réaménage les comportements généraux des unités au niveau de l'ensemble du corpus. Le texte aussi se positionne en se rapportant à ses voisins et en s'en différenciant. Chaque texte a sa façon de s'écarter légèrement de la norme (théorique). Considérer une suite de paragraphes comme texte délimite donc deux versants, un contexte interne (sa structure close, orientée), et un contexte externe, inter-textuel.

Une formule de pondération proposée par Ghislaine Chartron (Chartron 1988, § VII.1) est particulièrement éclairante sur ce point. La formule combine deux termes. Le premier mesure la *représentativité d'un mot par rapport au document* : s'agit-il d'un mot plus présent que les autres, dominant dans l'équilibre général du document. Le second terme retourne le point de vue, et évalue la *représentativité du document par rapport au mot* : une notion à laquelle le texte accorde une place secondaire peut se trouver revalorisée parce que ce texte se trouve être le seul à l'aborder.

Cette manière de faire a sans doute été introduite par Christian Fluhr (Fluhr 1977, §III.5), qui se sert de ces deux notions pour filtrer l'attribution des termes d'index aux documents[95] : on élimine un terme d'index pour un document si ce document fait partie des documents pour lesquels le terme est de moindre importance (contexte externe), ou si le terme d'index fait partie des termes de moindre importance pour le document (contexte interne).

Les unités descriptives ne forment donc pas directement la représentation du texte : il leur faut se doter, dans chaque contexte, d'une coloration particulière. La représentation se bâtit avec les unités descriptives qualifiées par un certain nombre d'informations sur la manière dont elles se manifestent dans le texte. Les unités descriptives procèdent en effet d'une opération de normalisation, puisqu'elles définissent un référentiel commun pour la description de tout un corpus (voire de plusieurs corpus). Les unités descriptives, en rendant les textes commensurables, ne doivent pas avoir un effet de banalisation, mais doivent servir à valoriser les contrastes intéressants. L'enjeu est de garder quelque chose de la spécificité du document qui permette le tri et le choix : la diffusion ciblée doit être assez fine pour éviter un envoi systématiquement par paquets, et au contraire faire déjà un travail de sélection.

C'est donc là la seconde phase qu'il fallait distinguer, et que nous avons appelé *élection* : les unités une fois construites sont réattribuées aux textes, qui déterminent chacun les unités qu'ils promeuvent.

b) Les unités caractérisantes

C'est l'unité descriptive enrichie d'informations particulières à un contexte que l'on appelle unité caractérisante.

Une unité caractérisante peut s'interpréter comme l'*occurrence* d'un *type*, l'unité descriptive correspondante. L'univers est un dictionnaire qui recense les types, les caractérisations des textes recèlent les occurrences pour les textes considérés.

5. Récapitulatif : deux étapes, trois unités

a) Déroulement : du texte d'entrée à sa représentation pour le calcul de proximités

La caractérisation d'un texte s'obtient par une succession d'opérations, qui permettent l'interdétermination des informations locales et globales. Le texte se présente à la machine comme une chaîne de caractères. Une analyse décompose cette chaîne en unités élémentaires. Les unités élémentaires servent à construire les unités descriptives : tel motif d'unités élémentaires forme telle unité descriptive. Un ensemble d'unités descriptives peuvent alors être associées au texte. Le texte jauge chaque unité au regard de sa réalisation dans le contexte qu'il instaure. L'unité peut alors devenir caractérisante.

chaîne de caractères →(analyse)→ *unités élémentaires* →(construction)→
unités descriptives →(élection)→ *unités caractérisantes*

[95] « Quelles que soient les améliorations apportées à la constitution de la liste des mots vides, la méthode d'indexation par conservation des mots non vides présente l'inconvénient d'être systématique. Un mot vide est éliminé quel que soit le document dans lequel il figure et inversement pour les mots non vides.
La méthode que nous proposons est plus souple dans la mesure où l'élimination d'un concept dans un document dépend non seulement de sa répartition dans le corpus, mais de la répartition des autres concepts dans ce document.
Une telle attitude essaie de minimiser le volume d'informations à conserver. » (Fluhr 1977, §III.5)

b) Comparaison avec l'ancien traitement

Le traitement jusqu'alors pratiqué dans DECID version 1 consiste à parcourir la chaîne de caractères d'un texte, à la décomposer en unités à l'aide d'un module de découpage ou d'indexation, et à renvoyer l'ensemble des unités ainsi trouvées comme représentation du texte. Une même unité peut être reconnue plusieurs fois : la représentation mentionne le nombre d'occurrences de chaque unité dans le texte.

Soit donc le schéma suivant :

chaîne de caractères →(découpage ou indexation)→ *unités*

La nature d'une de ces unités correspond à un amalgame entre unité élémentaire, unité descriptive, et unité caractérisante. Elle a la définition purement locale de l'unité élémentaire, mais n'en a plus l'ancrage contextuel : sa désignation est celle d'un *type*, de la même façon que l'unité descriptive. Sa fréquence dans le texte est une information qui la module du point de vue du texte, c'est une des informations que peut porter l'unité caractérisante.

D. LES UNITÉS ÉLÉMENTAIRES DE DECID

1. Un découpage

a) Impératif de robustesse

Pour DECID, il est essentiel que l'analyse soit robuste. De grands volumes de textes doivent être correctement couverts, en un traitement entièrement automatique. La constitution de la base des profils demande le traitement de plusieurs milliers de pages, dont le contenu évolue d'année en année.

Il n'y a pas à exiger des textes qu'ils soient rédigés selon une grammaire canonique (voire de surcroît avec un vocabulaire de référence –comme cela est imposé dans l'aéronautique pour la rédaction de certaines documentations techniques en vue de leur traduction automatique) : cela rendrait DECID incapable de traiter la quasi totalité des documents qui lui sont soumis.

De même, DECID ne peut demander un précodage manuel des textes : il prend les textes tels qu'ils viennent. Leur structuration est diversement apparente. Il ne faut pas s'arrêter à la syntaxe « concaténatoire » et a-verbale de quelques mots-clés libres entrés au clavier, ni aux distorsions ponctuelles issues de fautes de frappe ou d'erreurs d'OCR.

Le traitement d'un texte soumis comme requête doit si possible ne pas faire attendre l'utilisateur, pour que celui-ci affiner sa recherche sur un mode interactif. Et dans le cas où le traitement prend un certain délai, celui-ci doit pouvoir être annoncé et chiffré. En particulier, si l'on dépend d'un serveur, il faut pouvoir évaluer le temps d'attente dans la queue. Des solutions pour remédier aux attentes répétitives et prolongées sont toujours envisageables : gestion de priorités, parallélisation du traitement.

Cependant, une requête soumise à DECID consiste en un (petit) texte, alors que l'analyse de l'ensemble des textes pour former la base de profils est d'un autre ordre de grandeur. Le petit volume (un texte de requête) doit être traité en temps réel ; les gros corpus (base de profils), préparés de l'ordre d'une fois par an, peuvent être traités en batch : le temps de traitement n'est plus critique au niveau de la seconde, mais au niveau de l'heure ou de la journée.

b) Une analyse sur domaine ouvert

Clairement, le vocabulaire des textes que DECID traite est ouvert. Il ne cesse d'évoluer : aurait-on acquis la couverture complète d'un corpus de textes représentant une année, que l'année suivante déjà les préoccupations se décalent, quelques nouvelles façons de parler apparaissent, les noms des produits, des projets, des équipes, se renouvellent, des découvertes bousculent les précédents repères, etc.

La priorité est donc une analyse qui prenne en compte l'ensemble du vocabulaire. L'analyse peut éventuellement s'appuyer sur des connaissances lexicales (mots grammaticaux, mots courants dans le genre du corpus) mais ceci ne doit pas cerner dans le même temps le domaine de ses résultats.

La question de la portabilité à une autre langue assez proche, essentiellement l'anglais, joue en faveur d'une analyse pour laquelle la prise en compte spécifique de la langue est bien définie et limitée.

c) Proximité au texte

Les unités fournies par l'analyse doivent conserver une information de localisation, d'ancrage dans le texte, aussi précise que possible, pour au moins trois raisons.

La construction et l'identification des unités descriptives fait largement appel à la notion de contexte. Les *contextes* d'une unité doivent être accessibles à toute une gamme de niveaux, depuis le plus étroit (formes composées) jusqu'aux espaces plus larges comme le paragraphe. Le paradigme de la recherche documentaire, basé sur les mots-clés, a habitué à penser les unités hors contexte, indépendantes les unes des autres, et directement rattachées au niveau du document. Cette manière de voir n'est plus valable dès lors que l'on veut travailler au niveau du texte.

De même, les calculs de proximité requête-document font communément intervenir des pondérations qui considèrent la distribution des termes parmi les documents. Replacé dans notre cadre, cela signifie que la facette intertextuelle est reconnue et utilisée ; en revanche, la facette concernant la construction interne du texte, son déroulement, la *disposition* (concentrée, suivie, marginale,...) des manifestations d'une unité est laissée de côté. La prise en compte de la textualité nous invite à réintroduire cette facette, pour mieux caractériser les textes avant de les positionner les uns par rapport aux autres. On a donc besoin d'une information sur la position des unités par rapport au déroulement du texte.

Enfin, dans une application qui manie des textes, l'utilisateur est aussi un lecteur. Sa bonne compréhension des résultats suppose une mise en rapport possible avec le texte initial (par exemple, quelles unités, quels passages précis du texte ont motivé tel rapprochement), ou un parcours de textes qui ont été rapprochés (mise en valeur permettant de repérer très vite le rapport du texte avec la demande initiale). Les descripteurs fournis par l'indexation automatique que nous avons expérimentée ne sont pas satisfaisants sur ce point. En effet, leur formulation est souvent très générique et abstraite. A cela s'ajoute l'absence d'information sur les mots du texte qu'ils représentent. Enfin, l'éloignement sémantique est quelquefois très grand entre l'interprétation intuitive du descripteur, et les formes trouvées dans le texte : l'enchaînement de transformations non contextuelles rend cette indexation particulièrement vulnérable aux phénomènes de polysémie et d'homonymie. Un découpage a l'avantage de la clarté, quant à la provenance, dans le texte, des unités obtenues : leur formulation reste dans les termes du texte, leur *désignation* ne peut porter à confusion.

Ce dernier point n'est pas moins fondamental que les précédents : il ne se situe pas au niveau d'une simple mise en forme, mais du processus interprétatif qui conditionne l'exploitabilité des résultats. Au moment du passage de la version utilisant l'indexation à celle utilisant le découpage, les échos des utilisateurs ont été très positifs. Certains ont souligné le gain qualitatif manifeste sur les résultats. Il aurait été révélateur de vérifier si, comme nous le pensons, l'essentiel de la perception de ce gain s'est joué sur l'image beaucoup plus claire et plus représentative des termes en commun (entre le texte de requête et le texte du profil). Cela contribue directement à l'impression que la représentation du texte que fait la machine est meilleure. Par exemple, dans un texte concernant la gestion économique des services documentaires, où apparaît la préoccupation de veiller à leur rentabilité en se fixant un *taux de couverture*, la caractérisation par le descripteur `TOITURE` désoriente bien plus l'utilisateur, que le démembrement de l'expression en *taux* ou en *couverture*, où l'utilisateur voit les limites du système mais rétablit l'association avec le texte.

d) Une base autonome

D'un point de vue stratégique, il est intéressant que DECID ne soit pas dépendant d'un analyseur coûteux ou / et dont on ne soit pas assuré du suivi. DECID doit pouvoir offrir un service stable, et doit pouvoir être installé dans différents secteurs et filiales de l'entreprise sans être arrêté par des problèmes de cession de droits ou engager des coûts prohibitifs.

On choisit donc de doter DECID de son propre analyseur, conçu comme un module, et réalisant un premier niveau d'analyse simple. DECID peut alors fonctionner de manière autonome, mais on garde la possibilité de greffer un analyseur supérieur (plus fin, plus complet) lorsque les circonstances permettent d'en disposer.

e) L'héritage de la version 1

Les bonnes performances du découpage dans la première version de DECID encouragent à poursuivre cette voie. Toutefois, le module de découpage existant ne correspond pas bien à la modélisation développée pour DECID v.2. Il procède à certaines simplifications dans des cas où l'on voudrait garder des informations de relation, qui seraient traitées à un niveau global au moment de la construction des unités descriptives : suppression des accents, réduction de la casse (équivalence entre majuscules et minuscules). Mais surtout, les unités produites n'ont pas d'ancrage contextuel : on ne les situe pas dans le déroulement du texte, ni dans un paragraphe ; les ponctuations sont complètement ignorées.

Si on renonce donc à l'utiliser tel quel, le découpeur de DECID v.1 reste un précurseur marquant. Il a montré qu'un simple découpage pouvait déjà apporter des résultats dignes d'intérêt. Il comporte quelques raffinements, comme le repérage de sigles, ou de numéros d'année, dont les principes ont pu être repris.

2. Description de l'analyseur par son comportement

a) Principes généraux

L'entrée donnée à l'analyseur est une chaîne de caractères nettoyée des caractères de contrôle ; elle se constitue uniquement de lettres (minuscules, majuscules, diacritiques), chiffres décimaux, et caractères graphiques dont l'espace. C'est en fait un segment textuel entre deux balises d'un fichier au format de la DTD Corpus. Il n'y a pas de retours-chariot : ceux-ci ont été interprétés et codés, le cas échéant, par l'élément SGML <NLS>. De même, d'autres formes d'espacement, comme le blanc insécable ou la tabulation horizontale, ont été traduits en espace simple au moment de la mise au format Corpus.

La règle générale est la suivante : les espaces délimitent les unités élémentaires. Les unités élémentaires correspondent donc grosso modo aux mots, définis graphiquement comme les suites de caractères sans espace.

Illustration : si l'on se conforme à la règle générale, la phrase suivante du premier paragraphe : « Il n'y a pas de retours-chariot : ceux-ci ont été interprétés et codés, le cas échéant, par l'élément SGML <NLS>. » est segmentée comme suit : [Il | n'y | a | pas | de | retour-chariot | : | ceux-ci | ont | été | interprétés | et | codés, | le | cas | échéant, | par | l'élément | SGML | <NLS>.].

Ce découpage rudimentaire doit être corrigé par quelques règles complémentaires. En effet, comme c'est la forme graphique qui définit l'unité, les « mots » qui n'ont pas strictement la même forme donnent des unités élémentaires qui n'ont aucune relation, et ne peuvent donc pas être rapprochées pour la suite des calculs : *Il* n'a rien à voir avec *il*, ni *codés,* (avec virgule) avec *codé* (sans virgule), ni *l'élément* avec *élément*. Sans compter que tout ceci multiplie nuisiblement les unités élémentaires.

Discussion

Typographie riche ou typographie pauvre

On a donc choisi ici de garder la typographie riche, à savoir l'information apportée par la casse (minuscules vs majuscules) et les marques des caractères diacritiques (accents, cédille, tréma).

L'inconvénient de ce choix est la sensibilité aux majuscules contextuelles et aux fautes de frappe ou d'orthographe. En particulier, une occurrence d'un mot en début de phrase est enregistrée comme une unité différente de celle du même mot en cours de phrase ; ou encore, chaque mot d'un gros titre en capitales n'est pas immédiatement rapproché de ses réalisations en minuscules dans le développement du texte. De même, cela disperse la description si une partie du corpus à considérer est en typographie pauvre (par exemple, des mots-clés en majuscules, ou des fiches de travail saisies avec une interface de type Minitel). Il faut être conscient de ces formes de dédoublements pour la suite du traitement.

En revanche, le gain attendu est de distinguer des formes fréquentes (*a* vs *à*, *la* vs *là*, et le présent de l'indicatif du participe passé pour les verbes du premier groupe)[96], mais surtout d'avoir des indices sur les formes représentant des noms propres ou des sigles.

[96] L'évaluation chiffrée donnée par (Beauchemin 1986) montre que le tréma et la cédille ne distinguent quasiment aucun homographes (*maïs* est absent du corpus) ; l'accent circonflexe distingue plusieurs dizaines de mots peu à moyennement fréquents (ex. : *boîte*, *côté*, *croît*, *faîte*, *jeûne*, *mûr*, *sûr*, *tâche*) ; l'accent grave sépare *à* de *a*, *là* de *la*, et *près* de *prés*, et parmi les « autres formes marquées d'un accent grave, nous n'avons pas repéré de cas où l'absence d'accent créerait un homographe, sauf si l'on supprimait aussi l'accent aigu » ; quant à l'accent aigu, il joue un rôle essentiel pour départager présent de l'indicatif et participe passé dans la conjugaison des verbes du premier groupe, et de ce fait est quantitativement d'importance.

En fonction des utilisations (et notamment des corpus à traiter), ce choix de la typographie riche pourrait être revu.

Les morphèmes

Le morphème, en deçà du mot (dont la définition n'est même ici que graphique, et pas linguistique), serait une unité de base préférable au mot.

> Sans nier la réalité –très discutée– du *mot*, reconnaissons que cette unité n'est pas simple : son contenu peut être constitué d'un ou plusieurs sémèmes. Mieux vaudrait prendre pour base de réflexion le morphème, dont le contenu –dans un contexte univoque– consiste en un seul sémème. Par exemple, dans le cliché *Les femmes sont des fleurs*, la métaphore est établie à strictement parler entre le contenu des morphèmes *femme-* et *fleur-*, plutôt qu'entre celui des mots *femmes* et *fleurs*. (Rastier 1987, §VIII.1.2.B, pp. 175-176)

L'obstacle majeur que nous rencontrons est la difficulté (voire la faisabilité) d'un repérage automatique des morphèmes. Un analyseur morphologique pourrait nous en rapprocher. L'unité que nous avons adoptée (la forme graphique) est de l'ordre d'un regroupement de morphèmes, elle (ne) perd (qu') un rang d'analyse. La construction des unités descriptives pourra en partie rattraper la confusion, par des effets de 'réunion' et d''intersection'.

b) Détachement des ponctuations

Les règles typographiques veulent que les ponctuations, à l'exception des ponctuations doubles, soient accolées au mot qui précède. Ainsi, les ponctuations doubles (point-virgule, deux-points, point d'interrogation, point d'exclamation) devraient être généralement bien traitées par la règle générale, mais quasiment jamais les virgules et les points.

Pour chaque chaîne découpée, on détache donc une à une les ponctuations accolées à la fin, voire au début (par erreur). On s'assure aussi que les points de suspension sont reconnus comme une seule unité.

Par ponctuation ici, on entend les ponctuations fortes, semi-fortes, faibles, ainsi que les signes qui fonctionnent par paire ouvrant / fermant comme les parenthèses, les guillemets, les accolades et les crochets.

c) Le point : fin de phrase ou / et abréviation

Le repérage des phrases nous intéresse, car elles forment des zones de localité que nous voulons utiliser pour définir certaines unités descriptives. Comme nous ne disposons pas ici d'analyse syntaxique, les ponctuations fortes sont les indications essentielles pour délimiter les phrases.

Dans plusieurs cas, le point à la fin d'un mot n'indique pas nécessairement une fin de phrase : les sigles transcrits par une suite d'initiales séparées par des points (*E.D.F.*), les abréviations par troncature de la fin du mot (*par ex.*, *M. L. Dupont*, *Ph.D., p. 81, Dupont et al.*). Le point terminal peut être à la fois la notation que la forme est écourtée, et la fin de la phrase : car si la forme se trouve en fin de phrase le point n'est généralement pas doublé.

Une unité élémentaire spéciale, dite « trace », est prévue pour représenter le point de fin de phrase éventuellement implicite après un point d'abréviation. Le point d'abréviation reste solidaire de la forme abréviée. Le cas des lettres seules doit être traité un peu différemment. Une lettre majuscule est tantôt une forme abréviée et notamment l'initiale d'un prénom (le point fait alors partie de l'unité), tantôt une désignation ou un nom (*les ensembles A et B*, *le jour J*, *un bac E*) et à ce moment-là le point ne note que la fin de phrase.

L'unité élémentaire trace est interprétée comme fin de phrase ou non, une fois tout le texte analysé, en fonction de critères globaux. En particulier, on examine si ce qui suit ressemble à un début de phrase : c'est le cas par exemple si le premier mot commence par une majuscule, si on ne le rencontre sous cette forme qu'en début de paragraphe ou après une ponctuation forte, et si il apparaît sans majuscule à d'autres endroits dans le corpus.

Ces règles ont pour but de traiter convenablement des cas du genre de ceux soumis par Max Silberztein[97] : repérer la fin de phrase dans « ...le langage C. Beaucoup... » , et pas dans « T. C. Elliot est un poète qui... ».

d) Des séparateurs particuliers : l'apostrophe et le tiret

L'apostrophe n'est conventionnellement ni précédé ni suivi d'espace. Pourtant, hormis pour quelques cas connus comme *aujourd'hui*, c'est un signe délimiteur d'unités. Cela se démontrerait linguistiquement par des tests de commutation ; cela se trouve réalisé dans les dictionnaires ; et l'ignorer conduirait à une explosion combinatoire[98] du nombre des unités.

En français, l'apostrophe marque l'élision, à savoir l'effacement d'une voyelle devant une autre (éventuellement précédée d'un *h* muet). L'apostrophe appartient au premier terme, élidé, et le sépare du second, qui s'ouvre par une forme vocalique : c'est le cas le plus représenté, notamment avec les formes élidées de mots grammaticaux (*l'*, *s'*, *qu'*, etc.).

Il arrive que l'élision porte sur une voyelle initiale, c'est massivement le cas en anglais pour des formes conjuguées de *to be* et *to have*, et de la marque du possessif (*'s*). Dans le cas de *to be* et *to have*, c'est la terminaison vocalique des pronoms personnels qui est préservée, et c'est l'initiale vocalique des verbes qui tombe. On a donc ici l'enchaînement *sujet - forme élidée*, celle-ci commençant par une apostrophe suivie d'une consonne (phonétiquement). Cette règle s'extrapole, pour la troisième personne, au pronom *it* et aux sujets autres que les pronoms personnels. Heuristiquement, elle fonctionne encore pour le possessif singulier et pour certaines contractions (*'n* pour *and*).

En revanche, cela ne couvre pas le cas, plus complexe, des négations contractées (*don't*, et de même toutes les formes verbales en *-n't*). Pour traiter notamment les noms de famille d'origine irlandaise de la forme *O'Neill*, une règle un peu brutale mais à première vue efficace rattache l'apostrophe qui sépare une lettre d'une suite de plusieurs lettres à la lettre initiale, indépendamment de considérations vocaliques.

Les formes figées telles que *aujourd'hui*, *d'abord*, que l'on souhaite voir décrites comme une seule unité, sont déclarées *a priori* (elles font partie des unités distinguées) ou/et mises en évidence par les indicateurs statistiques (genre calcul de lien).

Le cas du tiret est analogue en ce qu'il n'est en général ni précédé ni suivi d'espace.

Le tiret sépare souvent deux éléments qui existent aussi comme unités simples dans la langue. Certaines constructions grammaticales introduisent un tiret : interrogation (*est-ce que*, *va-t-on*), impératif des verbes pronominaux, adverbes (*ci*, *là*). Une exploration systématique de quelques corpus représentatifs doit permettre de repérer les constructions utilisées, et d'enregistrer les unités correspondantes via des unités distinguées.

Pour les noms composés, le tiret traduit une association forte entre les deux constituants, telle qu'elle devient une nouvelle unité lexicale à part entière ; elle acquiert un sens propre, s'écartant de celui de ses composants. Le tiret est en ce sens une étape vers des procédures linguistiques diachroniques d'intégration. Pour respecter cette analyse, le tiret et les constituants simples qu'il relie sont enregistrés comme des unités élémentaires ; et sauf si des occurrences attestent le contraire, une seule unité descriptive est créée, celle de la forme composée.

Ces quelques règles permettent de gagner par rapport à un découpage très simple comme celui utilisé pour DECID v.1. Pour ce genre de découpages, l'apostrophe est un délimiteur, au même titre que l'espace. On y perd en lisibilité (*l* au lieu de *l'* par exemple) et en justesse (éclatement en formes indépendantes *aujourd* et *hui*). Quant au tiret, la nouvelle analyse permet à la fois de respecter la délimitation des constituants et l'intégration lexicale forte de l'ensemble

Bien sûr, les choix que traduisent ces règles d'analyse restent relativement sommaires. Ils se justifient dans le cadre de la constitution d'un premier outil d'analyse, permettant l'expérimentation du modèle notamment dans l'étape de construction des unités descriptives. Mais il faudrait davantage

[97] Séminaire du LADL du lundi 20 octobre 1997 : *Reconnaissance automatique des phrases*, animé par Max SILBERZTEIN, à partir de l'expérience du développement de l'analyseur INTEX.

[98] Comme l'articulation entre deux unités n'est pas perçue, c'est bien toute la combinatoire de ces unités qui gonfle l'inventaire.

tirer partie de l'expérience acquise dans ce domaine, qui a accumulé des observations plus nombreuses et plus systématiques : pour le français, des contributions dans le domaine de la lexicométrie (Muller 1977) ou de la recherche documentaire sur le texte intégral recensent méthodiquement les cas de figure qui se présentent au traitement automatique. L'analyse en unités élémentaires est destinée à être affinée ou remplacée par des outils fondés sur des études plus approfondies.

e) Des atomes non linguistiques : chiffres et symboles

La langue française possède un système d'écriture alphabétique, grâce auquel elle se transcrit à l'aide d'un ensemble de lettres fixé. Cependant, certaines notations introduisent dans les textes des « mots » utilisant d'autres symboles. Le repérage de ces formes graphiques particulières est intéressant à double titre.

D'une part, il s'agit souvent d'éléments de nature particulière (sigle, donnée chiffrée, date, adresse électronique) qui prennent un rôle différent dans des calculs de similarité entre textes. Certaines formes signalent des désignations précises qui participent significativement à la caractérisation du texte. Les cas de *C++* ou du *G7* sont des classiques dans les comparatifs des moteurs de recherche par index sur Internet. D'autres, comme les données chiffrées, sont généralement trop particulières pour participer à des rapprochements (Fluhr 1977, p. 233).

D'autre part, ils forment des micro-contextes dans lesquels les signes de ponctuation ne sont plus interprétés comme des ponctuations rythmant la phrase. D'où le fait que nous les ayons annoncés comme des atomes : on ne souhaite pas que les points, deux-points et tiret d'une adresse Web comme `http://www.msh-paris.fr/texto` , ou que les points d'un *tirage à 2.000.000 d'exemplaires* (qui deviendraient des virgules dans une version anglophone), soient interprétés comme des ponctuations ou par les autres règles vues précédemment.

Selon la manière dont s'orientent les usages de l'application, des analyses spécifiques peuvent être décrites pour identifier des unités internes dans certains cas : repérage de l'année dans les divers formats d'expression d'une date (*en 97*, *le 13 juin 1997*, *au 13.06.97*, *(06/97)*)[99] ; repérage du code identifiant l'équipe qui a émis la Note, dans la référence documentaire d'une Note interne EDF (*N46* dans *HN-46/96/011*, mais sachant que dans les faits, les numérotations attestées ont des formes très irrégulières, surtout au niveau des caractères non alphanumériques qui structurent la référence[100]).

f) Majuscules de circonstance ou initiale(s)

Le traitement le plus simple est insensible à la casse (et, cela est en partie lié, aux accents). L'avantage est de bien résoudre ainsi les majuscules purement contextuelles : l'initiale en début de phrase, le titre écrit en capitales d'imprimerie. En revanche, on génère une représentation tout à fait artificielle des sigles (*onu*), des noms propres (*pincemin*), et parfois on écrase des graphies spéciales identifiantes, mêlant majuscules et minuscules. Et bien sûr, on perd une information qui s'avère discriminante quand une même forme est à la fois nom propre et nom commun.

Le découpeur de DECID v.1 apporte une seule amélioration, simple mais efficace, par rapport au traitement le plus simple. Une chaîne de plusieurs caractères tout en majuscules et constituée uniquement de consonnes est considérée comme un sigle (formé d'initiales) et n'est pas réduit à la graphie en minuscules : ainsi, *SGML* par exemple est correctement représenté.

Le nouveau modèle, avec la distinction entre unités élémentaires et unités descriptives, permet d'éviter la réduction immédiate et myope de la casse. Les unités élémentaires seront de préférence les formes telles qu'elles figurent dans les textes. En fonction de la manière dont apparaissent les majuscules dans un mot (aucune majuscule, tout en majuscules, la première lettre seulement en majuscule, mélange de majuscules et de minuscules), différentes relations avec d'autres formes sont

[99] Ce repérage peut se faire à l'aide d'automates décrivant des grammaires locales, plus ou moins élaborés (mais jamais exhaustifs) (Maurel 1996).

[100] Cette observation est fondée sur les travaux de Véronique JOLLY, dans la mise au point d'algorithmes pour gérer SPHERE, la base de textes électroniques de la DER d'EDF.

possibles. Avec la vue globale donnée par l'analyse de l'ensemble du corpus, des unités descriptives sont construites, représentatives des relations attestées.

Voici par exemple quelques comportements typiques et les interprétations (réductions) associées :

Formes attestées :										
tout minuscules	+									
majuscule initiale en début de phrase	(+)		+		(+)					
majuscule initiale en cours de phrase					+					
tout majuscules	(+)		(+)		(+)		+			(+)
points intercalés entre (ou après) chaque lettre							(+)		+	
mélange										+
Interprétation (nature)	unité lexicale en minuscules (majuscules contextuelles)				nom propre	dénomination				
						sigle				
Exemple de liste de formes attestées	*Avec*, *avec*		*Primo*		*Rastier*	*EDF*, *E.D.F*, *E.D.F.*				*SdT*

Ces interprétations n'épuisent pas les combinaisons théoriques : cela est normal, car ce genre de combinatoire n'est pas libre (certains regroupements sont réguliers, d'autres rares ou inexistants). Elles ne sont pas non plus exclusives, du fait des homonymies. On créera donc autant d'unités descriptives qu'on a de natures distinctes.

Si l'on adopte le tableau ci-dessus, chaque type de forme est associé de manière privilégiée à une interprétation : les + (forme obligatoire) forment une diagonale (d'où bijectivité), toutes les autres associations, indiquées par un (+), étant facultatives. Ceci facilite la détermination de l'interprétation liée à une liste de formes. La procédure de sélection de la ou des nature(s) associée(s) à une liste de formes attestées est en effet la suivante : *(i)* déduire de l'ensemble des formes attestées l'ensemble des interprétations privilégiées associées ; *(ii)* pour chaque interprétation, examiner si elle peut couvrir toutes les formes attestées ; si c'est le cas, c'est la seule possible (étant donnée la structure de notre table d'association), la nature est déterminée ; sinon, *(iii)* les natures associées sont celles de la liste des interprétations privilégiées, en fusionnant *nom propre* et *sigle* sous *dénomination* le cas échéant.

La nature *dénomination* sert à synthétiser au besoin les interprétations plus précises *nom propre* et *sigle*. On maintient néanmoins la distinction *nom propre vs sigle* car elle permet par la suite des analyses spécifiques : reconstitution des noms de personne pour les noms propres (titre en *M.*, *Mme*, *MM.*, *Pr.* etc., prénoms, particules, initiales), association avec une forme développée pour les sigles (correspondance des lettres en majuscule et existence d'une mise en relation au moyen d'une tournure définitoire : parenthèse qui suit immédiatement, apposition).

E. LES UNITÉS DESCRIPTIVES DE DECID

1. Des unités typées

a) Pas d'architecture neutre

Il n'y a pas d'unité descriptive sans type. En effet, ce sont des unités construites, et à visée sémantique : elles ont donc une forme et un usage. Le type d'une unité descriptive définit :

- la *structure interne* de l'unité, autrement dit la nature de l'information qu'elle enregistre, et
- les critères de reconnaissance de la *manifestation* de l'unité dans un texte.

Plus fondamentalement, on peut voir là la définition de contraintes qui préparent des parcours interprétatifs[101].

b) Une typologie ouverte

Les types que nous décrivons ci-après sont ceux dont la définition nous est apparue utile pour notre système et linguistiquement fondée. Ils sont complémentaires et équilibrés, mais rien ne permet de dire que c'est un système complet, fermé.

Il n'est pas exclu non plus de pouvoir définir un tout autre système de types d'unités descriptives, en ayant adopté le même principe initial d'étapes qui distingue unités élémentaires, unités descriptives et unités caractérisantes.

2. Les unités initiales

a) Caractéristique : le lien direct avec les unités élémentaires

Les unités descriptives sont construites à partir des unités élémentaires, soit directement, soit indirectement en s'appuyant sur des unités descriptives existantes. Les trois types d'unités par lesquels commencent la présentation (les *Solidarités*, les *Assimilations* et les unités *Simples*) sont les seuls types, dans le modèle que nous avons forgé, à utiliser directement les unités élémentaires. Les unités descriptives de tous les autres types sont définies à partir d'autres unités descriptives, elles n'ont pas accès aux unités élémentaires.

En réajustant la discrétisation opérée par les unités élémentaires, les unités initiales (Solidarités, Assimilations, unités Simples) constituent les *premières* unités sémantiques. En tant que telles, leur structure interne peut n'être pas perçue comme sémantique, puisque cette structure ne se laisse pas penser comme des éléments sémantiques qui entrent en relation. Solidarités, Assimilations et unités Simples affichent donc une égale valeur sémantique. En revanche, la nature de la structure interne, qui correspond au type de l'unité, a un caractère opératoire pour la reconnaissance de l'unité à partir des unités élémentaires.

Les unités de ces trois types initiaux médiatisent la relation aux unités élémentaires. Elles déterminent la couverture d'un texte par un ensemble d'unités descriptives. Le rôle, complémentaire, des autres unités, sera de déployer des possibilités interprétatives à partir de ces descriptions de base.

b) *Les* Solidarités

Les Solidarités servent à représenter les locutions et les formes composées figées que l'analyse en unités élémentaires a éclatées. Elles reconstituent des unités apparemment en plusieurs

[101] « Si le principe de compositionnalité est invalide en sémantique linguistique et si l'interprétation échappe ainsi de droit au paradigme du calcul, la détermination du global sur le local s'exerce par les contraintes initiales imposées ou proposées au parcours interprétatif. Elles unifient l'analyse sémantique en prédéfinissant le type de pertinence attendu, et par là le type des traits sémantiques à sélectionner ou à construire. » (Rastier, Cavazza, Abeillé 1994, §II.5, p 37)

mots, mais qui fonctionnent en fait comme un bloc ; il arrive d'ailleurs souvent que ces expressions aient des équivalents sous forme d'un seul mot.

La Solidarité assure la représentation de la non compositionnalité du sens des constituants, elle représente des formes complexes pour lesquelles le sens de l'unité complète n'est pas (ou plus) motivé par le sens des constituants. Concrètement, quand une solidarité est définie, toutes les occurrences de la suite d'unités élémentaires à laquelle elle correspond sont reconnues comme des occurrences de la solidarité ; la reconnaissance d'une solidarité inhibe la reconnaissance d'unités correspondants à ses constituants.

Il faut se garder d'abuser des Solidarités. Elles représentent un cas réel de relation linguistique, qui a la particularité de poser des contraintes très fortes. Le *choix interprétatif* marqué par une Solidarité est celui de renoncer à (ou de s'interdire de) remotiver les constituants et défiger la locution[102]. D'autres enchaînements syntagmatiques significatifs existent, pour lesquels d'autres types d'unité descriptive sont prévus : ce qui ne serait pas représenté adéquatement par une Solidarité peut l'être par une Séquence (voir plus loin).

Voici des exemples et contre-exemples de Solidarités pour illustrer cette mise en garde. *a priori* est un bon candidat pour être une Solidarité : son constituant *priori* n'a d'ailleurs aucune autonomie[103]. Même appréciation pour *d'abord* et *d'ailleurs* : on peut trouver d'autres contextes d'occurrence de leurs constituants ; les définir comme Solidarités, c'est indiquer que pour les calculs on considère que sémantiquement *d'abord* n'a rien à voir avec *abord*, ni *d'ailleurs* avec *ailleurs*. Avec ce qu'on appelle communément les mots composés, l'enregistrement sous forme de Solidarité peut devenir plus litigieux, même si le tiret est un indice de figement : on ne voit pas *forme* dans *plate-forme*, mais choisira-t-on de lire *circuit* dans *court-circuit* ? Plus délicat encore, celui de dénominations composées soulignées par des initiales en majuscules (*Faits Marquants*, titre d'une publication DER), et dont parfois l'intégration semble confirmée par la saisie dans un sigle acronyme (*Langage Naturel* / *LN*, *Assurance Qualité* / *AQ*). Certains syntagmes[104] comme *cahier des charges* fonctionnent comme des Solidarités. Avec *World Wide Web*, il semble que l'on sorte du champ des Solidarités : en effet, on ne veut pas affirmer l'indépendance entre des occurrences du seul mot *Web* et celle de la forme complète en trois termes.

Tous ces exemples convergent pour rappeler qu'une Solidarité n'est pas une définition *linguistique* (*i.e.* au niveau général de la langue), mais une décision *interprétative* qui, comme telle, n'interdit pas que dans d'autres contextes une décision inverse soit prise.

Le dictionnaire reflète les différents degrés de figement que nous venons de parcourir : la forme composée a sa propre entrée, sous laquelle se trouve sa définition (*a priori*) ; la forme composée se trouve sous l'entrée d'un de ses composants, mais elle a son propre sens, non apparenté aux autres sens (*d'abord*) ; la forme composée se trouve sous l'entrée d'un composant et se présente

[102] Une typologie de mécanismes de défigements est exposée dans (Rastier 1997). Le défigement peut être motivé par des facteurs syntagmatiques (reprise proche d'un élément de la locution, présomption d'isotopie se propageant aux composants) ou par allusion paradigmatique (la locution transparaît derrière une occurrence qui la déforme : inversion, intercalation, dislocation ; substitution paradigmatique d'un constituant). Mais ces passages possibles décrits par la sémantique ne préjugent pas de leur validité interprétative. Il y faut certaines conditions herméneutiques (par exemple, le titre, conventionnellement lourd de sens, est une zone favorable au défigement), et le *genre* est un indicateur primordial, en ce qu'il détermine la position de l'énonciateur et son rapport au lecteur, et la prévisibilité de son propos (Rastier 1997, p. 325). Pour les premiers corpus envisagés pour DECID, les genres considérés (en particulier les textes d'Action) ne se prêtent guère au défigement. La possibilité de décrire, pour les Solidarités, des mécanismes de défigement, reste une extension intéressante pour une application du modèle des unités descriptives à d'autres corpus et dans d'autres contextes.

[103] Les expressions composées et locutions dont un constituant ne s'emploie dans aucun autre contexte (sauf jeu littéraire ou pratique très particulière) ne sont pas si rares : *a fortiori*, *d'emblée*, *d'ores et déjà*, *tandis que*, *parce que*, *belle lurette*, *au fur et à mesure*, *à la queue leu leu*...

[104] L'observation du corpus est déterminante pour la détection de syntagmes qui peuvent être interprétés comme des Solidarités. Par exemple, il est souvent question de la *durée de vie* d'un matériau, de l'étude de son *vieillissement*. On relève l'occurrence (maladroite ?) *vieillissement de vie*, qui témoigne que les constituants de *durée de vie* sont encore « sensibles » par delà l'expression, et qu'il serait dangereux de retenir *durée de vie* comme Solidarité, dans le cadre des descriptifs de l'activité de la DER d'EDF.

comme une acception ou un emploi dans un des sens du composant. Toutes ces lexicalisations ont une importance sémantique propre, qui justifie leur enregistrement (Martin 1994, §II.A.3, pp. 100-101).

Ces exemples esquissent déjà les méthodes qui peuvent être mises en œuvre pour repérer et définir des Solidarités :

- la déclaration *a priori*, en tant qu'unité distinguée, en se basant par exemple sur un dictionnaire de locutions.
- l'utilisation de scores distributionnels : par exemple, telle unité élémentaire qui a un nombre d'occurrences significatif dans le corpus, et qui est toujours précédée ou toujours suivie de telle autre unité élémentaire.
- l'utilisation d'indices morphologiques : suite de mots commençant avec une initiale majuscule, avec éventuellement des prépositions ou des articles définis intercalés. Ce critère constitue une présélection intéressante sur nos corpus techniques, et complètement inopérante en application sur un corpus de sciences humaines.

Ces deux dernières pistes gagnent évidemment à être utilisées conjointement et à se renforcer mutuellement. De plus, l'investigation ne doit pas se cantonner aux enchaînements de deux unités élémentaires ou aux groupes nominaux, comme le montre l'intérêt des approches de type segments répétés (Lafon, Salem 1983) (Salem 1984)[105].

c) *Les* Assimilations

Les Assimilations désignent les ensembles d'unités élémentaires que l'on choisit de ne pas distinguer pour la suite de l'analyse.

Ce type d'unité a pour vocation d'annuler les distinctions non significatives entre unités élémentaires. L'Assimilation convient pour regrouper les variantes insensibles à la lecture. A travers une Assimilation, on considère qu'il n'y a absolument aucune différence entre telle et telle unité élémentaire, et tout ce que l'on dit de l'occurrence de l'une peut être dit de l'occurrence de l'autre.

Là encore, comme pour les Solidarités, il faut se garder de sous-estimer la force de la relation exprimée : l'Assimilation pose davantage que l'*équivalence* des unités élémentaires qu'elle rassemble, elle en affirme l'*identité*. Dès que de légers distinguos s'introduisent, ce n'est plus à l'Assimilation qu'il faut recourir, mais peut-être plutôt à l'Association. [106]

Une Assimilation est le moyen d'effectuer les fusions entre unités élémentaires suspendues dans l'attente d'une validation globale. On bénéficie ainsi des effets de réduction comparables au passage en typographie pauvre (on ne considère plus les accents) ou à la réduction de la casse (on ne considère plus les majuscules), tout en évitant les confusions et quelquefois en apportant une information interprétative. Sécurité supplémentaire, si une Assimilation n'est pas justifiée, elle n'arrivera pas à s'intégrer dans des descriptions complexes : sa non-utilisation manifeste remet en cause la décision interprétative de regroupement, celui-ci peut être dissout, l'information de variation d'une unité élémentaire à l'autre est alors récupérée, réactivée (elle aurait été définitivement perdue par une réduction prématurée).

Les illustrations d'usage les plus simples sont donc, pour un découpage, l'assimilation des formes en début de phrase et en cours de phrase (*Le* et *le*), des variantes de transcription (*EDF*, *E.D.F*

---

[105] « les segments répétés peuvent représenter soit des locutions fonctionnant comme un seul mot (ex. : *mettre à l'ombre* pour *emprisonner*) soit des équivalents en plusieurs mots de formes fléchées (*de la ville* en français pour *urbis* en latin) soit les produits d'une rhétorique de la répétition (*faudrait-il leur rappeler que*... (bis) ; (ter)...). La distinction entre segment répété et mot unique étant souvent conventionnelle dans la mesure où la segmentation précise du discours en mots est par essence postérieure au langage. » (Salem 1983, p. 493)
Sur la nature et la composition des segments répétés, voir aussi (Fiala 1986), (Lessard & Hamm 1991).

[106] Dans la Sémantique Différentielle Unifiée de François Rastier, le terme d'*assimilation* est utilisé pour rendre compte d'opérations interprétatives, dans le registre d'une conception de la sémantique comme perception (c'est une « loi de la perception sémantique », (Rastier 1989, §I.1.C, p.20)). L'assimilation est le mouvement perceptif qui rapproche et unifie, et le mouvement inverse, qui différencie, contraste, oppose, et positionne en complémentarité, reçoit le nom de *dissimilation*.
Nous donnons ici à *assimilation* son sens le plus fort, qui sort vraisemblablement de l'acception précédente. Les unités élémentaires assimilées perdent définitivement leur identité propre au niveau descriptif.

et *E.D.F.*), d'une écriture en lettres capitales dans le titre avec l'écriture dans le texte courant (*PRODUCTION* et *production*). Les Assimilations sont aussi le moyen d'exploiter l'information d'une lemmatisation : regroupement des unités élémentaires trouvées comme les flexions d'un même lemme (*étape*, *étapes*) ; regroupement dans certain cas des formes d'une même catégorie, quand seule celle-ci importe (par exemple *auxiliaire*, sans faire la part entre *être* et *avoir*).

Dans ce cas précis de la lemmatisation, la réversibilité permise apporte ce juste milieu entre la cécité du système quant aux relations sensibles entre les différentes formes d'un même mot (*chat* n'ayant pas plus à voir avec *chats* qu'avec *contribue*), et l'écrasement brutal du signifié sur le signifiant, la continuité du signifiant à travers le système des flexions n'étant pas nécessairement doublée par celle du signifié (une seule unité lexicale *travail* / *travaux*, mais deux contextes d'usage disjoints : le *temps de travail*, les *travaux de réfection du bâtiment*).

Un exemple complémentaire concret d'unités représentables par des Assimilations sont les familles de mots, regroupant les variantes par flexion ou abréviation, utilisées pour la description des titres d'un corpus de commandes (transactions commerciales) (Bommier 1993). Après un découpage des titres, un balayage global opère automatiquement le regroupement d'unités comme PHOTOCOP, PHOTOCOPIES, PHOTOCOPIEURS, PHOTOCOPIEUSE, ou encore CLIMATISATION, CLIMATISAT, CLIM. La formation des classes est guidée par la connaissance des règles morphologiques de formation du pluriel en français, et par celle des mécanismes d'abréviation (essentiellement par troncature) à l'œuvre dans la rédaction des titres, pour le genre considéré. La chaîne de traitement mise en place est antérieure à la modélisation en unités élémentaires *vs* unités descriptives, mais elle s'inscrit déjà dans la même logique : analyse locale (découpage), relecture globale pour la formation d'une nouvelle génération d'unités (traitement des abréviations des titres), utilisation de ces nouvelles unités pour les calculs.

La construction automatique des Assimilations envisagée ici s'ancre dans les informations sur les unités élémentaires produites par l'analyse. Le travail supplémentaire réside dans le choix des informations à utiliser : il doit être guidé par l'évaluation de ce qui est une lecture utile pour la suite des traitements.

Les visées du traitement amènent aussi la définition d'unités distinguées : par exemple, si l'on fait l'hypothèse qu'il y a une classe de mots, les mots-outils, qui n'ont pas à entrer dans la description, il suffit de les assimiler en en donnant la liste *a priori*. Ensuite, toute occurrence d'une des unités élémentaires de la liste est perçue, sur le plan descriptif, comme « un mot-outil ».

d) *Les unités* Simples

Une unité simple est le correspondant descriptif d'une unité élémentaire.

L'unité Simple correspond au cas où l'unité élémentaire fournit déjà, telle quelle, une bonne unité descriptive. Elle permet donc d'enregistrer tout élément d'information, tel qu'il se présente au premier abord.

Quand les unités descriptives sont définies à partir d'un corpus, l'ensemble des unités descriptives obtenu donne par construction la couverture complète du corpus. Les unités Simples sont alors la traduction de toutes les unités élémentaires qui n'entrent pas dans la définition d'une Solidarité ou d'une Assimilation, mais pas seulement. De la même manière qu'il peut y avoir des constituants utilisés à la fois par des Assimilations et par des Solidarités, des unités Simples sont aussi définies pour représenter les occurrences des constituants d'une Solidarité, pour les contextes où celle-ci n'est pas réalisée (par exemple on pourrait avoir : *forme* et *plate-forme*, *éléments* et *éléments finis*, *centrale* et *unité centrale*).

Les unités Simples n'ont pas une définition purement négative, comme ce qui n'est ni Assimilation ni Solidarité. Avec les Assimilations et les Solidarités, elles cernent le champs descriptif d'un univers : quand un texte confronté à cet univers présente une nouvelle unité élémentaire, celle-ci reste stérile pour la description. Les unités Simples contribuent de façon majeure à la définition de l'ontologie sous-jacente à l'analyse, autrement dit aux éléments qui peuvent être perçus, et à la manière de découper le réel. Elles ont aussi un rôle positif en ce qu'elles apportent des unités de description minimales, n'imposant pas d'autre contraintes que leur présence pour pouvoir entrer dans

la représentation : elles ne sont pas sans rapport avec le caractère compulsif de l'interprétation, son caractère irrépressible –quel que soit le texte qu'on lui présente, le lecteur s'en fait toujours une idée.

3. Les unités paradigmatiques et syntagmatiques souples

a) Intérêt : relayer les Solidarités et les Assimilations pour des relations nuancées

Les unités décrites jusqu'à présent ont une réalisation entière, sur le mode présence *vs* absence : en un point donné du texte, soit l'unité est manifestée, soit elle ne l'est pas. Il n'y a pas d'appréciation qui évaluerait le degré de réalisation de l'unité par l'occurrence, il n'y a pas de variation de réalisation d'une occurrence à l'autre.

Ce mode de réalisation ne permet pas de rendre compte d'une grande partie des relations linguistiques, dont la manifestation est plus nuancée et plus libre. Les Solidarités ne traduisent qu'un cas limite des relations d'ordre syntagmatiques (*i.e.* concernant la succession des unités dans le texte) : de fait, elles reconstituent une unité plutôt qu'elles expriment une relation entre unités. De même, les Assimilations, dont la définition est paradigmatique (*i.e.* elle concerne les familles d'unités vues comme autant d'alternatives équivalentes pour s'insérer en un point du texte), corrigent l'analyse locale en unités élémentaire sans réellement déployer une gamme d'unités en interrelation.

Les deux types d'unités qu'il nous manque vont donc poursuivre la description selon les axes syntagmatiques et paradigmatiques : ce sont respectivement les Séquences et les Associations.

b) Les Séquences

Les Séquences traduisent des formations syntagmatiques à la fois stables, et alors porteuses d'une sémantique précise particulière, et souples, c'est-à-dire admettant une certaine marge de variation dans leur réalisation.

Les Séquences se situent au niveau des syntagmes. On sait que le syntagme est une zone d'associations fortes sur le plan sémantique (l'actualisation et la propagation de sèmes sont facilités). Les Séquences recouvrent ces termes, ces formules, que le lecteur identifie dans leur unité, alors qu'elles n'ont pas toujours la même apparence[107]. Une partie peut être effacée, dans une reprise elliptique : *Le World Wide Web... Le Web...*. Des variantes modifient une préposition, font disparaître un déterminant, insèrent une qualification (*diffusion ciblée d'informations*, *diffusion électronique ciblée de l'information*). Des constructions dissocient les unités constituantes, comme la coordination qui peut ouvrir le paradigme d'un seul des composant (*à court et moyen terme*). Une observation et une description systématique des variations de termes dans un corpus médical (Jacquemin, Royauté 1994) montre la diversité et la régularité des formes que l'on souhaite reconnaître par une même unité. Ce genre de variations est relevé comme une des principales caractéristiques linguistiques qu'il faut pouvoir prendre en compte dans l'analyse de documentations techniques (Assadi 1998, §I.5.2, p. 67).

La frontière qui sépare Solidarités et Séquences reprend les tests forgés par les linguistes pour décider de ce qui constitue une unité lexicale, et faire la part par exemple entre un nom composé (*pomme de terre*, *chemin de fer*) et une dénomination complexe analysable.

En revanche, les Solidarités et les Séquences ont ceci en commun : ce sont les deux seuls types d'unités descriptives proposés ici qui intègrent une *relation orientée*[108]. Si les unités

[107] Unifier ces diverses manifestations de la même unité permet par exemple d'en mettre en évidence la forte concentration locale ; (Paice 1990) y voit alors l'indice d'une thématique significative, à consigner dans un index.

[108] (Lafon 1981b) fait une recherche de paires et de couples de mots statistiquement significativement cooccurrents (dans une même phrase ou dans deux phrases successives). La notion de couple lui sert à distinguer des associations orientées, c'est-à-dire ici où l'ordre des cooccurrents importe. Lafon calcule également, à titre indicatif, pour chaque association de deux mots, la distance moyenne entre ces deux mots.
Sur ses résultats, on peut observer qu'une bonne proportion des *couples* sélectionnés reflètent des associations à l'échelle du syntagme (ou de la lexie) plutôt que de la phrase. On y reconnaît des groupements nom - adjectif

élémentaires sont issues d'un découpage, l'information d'ordre apportée par l'analyse est celle de la succession linéaire. Les relations de dépendance fournie par une analyse syntaxique locale sont également une information orientée, davantage fondée linguistiquement[109]. Les Séquences s'intéressent à ces interactions de proximité, fortes, et non purement et entièrement décrites comme des relations syntaxiques : la retranscription d'une prise de parole rend sensible la désarticulation syntaxique des relations définissables au niveau de la phrase, mais le maintien de régularités au niveau local des expressions, locutions, groupes de mots.

La construction automatique de Séquences peut s'inspirer des travaux sur la sélection de cooccurents proches, dans une relation orientée[110], des travaux sur l'extraction de termes complexes (Sta 1997), et des jeux de règles d'appariement de variantes expérimentés et pratiqués (Jacquemin, Royauté 1994). Elle s'outille du côté des extracteurs de terminologie.

Les très courtes distances des relations syntagmatiques amènent à leur limites les mesures de longueur du type *fenêtre* (une fenêtre centrée sur un mot donné sélectionne les *n* mots qui précèdent et les *n* mots qui suivent le mot considéré). En effet, la fenêtre devrait être assez courte pour ne pas déborder des syntagmes simples, mais aussi assez longue pour tolérer les variantes qui multiplient les mots intercalés. D'où la mise au point de fenêtres qui ne font pas entrer les mots-outils dans le décompte de la longueur (Haas, Losee 1994). Cela peut s'élargir à d'autres catégories ou structures traduisant des compléments, syntaxiquement facultatifs : on fait ainsi abstraction de rattachements syntaxiques qui ne font que se greffer sur la structure étudiée, afin de faire ressortir les liens principaux. Les ponctuations, qui marquent des ruptures dans le déroulement du texte, font également office de limites et modulent l'étendue des fenêtres. Certains usages des virgules, en particulier dans les incises et les coordinations, constituent peut-être un cas particulier, réalisant une intercalation plutôt qu'une rupture.

La pertinence de combiner les critères linguistiques et statistiques a été maintes fois soulignée, dans le cadre de la construction de terminologie assistée par ordinateur. Les travaux dans ce domaine sont nombreux et très riches : approfondissement de la réflexion linguistique sur le statut des termes, réalisations opérationnelles avancées, etc. Ceci ne doit pas occulter pour nous l'intérêt des formes non centrées sur le nom : elles entrent dans la construction d'unités au même titre que les autres. Les travaux autour des segments répétés montrent la part dominante, mais pas exclusive, des formes nominales dans les affinités syntagmatiques (Lebart, Salem 1988).

Concernant l'homologation des variantes, Richard Quatrain s'est peu à peu forgé un ensemble de règles, qu'il a d'abord appliqué au dédoublonnage de descripteurs issus de référentiels terminologiques différents. Les règles repèrent les variations de nombre (singulier / pluriel), les variantes jouant sur les mots grammaticaux (cela revient à les cacher), les variantes sur les

---

qualificatif, nom - complément de nom, déterminant - nom, préposition - nom, quelques expressions coordonnées « figées » dans le corpus, des locutions.

Quand la distance moyenne entre deux mots cooccurrant *de façon orientée* est importante (dix mots ou plus), le lien statistique trouvé entre les mots a souvent du mal à trouver une interprétation claire, surtout si le couple comporte un mot grammatical (au moins).

Inversement, les paires sélectionnées pour lesquelles la distance moyenne entre les deux mots est faible (mettons inférieure à cinq mots) correspondent quasiment toutes à *un* couple sélectionné, dont à un lien orienté.

Tout ceci conforte le choix de n'enregistrer une information d'orientation qu'à l'échelle du syntagme ou de la lexie, sans aller jusqu'à la phrase ou à la période (cf. plus loin, les unités descriptives appelées Associations).

[109] L'utilisation de l'ordre linéaire vaut comme reflet de familles de constructions syntagmatiques stables. Cette simplification qui peut dans la plupart des cas convenir pour le français, serait inadaptée pour l'anglais : « Pour une langue comme l'anglais où le phénomène d'inversion est fréquent, la prise en compte de l'ordre des mots distinguera les expressions *A of B* et *B A* alors qu'elles ne sont que des variations d'une même unité terminologique. » (Sta 1997, 6.3.1.3)

Mais aussi, l'information syntaxique n'élude pas toute considération d'enchaînement linéaire. Il arrive en français que l'ordre linéaire ait un rôle sémantique décisif, pour un rattachement syntaxique analogue, en particulier pour des constructions d'adjectif (antéposé, postposé), comme *ancien*, *brave*, *propre* ou *seul* (Martin 1994).

[110] Par exemple (Chartron 1988, §V, p. 58 sq.), mais qui reconnaît que sa mesure statistique (un 'coefficient d'implication réciproque') laisse échapper les variantes par ellipse, coordination, insertion (adjectif qualificatif ou complément de nom), anaphore.

terminaisons, les inversions dans l'ordre des mots. Conçues à l'origine pour mettre en relation des descripteurs (termes d'indexation), Richard Quatrain observe que ces règles s'extrapolent, mais sont alors peut-être un peu trop tolérantes, pour des termes relevés dans les textes.

c) Les Associations

Les Associations regroupent des expressions différentes d'un même élément sémantique.

Les Associations correspondent au phénomène linguistique de synonymie, voire de parasynonymie (synonymie lâche). A la différence des Assimilations, les Associations permettent de reconnaître la richesse de vocabulaire pour cerner un même objet sémantique. Tout en reconnaissant qu'il peut y avoir des nuances d'une expression à l'autre, on pose qu'il peut être légitime de suspendre ces écarts, dans le but de mieux faire ressortir le point de convergence des unités ainsi rassemblées.

Un peu plus largement, les Associations peuvent s'étendre aux associations d'idées les plus immédiates, à savoir les unités qui se font écho. Des expérimentations de psycho-linguistique montrent que la relation d'antonymie est plus prégnante que celle de synonymie : le premier mot qui vient à l'esprit en répondant à un qualificatif sera généralement le qualificatif contraire. Dans la mesure où des antonymes indiquent des pôles remarquables sur un *même axe* sémantique, leur regroupement en Association peut se justifier[111]. Cela revient à saisir, dans une même unité descriptive, la gamme de valeurs ou d'appréciations disponibles dans certains contextes[112].

Dans les termes de la Sémantique Descriptive Unifiée (cf. les travaux de Rastier), les Associations se situent au niveau des taxèmes. Un taxème est une classe ensembliste (*i.e.* tous ses

---

[111] Une étude sur corpus du phénomène d'antonymie (Justeson & Katz 1991) confirme qu'il y a une association forte entre adjectifs antonymes. Il apparaît que cette association n'est pas purement paradigmatique (elle se traduit par de fortes cooccurrences, mesurées au plan syntagmatique). Ce résultat n'est pas retrouvé par Etienne BRUNET, qui, dans l'étude des expressions de sentiments sur un corpus de romans, note : « La paire *vice-vertu* est l'un des seuls couples antithétiques que l'analyse [des cooccurrences dans les phrases] ne désunisse pas » (Brunet 1995, note 7, p. 51).

Selon (Justeson & Katz 1991), d'autres études auraient montré que l'ordre d'apparition des deux adjectifs antonymiques, lorsqu'ils cooccurrent, n'est pas fortuit : l'évaluation positive précéderait l'évaluation négative, le terme plutôt neutre serait avant le terme plus marqué, et le terme le plus fréquent avant le terme le moins fréquent. Notre modèle des Associations ne rend pas compte de cette asymétrie : une étude plus approfondie pourrait donc conduire soit à affiner la structure des Associations, soit à enrichir la liste des types d'unités descriptives.

[112] A rapprocher de nos Associations, les *rapports associatifs* à l'intérieur des *champs contextuels*, dans la présentation des résultats d'une analyse réalisée avec l'outil ALCESTE. Max REINERT en souligne bien le caractère relatif :

« Nous nous sommes servis de la distinction faite par F. de Saussure (1972), entre les *rapports syntagmatiques* et les *rapports associatifs* (nous préférons ce terme à « paradigmatique », dont le sens est plus formel). Dans la chaîne du discours, les éléments en rapport syntagmatique se coordonnent dans un même énoncé, alors que les éléments en rapport associatif sont susceptibles de se substituer les uns aux autres. Cette distinction n'est pas spécifique de la langue et peut être utilisée pour appréhender n'importe quelle représentation. F. de Saussure utilise d'ailleurs l'image suggestive suivante pour faire comprendre son propos : *A ce double point de vue, une unité linguistique est comparable à une partie déterminée d'un édifice, une colonne par exemple ; celle-ci se trouve, d'une part, dans un certain rapport avec l'architrave qu'elle supporte ; cet agencement de deux unités également présentes dans l'espace fait penser au rapport syntagmatique ; d'autre part, si cette colonne est d'ordre dorique, elle évoque la comparaison mentale avec les autres ordres (ionique, corinthien, etc.), qui sont des éléments non présents dans l'espace : le rapport est associatif.*

[...] La procédure utilisée pour l'approche d'un champ contextuel consiste, dans un premier temps, à regrouper le vocabulaire dans des classes associatives. Cette procédure rappelle celle utilisée par J. Dubois (Mounin 1975, p.69), dans son étude sur *le vocabulaire politique et social en France de 1869 à 1872*, quand il réunit, par exemple dans une même classe, les termes *ouvriers*, *travailleurs*, *salariés*, *pauvres*, *déshérités*.

Si cette procédure s'apparente nettement avec la manière dont on construit un *champ lexical*, elle en diffère par le fait essentiel que ces associations n'ont pas de sens « absolu » mais relativement à un champ contextuel particulier (dépendant d'une classe d'énoncés particulière à l'intérieur d'un corpus précis) dont le vocabulaire est fixé préalablement à l'aide d'une analyse statistique. » (Reinert 1990, §2.6)

éléments ont le même statut), représentant un paradigme minimal (Rastier 1987, p.50)[113]. Les éléments du taxème sont réunis par le sème microgénérique qu'ils ont en commun : en ce sens, l'unité que forme le taxème a bien une consistance sémantique.

Selon la perspective de l'application, les écarts que l'on choisit de négliger sont de nature très diverse. Les différences de signifiant, et de phonétique, qui sépareraient deux manières équivalentes de désigner la même chose dans un texte scientifique, ne sont plus toujours assimilables dans l'étude d'un texte poétique, pour lequel les sonorités font sens. Les écarts de registre de langue pourraient être ignorés dans le cadre d'une opération de représentation des connaissances, se basant sur des enregistrements en situation de travail et sur des manuels de référence ; en revanche, une étude sociologique pourrait au contraire y relever des oppositions fondamentales pour sa description. Les glissements de sens se voient aussi assigner des marges de tolérance.

Le point d'appui des procédures automatiques est principalement le test de commutation : il s'agit de faire apparaître les familles d'unités attestées dans un ensemble de contextes locaux[114].

Si l'on se focalise sur une unité et que l'on cherche les unités avec lesquelles elle pourrait s'associer, c'est le résultat d'un calcul en deux temps (ou *2n* temps pour la version itérative) : la première phase détermine les cooccurrents significatifs de l'unité étudiée ; la deuxième phase explore les cooccurrents de ces cooccurrents (Schütze, Pedersen 1993). La première phase s'en tient aux unités en relation *in praesentia*, la seconde ouvre l'espace des relations *in absentia*. Par construction (principe de commutation), il y a une affinité certaine entre les constituants d'un paradigme et les unités en relation *in absentia*[115].

Très proches de nous, à la DER, les travaux sur la constitution de classes d'adjectifs (Assadi, Bourigault, Gros 1995) (Assadi 1998) donnent une méthodologie et des outils (*Lexiclass*) pour la génération d'Associations. Dans le contexte du *Guide de planification des réseaux électriques*, on obtient notamment :

C_1 = {FAIBLE, FORT} - contexte :
HYDRAULICITE, OCCURRENCE, PUISSANCE, SECTION

C_3 = {EQUIVALENT, MONOPHASE, TRIPHASE} - contexte :
APPAREIL, CHARGE, COURANT, COURT-CIRCUIT, DEFAUT, IMPEDANCE, LIAISON, RECEPTEUR, RESEAU, SCHEMA, SOUDEUR, STRATEGIE, SYSTEME, TRANSFORMATEUR

---

[113] Les classes sémantiques d'ordre supérieur, les *domaines*, ont une structure méréologique (*i.e.* partie / tout), et donc ne donnent pas aux taxèmes qui les composent un statut équivalent, mais une organisation en complémentarité. Le taxème (ensembliste) a des affinités avec le OU booléen, le domaine (méréologique) serait plutôt du côté du ET. Les unités descriptives que l'on pourra mettre en relation avec les domaines sont les Communautés, présentées un peu plus loin.

[114] Et d'ailleurs, c'est là le seul véritable point de départ d'une linguistique descriptive : « Les paradigmes ne sont pas des classes déjà données. La plupart ne sont pas des classes finies ; seul un petit nombre d'entre eux –qui regroupent des grammèmes– relèvent du système fonctionnel de la langue.
La notion de projection oblitère quelque peu le fait que l'axe paradigmatique ne peut en aucune de ses parties être projeté tel quel sur l'axe syntagmatique. Les grandeurs situées sur ces deux axes n'ont pas le même statut épistémologique : une chaîne syntagmatique est pour le linguiste un objet empirique, alors qu'un paradigme relève d'un modèle descriptif. En bonne méthode, ce sont les relations syntagmatiques d'équivalence entre sémèmes qui permettent d'identifier les relations associatives sur l'axe paradigmatique, et non l'inverse. » (Rastier 1987, p.96)

[115] Affinité certaine n'est pas nécessité (nous y reviendrons à la fin de cette partie sur les Associations) :
« Des lectures trop cursives ou trop orthodoxes de Saussure ont accrédité l'idée que la langue serait le site des relations paradigmatiques ; et le discours, celui des relations syntagmatiques. La réalité paraît plus complexe.
a) La représentation du sémème en langue inclut des valences syntagmatiques. Et l'on peut même dire que chacun de ses composants détermine de telles valences.
b) Par ailleurs, dans le texte même, le sémème continue d'entretenir des relations paradigmatiques, puisque ses sèmes inhérents sont définis relativement à une classe de sémèmes dont les autres membres ne sont pas ordinairement présents en contexte.
c) Enfin un texte peut manifester des paradigmes pour ainsi dire *in praesentia* : par exemple une énumération peut présenter dans le même syntagme les membres d'une même classe sémantique. Dans ce cas, la classe contextuelle et la classe systématique contiennent les mêmes membres, et définissent, mise à part la relation d'ordre, les mêmes relations entre ces membres. » (Rastier 1987, p. 77)

C_6 = {ADMISSIBLE, MAXIMAL, MAXIMUM, NOMINAL, SUPERIEUR} - contexte :
CHARGE, COURANT, COURANT DE COURT-CIRCUIT, INTENSITE, LONGUEUR, NIVEAU, NOMBRE, PUISSANCE, TEMPERATURE, TEMPS, TENSION, TRANSIT, VALEUR
(Assadi, Bourigault, Gros 1995)

L'accent est mis sur la *qualité* des groupes d'adjectifs obtenus, en s'en tenant à contextes linguistiquement bien définis, à savoir les candidats termes obtenus par LEXTER, avec leur structuration interne en Tête / Expansion. Sur le plan *quantitatif*, les critères extrêmement sélectifs pour retenir le sous-ensemble des contextes de travail, sur lesquels les traitements statistiques à appliquer sont valides, limitent drastiquement la proportion des adjectifs étudiés.

Bien sûr, rien n'empêche un paradigme de se manifester aussi *in praesentia*, par exemple en se déroulant sous forme d'une énumération[116]. A ce titre, le repérage des marqueurs typographiques et linguistiques des énumérations (ponctuations faibles et semi-forte, mais aussi formules introductives, construction de coordinations longues, listes de mentions, *etc.*) trouverait ici tout à fait sa place.

4. Les *Communautés*

a) Le dessous des isotopies

L'*isotopie* est en quelque sorte un accord sémantique. L'écolier apprend que l'adjectif épithète s'accorde en genre et en nombre avec le nom qu'il qualifie ; cet accord morpho-syntaxique se double d'un ajustement sémantique, sous peine d'une suite de mots désarticulée. C'est le facteur d'homogénéité, de cohérence sémantique d'une phrase, d'un paragraphe, d'un texte : c'est donc un élément-clé de la description textuelle[117].

---

[116] « Les énumérations linéarisent souvent des taxèmes. » (Rastier, Cavazza, Abeillé 1994, p. 62), et favorisent la propagation des traits sémantiques (Cavazza 1996, p. 60) ; si bien que l'énumération est un mode d'expression (*résultat*) mais aussi de constitution (*source*) d'un lien sémantique : « le parcours interprétatif des énumérations procède par assimilation, et conduit à la construction d'isotopies génériques locales, même en présence de coq-à-l'âne apparents. » (Rastier 1987, p. 79)
En pratique, (Maingueneau 1991, p. 34 sq) s'appuie sur les relations manifestées dans le texte sous forme de chaînes parasynonymiques, pour faire apparaître deux classes d'adjectifs qui structurent le discours.
En revanche, Barakat-Barbieri juge les énumérations parasites, lors du calcul de relations sémantiques entre les mots par le biais de cooccurrences. Malheureusement, il ne cite aucun exemple, car il est difficile de comprendre son rejet si farouche :
« Un phénomène se produit régulièrement lorsque dans le texte des documents apparaît une structure de liste, c'est-à-dire une énumération de concepts liés à un concept principal mais sans que ces derniers aient de liens entre eux. Nos méthodes reposant sur la notion de cooccurrence, lorsque ces structures de liste sont répétitives (dans un nombre de cas non négligeable) cela fausse la statistique et donne naissance à des anomalies dans le graphe. » (Barakat-Barbieri 1992, p. 104)
Il nous semble au contraire qu'une énumération souligne –si elle n'induit, interprétativement et dynamiquement,– une relation sémantique entre les termes (cf. les remarques introductives de cette note). Le résultat visé, de construction d'un graphe de concepts, impose sans doute d'autres contraintes, qui neutralisent certains effets interprétatifs. Par exemple, des notions qui sont rassemblées par une pratique (mettons : *papier*, *crayon*) ne se laissent pas nécessairement décrire par des rapports canoniques de *généricité / spécificité* ou *partie / tout*.
Il reste qu'il est difficile de cerner ce qu'entend Barakat-Barbieri par énumération (« rien n'empêche ces énumérations de s'étaler sur plusieurs phrases » (Barakat-Barbieri 1992, p. 104)), et que son corpus ne nous est connu que par simple mention :
« - 3 500 documents concernant la gestion du personnel d'une grande entreprise française. C'est le plus important corpus que nous possédions.
- 1 000 documents CEE concernant les arrêts de la Cour de Justice des Communautés Européennes.
- 300 documents AFP qui sont des dépêches d'actualité.
- 1 245 documents de revendications dans le domaine de l'électricité et du nucléaire de l'Office Européen des Brevets. » (Barakat-Barbieri 1992, p. 52).

[117] La réalité de cet accord sémantique n'est que confirmée par la possibilité d'un jeu comme S + 7 ou le *cadavre exquis* des oulipiens.
Le cadavre exquis : on se donne un schéma de phrase, par exemple *nom précédé d'un déterminant - adjectif - verbe (transitif) - déterminant nom - adjectif*. Chaque joueur choisit un mot pour une des places prévues, sans

Une isotopie se traduit par la réitération d'un élément sémantique, que l'on appelle *sème*. Les unités qui composent le texte partagent des éléments sémantiques (des sèmes) : par exemple, dans le paragraphe précédent, les mots *sémantique*, *morpho-syntaxique* appartiennent au vocabulaire de la linguistique ; cette indication sémantique (« il s'agit de linguistique ») apporte une coloration aux termes *genre*, *nombre*, *cohérence*, etc. qui à leur tour viennent renforcer et développer le sujet. On note alors la récurrence (le retour) d'un sème que l'on peut désigner par l'étiquette */linguistique/*.

Pour mémoire, les unités sémantiques ne se limitent pas à des unités thématiques ; elles comprennent aussi des évaluations, comme */technicité, expertise/* que l'on peut vouloir attribuer à *isotopie* et *morpho-syntaxique*. C'est le début d'une deuxième isotopie, entrelacée avec la première.

La description sémantique à visée d'automatisation part traditionnellement d'une analyse des unités du texte en sèmes (éléments sémantiques), puis suit l'apparition répétée d'un sème pour tracer l'isotopie correspondante. Une isotopie, pressentie à partir d'un sème *inhérent* à certaines unités, est consolidée par les possibilités d'attribuer ce sème à d'autres unités, à la description sémantique desquelles il s'intègre (on parle alors d'*afférence*, par opposition à *inhérence*). En revanche, un sème isolé et incompatible avec une des isotopies est *inhibé* : il passe inaperçu, ne rentre pas dans la construction de l'interprétation du texte.

Cette présentation très sommaire[118] laisse déjà percevoir deux difficultés majeures : *(i)* disposer d'une analyse en sèmes complète et pertinente de toutes les unités du texte ; *(ii)* rendre compte de l'interaction entre les sèmes qui détermine dynamiquement l'interprétation : effacement de certains éléments, extension d'autres. C'est surtout (ou déjà ?) la première qui est redoutable pour une application comme la diffusion ciblée : le champ des informations potentiellement intéressantes pour générer un rapprochement est vaste, ouvert, et en constante évolution. On ne peut prévoir, et on ne veut pas non plus fixer, les éléments sémantiques pertinents *a priori*.

Au lieu de partir du lexique, retournons le problème et partons du texte[119]. Nous cherchons une caractérisation sémantique du texte. Où trouver des éléments sémantiques ? Chaque isotopie est porteuse d'un élément sémantique, celui qui se répète, celui qu'ont en commun les occurrences qui se raccordent à l'isotopie. Cet élément sémantique, au cœur de l'isotopie considérée, peut connaître plusieurs occurrences, c'est-à-dire se manifester à différents endroits du corpus. A ces endroits, l'isotopie se forme également à partir d'unités du texte. Par définition de l'isotopie, ces unités font toujours partie d'une même ensemble, celui des unités du texte qui comportent l'élément sémantique au cœur de l'isotopie considérée.

Retourner l'approche sémantique des textes, c'est donc remplacer un dictionnaire-lexique, qui aux unités du texte associe une liste de sèmes, par un univers d'isotopies, définies par les unités du texte susceptibles d'y contribuer. Les nouvelles unités descriptives, complétant les précédentes et construites à partir d'elles, représentent donc chacune le sème au cœur d'une isotopie. Le sème est le foyer sémantique vers lequel convergent les unités constituantes, c'est leur point d'intersection, leur point commun. D'où le nom générique de ces nouvelle unités descriptives : *Communautés*, pour rappeler que leur essence est le point commun partagé par leurs constituants.

La structure des Communautés, modelée sur la description des isotopies[120], est indifférente à un ordre des composants ou à un ordre de leur réalisation au fil du texte. La relation sous-jacente est

---

connaître les choix des autres pour les autres places. Le jeu devrait son nom à l'un des premiers enchaînements pour le moins inhabituels qui se révèlent quand les choix sont dévoilés : *Le cadavre exquis a bu le vin nouveau*.
S + 7 : on se donne une phrase de départ « ordinaire ». Chacun de ses mots est remplacé par le septième mot de la même catégorie qui vient après lui dans le dictionnaire. Là de même, on maintient la structure syntaxique et on joue sur l'accord sémantique.

[118] L'ouvrage de référence à consulter pour en savoir plus est (Rastier 1987).

[119] « [Une conception textuelle de l'isotopie] conduit à un déplacement de problématique. En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire les sèmes. » (Rastier 1987, Introduction §II.A, pp. 11-12)

[120] Un approfondissement théorique trouverait sans doute derrière nos Communautés un mélange d'isotopies et de *paratopies* :

l'identité du sème commun, c'est donc une relation symétrique, commutative (Rastier 1987, §IV.2.5.A). Il est donc valide de se représenter les composants comme un ensemble, une classe.

Pour reprendre l'exemple précédent, on remplace le recours à un dictionnaire sémantique de la forme :

sémantique = /linguistique/ + ...
morphosyntaxique = /linguistique/ + ...
nombre = ... (+ /linguistique/)...

par des unités descriptives :

$n^{\text{ième}}$ unité descriptive = sémantique ∩ morphosyntaxique ∩... (∩ nombre) (∩...)

Si l'on fait de cette $n^{\text{ième}}$ unité descriptive une unité distinguée, on peut lui donner le nom *linguistique* par exemple.

Un doute reste à chasser : n'a-t-on pas simplement substitué un catalogue (isotopies définies par des unités du texte) à un autre (unités du texte définies par des sèmes), sans rien résoudre des problèmes de complétude et de fermeture a priori de ce genre d'inventaires ? Tout se joue dans le retournement opéré par la seconde approche : le renversement de la perspective lexicale à la perspective textuelle, le primat des types prédéfinis cédant la place au primat des occurrences en contexte. Forgées par la matière même de leurs occurrences, les Communautés se construisent directement à partir des textes. Elles court-circuitent le passage par un référentiel méta-linguistique, le répertoire des sèmes, élaboration abstraite détachée des textes. Cela ouvre la perspective de trouver des méthodes pour construire les Communautés à partir des corpus, et ainsi recueillir dynamiquement des éléments sémantiques des textes considérés.

Concrètement, une recherche automatique de Communautés adopte le raisonnement suivant : si l'on repère des unités du texte qui apparaissent régulièrement ensemble en divers endroits du corpus, on tient peut-être les différentes occurrences d'une même isotopie, et donc l'élément sémantique au cœur de l'isotopie, et donc une (nouvelle) unité descriptive, synthétique, signifiante, et reflétant le corpus.

Même peaufinée, cette procédure n'assure pas, à l'évidence, le repérage de toutes les isotopies qu'un lecteur pratiquant la sémantique interprétative voudrait relever. Notamment, une isotopie qui ne se manifesterait qu'une seule fois dans le corpus ne pourrait être mise en évidence. En revanche, on peut espérer obtenir les isotopies parmi les plus saillantes, et parmi les plus actives (au sens où elles fonctionnent déjà, par leur répétition, comme des points de rapprochement)[121]. Pour le reste, les autres unités descriptives restent présentes pour compléter la description.

« L'actualisation d'un trait favorise sa réitération. En ce cas, et selon le statut de ce trait, cela constitue une isotopie générique ou spécifique. La production des antonymes, massivement attestée par les associationnistes du siècle dernier, est un exemple d'activation au sein d'un même taxème, par la constitution d'une isotopie générique minimale.

L'actualisation d'un trait favorise aussi la réitération des traits voisins dans la même molécule sémique : c'est pourquoi des lexicalisations partielles d'un même thème sont fréquemment cooccurrentes dans la même période, voire dans le même syntagme. Ce phénomène pourrait être appelé *paratopie*. Il est à l'œuvre dans ce que l'on nomme les anaphores associatives [...]. Ces diffusions d'activation sont le corrélat sémantique des phénomènes que la *Gestalt* nommait *lois de bonne continuité*, et que la psychologie cognitive étudie sous le nom général d'amorçage (*priming*). Elles justifient sémantiquement l'étude statistique des cooccurrences lexicales pour l'analyse thématique. [...]

L'hypothèse qui fonde la transformation de la cooccurrence [statistique] en corrélation [sémantique] est celle-ci : le contexte proche est structuré par des isotopies qui marquent l'appartenance à un même fond sémantique, ou des paratopies –qui marquent l'appartenance à la même forme sémantique. » (Rastier 1995a, §I.b & II.3.b, pp. 229 & 241)

[121] La description est opportune. A quoi bon construire des regroupements complexes qui n'auraient pas une pertinence et une utilité descriptive, dans notre cas concret ou dans plus largement le cadre de l'application ? Ludovic Tanguy ne nous démentirait pas :

« En nous risquant à une formule trop facile, un taxème n'est qu'une isotopie potentielle. Dès lors que nous nous intéressons plus au discours qu'à la langue, nos taxèmes ne seront utiles que si leur extension est réellement, ou du moins en partie, présente dans le texte analysé. » (Tanguy 1997, §III.4.3.2, p. 81)

b) Des propriétés qui sont autant de nouvelles exigences et de nouvelles libertés

L'utilisation de classes de termes pour la caractérisation de textes n'est pas absolument neuve. Les unités conceptuelles, par exemple celle du moteur de recherche Excite sur Internet, semblent aussi se baser sur des regroupements de termes. Il faut s'en tenir ici à des conjectures : les diverses unités conceptuelles attisent d'autant plus la curiosité que leur constitution s'entourent d'un halo mystérieux, un précieux secret de fabrication. Pour autant, rien ne laisse présager que l'on n'applique, avec un certain succès, les techniques éprouvées : classification automatique, classement au moyen d'un indice comme l'information mutuelle.

Concevoir les groupements de termes sous l'angle des isotopies précise un certain nombre de propriétés pour ces groupements de termes. L'explicitation de ces propriétés est capitale, car elle oblige à réviser les approches par les algorithmes classiques de classification et de classement.

- *Interdéfinition globale interne* des éléments d'une classe : les unités du texte regroupées comme participant de la formation d'une même isotopie sont embrassées simultanément, dans une même saisie.

  Il n'y a pas une unité fondatrice, qui catalyse le rattachement des autres (*irréductibilité à un pôle originel*). L'indifférence par rapport à l'existence éventuelle d'une unité centrale, parangon, est justement le ressort pour décaler le plan de l'expression par rapport au plan du contenu. Or ce degré de liberté est crucial : la linguistique considère « la non-conformité [des plans du contenu et de l'expression] comme [...] [un] trait fondamental de la structure de base du langage » (Hjelmslev 1968, p.230).

  L'ensemble n'admet pas non plus une décomposition standard en sous-parties autonomes, notamment en paires : rien n'oblige les relations qui réunissent les unités à avoir une arité de deux, certaines relations ne prennent consistance qu'activées par trois ou quatre unités (*irréductibilité à des paires*)[122].

  Autrement dit, nos Communautés sont démocratiques, chaque unité ayant voix au chapitre, sans préséances, privilèges ou morcellement. Tout au plus distingue-t-on deux statuts : les unités pour lesquelles le sème est plutôt inhérent, celles pour lesquelles il est plutôt afférent. Les premières ont le pouvoir (exécutif), mais non sans l'appui des secondes. La Communauté comporte donc un *noyau*, composé des unités qui signalent l'isotopie ; les autres unités, appelées *satellites*, viennent confirmer et renforcer l'effet.
- *Possibilité de multi-appartenance* : une unité du texte présente généralement plusieurs sèmes (et même plusieurs sèmes inhérents). Elle peut donc tout à fait contribuer à plusieurs isotopies différentes. Ceci ne rend pas satisfaisante une représentation des isotopies par des classes d'unités *disjointes* : il est plus juste de travailler à partir d'ensembles, ayant potentiellement des *recouvrements*.

  De plus, une unité ne doit pas être dévalorisée, voire *neutralisée*, parce qu'elle participe à plusieurs Communautés : il faudrait éviter le biais qu'apportent alors par exemple les analyses factorielles, qui font ressortir les unités discriminantes et univoques, et reversent les autres au centre, dans un magma confus.[123]

  La multi-appartenance possible d'une unité traduit aussi le fait que les isotopies ne sont pas des classes d'équivalences sur les unités, au sens mathématique du terme. En effet, la propriété de transitivité ne se vérifie pas en linguistique : si une unité A et une unité B participent d'une même isotopie, et que l'unité B est également indexée sur une isotopie commune avec l'unité C, rien

[122] La notation fléchée des réseaux en tous genres contribue sans doute à l'oubli qu'il puisse exister des relations plus que binaires.

[123] Dans le cadre de la recherche documentaire ou de la construction de typologies, domaines où se concentrent la plupart des travaux sur l'analyse automatique de données textuelles, il est coutume d'écarter et de faire peu de cas de ce qui n'est pas discriminant, ces unités étant considérées comme inefficaces voire perturbatrices.
« toute classification typologique implique que l'on dispose d'indicateurs *discriminatifs*, si possible spécifiques de chaque type. Nous avons été ainsi amenés à écarter des morphèmes ou lexèmes *plurifonctionnels*, qui jouent des rôles divers dans la phrase ou dans le texte. » (Bronckart & *al*. 1985, §V.A.3, p. 69)

n'oblige que l'unité A et l'unité C appartiennent à la même isotopie, qui plus est que toutes ces isotopies soient une unique isotopie regroupant A, B, et C.

Il faut déjà faire place à la polysémie :

> Quand un signifiant morphémique a plus d'un sens, il a pour contenu plus d'un sémème, il est polysémique. Pour une unité d'un rang supérieur, avoir plus d'un sens c'est présenter plus d'une isotopie. (Rastier 1987, §VIII.1.1.3.A)

Mais aussi, de multiples isotopies peuvent entrer en interrelation dans un texte, jusqu'à se superposer : François Rastier n'hésite pas à parler à ce propos de *polyphonie sémantique* et de *contrepoint sémantique* (Rastier 1989, §I.7). Le cas d'unités qui présentent les sèmes de plusieurs isotopies est étudié et attesté (voir par exemple (Rastier 1989, §II.4)), c'est même le cas inverse qui ferait figure d'exception (Rastier 1987, §VIII note 16).

- *Possibilité de non-appartenance* : une unité n'a pas à être enrôlée de force dans une Communauté. Il peut très bien se trouver qu'une unité ne participe pas aux isotopies les plus pertinentes dans le cadre du corpus et de l'application[124]. Dans notre modèle, les unités Simples coexistent avec les Communautés, et relayent celles-ci pour les aspects complémentaires aux isotopies dominantes.

  Il est d'ailleurs bien connu que, dans bien des cas, les résultats d'une classification automatique comportent (au moins) une classe « divers » ou « poubelle », dans laquelle s'accumule le « reste », les éléments originaux ou déviants par rapport au système de classe établi. Cette classe n'a évidemment pas la cohérence des autres, elle est purement artificielle. Il n'y a pas lieu, même si elle est produite, de l'instaurer en tant que classe au même titre que les autres.
- *Lexicalisation libre* : une Communauté est définie par ses constituants[125]. On peut la doter d'un nom, mais cela ne détermine en rien son contenu sémantique[126]. Le nom a plusieurs fonctions : rôle

---

[124] « Un lexème peut ne lexicaliser aucun thème [...]. Mais il peut aussi en lexicaliser plusieurs. Enfin, son lien avec le palier thématique est relatif à un discours (littéraire, médical, etc.), un genre, et un corpus. » (Rastier 1995a, §II.1, p. 235)

[125] Nous rejoignons (Tanguy 1997), qui définit les sèmes non par leur nom (*ibid*., §II.1.4, p. 45), mais par les isotopies qui les manifestent (*ibid*., §III.4 & 5, p. 92 sq.). La *fonction isotopie* associe à un sème l'ensemble *(i)* des couples de sémèmes pour lesquels le sème est un sème spécifique qui oppose le premier membre au second, *(ii)* des taxèmes, définis eux mêmes comme des ensembles de (plusieurs) sémèmes, et pour lesquels le sème est un sème générique réunissant les éléments du taxème, *(iii)* des sémèmes, qui reçoivent directement le sème par afférence, sans médiation d'une structure de classe.

« La présence d'un sémème dans une telle isotopie traduit l'attribution du sème de cette isotopie à ce sémème. Le type d'entité au travers de laquelle ce sémème est présent dans l'isotopie (respectivement spécème, taxème ou directement le sémème) traduit le type de sème pour le sémème considéré (respectivement spécifique, générique ou afférent). » (Tanguy 1997, §III.5.1, p. 94)

[126] Cette dernière propriété représente un enjeu plus important qu'il n'y paraît, sachant qu'une Communauté représente un (motif de) sème(s), et que nommer une Communauté c'est désigner ce(s) sème(s) :

« La dénomination des sèmes se trouve être un problème majeur pour toutes les théories [...]. En effet, si le sème peut être décrit comme une « périphrase à vocation métalinguistique », on oscille entre deux problèmes : une exigence de lisibilité du sème [...] et la nécessité d'un intitulé simple, qui facilite également la vérification de la cohérence d'une description manuelle sur des vocabulaires ou des corpus étendus. L'enjeu théorique de la dénomination des sèmes est celui de la *sémiosis illimitée* [...] ; mais l'enjeu pratique est celui de la description manuelle de lexiques sémantiques de taille significative, qui impose une bonne lisibilité tant pour des raisons de maintenance que d'échange entre différents intervenants. » (Cavazza 1996, p. 66)

Dans ce paragraphe, l'auteur revient à deux reprises sur le contexte dans lequel il se situe : il s'agit de la constitution *manuelle*, et de l'entretien *manuel*, d'un lexique sémantique. La dénomination est le point d'entrée interprétatif, l'ancrage de référence, qui détermine le sens.

La démarche est ici inverse : c'est la constitution interne de la Communauté qui est première, et qui se dote de sens ; la dénomination explicite une lecture, utile pour poursuivre l'analyse (cf. les unités distinguées) ou l'interprétation (étiquette apposée par un utilisateur comme point de repère).

Une Communauté est conçue comme une unité interne au traitement, et de ce fait ne présente pas les mêmes exigences de lisibilité et de concision qu'un lexique destiné à être lu et travaillé par des êtres humains. La question du soin à apporter à sa dénomination est contournée, par le fait même que ce qui détermine le comportement du système et les interprétations que chacun peut construire à partir des résultats, c'est la constitution de la Communauté, non sa dénomination. Plus encore : il n'y a aucune raison ici de restreindre la possible diversité des interprétations d'une même Communauté.

mnémonique (rappelle en un terme synthétique l'interprétation donnée à l'ensemble des constituants) ; *étiquette* (permet de mentionner brièvement l'unité dans un relevé ou dans un graphique) ; *identifiant* (introduit l'unité parmi les unités distinguées, ce qui la rend accessible pour jouer un rôle particulier dans le traitement). Le choix du nom se porte de façon naturelle sur une des lexicalisations possibles du sème. Cette lexicalisation peut faire partie, ou non, des unités trouvées dans le texte et entrant dans la Communauté. Il n'y a pas de lexicalisation canonique associée à un sème[127], aussi une même Communauté peut recevoir un nom différent selon le contexte (corpus considéré), selon l'utilisateur (chacun se repère à travers son propre jeu d'étiquettes), etc.

c) *Structure interne d'une* Communauté

La répartition des unités constituantes d'une Communauté en tant qu'élément du *noyau* et que *satellites* a été évoquée précédemment[128]. Les unités du noyau sont celles qui comportent le sème de façon stable ; une unité occasionnellement porteuse du sème est un satellite pour cette Communauté. Le rôle de noyau ou de satellite est, à une unité dans une Communauté, ce qu'est la qualification d'inhérent ou d'afférent, au sème dans une unité.

La distinction noyau *vs* satellite est opératoire pour reconnaître la manifestation d'une isotopie. Tout d'abord, l'isotopie suppose la répétition, donc la présence d'au moins deux unités porteuses du sème. Pour prévenir les cas litigieux, on ajoute une condition supplémentaire : qu'il y ait au moins une unité du noyau, qui assure l'ancrage de l'isotopie. (Cette règle, heuristique, pourrait évoluer en fonction des résultats expérimentaux.) Ensuite, le nombre d'unités qui concourent à une réalisation donnée de l'isotopie, la proportion d'unité du noyau, etc. sont autant d'indicateurs qui

---

[127] Une Communauté peut être aussi vue comme une formation thématique (« Nous nommerons thème une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses » (Rastier 1995a, §I.a, p. 224)), les thèmes se caractérisant également par la multiplicité ou l'absence possible de lexicalisation, et la non existance d'une lexicalisation canonique :
« Un thème, défini comme molécule sémique, peut recevoir des expressions diverses, par des unités qui vont du morphème au syntagme. Nous les nommerons, pour simplifier, *lexicalisations*. On peut distinguer des lexicalisations synthétiques qui manifestent au moins deux sèmes, et des lexicalisations analytiques, qui n'en manifestent qu'un. Ainsi, un thème peut être manifesté de manière diffuse, par exemple dans un paragraphe où divers sèmes seront lexicalisés tour à tour. La lexicalisation la plus synthétique ne jouit d'aucune prééminence théorique par rapport aux autres lexicalisations : elle n'est pas le 'mot juste' dont toutes les autres expressions ne seraient que d'imparfaits avatars. [Note : Par exemple, dans cette étude sur les sentiments dans la littérature française, il apparaît bien que] un thème peut avoir une lexicatisation privilégiée (ex. *ambition*), ou plusieurs (*pitié*, *commisération*, *compassion*). Il peut s'agir d'une lexie (*amour paternel*) ou n'avoir pas de nom retenu par l'usage (*sentiment du beau*, *amour de l'art*). »(Rastier 1995a, pp. 227 & 246)

[128] Le choix des dénominations de toutes ces nouvelles structures n'a pas été sans hésitations.
Communauté rappelle que les éléments sont réunis pour ce qu'ils révèlent avoir en commun.
Autour du *noyau*, central, au cœur de la Communauté, gravitent des *satellites*, pas toujours proches et moins stables vis-à-vis de l'unité représentée.
Nous aimions aussi la désignation de *constellation* (qui a inspiré aussi (Berni Canani 1986), et qui aurait pu remplacer celle de Communauté), pour sa consonance poétique, comme pour le type de réalité qu'elle définit : une délimitation « interprétative » sur la voûte céleste, les motifs qui organisent les étoiles, et la reconnaissance de l'importance de la perspective (ce qui apparaît comme une étoile double peut très bien concerner deux objets célestes séparés par une grande distance selon l'axe de visée). On comprend que l'utilisation de 'constellation' pour désigner des groupements de mots et des motifs linguistiques ait des précédents...
La notion de *molécule* (cf. les *molécules sémiques* chez François Rastier) rend également compte d'une structuration interne constituante (il y a des énergies en jeu), tout en sachant que les représentations pédagogiques sous formes de boules reliées deux à deux par des bâtonnets sont simplificatrices.
Quant aux *noyaux*, ils auraient aussi pu s'appeler *catalyseurs*, ce qui aurait souligné que ce sont eux qui amorcent la reconnaissance de l'isotopie, sans pour autant détenir une position supérieure aux autres éléments dans le déploiement de l'isotopie dans le texte.
Pour les *satellites*, nous avions aussi trouvé *acolytes*, qui nous sortait de la mécanique céleste, mais à consonance peut-être un peu trop spécialisée ou péjorative...

peuvent qualifier l'isotopie et qui contribuent à définir l'unité caractérisante. (Cela relève donc d'un autre chapitre.)

> [La dimension d'une isotopie] est définie par un nombre de positions syntagmatiques occupées par des morphèmes. L'isotopie minimale indexant au moins deux sémèmes, elle s'établit entre les contenus de deux morphèmes. Le groupe de morphèmes le plus simple est le mot. [(Ceci est un cas-limite). Notes :] Il n'est pas exclu qu'un sémème puisse comprendre au moins deux sèmes identiques. Par exemple 'épouser' comprend le trait /humain/ dans chacun de ses deux actants internes, tels qu'ils sont représentés dans son sémantème. Toutefois, il ne s'agit pas là d'une récurrence, mais d'une seule occurrence, occupant une seule position syntagmatique. (Rastier 1987, §VI.1)

Pour les unités effectivement manifestées, leur enregistrement sous forme d'une Communauté opère une réduction par regroupement et allège la représentation. Pour les unités absentes du texte, l'attribution de la Communauté au texte leur donne une présence latente, implicite. En effet, on peut par la suite concevoir un rapprochement, via la Communauté, avec un autre texte qui, lui, manifeste ces termes latents.

Jusqu'à présent, une dimension fondamentale des isotopies a été passée sous silence : leur nécessaire localisation. La dénomination *isotopie* s'analyse en fait en *iso-*, « même », et *-top(os)*, « lieu ». Derrière la notion de *lieu*, il faut entendre aussi bien le *lieu conceptuel*, *i.e.* le sujet exploré, au centre du discours, et le *lieu spatio-temporel* : la zone du texte, le moment du discours ou de la conversation. En effet, pour pouvoir constater une répétition, la récurrence qui forme l'isotopie, il faut enregistrer le retour d'un *même* élément sémantique dans un *même* espace textuel. La définition de la zone de récurrence est donc constitutive de l'isotopie. Elle fait partie intégrante de la définition d'une Communauté, au même titre que les unités constituantes qui pointent son « lieu sémantique ».

La localisation d'une isotopie se définit en termes de zone plutôt que de contiguïté :

> Les relations constitutives d'une isotopie ne sont pas nécessairement liées à la contiguïté (qui ne peut être appréhendée qu'au niveau de l'expression). Elles obéissent à un principe de *localité* (cf. ch.V), mais en général les morphèmes qui contiennent des sémèmes indexés sur une même isotopie ne sont pas contigus. (Rastier 1987, §IV.2.4.D)

Les trois types d'unités descriptives qui se déclinent sous le chapeau de Communauté, se distinguent par la nature des zones de localité dans lesquelles se réalise l'isotopie. Les zones s'inscrivent toutes dans la linéarité du texte, ce qui les rend commensurables en termes de longueur par exemple. Cependant, un point de vue linguistique fait apparaître des dénivellations, qui séparent des paliers de description.

> On distinguera trois paliers de description du contexte, qui sont autant de zones de localité : le syntagme minimal (ou mot), l'énoncé, et son au-delà textuel. Cette distinction paraît nécessaire, car des relations contextuelles sur un palier peuvent être incompatibles avec d'autres relations sur un autre palier. (Rastier 1987, §III.2.1.2)

Passer d'un palier à l'autre est un peu comme changer d'ordre de grandeur. Du point de vue des isotopies, bien que formellement des isotopies associées à différents paliers se réalisent de la même façon, par réitération d'un sème dans une zone, les isotopies d'un palier n'ont pas la même interprétation que les isotopies d'un autre.

D'autres formes de localité et de contextes se détaillent dans des réalisations opérationnelles :

> [Pour spécifier les contextes auxquels s'appliquent les règles de reconnaissance des sèmes,] on définit [...] des distances positionnelles (exprimées en nombre de mots) et des distances syntaxiques qui conditionnent l'application des règles (distance 1 : même syntagme ; 2 : même énoncé ; 3 : même période ; 4 : période adjacente). (Rastier, Cavazza, Abeillé 1994, §III.7)

La zone de localité du syntagme minimal correspond aux unités descriptives initiales, et notamment aux Solidarités. Le syntagme, en tant que premier groupement syntaxique, est aussi une zone de localité significative (Bonhomme & *al*. 1996, § V.II). Le syntagme présente une organisation plus structurée que les isotopies en général (notamment la dissymétrie des relations entre composants, leur orientation) : cette zone est décrite par les Séquences. Les Associations quant à elles englobent dans leur champ ces zones particulières que sont les énumérations. Les Communautés vont considérer les isotopies des paliers supérieurs, donc celles au palier de l'énoncé, et celles au palier du texte. Le troisième type de Communauté ajoute une zone de localité intermédiaire, dans l'esprit de la « période adjacente » mentionnée dans la citation précédente..

d) Les Relations

La zone de réalisation (ou portée) d'une Relation est la *période*.

La *proposition* grammaticale s'organise à partir d'un verbe et de l'ensemble de ses relations avec ses arguments (sujet, compléments réalisant diverses fonctions). La *phrase* peut articuler plusieurs propositions (proposition principale *vs* subordonnées, propositions indépendantes). Dans le cas des phrases nominales, cas certes marginal en soi mais largement attesté (notamment dans la presse), la phrase échappe à une définition grammaticale fondée sur les propositions. Ce qui ressort alors est son caractère rythmique, propre à détailler le texte en parts à la mesure de la perception et des capacités cognitives humaines. La typographie souligne doublement ce découpage, par le point final de la phrase et par sa majuscule introductive. L'art oratoire s'est appliqué à ciseler ces phrases, ou *périodes*, balancées avec adresse et portées par le souffle de la respiration.

Proposition, *phrase* et *période* gravitent autour d'un type d'unité de localisation, dont il ressort deux caractéristiques. Cette unité de localisation se mesure à l'aune des capacités cognitives et perceptives humaines, qui ajustent et régulent sa portée[129]. Dans l'empan de ce cadre, s'explicitent, par la voie de la grammaire, les relations entre les unités qu'elle embrasse. De façon simplifiée, il y a toujours une *relation*, directe ou indirecte, entre deux éléments d'une phrase : qualification, dépendance, rôle complémentaire autour d'un même objet, etc.[130]. L'arrêt, marqué par un point ou une ponctuation forte, est aussi une délimitation sémantique : la propagation de certains sèmes est inhibée.

La pertinence de cette zone de localité étant admise, nous choisissons de la désigner sous le nom de *période*. *Proposition* et *phrase* seraient des choix moins souples et plus dangereux, par leur lourde charge de définitions et de controverses.

Période a l'avantage de souligner tout autant la mise en relation des unités incluses, et la régularité du « volume » d'information saisi[131]. Que les résonances mathématiques de « volume d'information » ne nous entraînent pas sur une pente dangereuse ; on veut simplement prendre acte de

---

[129] Explication proposée par (Greimas 1966, §VIII.2.c) : « Nous ne disposons, au départ, que du modèle syntaxique pour nous donner une première idée de la façon dont il faut concevoir l'organisation des contenus à l'intérieur de l'univers manifesté. Le modèle syntaxique nous frappe d'abord par sa simplicité, c'est-à-dire à la fois par le petit nombre d'éléments constitutifs du message et par les dimensions très limitées assignées au message dans le déroulement du discours [...]. A y regarder de plus près, on ne voit qu'une [explication] possible : la limitation de l'activité syntaxique ne peut provenir que des conditions que lui impose objectivement la réception de la signification. Bien que le message se présente, à la réception, comme une succession articulée de significations, c'est-à-dire avec un statut diachronique, la réception ne peut s'effectuer qu'en transformant la succession en simultanéité et la pseudo-diachronie en synchronie. La perception synchronique, si l'on en croit Brøndal, ne peut saisir qu'un maximum de six termes à la fois.

[...] L'univers sémantique éclate ainsi en micro-univers, qui seuls peuvent être perçus, mémorisés et « vécus ». En effet, si nous pensons quelque chose à propos de quelque chose, nous projetons ce quelque chose devant nous comme une structure de signification simple, ne comportant qu'un petit nombre de termes. Le fait que nous pouvons, ensuite, « approfondir » notre réflexion [...] ne change rien à cette saisie première. »

[130] Greimas accorde une grande place aux conséquences sémantiques du découpage et de la structuration opérés par la syntaxe (Greimas 1966, §X.2) : « Nous avons été frappé par une remarque de Tesnière –qu'il ne voulait, probablement, que didactique– comparant l'énoncé élémentaire à un spectacle. Si l'on se rappelle que les *fonctions*, selon la syntaxe traditionnelle, ne sont que des rôles joués par les mots –le sujet y est « quelqu'un qui fait l'action » ; l'objet, « quelqu'un qui subit l'action », etc.–, la proposition, dans une telle conception, n'est en effet qu'un spectacle que se donne à lui-même l'*homo loquens*. Le spectacle a cependant ceci de particulier, c'est qu'il est permanent : le contenu des actions change tout le temps, les acteurs varient, mais l'énoncé-spectacle reste toujours le même, car sa permanence est garantie par la distribution unique des rôles.

[...]

A partir de là, nous avons pu tenter l'extrapolation suivante : puisque le discours « naturel » ne peut ni augmenter le nombre des actants ni élargir la saisie syntaxique au-delà de la phrase, il doit en être de même à l'intérieur de tout micro-univers ; ou plutôt le contraire : le micro-univers sémantique ne peut être défini comme univers, c'est-à-dire comme un tout de signification, que dans la mesure où il peut surgir à tout moment devant nous comme un spectacle simple, comme une structure actancielle. »

[131] Pour son système ALCESTE, Max REINERT conjugue également ces deux critères –repérage de ponctuations fortes et longueur (250 caractères, puis nombre de mots différents analysés)– pour construire ses *unités de contexte*, de l'ordre de l'énoncé (Reinert 1990) (Reinert, Piat 1995).

la réalité suivante : une succession de phrases (typographiques) très courtes peut, sur le plan interprétatif, être reçu comme une seule unité. La juxtaposition, même hachée par des points, est une marque de relation (sur le plan interprétatif). Inversement, une phrase extrêmement longue n'est pas, dans un premier temps, saisie dans sa totalité : la lecture se déroule au rythme de la construction syntaxique et des ponctuations, qui rééquilibrent l'ensemble en y retraçant des zones.

En ce sens, la recherche de « corrélats sémantiques » à un mot donné (Bonhomme & *al*. 1996) s'intéresse à des zones de localité de l'ordre de la période. On explore un contexte de dix mots avant et dix mots après le mot considéré, ce contexte s'arrêtant avant dix mots, avec la première ponctuation forte rencontrée, le cas échéant. Les deux critères, longueur et ponctuation, sont utilisés ; on ne se donne pas de longueur minimale cependant. De plus, la fenêtre (dix mots) peut très bien interrompre la phrase « en plein élan ». L'appui sur la ponctuation serait préférable, il évite l'arbitraire des fenêtres rigides. Il n'y a pas la même coupure brutale et artificielle du contexte lorsque l'on respecte les seuils significatifs marqués par la ponctuation.

Quant à la manière d'utiliser les relations internes, dont l'existence vient d'être postulée, elle est ici tout bonnement *implicite*. Trois motivations y conduisent ; énonçons-les avant de les examiner une à une : la difficulté de mise en œuvre, le gain réel prévu non significatif pour la caractérisation d'un texte relativement à d'autres, l'indépendance du calcul de rapprochements par rapport aux (petites) variations de relations.

Les relations syntaxiques sont décrites par la grammaire au moyen d'une panoplie de fonctions ou de cas. Les études sémantiques s'appliquent à décrypter, au fil des phrases, derrière les schémas grammaticaux et les rapports lexicaux, les cas sémantiques qui structurent l'interprétation (il est bien connu que le sujet grammatical ne désigne pas toujours celui qui fait l'action). La difficulté est de trouver un système de cas satisfaisants[132], et de lui donner une efficacité opératoire. Quelle pertinence cherchée à travers un système de cas, pour quelle application ? Et comment en assurer la correcte mise en œuvre ? Notamment, pour un système automatique, il faut sans cesse se méfier de ne pas être *trop* influencé par les cas syntaxiques, informations accessibles, mais qui ne traduisent que les possibilités de schémas de phrase prévus par la langue.

Nous faisons l'hypothèse que les valences casuelles et les restrictions de sélection sémantiques limitent la combinatoire théorique à un ou très peu de cas de réalisation vraisemblables. C'est d'ailleurs un des principes de ces jeux grammaticaux que l'on donne aux écoliers : « voici des mots (dans le désordre), reconstituez la phrase ». Ceci est comme démultiplié à l'échelle du texte, où le contexte abonde d'éléments dans le sens de l'interprétation souhaitée : l'indexation structurée (Coret & *al*., 1994) n'illustre sa nécessité qu'à l'échelle d'une requête de quelques mots (Lefèvre 1997) –ou pour un système qui raisonne au niveau des mots, sortis de leur contexte textuel (Maniez 1983)[133].

---

[132] Les travaux de Greimas, au cœur de la linguistique structurale, ont déjà été évoqués : il propose un système de catégories actancielles, organisées en axes bipôlaires (Greimas 1966). Pottier s'appuie sur un système d'axes (actance et dépendance) pour présenter les cas comme des *zones casuo-conceptuelles* fondamentales (Pottier 1974, §50sq.) (Pottier 1987, §X). Tout ceci permet de comprendre l'ensemble des cas comme un tout cohérent, complet et équilibré, et de situer la réflexion au niveau d'une linguistique générale (s'intéressant au langage, à travers et par-delà les langues particulières).

Dans les traitements automatiques, Fillmore inaugure les grammaires de cas (Sabah 1988, §3) et introduit explicitement la notion de *cas sémantique* en 1968. En 1972, Schank, avec les *dépendances conceptuelles*, développe toute sa description à partir de primitives représentatives d'actions élémentaires ; la nature des liens qui intègrent ces primitives s'identifie à l'une des six rôles conceptuels prévus. Initialement, la représentation décrit les rôles respectifs des éléments dans la phrase ; elle se déporte vers le niveau lexical, en explicitant les rôles des différents arguments d'un prédicat (verbe), éventuellement analysable en primitives (autre ex. : Desclés 1990, §11.7).

Rastier rappelle que les cas sémantiques sont structurants à tous les paliers de la description linguistique : mot (plus exactement *lexie*) (Rastier, Cavazza, Abeillé 1994, §III.3.2), phrase (syntagme et période) (*ibid.*, §V.7), texte (*ibid.*, §VII.4.3). Le choix et l'utilisation des cas sémantiques sont surtout abordés dans la présentation de la composante *dialectique* de la description textuelle (Rastier 1989, §I.5).

[133] (Maniez 1983) souligne la polyvalence de l'opérateur booléen ET, source d'imprécision ou d'erreurs dans les recherches ; en revanche, sur le texte intégral, les contraintes de distance (champ de la notice, paragraphe, phrase,

Nous considérons aussi que le rôle sémantique précis de telle ou telle unité n'est en général pas pertinent pour décider d'un rapprochement. L'application DECID met en relation des textes sur le plan de leur thématique : les éventuelles variations de rôles dans un même cadre général ne sont pas moins intéressantes à rapprocher.

Les Relations se comportent donc comme des Communautés, pour les isotopies « serrées », à l'échelle de la période.

e) Les Voisinages

Les Voisinages sont des Communautés, dont la zone de réalisation associée est le *paragraphe*. Là encore, la pertinence de cette zone de localité doit être éclairée[134].

La formule « un paragraphe, une idée » fait partie des principes rédactionnels courants dans notre culture. Elle résume un premier aspect du paragraphe, concernant son unité sémantique. L'unité est à la fois interne et externe : le paragraphe ne développe qu'une seule idée, et le changement de paragraphe marque le passage à une autre idée. Tout cela ne relève pas d'un état de faits, surtout dans cette formulation abrupte : on trouverait des textes, voire des genres, pour lesquels ce principe serait tout à fait discutable. L'intérêt est d'y reconnaître une attente interprétative. Le découpage en paragraphes guide l'interprétation, en annonçant un point nouveau, une étape dans le cheminement du texte, et en induisant une présomption de participation de ses éléments à une même situation. La typographie (dont une des principales raisons d'être est d'accompagner la lecture) souligne nettement ces transitions : interruption de la ligne en cours, retrait, espacement plus large des lignes.

A l'intérieur d'un paragraphe, les éléments ne sont pas nécessairement en relation directe, comme à l'échelle de la période ; il s'agit plutôt d'une relation de coexistence, de voisinage –d'où le nom choisi pour ces unités descriptives.

Le voisinage est aussi un voisinage visuel sur la page. Ainsi, le lecteur peut considérer d'un seul tenant une série de points brefs, entre deux paragraphes plus développés. D'autre part, le début d'un paragraphe « voisine » aussi la fin du précédent, c'est le lieu où le rédacteur « ménage la transition » et assure le fil de son discours.

L'implémentation tient compte de ces deux points. D'une part, les *alinéas* du texte, qui correspondent au marquage matériel de la structure du texte (généralement le retour à la ligne), sont si nécessaire regroupés pour reconstituer une zone de localité plus conforme aux propriétés du paragraphe. En revanche, même relativement long, un alinéa n'est pas redécoupé. Ainsi, on respecte la forme du texte (alinéas) tout en gardant une liberté de réinterprétation plus sémantique (paragraphe). Par exemple, un alinéa constitué d'une seule période, et qui ne correspond pas à un (inter)titre, est regroupé avec l'alinéa suivant pour former un paragraphe, sauf si le second alinéa correspond à un (inter)titre ; dans ce cas, il est regroupé avec l'alinéa précédent.

La perméabilité des frontières des paragraphes est également modélisée. L'isotopie associée à un Voisinage se réalise au cœur d'un paragraphe, mais elle peut aussi s'initier ou / et se conclure dans les marges des paragraphes connexes. La représentation est différente lorsque l'isotopie est développée dans plusieurs paragraphes consécutifs, ou lorsqu'elle reste en marge des paragraphes. Les tableaux ci-après précisent le comportement choisis dans chaque cas.

- Chaque + représente la présente d'au moins une unité pouvant contribuer à l'isotopie dans la zone considérée ;
- un 1 représente la présente d'au plus une unité pouvant contribuer à l'isotopie dans la zone considérée ;
- un 0 indique l'absence d'unité contributrice dans la zone ;
- (les cases laissées blanches signifient que la présence et le nombre d'unités contributrices sont indifférents).

---

fenêtre de *n* mots, adjacence) sont un procédé « simple et efficace » (*ibid*., §4, p. 56). Jacques MANIEZ conclut à la pertinence de rechercher un juste équilibre : la syntaxe n'a pas à être négligée mais n'a pas non plus besoin de trop de finesse ; elle ne doit pas supplanter l'attention primordiale à la terminologie et au lexique.

[134] La consistance et la valeur de la notion de paragraphe ne sont pas des acquis, et une définition sémantique et herméneutique du paragraphe se cherche encore (Laufer 1985).

- les cas de figure symétriques par rapport à l'intérieur du paragraphe *n* reçoivent la même interprétation.

Paragraphe *n - 1*	marge initiale					0	0			
	intérieur					0	0			
	marge finale					+	+		0	
Paragraphe *n*	marge initiale		+	++	+			0	0	1
	intérieur	++	+			+		0	1	0
	marge finale				+		+	0	0	0
Paragraphe *n + 1*	marge initiale								0	0
	intérieur						0			
	marge finale						0			
Réalisation du Voisinage dans le paragraphe *n* :		oui						non		

Pour l'implémentation on a choisi de fixer la marge initiale d'un paragraphe à sa première période, sa marge finale à sa dernière période.

On accorde donc au paragraphe, en tant que zone de localité, les caractéristiques suivantes : unité de contenu, taille d'ordre supérieur à celle de la période, délimitation marquée typographiquement, recouvrements possibles aux frontières. A notre connaissance, tous les traitements automatiques qui s'intéressent à des *passages*, de l'ordre du paragraphe, s'appuient sur l'une de ces caractéristiques, plus rarement sur plusieurs. Il y est fait usage de fenêtres (découpage du texte tous les n mots), de fenêtres chevauchantes, ou des délimitations physiques du texte (retour à la ligne) (Callan 1994).

f) Les Arrière-Plan

La zone de localité associée aux Arrière-Plan est le texte.

Les Arrière-Plan correspondent aux isotopies à maillage large, peu ou pas du tout discernables au niveau des zones de localité d'ordre inférieur. Les Arrière-Plan sont aussi un fond qui souligne l'unité du texte, et qui peut le contraster d'autres textes.

Les interactions à longue distance ne sont évidemment pas de source syntaxique (comme les Relations). Elles traduisent des thématiques générales, des domaines, notées de façon diffuse (sinon c'est le registre des Voisinages). Elles peuvent aussi regrouper des unités caractéristiques d'un genre, qui se font écho à l'échelle du texte : d'une partie à l'autre, d'une extrémité à l'autre.

L'utilisation du texte comme contexte pour repérer des interactions entre unités du texte a été utilisée avec un certain succès sur les textes d'Action.

- Une classification automatique de termes (groupes nominaux désignant un concept), en fonction des documents dans lesquels ils apparaissent, fournit des classes, présentables à l'utilisateur comme points d'entrée pour une recherche documentaire (Quatrain, avril 1996).
- On veut sélectionner, dans un ensemble de termes, ceux que l'on va proposer comme descripteurs pour enrichir le thesaurus. L'informativité d'un terme, et son efficacité pour une application de recherche documentaire, sont évaluées sur sa capacité à discriminer les textes d'Actions les uns des autres. Parmi les trois indicateurs statistiques retenus, le critère de densité locale est contextuel : il retient un terme si les documents dans lesquels il apparaît se ressemblent (Sta 1998). On retient donc un terme pour sa stabilité contextuelle.

Cependant, dans ces expérimentations, les différents paliers de contexte correspondant aux différentes natures de Communauté ne sont pas distingués ; de plus, les textes d'Action représentent un genre assez bref (textes d'une à deux pages). Les regroupements trouvés (et en fait cherchés ?) s'apparentent donc plus à des Voisinages qu'à des Arrière-Plan.

F. DISCUSSION : CONFRONTATION DE LA TYPOLOGIE DES UNITÉS DESCRIPTIVES AUX APPROCHES PRÉCÉDENTES

1. L'indexation : une désignation unique et des réalités très diverses

a) Le thesaurus, comme référentiel conceptuel hiérarchique

Les thesaurus représentent l'aboutissement d'un mode de recherche documentaire (Lefèvre 1997) (Zizi 1995, §3.1)[135]. Les descripteurs fixent une manière univoque, normalisée, de représenter un concept (*vocabulaire contrôlé*) : ainsi, on évite qu'une recherche documentaire échoue faute d'utiliser les bons termes, ceux qui correspondent à la description du document dans le fonds. D'autre part, leur structuration hiérarchique se prête à un *parcours méthodique* et efficace, en précisant peu à peu le concept recherché. Cela permet également de rendre compte de différents *niveaux de détail*, des concepts généraux aux concepts très pointus. Faire varier le niveau de détail au cours d'une recherche documentaire est un moyen direct pour élargir ou focaliser le champ d'investigation. La typologie des relations présentes dans le thesaurus calquent ces opérations : *voir aussi*, *terme spécifique vs terme spécifique*.

En explicitant l'ensemble des concepts qui peuvent être utilisés et en les organisant, le thesaurus correspond à une vision du monde. Un thesaurus n'est *jamais universel*, au sens où il saurait refléter tous les thèmes d'intérêt. Cela tient au choix et à la délimitation des concepts, mais aussi aux partis pris dans l'organisation même de ces concepts. Si par exemple le thesaurus suit les divisions des disciplines académiques, un thème d'intérêt transdisciplinaire est mal représenté. La possibilité de combiner plusieurs concepts pour exprimer un concept nouveau (*langage post-coordonné*) ne résout pas tout, car les concepts ne sont pas (et ne peuvent être) comme les primitives d'un langage formel. Les deux points de fuite sont : l'incomplétude patente des concepts exprimés ; l'irréductibilité des concepts qui ne sont pas déjà exprimés à une composition de concepts précédents. Moins un concept a sa place prévue par la structure, plus son expression devient approximative. Un thesaurus doit donc toujours être compris comme relatif à un contexte : il est équilibré vis-à-vis du fonds qu'il décrit, il est orienté et ajusté en fonction des attentes des lecteurs. Une bibliothèque publique générale, une bibliothèque d'archives régionales, une bibliothèque d'un industriel dans le domaine de la chimie, etc., n'ont clairement aucun intérêt à adopter le même thesaurus. L'utilisation d'un thesaurus est tout autant inadaptée dans un contexte évolutif et changeant.

Une très importante base documentaire comme INSPEC ménage d'autres modes d'indexation parallèlement à son thesaurus. Un document se voit classiquement affecter trois à huit descripteurs (le plus souvent quatre ou cinq), et une dizaine à une vingtaine d'*identifieurs* ou *termes libres*. L'observation montre que ces identifieurs reprennent, comme le ferait un surlignage, tous les *termes* du résumé qui sont représentatifs et caractéristiques des sujets abordés. Le complément au thesaurus est donc une *description très proche du texte*, rendant compte des formulations développées et précises qu'il choisit, des entités particulières qu'il mentionne (organismes, marques), des nouveautés qu'il définit et nomme. La description par le thesaurus, standardisante, a pour effet d'élaguer tout ce qui n'entre pas dans le cadre général du thesaurus. La description ajoutée par les identifieurs contribue à rétablir ce qui fait la *spécificité* du document, et par là son intérêt propre, sa valeur distinctive.

---

[135] (Picard & Poibeau 1997) identifient trois moyens d'accès à l'information dans le cadre de la documentation électronique : *(i) la table des matières*, donnant une image de la structure globale du texte, mais soumise à l'ordre et aux localisations linéaires ; *(ii) la recherche en texte intégral*, sur chaînes de caractères combinables au moyen d'opérateurs, et qui a pour elle sa simplicité d'accès ; *(iii) la recherche par index* –où l'index est structuré comme un thesaurus–, « plus fine que la recherche plein texte car elle est normalisée, structurée et contextualisée » (réduction des polysémies).
A l'échelle d'un fonds documentaire, l'accès aux documents est aussi médié par un plan de classement, qui peut s'apparenter au *(i)*.

L'indexation d'un document à l'aide d'un thesaurus s'affranchit du media : le document peut être un texte ou un film par exemple. En se positionnant comme conceptuelle, l'indexation occulte le texte et ses facettes propres[136]. Le texte est notamment objet de lectures et d'interprétation. La représentation par des descripteurs s'arrête nécessairement à l'une ou quelques-unes de ces lectures, et ne trouve rien qui n'était déjà prévu, directement (descripteur) ou indirectement (combinaison de descripteurs).

C'est ici que l'on perçoit l'ambivalence du « texte intégral »[137]. D'une part, reste consultable et interrogeable en texte intégral *ce qui ne vaut pas la peine* de faire l'objet d'une analyse documentaire et d'être indexé : l'indexation reste un processus complexe, coûteux, qui demande un certain temps ; elle mobilise un savoir-faire de retranscription pour passer dans le référentiel du thesaurus. On n'indexe que ce qui n'est pas périssable, ce que l'on a sélectionné, ce que l'on veut conserver et pouvoir retrouver. Mais d'autre part aussi, le texte intégral, c'est ce qui respecte le plus fidèlement le texte que l'on ne saurait réduire à quelques descripteurs ; c'est s'accorder un volume de données beaucoup plus grand, pour conserver toute la richesse du document original, *dans le cas où une indexation ne peut être satisfaisante*.

La voie choisie pour DECID se passe donc de thesaurus. Les unités descriptives peuvent évoluer dynamiquement avec le corpus, tout en gardant quelques repères stables (unités distinguées). Ancrées dans les textes, elles sont sensibles aux écarts entre diverses pratiques de lectures (notamment degré de spécialisation) et se situent de plain pied avec les pratiques effectives[138], tout en ménageant des relais entre diverses formulations qui se répondent effectivement dans les textes (Associations et Communautés). Enfin, les unités descriptives traduisent certains points d'appui de l'interprétation (indices sémantiques) plutôt qu'une interprétation aboutie (concepts).

b) L'indexation automatisée, une autre forme d'indexation

Ghislaine Chartron développe un logiciel qui indexe automatiquement des documents (à partir de leur titre et de leur résumé). Le fonds documentaire auquel s'applique son expérimentation (base EDF-DOC) est déjà indexé manuellement. C'est l'occasion de comparer ces deux modes[139] d'indexation, et de montrer les points forts de chacun (Chartron 1988, §VIII.4.2).

Les points forts de l'indexation automatique sont :

- sa régularité : l'indexation est déterministe. Un même document est toujours représenté de la même manière, une notion repérée dans différents textes est toujours rendue par le même terme d'index. *A contrario*, l'indexation manuelle est connue pour les disparités entre les indexeurs, et pour un même indexeur au cours du temps[140].
- sa spécificité et sa proximité au texte : l'indexation automatique rend compte de l'information telle qu'elle se présente dans les textes.

---

[136] Ce sont les quatre facettes du texte proposées dans cette thèse : son matériau linguistique, son organisation interne, l'intertextualité, la dynamique de la lecture. La fiche bibliographique décrivant un ouvrage enregistre d'ailleurs d'autres informations que son indexation. La langue du document, son titre, son auteur, son nombre de pages, etc., font échos aux trois premières facettes.

[137] L'observation qui suit revient à François RASTIER.

[138] François RASTIER remarque que le superordonné dans une terminologie n'intéresse pas l'expert : dans le champ de sa pratique, le détail des entités considérées s'impose.
Du côté des réalisations logicielles, l'écho est le même. Bruno BARAKAT-BARBIERI, cherchant à produire automatiquement des graphes de concepts (pour enrichir le système documentaire SPIRIT), renonce à construire « le haut du graphe » : force est de constater que « les concepts qui permettraient d'englober de façon très large une grande majorité de notions ne figurent pas forcément de manière explicite dans les documents. » (Barakat-Barbieri 1992, p. 49).

[139] Nous ne reprenons pas ce qui est lié au fait que l'indexation manuelle est faite ici en vocabulaire contrôlé (thesaurus) alors que l'indexation automatique est libre. Ce sont les différences spécifiques automatisme / manuel qui nous intéressent. Le débat entre indexation libre ou contrôlée est étudié ailleurs. Ces deux dimensions ne sont pas complètement décorrélées : l'indexation manuelle est plus naturellement une indexation contrôlée, l'indexation automatique une indexation libre, mais l'inverse existe dans les deux cas.

[140] Une forme d'aide semi-automatisée à l'indexation manuelle serait la possibilité de retrouver efficacement certains choix d'indexation précédents, pour poursuivre l'indexation en cohérence avec ces choix déjà effectués.

- son évolutivité : reprendre l'indexation générale de tout un fonds documentaire déjà indexé automatiquement ne se chiffre qu'en temps de calcul ; en revanche, une réindexation rétrospective manuelle mobilise de nombreux documentalistes, pour un temps beaucoup plus long, avec le risque que les manières d'indexer soient disparates et dérivent, et de surcroît qu'au bout de cet effort la nouvelle indexation soit déjà dépassée.

Les points forts de l'indexation manuelle sont :

- la prise en compte de l'intégralité du document, et notamment l'indication d'éléments précis non mentionnés dans le titre ou le résumé ;
- l'introduction de désignations synthétiques, génériques, qui servent à « 'traduire' le contenu des documents pour des interrogateurs travaillant dans des disciplines variées ».Elles reflètent le vocabulaire usuel des interrogations (c'est le niveau conventionnel du « bon » mot-clé, tel un point de rendez-vous tacite, spontanément adopté dans la pratique documentaire). « Les fortes occurrences des termes 'intelligence artificielle', 'reconnaissance des formes', 'système expert' traduisent une habitude à indexer par ces termes même si un degré plus grand de spécificité peut être atteint ».
- une lecture orientée, qui met en valeur les aspects correspondants aux usages prévus des documents. Pour le fonds EDF, c'est « une amplification de certains éléments des documents en rapport avec une volonté de mettre en valeur les applications mentionnées dans les documents. Les documentalistes du centre [...] [opèrent souvent] en deux temps à savoir une première indexation relative à la technique décrite et une autre indexation relative aux applications. [...] [Ces] orientations [sont] liées aux demandes des utilisateurs de la base ».

Le traitement recherché pour DECID, tout automatique, bénéficie des points forts des approches automatisées. La modélisation proposée pour DECID réintègre par ailleurs des qualités qui semblait échapper à l'automatisation. DECID prend en compte dans leur entier les textes qui lui sont soumis. Le fait que la caractérisation des destinataires et l'interrogation pour l'envoi d'un document se fasse de part et d'autre par l'intermédiaire d'un texte, permet de faire le calcul de mise en relation sur la base de représentations proches des textes. Les unités descriptives favorisent la reconnaissance de liens, que ce soit entre des formulations différentes, ou entre des concepts d'un même domaine. Les unités caractérisantes et la possibilité offerte par l'interface de surligner un texte de requête, donnent les moyens de nuancer et d'orienter la sélection par rapport aux points les plus saillants pour l'utilisateur.

En revanche, la représentation d'un texte calculée par DECID, destinée à un traitement automatique interne, n'a pas la lisibilité (et en particulier la concision) d'une liste de descripteurs choisis manuellement. C'est là que nous préférerions situer la démarcation entre caractérisation automatisée d'un texte et indexation manuelle d'un document : la première est une représentation interne, forgée par la machine et utilisée par elle ; la seconde est une représentation externe, une traduction humaine, la réécriture d'un texte dans le genre « liste de mots-clés », genre se prêtant mieux ensuite aux recherches documentaires, et faisant partie de la présentation du document donnée au lecteur, au même titre que les autres caractéristiques bibliographiques.

c) Les « mots-clés » du texte intégral

Avec le développement des techniques documentaires, l'appellation générale *mot-clé* renvoie à des réalités extrêmement différentes. L'habit ne fait pas le moine : la liste de groupes nominaux ne fait pas l'ensemble de descripteurs ou de mots-clés. Or on constate plutôt une confusion entretenue, qu'il est temps de dissoudre.

Abstraction et généricité

Destinés à traduire un document en une liste de quelques mots, les descripteurs documentaires et mots-clés adoptent naturellement un certain degré de généralité, apte à cadrer une large part d'un texte et de ses développements. Ils rallient à une discipline, une famille de méthodes, un secteur d'application. L'expression succincte d'un thème de recherche adopte le même ton, puisqu'elle s'efforce de situer en quelques mots les documents qui pourraient convenir.

En revanche le texte, destiné à une certaine communauté de lecteurs, adopte le ton correspondant à cette communauté. Pour un lectorat expert, il choisit avec soin des termes précis pour les notions importantes, centrales ; avec une visée pédagogique, le vocabulaire peut être plus imagé, multiplier les manières de présenter, et l'introduction des noms des concepts sous-jacents peut être retardée ou marginalisée (dans des notes, des appendices). La formulation générique des thèmes n'apparaît pas toujours (Renouf 1993a). Quant au mot extrait d'un texte, il fait partie d'un développement, est un élément d'expression en relation avec d'autres. Au cœur du texte, il n'a pas naturellement une portée méta-textuelle, c'est-à-dire celle d'une description du texte, qui prend du recul par rapport à lui.

Contextualisation

Le thesaurus, en situant ses concepts les uns par rapport aux autres, les dote d'un contexte. C'est ce qui fait que si, malgré les précautions prises dans le choix du terme pour désigner le concept, la signification d'un descripteur pose question, la consultation de son rattachement et des autres descripteurs auxquels il s'oppose permet de cerner ce qu'il recouvre. Les descripteurs d'un thesaurus trouvent donc bien leur place dans une liste qui égrène individuellement les concepts attribués à un document.

Le mot, extrait du texte, et ajouté à une liste de mot-clés obtenus de la même façon, perd son contexte textuel, et ne retrouve d'autre contexte que celui des autres mots de la liste. Si la liste est pauvre ou dispersée, et surtout si elle n'est exploitée que terme à terme, alors le mot, privé de son « milieu naturel », « dépérit » : il se vide de ce qui lui donnait tout son sens, des multiples reflets reçus de ses interrelations en contexte. Même dans un dictionnaire, le mot qui est défini ne prend consistance que dans les différents contextes donnés par sa définition.

2. Les systèmes de recherche sur les documents textuels : les tactiques pour passer de l'expression à l'idée

a) Synonymes, expansion, reformulation

Dans notre système d'unités descriptives, les Associations sont le mode d'expression de liens de synonymie.

Une différence de nature est quelquefois posée entre des synonymies parfaites, synonymies hors contexte (textuel) (mais néanmoins dont la validité correspond à un certain point de vue interprétatif), et des synonymies en contexte. Ce qui sépare ces deux formes de synonymies, c'est qu'un ensemble de synonymes de la première forme sont toujours interchangeables, alors que dans le second cas le remplacement d'un mot par son synonyme fonctionne dans certains contextes, pas dans d'autres. La synonymie hors-contexte considère une superposition entière de tous les sens pour toutes les unités synonymes : chaque sens d'une unité est aussi transcrit par l'autre. La synonymie en contexte est celle du recouvrement de certains sens, mais pas de tous.

La notion de synonymie hors contexte a trouvé un appui considérable dans le développement des logiciels de recherche documentaire, avec l'interrogation en langage libre (c'est-à-dire sans passer par une syntaxe et un lexique spécialisés, univoques). Conçue comme un remède à la non correspondance malheureuse des termes de la requête avec ceux qui caractérisent le document, chaque terme de la requête appelle son éventuel ensemble de synonymes, et la requête ainsi étendue et enrichie prévoit toutes les alternatives de formulation auxquelles pourraient avoir recours les documents[141]. Comme dans l'analyse de la requête chaque terme est une unité indépendante et que la notion de contexte est inopérante, les ensembles de synonymes dont on munit le système sont donc

[141] (Radasoa 1988) fait le point des tactiques de reformulation, au sens le plus large : morphologie (lemmatisation, mots de la même famille), utilisation d'un thesaurus, correcteurs de fautes de frappe ou d'orthographe, apprentissage de règles de reformulation enregistrées dans un profil... et interrogation par une partie d'un document.

des synonymes hors contexte. L'expansion de la requête est en fait l'expansion de chaque élément de la requête, isolé de son contexte.[142]

Cette opération est évidemment dangereuse, car riche de dérives et de contresens potentiels, vu l'absence de prise en compte du contexte. C'est ce qui peut expliquer que l'opération duale ne lui soit pas préférée : au lieu d'expanser la requête en ajoutant les équivalents, il s'agirait de réduire les représentation des documents en regroupant les mots par groupe de synonymes. Cela serait incontestablement avantageux en termes de volume mémoire occupé et de rapidité de calcul... mais la déformation engendrée par ces synonymies hors contexte est trop grande pour être généralisée et enregistrée au niveau même de la base.

La définition d'une Association n'est plus liée à cette distinction, somme toute assez délicate, entre synonymie hors *vs* en contexte. Cette unité descriptive n'acquiert toute sa consistance que réutilisée par des unités descriptives supérieures, qui la contextualisent. C'est *a posteriori*, au vu de sa reprise plus ou moins marginale ou générale par d'autres unités descriptives, que l'Association se situe comme plus ou moins sensible au contexte. Une Association pertinente est utilisée dans les descriptions, et son domaine d'utilisation est exprimable par son intégration dans une unité de type Communauté.

b) Les champs sémantiques

Le champ sémantique d'une unité servant à indexer les documents (en général, un mot ou une expression extrait des textes) est constitué par l'ensemble des unités qui apparaissent dans les mêmes contextes que cette unité[143]. Chaque unité faisant partie du champ sémantique est *pondérée* par sa fréquence moyenne d'apparition sur l'ensemble des contextes ; certaines unités peuvent au besoin être *éliminées* (champ sémantique trop grand, rôle marginal d'une unité qui n'a qu'une cooccurrence faible).

La définition des contextes a été fixée soit au document, soit au paragraphe, soit à la phrase.

Les champs sémantiques peuvent être utilisés pour déduire des relations entre les unités, et les différencier ou les regrouper. Ces manipulations reposent sur les deux idées suivantes : *(i)* les rapports d'*inclusion*[144] au niveau des champs sémantiques expriment des rapports de généricité / spécificité au plan des unités ; *(ii)* la *similarité*[145] (resp. la dissimilarité) entre deux champs sémantiques traduit l'équivalence sémantique (resp. la distinction au plan sémantique) des unités ayant ces champs sémantiques. La similarité sert par conséquent à différencier des homographes, à reconnaître des expressions composées (qui ont un sens autonome vis-à-vis de celui de leurs constituants), à

[142] (Voorhees 1994) fait une analyse complète et critique des techniques et résultats liés à l'expansion de requête. Cette expansion serait à concevoir avec parcimonie, si l'on ne veut pas dégrader les résultats au lieu de les améliorer. Ainsi, il faut en limiter l'usage : aux requêtes vraiment courtes, en ne recourant qu'aux relations lexicales les plus étroites et les plus directes (si l'on utilise un thesaurus ou un réseau lexical), et si possible en s'appuyant sur des données relatives à la requête elle-même, à la base, aux résultats antérieurs.
Observation importante : la tactique qui consisterait à n'ajouter un terme que s'il a un lien à plusieurs termes de la requête initiale ne se révèle pas satisfaisante. On peut penser que cela est dû à la fois à la pauvreté des requêtes considérées (très courtes) et à l'inadéquation de la structure et de la composition du thesaurus pour cela. En effet, les termes candidats à l'expansion sont des hyperonymes, jugés trop généraux et polysémiques.
De toutes façons, il ressort que même le recours à la meilleure procédure d'expansion se montrerait inférieur à l'utilisation d'une indexation initiale plus contextuelle : « Note however, that methods that exploit statistical relations but do not expand the query, such as Latent Semantic Indexing [...] have been more successful. » (*ibid*., p. 61)

[143] C'est ici la définition au sens de (Fluhr, 1977, §III.4.2.3).

[144] Le degré d'inclusion du champ sémantique A dans le champ sémantique B est caractérisée par la vérification de deux conditions sur les tailles (la taille est donnée par le cardinal au sens des sous-ensembles flous) : *(i)* la taille du champ sémantique A est significativement inférieure à celle de B ; et *(ii)* la taille de la partie de A regroupant les éléments qui ne figurent pas dans B est négligeable devant la taille totale de A.

[145] Christian Fluhr propose de prendre la mesure de Tanimoto (rapport de l'intersection à l'union), au sens des ensembles flous (le degré d'appartenance d'un élément au champ sémantique est donné par son poids dans ce champ sémantique) (Fluhr 1977, §III.4.2.3).

introduire des liens de synonymie et des liens d'association plus large qui ne relèvent pas du rapport générique / spécifique.

Les champs sémantiques peuvent être encore davantage intégrés à la représentation, en devenant la traduction de chaque unité relevée dans le texte pour la caractérisation du texte. Concrètement, dans une représentation vectorielle, le vecteur décrivant les unités relevées dans le texte est multiplié par la matrice des co-fréquences, pour devenir le vecteur représentatif du texte. L'effet recherché est celui d'une *explicitation* et d'un *renforcement* du contexte vers lequel converge l'ensemble du texte. Si le texte n'est en fait que quelques mots-clés, le passage par les champs sémantiques introduit dans la représentation ce que ces mots-clés évoquent dans le cadre du système. Si le texte est plus long, sa cohérence (présumée) permet de sélectionner comme interprétation du texte les contextes partagés par la plupart de ses unités.

3. Une lecture des opérateurs documentaires (TOPIC) comme explicitation de structures linguistiques et artéfacts dus à la modélisation

a) Le choix de TOPIC comme référence

Les opérateurs de TOPIC sont à l'état de l'art de l'interrogation par équation de recherche

Le système TOPIC est un logiciel de recherche documentaire très connu, très utilisé, et très complet.[146]

TOPIC fait partie des logiciels les plus cités dans son domaine. C'est un outil commercialisé, sur le marché international, et qui est en mesure de faire valoir son utilisation de longue date dans des grandes entreprises et des organismes prestigieux Il est intégré, comme moteur de recherche, dans la plupart des offres de gestion électronique de la documentation (GED). Il se positionne différemment des moteurs Web. Ces derniers mettent en œuvre les techniques les plus frustes, tant pour l'indexation que pour le calcul des documents répondant à une requête ; leur force est dans leur simplicité (de fonctionnement et d'utilisation) et dans leur efficacité sur les bases textuelles extrêmement volumineuses. TOPIC, lui, offre toute une palette de modes d'interrogation, classiques et avancés, correspondant mieux aux attentes des professionnels de la documentation et de l'information. TOPIC peut également intégrer un dictionnaire de synonymes « maison » et des traitements linguistiques, et séduit ainsi au plan technique par sa compatibilité avec ces traitements considérés comme les derniers perfectionnements dans le domaine.

Ce qui est mis en avant quand il est question de TOPIC, mais qui ne nous intéresse pas en tant que tel ici, ce sont les concepts[147] (*topic*), qui sont des (éléments de) requêtes, conçus et enregistrés par certains utilisateurs, et réutilisables par tous[148]. Cette approche, qui demande un lourd investissement pour disposer d'une base de *topics* efficaces et cohérents, et qui prédétermine un ensemble de points de vue et de connaissances *a priori*, n'entre pas dans la logique de DECID.

En revanche, toutes les formes d'interrogation dans TOPIC sont basées sur l'utilisation d'opérateurs[149], ce qui explique la richesse du jeu d'opérateurs proposés. Grosso modo, TOPIC a tous les opérateurs déjà imaginés, c'est-à-dire non seulement les classiques opérateurs booléens de

[146] Pour une présentation introductive et peu technique, voir (Collas & Chartron 1994).

[147] Le terme de *concept*, fort bien reçu actuellement, est utilisé à toutes les sauces : ici élément de requête documentaire, ailleurs expression ou ensemble de synonymes...

[148] (McCune & *al*. 1985) explique la démarche des concepteurs de TOPIC, qui s'appelait alors RUBRIC : concevoir les requêtes comme un assemblage (donc réutilisabilité et partage de stratégies d'interrogation) de règles logiques (donc clarté, facilité pour faire évoluer) manuellement pondérées (donc booléen 'flou', permettant une correspondance partielle requête - document, et reflétant le caractère vague et nuancé des intérêts).

[149] Même la requête simple qui ne consiste qu'en un seul mot fait intervenir au moins un opérateur : car soit le mot en question désigne un concept préenregistré, lui-même structuré avec des opérateurs, soit TOPIC interprète la requête comme s'il y avait la présence implicite de l'opérateur `RACINE`.

conjonction, de disjonction et d'exclusion (respectivement ET, OU, NON ou SAUF), mais aussi des opérateurs avec pondérations, des opérateurs sur les chaînes de caractères (troncatures), des contraintes de proximité (distance entre les mots dans le texte), etc.

Approche adoptée

Délimitation de l'étude

C'est sur le lexique et la syntaxe de la représentation fondamentale des requêtes[150] dans TOPIC que se porte notre attention. En effet, cela détermine entièrement la nature des représentations des documents et des requêtes, et l'information gardée pour les calculs de mise en relation. On ne considère ici que les requêtes (et parties de requêtes) qui portent sur le texte intégral, accessoirement sur des champs textuels ; les opérateurs de type relation d'ordre sur des valeurs chiffrées contenues dans des champs factuels ne nous intéressent pas dans le cadre de cette étude.

Déroulement et objectifs suivis

En suivant une typologie des opérateurs (en fonction de leur comportement et par niveau croissant de détermination), il s'agit d'interpréter la signification prêtée à chaque opérateur, de recenser ses cas typiques d'utilisation (notamment à partir des exemples du Guide de l'utilisateur), de voir si cela traduit des caractéristiques linguistiques ou textuelles –et lesquelles–, enfin de vérifier que toutes les propriétés linguistiques et textuelles ainsi communément reconnues comme pertinentes dans le cadre de la recherche documentaire ont été prises en compte dans le système des unités descriptives mis au point pour DECID. L'enjeu n'est pas d'intégrer systématiquement tout ce qui existe dans TOPIC, mais d'avoir examiné, avec un recul critique, un modèle largement représentatif des systèmes de recherche documentaire classiques évolués et fonctionnant sur du texte intégral.

Remarques

Une précision de vocabulaire nécessaire : nous appelons ici *opérateur* toute fonction utilisée pour la construction de la requête. Cela recouvre à la fois les *opérateurs* et les *modificateurs* de TOPIC. TOPIC appelle *modificateurs* des opérateurs qui ne peuvent être utilisés que dans le contexte de certains opérateurs. Cette distinction, qui s'avère plus syntaxique que sémantique[151], n'est pas pertinente pour notre description.

Les exemples de requêtes sont tous tirés du Guide d'utilisation (TOPIC 3.1). Pour la clarté de l'exposé, ils ont été simplifiés (ce sont en général des parties de requêtes plus développées) pour se centrer uniquement sur la relation entre un opérateur et ses arguments.

b) L'exclusion : NON

L'opérateur d'exclusion se traduit conventionnellement en anglais par NOT, en français par NON, ou SAUF, qui traduit plutôt la valeur composite de ET + NON[152]. C'est un opérateur hérité des systèmes de recherche documentaire interrogeant par quelques mots-clés ou par des descripteurs issus d'un référentiel général (thesaurus). Il sert à exclure un pan de l'ensemble des résultats. Il s'agit d'éviter un ensemble de documents soit qui interfère en partie avec l'ensemble défini par la requête, et correspond à un *autre* domaine rejeté comme indésirable *de toute évidence*, soit qui représente une sous partie incluse dans cet ensemble et jugée *à coup sûr trop* spécifique.

```
computer crime = ET (0.50 « computer crime », 0.50 « computer crime
indicators », NON 0.50 « reported crimes »)
computers = ET (apple, dec, NON ibm)
ET (coke, NON coal)
```

(on veut *coke* au sens de *Coca-Cola* et non du produit intermédiaire créé lors de la production du charbon)

[150] Cela inclut donc les fameux concepts, puisque ce sont des éléments de requête.

[151] Les restrictions contextuelles semblent plus liées à une limitation volontaire des calculs qu'à une impossibilité de donner un sens à ces modificateurs dans les contextes exclus.

[152] La documentation française de TOPIC ne permet pas de savoir quel est le nom choisi pour cet opérateur, puisque dans le texte elle parle de l'opérateur SAUF, et dans les exemples elle le représente par NON.

Du bon usage de l'opérateur d'exclusion : les abus dangereux

Les professionnels de la documentation et de l'information sont beaucoup plus prudents et réticents quant à l'usage de cet opérateur que le néophyte. Ils en connaissent les effets sournois et irréversibles : ce qui est exclu disparaît des résultats sans laisser de trace, c'est comme si les documents correspondants n'existaient pas, et cela conduit fréquemment à une image déformée de la réalité (du moins celle du contenu du fonds documentaire). En termes documentaires, on dira que la négation est source de *silence*. En somme, on *maîtrise* mieux l'impact d'une expression positive que d'une expression négative d'un thème de recherche. Du côté de l'utilisateur néophyte, l'utilisation de la négation apparaît dans bien des cas comme une échappade, un raccourci, une solution de facilité, préférée au développement plus explicite et plus précis d'une requête trop générale. Son abus vient d'une difficulté de formulation : sans trop d'efforts, on réduit ainsi un volume de résultats sinon intraitable.

Deux dimensions régissent les situations favorables ou non à l'emploi d'un opérateur de négation.

Nuisible s'il s'agit d'explorer systématiquement les documents d'une base et de faire une recherche bibliographique aussi exhaustive que possible, la négation est indéniablement efficace pour cerner rapidement une information, dont on souhaite retrouver une mention dans un document quelconque. L'opérateur de négation se justifie quand on renonce à interpréter le résultat de la requête comme une image du nombre et de la diversité des documents de la base concernés par le thème de recherche.

Tout dépend également du mode d'indexation, qui détermine le type d'indicateurs sur lequel porte l'opérateur. La négation s'emploie très bien pour des systèmes de description *fermés* et correspondants à une description *synthétique* de chaque document. Par exemple : code(s) de classement (renvoyant à un plan de classement) ; descripteur, issu d'un thesaurus, et affecté au terme d'une analyse documentaire, indiquant qu'il reflète un aspect « important »[153] du document. Le fait qu'il s'agisse d'un système de description fermé permet de cerner au mieux le domaine couvert par le code ou du descripteur, grâce au contexte formé par les autres valeurs possibles : tout choix se présente comme une préférence par rapport aux autres valeurs disponibles et non retenues. S'il s'agit d'alternatives et qu'une seule valeur puisse être choisie (par exemple : classement dans un seul rayon de la bibliothèque), la négation est l'envers d'une formulation positive, et perd son caractère d'indétermination. A la condition de fermeture du système de description, nous avons ajouté celle de description synthétique. En effet, sans cette condition, rien n'exclut qu'un document répondant bien au thème recherché ne disparaisse parce qu'il fasse cas du descripteur interdit, même ponctuellement et marginalement. Toutefois, même dans ce cas de figure le plus favorable à l'usage de la négation, on perd les documents pour lesquels l'aspect rejeté est un aspect dominant, mais qui comporteraient néanmoins un petit développement sur la question considérée.

Le cas de figure opposé –système de description ouvert, représentation non synthétique–, et qui se trouve donc être celui où l'opérateur de négation est le plus fortement générateur de *silence*, correspond à la recherche sur le texte intégral. Les facteurs d'exclusion malheureuse prolifèrent : perception souvent incomplète des cas d'usage d'une expression (on croit exclure tel aspect, on exclut en fait beaucoup plus) ; mise à plat du texte (rejet alors que le terme est utilisé dans un exemple, ou est justement mentionné par l'auteur comme « ce qui ne sera pas abordé dans ce document ») ; décontextualisation des mots qui servent de critère, etc.

Pour DECID : un renoncement justifié

DECID travaille à même les textes, ne s'enferme pas dans un référentiel clos *a priori*, et cherche à ne masquer aucune possibilité de mise en relation, aussi discrète et inattendue soit-elle – c'est même précisément là sa raison d'être. D'après ce que nous venons de voir, cela est de mauvais augure pour l'introduction d'un opérateur de négation. Mais surtout, pour DECID qui s'appuie sur

[153] Un aspect important, ce peut être tout à la fois : un aspect central, dominant (point de vue interne du document) ; original, novateur (point de vue de la base, intertextuel) ; correspondant aux préoccupations des personnes interrogeant ce fonds documentaire (point de vue des besoins, des pratiques dans lesquelles les documents sont utilisés) ; etc.

une interrogation par un texte, disparaît l'essentiel des motivations d'emploi d'un opérateur de négation. La difficulté d'explicitation de la requête n'a plus cours : le texte fournit d'emblée une requête riche. Le besoin d'éliminer un domaine voisin avec lequel il y a quelques interférences de vocabulaire ? Le texte de requête fournit un contexte qui dessine le domaine concerné. Quant à la restriction du domaine, en écartant des parties trop spécifiques, le domaine n'est plus déterminé à l'échelle d'une discipline, trop large, et qu'il faut ensuite restreindre ; il est représenté par un texte, qui reflète ce qui est pertinent dans une pratique donnée, et « donne le ton » quant au niveau de détail attendu. Bref, pas d'opérateur de négation dans DECID, mais DECID n'y perd rien, s'il n'y gagne.

c) *L'équivalence : QUELCONQUE, et tous les opérateurs de reformulation*

L'opérateur QUELCONQUE de TOPIC est l'équivalent d'un OU booléen. Il sert à énumérer des unités que le moteur normalement distingue, mais dont on indique ainsi la parfaite équivalence. Certaines équivalences, au lieu d'être détaillées en explicitant tous les cas, peuvent être représentées de façon synthétique par la mention de la transformation qui permet de rapprocher toutes les formes équivalentes, qui instaure cette équivalence. C'est le fait des opérateurs suivants, dits *opérateurs d'extension de feuille*[154], qui ne s'appliquent qu'aux chaînes de caractères :
(cf. tableau page suivante)

[154] Les requêtes et les concepts (*topics*) ont une structure arborescente ; les feuilles en sont les chaînes de caractères à chercher dans le texte intégral.

Opérateur	Caractéristique dont les variations sont neutralisées	Exemples dans le Guide d'utilisation
par défaut (annulé par MAJ/MIN)	Casse (*i.e.* distinction minuscule / majuscule) Rq. : l'opérateur MAJ/MIN ne peut qu'être associé à l'opérateur MOT ou à l'opérateur TRONCATURE.	APPLE → APPLE, Apple, apple... MAJ/MIN MOT Apple → Apple (mais pas apple ni APPLE) MAJ/MIN MOT Bush → Bush MAJ/MIN MOT UNITED → UNITED (mais pas United ni united)
RACINE, entre apostrophes, et par défaut (annulé par MOT, ou entre guillemets)	Suffixe ou terminaison	finance → finance, financed, financing,... bank → bank, banks, banking,... 'paragraph' → paragraph, paragraphs,... RACINE assert → assert, assertion, asserted,... ank" → bank (mais aussi Bank, BANK,...) MOT boeing → boeing (mais aussi Boeing, BOEING,...)
TRONCATURE	Variations d'écriture dans un mot, sur : une lettre à une position donnée, une suite de lettres de longueur indéterminée (sauf en début de mot), une terminaison en excluant une terminaison interdite	(Exemples illustratifs de la syntaxe : !an → ran, pan, lan, wan,... corp* → corporate, corporation, corporal, corpulent,... c[auo]t → cat, cut, cot. c[a-r]t → cat, cbt, cct,...,cqt, crt. micro[^chip] → micron, micros,... mais pas microchip.) TRONCATURE mil* → mil, mil-spec, military... TRONCATURE air* → air, aircraft, airport, airspace,... TRONCATURE pharmac* → pharmaceutical, pharmacology, pharmacodynamics,...
CONSONANCE	Analogie phonétique... très large. Rq. : « Vous remarquerez que l'opérateur génère de meilleurs résultats lorsque les mots utilisés sont courts. » (TOPIC 3.1, p.12-10)	CONSONANCE airplane → airplane, airplanes, airflow, arrival, aeroflot... CONSONANCE preference → preference, preferential, preferable, preferred,... CONSONANCE capital → capital, capitalism, capitalization, capitalize,...
TYPO	Chaînes de caractères voisines (une lettre différente, une inversion, etc.) (calculées sur la base)	TYPO color → color, colour,... TYPO defense → defense, defence,... TYPO labor → labur, labro,... TYPO grey → grey, gray,...
SYNONYMES	Equivalence sémantique enregistrée dans un dictionnaire (en l'occurrence ici le *Random House Thesaurus* (!))	SYNONYMES altitude → loftiness, tallness, pitch, height, elevation. SYNONYMES purchase → purchase, buy, acquire, get, acquisition, procure,... SYNONYMES admire → admire, value, prize, esteem,... SYNONYMES negotiation → negotiation, arbitration, adjudication, debate,... SYNONYMES security → security, invulnerability, impregnability, immunity,... SYNONYMES protection → protection, defense, asylum, sanctuary,...
SUGGESTION	Equivalence distributionnelle (calculée sur la base, fondée sur la cooccurrence dans un même document)	SUGGESTION aviation → aviation, avionics, aerospace, flight,... SUGGESTION su → su, super user, root user,... SUGGESTION advocate → advocate, proponent, patron, activist,...

Des opérateurs qui s'interprètent linguistiquement

Ces opérateurs, qui représentent une série d'équivalents, s'interprètent comme des opérations de normalisation, par rapport aux dimensions linguistiques suivantes :

- la matérialité graphique, l'écriture sous forme d'une suite de lettres : la correspondance majuscule / minuscule pour chaque lettre, utilisée pour interpréter des variations contextuelles (majuscule en *début de phrase*, expression ou titre *mis en valeur* par le passage en capitales), mais bloquée pour l'initiale d'un *nom propre*, et parfois les lettres détachées qui forment un *sigle* ; la possibilité, en fonction du contexte, de rétablir une *erreur sur un caractère*, une faute de frappe, en détectant l'anomalie (mot qui n'existe pas, mot qui ne s'insère pas dans le contexte) et en trouvant la forme proche ajustée.
- l'expression phonétique associée : le phénomène d'*homophonie* (rendu possible par la possibilité d'écrire différemment des sons analogues), entre deux mots existants (confusion possible en utilisant une écriture pour l'autre), ou entre un mot tel qu'il est enregistré dans le dictionnaire et une transcription approximative de l'oral se soldant par une erreur orthographique[155].
- la morphologie, avec les variations flexionnelles (*accord*, *conjugaison*) et dérivationnelles (*mots de la même famille*) : TOPIC se place dans le cas des langues dont les variations morphologiques interviennent essentiellement sur la *fin des mots*.[156] Deux approches sont prévues. Pour plus de sûreté, on peut s'appuyer sur le fait que l'ensemble des terminaisons forme un système défini (décrit par la grammaire), et que l'on retrouve les terminaisons sur les divers mots de la langue. Pour plus de flexibilité, on peut se contenter de préciser juste la partie de mot qui reflète l'élément cherché, indépendamment des manière dont elle peut ensuite être prolongée pour former un mot.
- la sémantique : le phénomène de *synonymie*, tout au moins pour la manière dont il est perçu dans les systèmes de recherche d'information : des mots différents peuvent exprimer une même idée[157].

---

[155] Accorder à TOPIC la prise en compte de la dimension phonétique est peut-être généreux. De nombreux indices donnent à penser que les mots rapprochés par l'opérateur `CONSONANCE` (dont le nom oriente vers une interprétation phonétique des rapprochements) ne le sont que sur un critère de proximité graphique, entre chaînes de caractères, sans nécessairement tenir compte des voisinages de lettres ou de groupes de lettres en termes de sonorité.

Ainsi, `CONSONANCE` et `TYPO` sont régulièrement comparés, en soulignant le fait que ce qui caractérise chacun c'est un calcul sur la base courante pour `TYPO`, par opposition à des associations prédéterminées pour `CONSONANCE`. Cela conduit à abandonner, non sans une certaine déception, l'interprétation naturelle qui consiste à penser l'opérateur `CONSONANCE` comme concernant les sonorités, et l'opérateur `TYPO` comme concernant l'écriture graphique.

« L'opérateur `TYPO` est utilisé pour lancer des extractions similaires à celles lancées avec l'opérateur `CONSONANCE` pour des mots similaires à ceux figurant dans des documents de votre base de données TOPIC. L'opérateur `TYPO` diffère cependant de l'opérateur `CONSONANCE` dans le sens où les mots détectés à l'aide de l'opérateur `TYPO` sont *spécifiques* aux documents de votre base de données TOPIC, alors que ceux détectés à l'aide de l'opérateur `CONSONANCE` sont des mots de même consonance plus génériques. » (TOPIC 3.1, p. 7-19)

D'autre part, certains exemples, ainsi que les copies d'écran affichant les candidats sélectionnés pour leur affinité de type `CONSONANCE`, montrent que l'on est parfois très loin d'une analogie phonétique (et l'expansion proposée pour la requête est pour le moins effrayante) :

`Call → Colo, Cole, Cola, Col, coal, clue, Cl, Chili, Chile, call, Call`

Le lecteur informaticien repérera aussi que la fonction, sur l'interface anglophone, est intitulée `SOUNDEX` : la fonction utilisée serait-elle tout bonnement la commande UNIX de ce nom ?

[156] Certains moteurs de recherche du Web utilisent le fait que, pour les langues qu'ils considèrent (essentiellement l'anglais), la racine est de longueur limitée, et que les variations flexionnelles se positionnent à la fin des mots, pour développer leurs index en tronquant systématiquement la fin des mots de plus de dix caractères.

[157] Il est remarquable que, au vu de ses exemples, TOPIC ne s'en tienne pas à des équivalents de même catégorie morphosyntaxique. Il n'hésite pas à proposer comme « synonymes » d'un verbe d'autres verbes, mais aussi des noms, et pourquoi pas des adjectifs, etc. Il est vrai que le système dérivationnel « pauvre » de l'anglais favorise les passages insensibles d'une catégorie à une autre ; et que peut-être la procédure de construction du dit *Random House Thesaurus* (offert par TOPIC comme dictionnaire de synonymes) pourrait bien n'avoir mis en jeu quasiment aucune connaissance linguistique... Il n'empêche que la mise en œuvre de cette notion de synonymie, déliée de l'emprise de la morphosyntaxe et au service d'une sémantique unifiée, mérite d'être soulignée.

L'équivalence peut être soit établie *a priori* et *dans l'absolu*, hors contexte, enregistrée dans un dictionnaire ; soit déduite des usages observés dans la base : deux mots sont considérés analogues s'ils sont employés dans les *mêmes contextes*.

Usages de la relation d'équivalence

Le premier cas d'usage des opérateurs d'équivalence est pour exprimer des ensembles de *variantes*, *ad hoc* ou définies linguistiquement. Mathématiquement, l'opérateur QUELCONQUE correspond à une définition *en extension*, les autres opérateurs qui traduisent des modulations autour d'une chaîne de caractères correspondent à une définition *en compréhension*.

```
QUELCONQUE (text, document)
```

La documentation TOPIC illustre deux autres usages de l'opérateur QUELCONQUE. Ces usages se comprennent en opposition avec le recours à l'opérateur OU, qui permet d'introduire une pondération. L'utilisation de QUELCONQUE est par conséquent l'expression que chacune des alternatives qu'il réunit suffit pleinement à elle seule, que c'est une alternative que l'on veut considérer au même titre que les autres, et qu'il n'y a pas de raison de diminuer l'importance de telle ou telle qui serait moins fiable[158]. Les deux usages annoncés répondent à ces critères : c'est l'énumération d'une *liste de noms propres ou de désignations univoques* (références techniques), qui peuvent apparaître conjointement ou de façon indépendante, sans que cela joue sur l'intérêt a priori du document ; et le second usage est la fusion de deux (sous-)requêtes, portant sur des aspects indépendants, mais que l'on veut réunir au niveau des résultats pour ne pas avoir à *dédoublonner* indépendamment chaque recherche partielle.

```
Boeing companies = QUELCONQUE (boeing computer services, boeing aerospace,
boeing defense)
```

L'équivalence dans DECID

La modélisation en unités élémentaires et unités descriptives prévoit plusieurs formes d'équivalences. Il y a deux formes de représentation d'équivalences strictes, sachant que d'autres formes d'équivalence existent ; comme celles-ci correspondent davantage à d'autres opérateurs (en particulier CUMUL), elles seront abordées à ces occasions.

La construction des unités élémentaires peut déjà opérer des équivalences, mais qui correspondent à des réductions systématiques et irréversibles, ce qui est très rarement le cas. C'est la manière par exemple d'introduire un traitement en typographie pauvre, si le corpus est très irrégulier quant à l'usage des majuscules et des accents.

Au niveau des unités descriptives, les Assimilations permettent de tester une équivalence forte, telle que aucune différence n'est faite par la suite entre les différentes unités ainsi regroupées. Il est essentiel de voir qu'une Assimilation peut être proposée par un module du traitement, puis infirmée par un autre. Cela peut correspondre au fait, pour TOPIC, de prévoir des opérateurs qui inhibent les réductions sinon faites par défaut. Par exemple, on exprime que l'écriture toute en minuscules n'est pas pertinente pour un nom propre, ou que c'est précisément telle expression que l'on cherche, et non d'autres formes de la même famille.

En ce qui concerne les variations reconnues par TOPIC, DECID leur accorde des places diverses actuellement. *A priori* cependant, aucune de ces équivalences n'est à établir dès les unités élémentaires, puisqu'aucune n'est systématique. En effet, dans TOPIC, il faut choisir d'*ajouter* un opérateur, ou, quand il s'agit d'une réduction par défaut, elle peut toujours être inhibée (un opérateur est prévu).

Les équivalences phonétiques ou les proximités entre chaînes de caractères sont utiles pour les contextes où il faut prévoir des erreurs de type fautes de frappe ou d'orthographe (textes acquis par OCR, mauvaises conditions de saisie, etc.). Ces cas ne se présentent pas actuellement pour DECID. Autant éviter ces calculs d'équivalence quand le contexte le permet, car ces procédures alourdissent les temps de traitement. Si cependant il fallait introduire ces équivalences dans DECID,

[158] La disjonction non exclusive peut en effet être utilisée pour marquer l'*indifférence* (et unir deux ensembles de résultats), ou encore l'*incertitude* (l'utilisateur hésite sur ce qui va être effectivement utilisé par le système, dans l'indexation, dans les documents) (Denos 1997, §III.2.1.1, pp. 80-81).

la tactique envisagée serait de calculer des rapprochements phonétiques pour les unités élémentaires qui n'apparaissent que dans un seul document ou pour un seul rédacteur (présomption de faute d'orthographe), et des rapprochements graphiques pour celles dont l'occurrence est anormalement rare (une seule occurrence par exemple) (présomption de faute de frappe ou d'erreur de reconnaissance, qui a touché quasi aléatoirement telle occurrence). Ces corrections, que nous jugeons hypothétiques telles qu'elles sont proposées par un simple calcul de similarité non contextuel, seraient transcrites par des Assimilations (la correction doit être confirmée ; ensuite, on ne choisit plus voir aucune différence entre les formes erronées et les formes correctes[159]). Des algorithmes pour mettre en œuvre de tels rapprochements ont été mis au point et seraient directement exploitables[160].

L'équivalence sur la casse est un des motifs de création d'Assimilations dans DECID. Selon les usages observés sur le corpus, l'Assimilation est créée ou non (il n'y a pas d'Assimilation avec une forme en minuscules pour un nom propre ou un sigle).

L'introduction de réductions flexionnelles et dérivationnelles demandent la mise en place de modules, qui opèrent les rapprochements sur la base de considérations morphologiques et sur l'observation du corpus. Les rapprochements ainsi proposés se traduisent par la création d'Assimilations (réductions flexionnelles) et d'Associations (réductions dérivationnelles). En effet, les différences entre des mots d'une même famille sont suffisamment sensibles pour n'être pas complètement occultées. Dans tous les cas, pour ces regroupements calculés, c'est l'usage observé sur le corpus qui confirme ou infirme la réduction.

Aucune traduction d'équivalence stricte n'est prévue dans DECID pour des équivalences sémantiques. La substituabilité d'unités se traduit par des Associations, et les autres liens sémantiques trouvent un mode d'expression dans les Communautés.

DECID permet donc la prise en compte de toutes les variantes linguistiques reconnues dans TOPIC. DECID offre en outre une description plus nuancée. En effet, dans TOPIC, toutes les équivalences calculées sont des équivalences strictes. Dans DECID, l'identité et le rôle de chacune des unités regroupées peut être conservé. Cette possibilité de nuance est importante pour faire la part entre des équivalences sur des variations jugées non significatives (erreurs aléatoires, majuscule dictée par la syntaxe), et des équivalences sur des variations dont chaque forme peut, sans renier l'analogie commune, garder un rôle qui lui est propre (le rédacteur choisit tel mot plutôt que tel autre, adopte tel point de vue et non tel autre).

Dédoublonnage

L'interrogation par un texte (dans DECID) fournit des documents proches du texte soumis dans son ensemble, mais aussi des documents proches sur des aspects qui ne sont abordés qu'en certaines parties du texte considéré. On peut donc avoir dans les résultats des rapprochements calculés sur des caractéristiques différentes, et présentés dans un unique ensemble de résultats, structuré globalement.

---

[159] Cet effacement des erreurs est un choix interprétatif. Dans d'autres contextes, on pourrait vouloir par exemple garder trace d'une non maîtrise de l'orthographe de certains mots (« fautes d'usage »), et soit enregistrer les correspondances de formes en préservant l'identité de chacune (le type d'unité descriptive qui peut servir à cela est l'Association), soit même garder ces unités comme complètement distinctes.

[160] Ces algorithmes font appel : à des considérations de fréquence et de statistiques sur mots, digrammes ou trigrammes ; à des dictionnaires (phonétiques) associés à une métrique pour confronter type (en tant que référence) et occurrence (en tant que déformation possible du type) ; à des connaissances en morphologie ; à l'étude de l'ergonomie du clavier (disposition et utilisation des touches) ; au rôle dissymétrique des consonnes et des voyelles, du début et de la suite du mot ; à des critères de longueur, de simplicité, de répétition (Salton 1989, §12.2).
Sur cette question, nous renvoyons à (Sabah 1989, §6), (Gilloux, Lassalle, Prigent 1989, p. 46), (Courtin, Genthial, Menezo 1993), (André 1997), références intéressantes et instructives, mais sans que cette liste soit limitative.

d) La préférence : OU (pondéré)

L'opérateur OU de TOPIC rassemble des alternatives non exclusives. Comme pour QUELCONQUE, si plusieurs alternatives se trouvent présentes dans un texte, elles ne sont pas retenues dans leur diversité, elles ne comptent que comme une réalisation. L'opérateur OU ajoute une notion de préférence entre les diverses alternatives prévues (valeurs relatives des poids des alternatives les unes par rapport aux autres), et de fiabilité et d'importance (écart de la valeur maximale à la valeur plafond, 1). Chaque alternative est dotée d'une pondération, qui se comporte comme un facteur multiplicatif du poids de la réalisation effective correspondante[161]. Avec l'opérateur OU, c'est l'alternative qui réalise la pondération maximale qui est retenue.

L'opérateur OU et l'opérateur QUELCONQUE produisent les mêmes résultats de recherche à partir des mêmes éléments ; l'opérateur OU ajoute seulement un classement des résultats par ordre de pertinence[162].

Usages

L'opérateur OU remplace l'opérateur QUELCONQUE pour décrire un ensemble de variantes élargi, de valeurs ou d'importances jugées inégales. Une préférence lexicale rehausse alors certaines alternatives par rapport à d'autres, plus lointaines de l'idée que l'on souhaite représenter, ou moins sûres.

```
jargon PC = OU (  1.00 80286, 1.00 80386,
                  0.80 486, 0.80 386, 0.80 286,
                  0.40 clone)
```

> Les feuilles 80286 et 80386 (décrivant les microprocesseurs utilisés dans les PC) reçoivent automatiquement une pondération de 1,00. Il existe une forte probabilité que les feuilles 486, 386, et 286 se réfèrent à leur concept parent, ces feuilles reçoivent donc une pondération de 0,80. La feuille clone ne se réfère pas forcément aux clones de PC ; elle ne reçoit donc qu'une pondération de 0,40.
>
> (TOPIC 3.2, p. 7-51)

Un cas particulier important est l'équivalence entre une forme développée et une abréviation. Il arrive en effet que telle notation soit ambiguë mais pas telle autre. La notation moins sûre reçoit une pondération qui affaiblit son influence. Ainsi, si on ne trouve pas dans le texte la forme exacte cherchée, mais une forme qui pourrait traduire la même idée (sans que l'on puisse en être sûr), alors

---

[161] Ce n'est donc pas parce qu'une alternative a une pondération maximale pour l'opérateur OU correspondant, qu'elle sera celle retenue. Si elle n'est pas présente dans le texte, la pondération effective de l'alternative devient nulle. Si sa réalisation est elle-même dotée d'un poids très faible, le facteur multiplicatif même grand peut être insuffisant pour revaloriser l'alternative, et une autre alternative, avec une pondération moyenne pour l'opérateur OU multipliée par une forte pondération de réalisation, peut être l'alternative finalement retenue.

[162] L'utilisation de l'opérateur OU ne garantit pas une présentation ordonnée des résultats. Pour autant que nous ayons compris le fonctionnement des opérateurs TOPIC, les résultats avec OU pourront être ordonnés si et seulement si l'utilisateur a indiqué au moins une pondération, ou a utilisé au moins une fois l'opérateur PLUSIEURS ou l'opérateur CUMUL (explicitement dans sas requête, ou implicitement en appelant un concept (*topic*)) ; sans cela, les documents ont peut-être un score, mais ils sont tous *ex æquo*, aussi l'ordre n'est pas significatif.

L'explication fournie par le Guide d'utilisation, au paragraphe *Comparaison entre les opérateurs* OU *et* QUELCONQUE, n'est ni éclairante ni convaincante, et n'apporte ni démenti ni confirmation de notre analyse :

« alors que l'opérateur QUELCONQUE contraint TOPIC à affecter le même score (1,00) à chaque document extrait, l'opérateur OU permet à TOPIC d'affecter à chacun un score différent ; ainsi les documents les plus pertinents pour votre requête sont classés avec un score plus élevé alors que les moins pertinents ont un score plus bas. » (TOPIC 3.1, p. 12-42)

Il y a aussi des contraintes syntaxiques différentes qui opposent OU et QUELCONQUE. QUELCONQUE ne peut être utilisé sur des éléments obtenus par l'application d'un opérateur pondéré ; réciproquement, OU ne peut servir à calculer un élément qui sera réutilisé par un opérateur booléen strict (TOUT, QUELCONQUE). Ceci peut forcer à utiliser un OU alors que l'interprétation donnée à l'opération serait celle d'un QUELCONQUE, sans notion de préférence ou d'importance.

Toutes ces remarques s'appliquent pour les opérateurs ET et TOUT, à la place respectivement de OU et QUELCONQUE. En effet, ET et TOUT entretiennent entre eux et par rapport à l'ensemble des opérateurs les mêmes relations que OU et QUELCONQUE.

cette forme est tout de même prise en compte en remplacement, mais en lui conférant une influence plus limitée.

```
Video = OU (video, vcr)
```

L'opérateur OU peut également s'appliquer non pas à des alternatives lexicales, mais à l'élargissement contrôlé d'un domaine de recherche. La pondération traduit que tel domaine est le domaine central et privilégié à explorer pour la recherche, mais qu'à défaut tel autre domaine, secondaire ou d'intérêt plus inégal pour la recherche, peut être considéré pour poursuivre l'investigation.

```
OU (election, national elections, senatorial race)
```

L'opérateur OU peut être aussi préféré à QUELCONQUE non pas pour introduire des différences entre les éléments réunis, mais pour propager un calcul de pondération, et pour obtenir des résultats ordonnés. Cela est manifeste quand toutes les éléments réunis par l'opérateur OU sont dotés de la même pondération (c'est aussi le cas quand les pondérations sont laissées implicites).

```
Boeing planes = OU (0.50 707, 0.50 727, 0.50 737, 0.50 747, 0.50 757, 0.50
767)
OU (earnings, finance, profits)
```

Le relief dans DECID : influence décisive et utilité

Parmi les unités descriptives, seules les Communautés distinguent des composantes qui jouent un rôle central et des composantes qui jouent un rôle annexe.

Le cas de variantes lexicales, et notamment de variantes de notation, se prête à une représentation par une Association. L'influence inégale entre les membres vient alors de ce que certains peuvent garder une représentation sous forme d'unité Simple, et entrer dans la construction de d'autres unités. Si les comportements des membres sont trop contrastés, l'Association peut être dissoute : rien ne motive dans les faits son maintien.

Critique des pondérations

Dans tous les cas, DECID évite le recours à une pondération chiffrée, dont l'effet réel n'est pas évaluable par l'utilisateur qui doit la fixer, dont le sens se perd au fil des calculs, et qui oblige à ramener l'expression des différences à l'indication d'un ordre unique.

Dans le cas où plusieurs domaines sont parcourus lors de la recherche et qu'il convient de les distinguer, DECID les présente comme autant de pistes possibles, les indique à l'utilisateur. Celui-ci peut ensuite librement choisir d'approfondir ou non telle ou telle alternative, *a posteriori* et en connaissance de cause.

e) L'entière explicitation : CORRESPOND

L'opérateur CORRESPOND s'applique à la valeur d'un champ, pour une fiche structurée. L'opérateur spécifie que le champ doit contenir exactement la chaîne de caractères indiquée.

Ce cas, de réalisation littérale et sans reste, correspond à une indication factuelle qui prend la forme « linguistique » d'une chaîne de caractères, mais ne rentre pas dans le cadre d'une réalité textuelle.

```
Source CORRESPOND computer
```

f) Le renforcement par diversité : CUMUL

CUMUL est l'un des opérateurs les plus intéressants et les plus innovants de TOPIC. Il se situe en ce juste milieu entre la conjonction et la disjonction, entre le OU et le ET, et dont l'absence est si souvent regrettée dans les systèmes booléens. En effet, le OU ne permet pas de traduire que trouver plusieurs indices différents de pertinence dans un même document, c'est *a priori* mieux que n'en trouver qu'un seul ; et le ET, trop exigeant, anéantit sans rémission les documents qui réalisent *presque* tous les indices. L'utilisation de la logique floue, pour rétablir un continuum entre les deux pôles, OU et ET, de la logique booléenne, s'avère une proposition décevante, par la résolution trop purement numérique qu'elle impose. Dans ce contexte, CUMUL se présente comme une solution élégante et efficace.

Le principe de l'opérateur CUMUL, c'est donc de *n'imposer aucun* des éléments qu'il réunit (il suffit qu'un quelconque d'entre eux soit présent pour que l'unité représentée par l'opérateur soit réalisée), et de *valoriser* une réalisation dans un texte d'autant plus qu'elle concerne d'éléments *différents* parmi ceux mentionnés sous l'opérateur.

Comme le OU et le ET de TOPIC, l'opérateur CUMUL prend en compte des pondérations. L'idée est toujours de permettre le renforcement ou l'atténuation de certains éléments par rapport à d'autres, et d'indiquer une importance globale de l'unité représentée par rapport aux autres unités avec lesquelles elle est ensuite combinée.

Usages : l'opérateur de base

Par son comportement très intuitif –favoriser, sans imposer rigidement, la réalisation du plus grand nombre d'indices de pertinence–, CUMUL apparaît comme l'opérateur le plus souple et très souvent le plus adéquat, et est abondamment utilisé.

Il sert à représenter tout ensemble de termes, unis par une certaine affinité sémantique, mais apportant chacun sa contribution à l'évocation d'un thème. On trouve les exemples d'ensembles plus ou moins larges.

CUMUL est utilisé pour représenter des paradigmes (que l'on peut considérer comme des classes minimales, des taxèmes au sens de la Sémantique différentielle unifiée), qui détaillent un aspect central au thème de la recherche ; effectivement, un texte qui fait appel à une grande partie des éléments du taxème aborde le thème avec un certain niveau de détail. Il peut aussi déployer des paradigmes très généraux.

```
Performing-arts = CUMUL (ballet, musical, dance, opera, symphony, drama)
liberal-arts = CUMUL (literature, philosophy, languages, history, art)
```

Un peu plus largement, CUMUL peut regrouper des termes qui dans d'autres systèmes documentaires seraient reliés par des liens *terme générique / terme spécifique*. C'est dire que la mention d'un seul des termes est déjà intéressante (on n'impose pas à un document général d'aborder les notions les plus spécifiques, ni inversement à un document très spécialisé de rappeler explicitement dans quelle problématique générale il se situe), mais que la présence de plusieurs conforte la présomption de pertinence.

```
CUMUL (computers, laptops)
CUMUL (forecasts, economic trends)
```

Enfin, CUMUL peut rassembler des termes qui entretiennent entre eux des rapports de toutes sortes, pas seulement d'équivalence ou de généricité, et qui reflètent chacun une entité d'un même domaine. Là encore, il serait ridicule d'imposer l'apparition de tous ces éléments dans un texte, mais plus il y a de ces éléments présents, plus vraisemblable est l'adéquation de ce document au thème représenté.

```
Boeing = CUMUL (boeing companies, boeing label, boeing planes,...)
CUMUL (stock, forecasts)
```

Des caractéristiques qualitatives reprises par DECID

DECID adopte le critère de diversité (sans exigence d'exhaustivité) comme caractérisation de la réalisation des Associations et des Communautés. Ces différents types d'unités descriptives permettent une description plus fine que l'opérateur CUMUL, dont l'application est extrêmement large. Les Associations permettent de distinguer le cas où les unités rassemblées apparaissent comme équivalentes, et où elles forment un ensemble restreint et homogène ; les Communautés couvrent des unités qui entretiennent des relations de nature plus diversifiées et hétérogènes, mais qui partagent un même point commun sémantique.

DECID ne reprend pas le mécanisme des pondérations pour les raisons déjà évoquées à propos de l'opérateur OU. L'influence d'une unité se traduit par son appartenance ou non au noyau si elle fait partie d'une Communauté, et par l'étendu de ses contributions à la construction d'autres unités. Autrement dit, une unité peu intéressante ou peu fiable ne sert guère de point d'appui pour développer la description.

g) Le renforcement par la fréquence : PLUSIEURS

L'opérateur PLUSIEURS rend compte d'un critère complémentaire à la diversité : la densité de réalisation, mesurée comme le nombre d'occurrences rapporté à la « longueur » du texte. Il semble que ce critère serve *a posteriori*, une fois les documents sélectionnés (et en l'absence d'un score apporté par pondération ?), à ordonner les résultats.

PLUSIEURS computers

> Cette requête permet d'extraire des documents les occurrences du mot computers et d'afficher les résultats classés selon la densité de ce mot dans les documents extraits. (TOPIC 3.1, p. 12-36)

Limites de l'opérateur dans TOPIC

Curieusement, TOPIC limite l'application de l'opérateur PLUSIEURS à un mot (et ses variantes éventuelles), sans l'étendre à un ensemble de mots réunis par un opérateur comme CUMUL ou ET[163]. Il oblige en quelque sorte à répéter pour chaque mot si l'on veut faire entrer en ligne de compte le critère de fréquence, sans qu'il soit possible de l'introduire à un niveau plus global.

> Dans le cadre d'une extraction TOPIC, vous voulez voir apparaître les documents qui contiennent le plus grand nombre d'occurrences de concepts fils de l'expression computer-crime-indicators en tête de la liste de résultats, et les documents qui comportent moins d'occurrences ou qui contiennent les phrases définies par les concepts fils de computer-crime et reported-crimes en fin de liste. Pour ce faire, vous pouvez affecter le modificateur PLUSIEURS aux fils du concept « feuille » du fils computer-crime-indicators, comme indiqué ci-dessous[164] :

```
computer-crime = ET (   0.50 computer-crime,
                        0.50 computer-crime-indicators
                        0.50 reported-crimes)
```

> avec

```
computer-crime-indicators = QUELCONQUE ( PLUSIEURS RACINE virus,
                                         PLUSIEURS RACINE pirate,
                                         PLUSIEURS RACINE hacker)
```

> (TOPIC 3.1, p. 7-46)

Cela est peut-être un moyen de limiter l'usage de cet opérateur, que sa sensibilité rend à double tranchant : autant il peut renforcer un terme fréquent, autant il dévalue un terme dont l'apparition est rare. Or dans la majorité des cas quelques occurrences suffisent à rendre l'usage d'un terme significatif. En outre, on peut craindre, vue la méthode de calcul, que même les termes les plus fréquents se trouvent défavorisés, avec cet opérateur, devant les termes repérés sur leur simple présence (sans l'opérateur PLUSIEURS).

Enfin, prendre la longueur du texte comme facteur de normalisation n'est plus satisfaisant quand on a affaire à des textes de plus d'une page : les fréquences relatives deviennent toutes relativement faibles. Le critère de densité est connu pour ce biais en faveur des documents très courts, et son comportement inégalitaire sur des bases de documents de longueurs variées.[165]

---

[163] Le Guide affirme que PLUSIEURS est compatible avec (c'est-à-dire « peut s'appliquer juste au-dessus de ») tout opérateur à l'exception des opérateurs à pondération (ET, OU, CUMUL). Cependant, les illustrations données n'orientent que vers l'application de PLUSIEURS à un opérateur de feuille, mais jamais à un opérateur d'arité multiple comme les opérateurs booléens strict (proximité, TOUT, QUELCONQUE).

[164] Ce paragraphe est particulièrement embrouillé, non seulement parce que les choix syntaxiques de TOPIC sont mal définis, mais aussi suite à des erreurs manifestes de traduction –notamment le faux-ami *phrase* (anglais), qui signifie en français non pas *phrase* mais *syntagme*, *expression*. Voici quelle pourrait être une formulation un peu mieux adaptée, avec un minimum de changements (indiqués en italiques) :
« Dans le cadre d'une extraction TOPIC, vous voulez voir apparaître les documents qui contiennent le plus grand nombre d'occurrences de concepts fils *du concept* computer-crime-indicators en tête de la liste de résultats, et les documents qui comportent moins d'occurrences ou qui contiennent les *expressions* définies par les concepts fils de computer-crime et reported-crimes en fin de liste. Pour ce faire, vous pouvez affecter le modificateur PLUSIEURS *aux feuilles* du fils computer-crime-indicators, comme indiqué ci-dessous ».

[165] Il nous faut admettre que notre jugement porte sur une formule de calcul que l'on a déduit des indications (vagues) données dans la documentation, sans avoir la certitude que ce soit exactement la formule mise en œuvre. Voici la description la plus précise que nous ayons au sujet du calcul opéré par l'opérateur PLUSIEURS :
« Le modificateur PLUSIEURS évalue la *densité* d'un concept dans un document et génère un score basé sur la pertinence des documents extraits correspondant au concept associé. Plus le nombre d'occurrences du concept

DECID et les fréquences

DECID enregistre une fréquence de réalisation de chaque unité présente dans un texte ; il a aussi accès à la longueur « linéaire » du texte, représentant le nombre d'unités élémentaires initialement découpées. Ces informations sont disponibles pour le calcul de similarité entre textes. La manière de les utiliser est gérée à ce niveau, et peut évoluer avec les versions du moteur de recherche.

h) La dépendance : ET (pondéré)

Le ET est bien sûr le pendant du OU de TOPIC. De même que le ET booléen, il impose, pour être vérifié, que soient présents tous ses constituants. Comme le OU de TOPIC, il permet l'expression de pondérations. Le poids global du ET correspond au minimum des poids réalisés (le OU correspondait au maximum) : c'est en quelque sorte un énergique « nivellement par le bas », un alignement sur le composant le plus faible.

Usages

Le ET stipule parfois la présence nécessaire de toutes les composantes d'un paradigme. C'est un usage rare : il pose un niveau de contraintes très élevé, et de plus on en voit mal le sens d'un point de vue sémantique. Ce genre d'usage vient souvent de réflexes hérités de la pratique d'interrogations booléennes (où le choix des opérateurs est beaucoup plus limité). Le ET est dans ce cas avantageusement remplacé par un CUMUL.

```
Boeing military planes = ET (0.50 B52, 0.50 B47)
```

Le ET est quelquefois utilisé pour opérer l'intersection de deux domaines. Là encore, c'est l'habitude des interrogations booléennes qui induit cet usage, pour lequel l'opérateur le plus adapté est TOUT (qui fonctionne sans pondération).

```
ET (apple, ibm)
ET (pharmaceutical companies, stock)
```

Les usages intéressants du ET sont l'introduction de contraintes contextuelles. En effet, on peut lire dissymétriquement un ET, comme la prise en compte de tel mot si et seulement si tel autre est présent. A travers les exemples du Guide de l'utilisateur TOPIC, on analyse deux tels cas de figure.

Le premier est celui d'une ellipse : on ne va comptabiliser la forme elliptique que si la forme complète est attestée au moins une fois dans le document. Sinon, rien ne permet de penser que la forme elliptique (qui est beaucoup plus générale), reflète la notion recherchée.

Second cas de figure : le repérage d'un élément dans le contexte d'une partie standard, d'une rubrique d'un plan-type, d'une information dont la formulation est conventionnelle. L'un des éléments du ET assure que l'on est bien en présence du contexte où attendre l'information, l'autre contrôle la valeur trouvée.

Dépendance et contexte dans DECID

Ni les Associations, ni les Communautés de DECID n'imposent la présence de tous leurs constituants : car cette exigence n'est pas sémantiquement fondée. En revanche, l'unité n'est reconnue qu'à partir de l'apparition d'au moins deux éléments constituants (dont la présence de termes du noyau, pour une Communauté). La présence de plusieurs constituants assure un contexte minimal suffisant, sans avoir à en imposer à l'avance la teneur exacte. On s'aperçoit d'ailleurs que dans les cas de contrainte contextuelle relevés dans les exemples de TOPIC, un seul autre terme suffit à lever l'équivocité du terme douteux.

C'est ainsi que DECID peut rendre compte des ellipses (si on veut l'enregistrer indépendamment, on utilisera une Association) et des informations conventionnelles (les Communautés, avec de plus les conditions de localité qu'elles introduisent, sont mieux adaptées que le ET de TOPIC).

---

dans le document est important par rapport à sa taille, plus le score affiché dans la liste des résultats pour ce document est élevé. Etant donné que le modificateur PLUSIEURS tient compte de la densité pour l'extraction, un document plus long comportant plus d'occurrences peut avoir un score inférieur à celui d'un document plus court qui comporte moins d'occurrences ; la longueur du texte de chaque document étant prise en considération. » (TOPIC 3.1, p. 7-45 & 7-46)

i) L'exigence : TOUT

L'opérateur TOUT répond à l'opérateur QUELCONQUE. QUELCONQUE est le OU booléen, TOUT est le ET booléen. Il exige donc purement et simplement la présence de tous les éléments qu'il rassemble dans le document.

L'opérateur TOUT et l'opérateur ET produisent les mêmes résultats de recherche à partir des mêmes éléments ; l'opérateur ET ajoute seulement un classement des résultats par ordre de pertinence[166].

Usages

Les usages vus pour l'opérateur ET pourraient se retrouver pour l'opérateur TOUT.

D'une manière générale, l'opérateur de conjonction est une manière d'augmenter la précision d'une recherche. Il sert à pointer plus particulièrement une partie d'un domaine, ou à n'explorer que l'intersection de deux domaines.

```
Boeing label = TOUT (Boeing, Company)
TOUT (europe, foreign exchange)
TOUT (business, european)
```

Le document comme contexte

Le TOUT peut être vu comme une condition de cooccurrence dans un même contexte, qui se trouve être ici le document. La pratique montre que la condition imposée par l'opérateur booléen est rapidement trop forte si l'on augmente le nombre de termes, aussi n'est-elle utilisable que sur un petit nombre d'éléments (de l'ordre de deux ou trois). Cela conduit soit à limiter les indices de pertinence mentionnés, soit à développer une combinatoire fastidieuse (on remplace « A ET B ET C ET D » par « (A ET B) OU (A ET C) OU (A ET D) OU (B ET C) OU etc. »).

Les unités descriptives que sont les Arrières-plans ont les qualités de l'opérateur TOUT sans en avoir les inconvénients... Elles définissent un contexte textuel sans avoir à en expliciter *a priori* toutes les formes possibles.

En outre, les Arrières-plans, en tant que Communautés, manifestent leur analogie avec les autres Communautés qui s'intéressent aux autres types de contexte, alors que dans TOPIC rien n'indique cette analogie entre l'opérateur TOUT et les opérateurs suivants, PARAGRAPHE et PHRASE. Pourtant, l'opérateur TOUT s'inscrit dans cette série, comme un impératif de proximité à l'échelle du document.

j) L'impératif de proximité : PARAGRAPHE, PHRASE

Les opérateurs PARAGRAPHE et PHRASE introduisent des conditions de proximité : pour être validés, les éléments qu'ils réunissent doivent figurer dans la même zone de texte, paragraphe ou phrase.

Usages

Le contexte de la phrase rend plutôt compte de rapports syntagmatiques, celui du paragraphe de rapports paradigmatiques. Dans tous les cas, on cherche à déceler un passage qui établit une affinité sémantique, un lien explicite, entre les éléments indiqués.

```
PARAGRAPHE (boeing, defense)
PARAGRAPHE (yeltsin, commonwealth)
PHRASE (wavelets, information storage)
```

Les exemples trouvés illustrent trois cas de figure particuliers.

Le premier est la recherche d'une expression composée « déformable », c'est-à-dire en tolérant les variations comme la séparation de ses constituants par l'introduction d'un élément supplémentaire (par exemple l'ajout d'un adjectif ou d'un adverbe), ou encore l'éclatement de l'unité par la mise en facteur d'un de ses composants dans une coordination.

```
PHRASE (boeing, aerospace, electronics)
```

[166] Cf. remarque faite à ce propos pour l'opérateur OU (par rapport à l'opérateur QUELCONQUE).

Le second cas de figure est la recherche d'un terme générique accompagné d'un terme qui le qualifie ; on s'attend à ce qu'ils soient reliés par un lien syntaxique (donc dans la même phrase), sans pour autant être assuré qu'ils seront immédiatement voisins. Un exemple simple de qualificatif est un adjectif, qui est voisin du nom en construction épithète mais séparé de lui s'il est attribut ; mais ce peut être aussi différentes formes de compléments.

```
PARAGRAPHE (drug, cancer treating)
PHRASE (american, innovation)
```

Le troisième cas que l'on pourrait déceler est le rapport (paradigmatique) entre un mot générique et un mot plus spécifique, tel que l'on peut s'attendre à les trouver dans le même voisinage soit dans des formules définitoires (*Le X est un Y qui...*), soit dans des reprises anaphoriques (le générique sert à reprendre le spécifique en évitant une répétition littérale et en le situant).

```
PARAGRAPHE (drug, fda)
```

Localité et relations syntagmatiques souples dans DECID

DECID prévoit aussi les deux paliers descriptifs du paragraphe et de la phrase (reflétée par la *période*). Le paragraphe correspond à une zone de forte homogénéité sémantique. La phrase est le lieu des relations syntaxiques explicites, directes et indirectes. Une indication de cooccurrence à l'échelle du paragraphe se transpose dans DECID en un Voisinage, et pour la phrase l'unité descriptive correspondante est une Relation.

Cependant, quand il s'agit de la recherche d'une expression malgré quelques variantes, un voisinage plus étroit et une notion d'ordre pourrait mieux convenir. Dans TOPIC, il n'y a pas d'intermédiaire entre l'opérateur PHRASE et l'opérateur EXPRESSION (étudié un peu plus loin), qui précise l'ordre des composants mais exige également leur adjacence stricte. L'unité descriptive Séquence de DECID est adaptée à la représentation de groupements syntagmatiques, recherchant les composants dans un voisinage étroit (dépendance syntaxique immédiate), reflétant l'ordre défini par la syntaxe, et tolérant les constructions qui perdent la connexité de l'expression.

Enfin, DECID affine la définitions de ces zones de localité. Pour le paragraphe, DECID tient compte de la « perméabilité » aux frontières, qui d'ailleurs favorise l'enchaînement d'un paragraphe au suivant. Et plutôt que les phrases (dont la définition, linguistique, est difficile à automatiser), DECID s'appuie sur les périodes, délimitées à l'aide des ponctuations et de la longueur (répondant aux facultés perceptives et cognitives mises en jeu dans la lecture).

k) L'impératif de localisation remarquable : DEBUT, FIN

Les opérateurs `DEBUT` et `FIN` sont prévus pour s'appliquer aux valeurs d'un champ, pour des documents sous forme de fiches structurées avec des rubriques.

Bien que s'appliquant à des chaînes de caractères, ils ne sont pas conçus comme des opérateurs textuels.

```
Reporter DEBUT jack → Jack, Jackson, Jacks,...
Owner FIN ner → Milner, Wagner, Faulkner,...
```

l) L'impératif d'adjacence et d'ordre : EXPRESSION

L'opérateur `EXPRESSION` est le moyen de coder des termes composés dans TOPIC. Concrètement, il se traduit comme une double contrainte : ordre et adjacence.

Usages

`EXPRESSION` s'utilise donc pour spécifier des syntagmes stables (peu sujets à variations) : mots composés, locutions figées.

```
EXPRESSION (apple, computer) → Apple Computer
EXPRESSION (boeing, computer, services) → Boeing Computer Services
EXPRESSION (white, house) → White House
```

```
EXPRESSION (national, aeronautics, « and », space, administration) →
national aeronautics and space administration[167]
```

Les syntagmes et locutions figées dans DECID

DECID intègre bien sûr des types d'unités pour représenter les formes composées et les enchaînements syntagmatiques. Cependant, il s'efforce de faire la part entre les unités tellement intégrées qu'elles ont perdu leur motivation (le sens de l'ensemble s'est éloigné du sens des parties), et les unités dans lesquelles les constituants ont encore une certaine autonomie (par exemple, on trouvera des reprises elliptiques fondées sur un des constituants, isolé). A l'opérateur EXPRESSION de TOPIC répondent donc les unités descriptives Solidarité et Séquences, les secondes étant également plus tolérantes vis-à-vis des variations linguistiques.

m) Bilan

Le parcours opérateur par opérateur permet de tirer quelques conclusions générales.

Puissance expressive

Dans tous les cas examinés, DECID peut rendre compte des phénomènes linguistiques et textuels visés par les opérateurs de requête documentaire : variantes, voisinage sémantique, anaphore par reprise elliptique, expressions composées, majuscule de début de phrase, etc.

Finesse

DECID fait dans plusieurs cas une description plus fine que TOPIC.

DECID distingue par exemple deux formes d'équivalence : une équivalence qui efface les différences entre les éléments, et une équivalence qui préserve l'identité et l'autonomie de chaque composant. Avec les opérateurs, la seule façon de poser une équivalence en gardant une spécificité à chaque terme est de passer par des pondérations chiffrées, ce qui ne permet de penser la différence que sur le mode d'un ordre strict et de rapports de supériorité et d'infériorité. Le sens de ces pondérations se noie rapidement dans les calculs.

DECID prévoit une représentation des expressions (syntagmes, groupes de mots) qui connaissent des variantes de réalisation. C'est un intermédiaire entre l'opérateur EXPRESSION et l'opérateur PHRASE de TOPIC.

DECID se base sur une définition étudiée des zones de localité que représentent la phrase et le paragraphe. L'absence totale de considérations sur la construction, la délimitation et la signification accordée à ces zones dans la documentation de TOPIC trahit l'indigence de la réflexion sur ces points, et le traitement très fruste qui s'ensuit pour la mise en œuvre des opérateurs de proximité.

Artifices imposés par une conception encore booléenne

DECID se libère totalement des contraintes artificielles du booléen.

DECID se passe de l'opérateur d'exclusion (négation), inadapté au texte intégral et nuisible pour l'exploration approfondie d'un corpus.

Le caractère trop restrictif de la conjonction est connu, et conduit à des interrogations sur un contexte indigent, de deux ou trois termes au plus (au-delà, la requête s'avère très vite vaine : plus aucun document n'est sélectionné). Quant à la disjonction (non exclusive[168]), elle oblige à penser les

[167] Pour être exact, le « and » ne figure pas dans les arguments de l'opérateur EXPRESSION dans la documentation française. Serait-ce que la contrainte d'adjacence tolérerait l'insertion de « mots vides » (tels que la conjonction anglaise *and*) ? Le commentaire qui suit l'exemple nous détrompe, et donne à penser que l'exemple a été mal compris du traducteur qui a supprimé le « and » :
« Notez que vous devez placer des guillemets avant et après le mot *and* afin que TOPIC n'identifie pas ce dernier comme étant un opérateur. »
C'est donc que le « and » figurait dans l'exemple original anglais.
Donc l'adjacence serait stricte, sans tolérance pour des mots vides.

[168] La disjonction exclusive, intéressante au plan logique, est dépourvue de sens dans le cadre de l'interrogation documentaire. Absente des opérateurs de TOPIC, elle n'a pas à être considérée dans cette étude.

mots ou thèmes qu'elle réunit comme autant d'alternatives indépendantes, sans leur permettre de se renforcer et de se confirmer mutuellement. Elle multiplie les documents sélectionnés, nivelle des configurations pourtant hiérarchisables, noie les singularités et les présomptions de pertinence les plus fortes dans une confusion submergeante de documents.

Par conséquent, le booléen oblige une explicitation combinatoire et fastidieuse des contextes. Puisque (A et B et C et D et E) est trop restrictif, et que (A ou B ou C ou D ou E) sombre dans les généralités, il faudrait prévoir et décrire toute une gamme de motifs intermédiaires, par exemple : (A et B et (C ou D ou E)), (A et C et E), ((A ou B ou C) et D et E), etc. etc.

Même en introduisant des pondérations, qui transforment les opérateurs booléens QUELCONQUE et TOUT en ET et OU, la représentation reste tributaire soit de la présence obligatoire de tous les éléments (réunis par un ET), soit de la décontextualisation des alternatives du OU qui s'ignorent (dans le calcul des scores, c'est « chacun pour soi »).

Le seul opérateur qui échappe à cette fatalité, et qui a été le coup de génie de TOPIC, c'est CUMUL. Seul à gérer la combinatoire des contextes, il se trouve utilisé à toutes les sauces, et y perd en signification. DECID inscrit un grand nombre de ses types d'unités dans la lignée de CUMUL, mais en déclinant et en précisant les formes textuelles et interprétatives que peuvent prendre les combinaisons d'unités ainsi décrites.

Une syntaxe rigide, quelquefois contre-intuitive et limitante

La synthèse des contraintes d'emploi localement précisées pour chaque opérateur fait apparaître une organisation stratifiée. A partir des chaînes de caractères, les opérateurs s'appliquent successivement comme suit :

- 0 ou 1 opérateur de feuille (MOT, RACINE, TRONCATURE, CONSONANCE, TYPO, SYNONYME, SUGGESTION)
- 0 ou 1 MAJ/MIN si il y a l'opérateur de feuille MOT ou TRONCATURE
- 0 ou 1 PLUSIEURS
- 0 ou 1 opérateur de proximité (EXPRESSION, PHRASE, PARAGRAPHE)
- 0 ou *n* opérateurs booléens (TOUT, QUELCONQUE)
- 0 ou *n* succession de : 0 ou 1 opérateur d'exclusion (NON)
  1 opérateur à pondéré (ET, OU, CUMUL)
- 0 ou 1 opérateur d'exclusion (NON)

Cette séquence rigide fait apparaître bien des incompatibilités regrettables.

Les opérateurs de feuilles sont antinomiques (au sens où ils ne sont pas combinables : par exemple, on ne peut pas demander des variantes orthographiques pour des mots d'une même famille, ou demander les synonymes avec leurs formes dérivées, etc.).

La restriction de MAJ/MIN au contexte de MOT et de TRONCATURE est explicable : on vise les noms propres ou les sigles, qui n'ont *a priori* pas de dérivés (RACINE) ou de synonymes ; mais pourquoi avoir exclu les glissements orthographiques (CONSONANCE, TYPO), quelquefois bien réels pour des noms étrangers ou des sigles apparentés ?

L'opérateur booléen de conjonction (TOUT) fonctionne comme un opérateur de proximité au niveau document, il n'est donc pas gênant qu'il intervienne juste après les opérateurs de proximité plus précis. En revanche, on s'interdit de regrouper des variantes, avant d'imposer une contrainte de proximité, avec un QUELCONQUE, alors qu'on peut le faire, mais seulement pour des regroupements prédéterminés, avec les opérateur de feuille. Autrement dit, les opérateurs d'équivalence sont artificiellement séparés, QUELCONQUE d'une part, les opérateurs de feuille d'autre part, et sont inégaux devant les opérateurs de proximités.

Il y a aussi le rejet de l'opérateur d'exclusion au seul niveau des opérateurs pondérés, alors que paradoxalement la pondération du composant auquel s'applique l'opérateur NON ne joue absolument aucun rôle. Cette anomalie semble une conséquence du point de non-retour apposé à l'usage des pondérations : dès qu'une pondération est introduite, alors tous les opérateurs non pondérés sont interdits. On concevrait bien, pourtant, qu'un opérateur non-pondéré puisse s'appliquer, en ignorant tout simplement les pondérations des éléments qu'il rassemble. La pondération pourrait être utile dans une description locale d'un thème, sans avoir à contaminer ensuite toute la représentation globale. C'est d'ailleurs ce point de non retour qui force un dédoublement artificiel de

la disjonction en QUELCONQUE et OU avec pondérations implicites (neutres), et le même dédoublement artificiel de la conjonction en TOUT et ET avec pondérations implicites (neutres).

Ces rouages de la syntaxe des langages d'interrogation, que l'on peut si facilement gripper, montrent leurs limites dans TOPIC, mais les autres systèmes ne font pas mieux. En effet, l'interprétation des équations s'édifie par un calcul, que seules des limitations de ce type rendent opérationnel.

C'est pourquoi DECID explore une autre voie, celle des combinaisons globales très libres et peu évoluées, l'essentiel de la structuration significative se jouant au niveau de la construction des unités descriptives.

4. Définition de contextes pour le traitement automatique

a) Le choix d'un (et un seul) type de contexte

Un certain nombre de traitements font appel à la définition de contextes. Ainsi, le calcul de cooccurrences, visant à évaluer la force du lien entre certains mots, suppose la définition de zones dans lesquelles les unités (mots) apparaissent (« occurrent ») ensemble (« co- »). Ou encore, le découpage d'un corpus en segments fournit un tableau de contingence, recensant chaque occurrence en fonction du segment d'apparition et de l'unité lexicale (type) qu'elle représente ; ce tableau sert de base à une classification des unités ou / et des segments. Un troisième usage de contextes est celui de la caractérisation des unités lexicales : chaque unité est représentée par les unités qui apparaissent dans son voisinage ; deux unités ayant des voisinages analogues seront considérées comme quasi-synonymes[169].

Il y a aussi des usages « implicites » de contextes. Dans l'édition de concordances, le KWIC (*KeyWord in Context*) est un extrait centré autour d'une occurrence du mot-clé considéré, il tient habituellement sur une ligne et est donc de l'ordre de la *phrase*. Pour la recherche documentaire, la procédure de *relevance feedback*, qui habituellement utilise le *texte* d'un document jugé pertinent pour enrichir et ajuster la requête initiale.

Dans tous ces cas, on ne fait jamais intervenir qu'un seul type de contexte : la phrase (délimitée par une ponctuation forte), une fenêtre de *n* mots, le texte... Aussi, quand il n'y a pas d'usage fixé, soit on argumente pour démontrer que, parmi toutes les définitions de contexte que l'on pourrait envisager, l'une est plus pertinente que les autres ; soit on considère que la définition du contexte peut varier suivant les types de textes et les applications visées, et que c'est un paramètre à ajuster, souvent sur des considérations heuristiques (tel choix « marche mieux » que tel autre dans tel cas de figure). Notre proposition pour DECID, qui est de prendre en compte simultanément plusieurs types de contextes, ayant chacun leur valeur et leur signification propre, renouvelle la problématique en dépassant une alternative stricte (et réductrice).

Il reste instructif de recueillir les observations étayant les choix faits pour la définition de contextes dans différentes applications. Pour les contextes naturellement envisagés (phrase, paragraphe, etc. et les ordres de grandeur correspondants en terme de longueur), l'étude qui suit consigne les atouts et les points faibles avancés par les uns et les autres. Ceci n'est qu'un premier relevé : l'investigation (ici limitée faute de temps) mériterait d'être poussée plus loin.

b) La phrase, l'énoncé

Atouts

C'est l'échelle que choisit Max Reinert pour la définition d'unités de contexte dans son système ALCESTE. Ce choix est motivé à la fois par des considérations sémantiques et statistiques (Reinert 1990).

[169] cf. la présentation des *champs sémantiques*, au sens de (Fluhr 1977, §III.4.2.3), dans ce chapitre.

Une unité de représentation, incluant l'expression d'un point de vue

Du point de vue sémantique, l'énoncé correspond à l'expression de l'attitude d'un sujet à propos d'un objet. Il peut s'agir d'une affirmation, mais aussi d'une négation, d'une requête, d'un jugement. L'énoncé est représentation en ce qu'il a un double contenu, une description du monde objectif, et une expression d'un point de vue subjectif.

> C'est en cela que la trace linguistique de l'énoncé constitue la plus petite unité de texte susceptible de décrire, selon nous, la représentation sous-jacente d'un sujet. (Reinert 1990, p. 29)

L'expression de relations

Par les constructions syntaxiques qui la structurent, par le voisinage étroit propice aux influences et interactions sémantiques, la phrase est un lieu évident d'expression de relations.

> La phrase est un bon contexte de cooccurrence. Elle est le lieu idéal où l'auteur met en rapport, notamment syntagmatique, les unités lexicales. La cooccurrence de deux mots dans la phrase peut être l'expression d'une relation syntagmatique stable comme d'une relation paradigmatique. (Sta 1997, §6.3.1.2)

> [Pour observer les cooccurrences,] plusieurs approches sont possibles, dont l'une prend pour base l'unité large du roman, et l'autre la cellule étroite de la phrase. Dans la première perspective, la contrainte est lâche puisque l'on ne considère pas la distance qui s'établit sur le terrain entre deux variétés étudiées. La cooccurrence n'y est appréciée qu'au niveau global du texte et n'implique aucunement des relations de voisinage. C'est en revanche la cohabitation dans le même espace restreint qui est mesurée dans la seconde perspective et que nous voulons tenter. Ce faisant, c'est une sorte de combinatoire ou de syntaxe sémantique qu'on essaie de préciser. En dehors des contraintes de la syntaxe, comment les mots se marient-ils ? Par affinité ? par complémentarité ? par opposition ? Les synonymes ont-ils tendance à se donner la main ? Le marquage mutuel des antonymes est-il plus étroit ? (Brunet 1995, pp. 30-31)

Une taille suffisamment petite pour être très sélective et traduire des cooccurrences statistiquement significatives

L'énoncé, avec une taille de l'ordre de 20 mots, a également l'avantage de permettre aux statistiques de saisir des associations significatives :

> compte tenu de la dimension réduite de ces *unités de contexte* [(dorénavant notées *u.c.*)] [...], la simple apparition de mots dans une unité devient, en elle-même, très significative des liens pouvant unir ces mots. Par exemple, si nous considérons qu'une *u.c.* contient environ 20 mots, que le vocabulaire obtenu comprend 800 mots, et si nous codons par « 1 » l'apparition d'un mot dans une *u.c.* et par la valeur « 0 », sa non-apparition, alors le tableau des données associé comprendra plus de 97 % de zéros : c'est tout dire sur la signification statistique de la simple présence d'un vocable dans une *u.c.* (Reinert 1990, pp. 26-27)

(Bonhomme & *al*. 1996) font un comparatif de trois tailles de fenêtres (10, 20 et 100 mots) pour la sélection de corrélats au voisinage d'un mot pôle, et concluent à la supériorité de la fenêtre de 20, en ce qu'elle laisse échapper le moins de corrélats fortement pertinents (Bonhomme & *al*. 1996, §IV.V). L'explication proposée est celle de la localité des associations significatives recherchées :

> Qu'il n'y ait qu'une faible perte de corrélats si l'on fixe à vingt mots l'ouverture de la fenêtre de recrutement (par comparaison avec une fenêtre de cent mots), cela signifie que la cooccurrence est un phénomène local, et que les thèmes, acteurs et fonctions dialectiques sont bien des unités mésosémantiques. (Bonhomme & *al*. 1996, §III.7)

Points faibles

Difficulté de délimiter formellement des unités satisfaisantes

Mettre au point un algorithme de découpage satisfaisant n'a rien d'élémentaire[170]. Trouver *la* suite d'énoncés qui composent un texte donné procède d'un déterminisme illusoire. Aussi la mise en

[170] (Reinert, Piat 1995) commencent par délimiter des *segments ponctués* (taille inférieure à 256 caractères et coupure sur ponctuation moyenne ou forte) ; ces segments sont regroupés « en privilégiant les coupures associées à une ponctuation forte » (et toujours sans dépasser les 256 caractères) pour former les *segments de texte calculés* ou *segments de texte marqués* ; les *unités de contexte (élémentaires)* sont alors obtenues en concaténant ces

œuvre de contextes phrastiques suppose quelques aménagements, pour que l'horizon fixé par un découpage donné ne soit pas la seule référence.

> Du point de vue technique, la notion d'énoncé est une notion peu opérationnalisable car elle fait référence à plusieurs niveaux d'analyse : le niveau syntaxique ne suffit pas à la déterminer même si l'on sent qu'un énoncé a une relation avec la notion de proposition, de phrase ou de paragraphe. En effet, un énoncé, aussi bien en tant qu'acte de langage, qu'en tant que propos d'un sujet sur le monde, fait référence à un sujet et donc fait référence à un élément psychique.
>
> Aussi, plutôt que de cherche à obtenir un découpage rigoureux du texte en énoncés (auquel nous ne croyons pas vraiment) nous lui avons substitué un découpage plus arbitraire en unités de contexte, dont la définition peut varier dans certaines limites, et que nous faisons varier. De cette manière, les résultats stables, c'est-à-dire indépendants de ces variations, ne devraient pas dépendre de l'arbitrarité du découpage, mais uniquement de son ordre de grandeur qui est l'ordre de grandeur d'un énoncé (qui, chez un locuteur moyen, ne devrait pas dépasser quelques dizaines de mots, la mémoire à court terme étant très limitée).
>
> (Reinert 1991)

L'application à un corpus de volume important est un autre moyen de contourner cette difficulté : les associations contextuelles sont obtenues grâce aux régularités observées sur l'ensemble du corpus, l'influence des découpages malheureux est neutralisée si ces découpages sont en proportion négligeable. La force des statistiques est justement que les effets d'un critère formel, approximatif de la réalité à décrire, se compensent à l'aune d'un calcul général sur l'ensemble du corpus.

Relations identiques à celles obtenues au niveau paragraphe, et pertes résultant de la taille plus restrictive

Les calculs de cooccurrences au niveau des phrases ne révèlent pas de relations supplémentaires par rapport au calcul au niveau des paragraphes. La restriction plus forte qu'il impose semble même nuisible, car les relations qui apparaissent au niveau paragraphe et échappent au niveau de la phrase sont aussi pertinentes que les autres (celles qui ne sont sélectionnées qu'au niveau des phrases) (Barakat-Barbieri 1992, p. 86 et 107).

c) Le paragraphe

Atouts

(Barakat-Barbieri 1992, p. 85)[171] adopte le paragraphe pour le calcul des champs sémantiques des mots.

Motivation sémantique forte

Cette zone correspond conventionnellement à une unité au plan sémantique. Elle n'est pas assujettie aux limites quelquefois plus syntaxiques que sémantiques de la phrase, et qui peuvent être trop restrictives en ce qui concerne le développement d'une thématique.

Critères cognitifs : mémoire, perception

S'en tenir au paragraphe évite aussi l'excès qu'il y aurait à ne faire aucune différence entre des mentions que séparent plusieurs pages et des mentions qui voisinent au sein d'une même paragraphe : la perception qu'en a le lecteur n'est pas la même ; la mémoire travaille différemment, et le champ de vision englobe le paragraphe mais pas simultanément plusieurs pages.

---

segments de texte calculés jusqu'à atteindre une longueur suffisante donnée à l'avance (paramètre). La longueur de l'unité de contexte se chiffre en nombre de mots analysés différents qu'elle comporte ; elle s'affranchit du plafond à 256 caractères (mais doit rester en deçà de 3 000 formes analysées).

[171] « La taille de cette unité [(le paragraphe)] est suffisamment restreinte pour que l'on puisse raisonnablement exclure le fait que des sujets n'ayant aucun rapport entre eux y apparaissent. Elle est suffisamment étendue pour que des relations entre thèmes n'étant pas étroitement liés y soient mises en évidence tout de même. »

Points faibles

L'intégration dans le document

Considérer chaque paragraphe indépendamment, c'est l'isoler artificiellement, alors qu'il y a une certaine perméabilité d'un paragraphe à l'autre, et une unité d'ensemble des paragraphes d'un document.

> Nos résultats ont démontré que l'unité sémantique la plus riche [, entre la phrase, le paragraphe et le texte,] était bien le paragraphe. Ce découpage, dont le but est en fait de regrouper l'ensemble des phrases ayant une cohérence sémantique, ne reflète pas exactement la réalité. Il semble qu'un découpage en groupe de paragraphes serait plus approprié. (Barakat-Barbieri 1992, p. 142)

d) Le texte

Atouts

Le texte se présente comme une unité sémantiquement autonome, bien davantage que le paragraphe (ou la phrase). S'en tenir à une vision morcelée en paragraphes ou en phrases multiplie les entités à considérer. C'est aussi faire fi de la cohérence sémantique qui traverse le texte, et qui fait que des notions se font écho du début à la fin du texte, d'une partie à une autre.

Pour la plupart des traitements sur des genres « brefs » et « focalisés » (dépêches, annonces sur les forums électroniques, fiche descriptive d'activité des chercheurs EDF,...), le texte est l'unité considérée pour décrire les relations entre les mots (Sta 1995, §3.2).

Pour les documents plus longs, un découpage selon les parties logiques donne des contextes plus développés que le paragraphe, tout en ayant une certaine autonomie, régularité de longueur et cohérence interne forte.

> Une solution face à des documents longs est d'exploiter leur structure logique (note : c'est la solution adoptée pour l'expérience de constitution de l'index de ce document [thèse]). Un document est une suite de chapitres, sections et possède des titres, entêtes etc... Ce découpage logique du document et son exploitation sont favorisés par l'émergence de langages de structuration de documents : SGML, Hytime,... (Sta 1997, §6.3.1.2)

Une partie logique peut elle-même être considérée comme un texte, elle est du même ordre qu'un texte[172]. Elle a d'ailleurs un titre, et sa clôture est marquée par le passage à la partie suivante. Elle peut devenir un *extrait*, présenté indépendamment du reste du texte dans un autre ouvrage, et qui s'autonomise en tant que *texte choisi*.

Du texte à l'idiolecte

Le texte cerne un espace linguistique homogène et unitaire vis-à-vis des particularités de langage d'un rédacteur, d'un auteur. Il est ainsi, dans son entier, et par opposition à d'autres textes, le lieu de réalisation et de manifestation d'un *idiolecte*. La description est ainsi autorisée à rapprocher et homologuer des occurrences et contextes d'occurrences même éloignés, tant qu'ils figurent au sein du texte étudié.

> Cette clôture du texte par l'épuisement de l'information lui confère son *caractère idiolectal* : en effet, les dénominations contenues dans le texte sont déterminées par les définitions qui y sont présentes et uniquement par elles, de telle sorte que le texte constitue un micro-univers sémantique fermé sur lui-même. Cette propriété sémantique du discours rend légitimes les descriptions partielles, en établissant une sorte d'équation entre les textes finis et les univers signifiants clos. Elle n'offre pas, cependant, de solution définitive pour la description des univers sémantiques ouverts, caractérisés par l'afflux continuel d'informations. (Greimas 1966, §VI.3.c, p. 93)

Points faibles

Plusieurs propriétés des textes altèrent la régularité et la significativité du calcul de cooccurrences (Barakat-Barbieri 1992, p. 83-85) :

[172] Ce qui confère au texte son caractère fractal, cf. par exemple Louis TIMBAL-DUCLAUX.

- la *variation de taille* qu'il peut avoir entre les textes d'un même corpus (y compris au sein d'un type de textes donné) crée des déséquilibres au niveau des calculs statistiques ;
- le *vocabulaire lié à la structure* et à la présentation du document tendent à se mêler aux relations thématiques ;
- un même document peut traiter des *sujets différents* n'ayant aucun lien entre eux.

Les relations qui peuvent être calculées spécifiquement au niveau du document seraient plus liées au genre du texte, au type du document, qu'à sa thématique particulière. Ces relations sont intéressantes pourvu que l'on adopte un point de vue textuel.

G. UN CHANTIER À POURSUIVRE : LA CONSTRUCTION DES COMMUNAUTÉS A PARTIR D'UN CORPUS

1. Etude critique de techniques pour le groupement de mots, en vue de la construction automatique de Communautés

a) L'information mutuelle et autres coefficients d'association

Une fonction très largement utilisée évalue l'association de deux unités. Elle mesure la proportion de contextes où les deux unités cooccurrent, par rapport à leurs contextes d'apparition en général. Les unités sont d'autant plus liées qu'elles apparaissent ensemble dans une grande partie de leurs contextes, et qu'elles ont des fréquences d'usage proches.

Cette fonction, d'*information mutuelle*, ne se laisse pas aisément généraliser à plus de deux unités. Or on peut très bien concevoir des associations qui ne deviennent significatives qu'en trio. De plus, raisonner sur toutes les sous-parties d'un ensemble de deux, trois, quatre éléments voire plus, fait entrer dans une combinatoire énorme.

Bien sûr, un tel coefficient d'association peut être utilisé pour former des regroupements de mots de taille supérieure à deux : cela en a même été la première illustration, à laquelle il doit sa renommée[173]. Mais la constitution de ces regroupements reste très locale et « individualiste » : elle procède par étapes en ajoutant à chaque fois *un* élément à autre élément ou à un groupe en formation[174]. Cette forme de classification a par ailleurs d'autres propriétés qui ne nous semblent pas souhaitables : majoration *a priori* du nombre d'éléments dans une classe ; affectation d'un élément à une seule classe –il peut garder des liens externes avec d'autres classes mais ces liens n'ont pas le même statut.

L'utilisation d'un coefficient d'association pour faire apparaître des groupements de termes ayant une affinité sémantique globale ne semble donc pas appropriée, à la fois au vu des applications précédentes et de sa focalisation sur une analyse par paires.

Enfin, la formule de l'information mutuelle montre un biais : elle sélectionne surtout les associations exclusives, au détriment des associations remarquables mais multiples (diversifiées selon les contextes) d'un même mot (Sta 1997, §6.3.2.3). On obtient donc surtout des associations de mots peu fréquents, qui n'ont qu'un seul type d'usage dans le corpus considéré.

b) Les algorithmes connus de classification automatique

Plusieurs familles d'algorithmes existent pour classer des éléments.

Les classifications ascendantes hiérarchiques déterminent un ordre de regroupement des éléments et des classes d'éléments. Partant de l'ensemble des éléments, une classification ascendante hiérarchique désigne les deux éléments qui forment le premier regroupement (ce sont les deux plus proches) ; puis, le regroupement de deux autres éléments, ou celui d'un élément avec les deux précédents ; et ainsi de suite jusqu'au dernier regroupement, qui n'est autre que l'ensemble initial.

---

[173] Allusion à la recherche de Bertrand MICHELET (Michelet 1988), au Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris, et qui a été poursuivie dans cette équipe avec profit : logiciels *Leximappe* (dont on trouvera une présentation introductive dans (Courtial, Pochon, Vilain 1994)), *Lexinet*, (Chartron 1988), *Candide* (Teil 1991), etc.
Dans le même esprit, (Berni Canani 1986) mérite également mention : il présente toute une étude sur les graphes d'associations entre mots, et sur l'utilisation de leurs propriétés pour y déceler des groupements thématiques stables (ni purement syntagmatiques, ni purement paradigmatiques).

[174] Quelques raffinements peuvent être introduits : ne pas sélectionner les liaisons indépendamment les unes des autres (par ordre de force décroissante), mais sélectionner les termes centraux, à savoir ceux qui ont *n* liaisons fortes avec d'autres termes (Quatrain & Béguinet 1996, Annexe 4). L'approche est ainsi moins locale, mais reste néanmoins décomposée en interactions binaires.

Les algorithmes de classification hiérarchique se différencient par le choix de la distance, qui détermine à chaque étape les entités à regrouper (ce sont les entités les plus proches au sens de la distance choisie). La distance spécifie la distance entre éléments, mais aussi la distance d'un élément à une classe et la distance séparant deux classes. Ces deux dernières distances, faisant intervenir des classes, sont généralement définies à partir de la distance entre deux éléments. Chaque algorithme correspond à une manière de définir ces distances de classes.

En revanche, ils suivent tous la même démarche, si bien que le résultat d'une classification est toujours une suite ordonnée de représentations de l'ensemble de départ, chacune de ces représentations comportant une ou plusieurs classes d'équivalence sur tout ou partie des éléments. Autrement dit, cela pose plusieurs problèmes pour la définition de Communautés.

D'abord, le choix d'une étape, intermédiaire entre les deux extrêmes du « rien regroupé » et du « tout regroupé ». Ce n'est pas une question nouvelle et des propositions existent (par exemple, arrêt avant un saut de la valeur de la distance de regroupement suivante ; fixation d'un seuil sur la distance de regroupement ; limitation sur la taille de la plus grosse classe) ; mais les résultats ne sont pas toujours satisfaisants, pour des critères dont certains sont très complexes à mettre en œuvre.

Autre composante insatisfaisante de ces algorithmes de classification ascendante hiérarchique : leur instabilité. Rien n'exclue en effet qu'à une étape, plusieurs regroupements soient équivalents (du point de vue de la distance) et concurrents (un même élément ou un même groupe est équidistant de plusieurs éléments ou classes). Le choix se fait alors arbitrairement, pour n'adopter qu'une seule des alternatives. Or des expériences montrent que l'impact de ces choix peut être très grand, et avoir des répercussions majeures sur toute la suite des regroupements ultérieurs. Il y a donc là un biais que rien ne permet en l'état actuel de tempérer.

Concrètement, les algorithmes de classification hiérarchique mobilisent de grandes ressources de calcul : leur application à un important volume de données peut poser problème.

Le point véritablement dirimant est la construction de classes d'équivalences, donc non chevauchantes. Comme cela a été argumenté plus haut, cette contrainte ne convient guère aux Communautés.

Les classifications descendantes hiérarchiques procèdent également par étapes. Elles prennent comme point de départ une seule classe contenant tout l'ensemble de départ. A chaque étape, l'algorithme scinde une classe en deux, ou détache un élément d'une classe. Le choix de la scission s'effectue par un critère qui veille à maximiser le contraste entre les classes, et l'homogénéité interne des classes.

Ces classifications descendantes sont moins pratiquées que les classifications ascendantes. Elles comptent une application notable à des données linguistiques avec le logiciel ALCESTE.

Cependant, là encore, qui dit hiérarchique dit classes d'équivalences (disjointes), ce qui ne convient pas aux Communautés.

Quand la classification automatique n'est pas hiérarchique, elle fournit comme résultat une partition. Soit donc deux propriétés contraires aux attentes des Communautés : *(i)* tous les éléments son classés ; *(ii)* un élément ne peut être affecté à plusieurs classes. L'utilisation de ces algorithmes ne peut donc être envisagée sans réaménagements[175].

[175] (Renouf 1993c) semble avoir à sa disposition un outil de classification multiclasse non exhaustive, mais ne donne aucune précision sur l'origine de l'outil ni sur l'algorithme utilisé. Les groupements de mots ainsi calculés ne paraissent utilisés que pour être présentés à l'utilisateur, en complément au titre d'un texte analysé, et non comme unités exploitées par un traitement automatique subséquent.
Les travaux récents de (Gonzalez-Rubio & Guizol 1997) ont retenu notre attention, car cette collaboration franco-québécoise fait usage d'un module de classification floue, qui éviterait, par le biais des coefficients d'appartenance, les deux écueils du partitionnement. Un élément peut être reconnu dans plusieurs classes ; et si son appartenance est trop distribuée, alors il n'est de fait rattaché à aucune. Les questions qui se posent à nous sont alors les suivantes :
- l'algorithme ne semble pas si simple, et n'est appliqué qu'à un sous-ensemble des mots (« ayant des distributions non aléatoires et des indices de cohésion forts ») : comment gérerait-il l'ensemble général de toutes les unités descriptives ? Nous voulons en effet que la classification puisse aussi caractériser et regrouper entre elles les unités très réparties.

c) La tactique des classements indirects

Les algorithmes de classification recherchent classiquement une partition de l'ensemble des *individus* (éléments à regrouper en classes). Les proximités entre ces individus sont souvent exprimées en fonction de *variables*. Par exemple, on va classer des mots (qui prennent le rôle d'individus) en fonction des documents d'un corpus dans lesquels ils apparaissent (les documents prennent le rôle de variables).

Certains algorithmes tendent à produire simultanément un regroupement des individus et des variables, et considèrent simultanément et symétriquement ces deux plans. Autrement dit, intervertir les rôle des individus et des variables ne change pas le résultat. Cela se pratique dans le domaine de l'analyse de corpus, avec l'analyse factorielle des correspondances (Benzécri & *al.* 1973b), appliquée à des données linguistiques et textuelles (Benzécri & *al.* 1981), et de la sériation par blocs (Marcotorchino 1987) (Marchotorchino 1991), à la base de l'analyse relationnelle et du *Text Mining* chez IBM[176][177]. Le résultat vise donc une double partition, celle des individus, et celle des variables : or justement la structure de partition ne convient pas à la définition des Communautés, puisqu'elle impose des contraintes de classement exhaustif et de non recouvrement des classes.

On considère alors les algorithmes qui opèrent une seule partition (sur les individus) : si l'on associe à chaque classe les variables représentatives des individus qu'elle contient, on obtient, pour les variables, une série de classes qui vérifient les propriétés cherchées (non exhaustivité, recouvrements possibles). Reste que c'est au prix d'une partition sur les individus.

Ainsi, du classement de documents en fonction de mots qu'ils contiennent, on peut déduire des regroupements de mots sans la contrainte de répartition totale et d'attribution unique[178]. Mais

---

- les coefficients d'appartenance sont une mesure continue : soit l'on considère un (potentiellement) grand nombre d'appartenances pour chaque item, ce qui alourdit considérablement les données, soit se pose le problème du seuillage, qui introduit une part d'arbitraire et repose entièrement sur l'analyste. Une classification en soi directement multiclasse oblige à décrire le choix d'affectation ou non d'un élément au cœur de l'algorithme, et nous semble en ce sens plus significative.

[176] (Warnesson 1985) recherche ainsi sans sourciller des classes de synonymes formant des classes d'équivalences ; en 1989 (Bédécarrax, Warnesson 1989), un point de vue un peu moins formel la conduit à interpréter linguistiquement les écarts observés, et à préférer une extension de la sériation par blocs, appelée quadri-décomposition, ménageant la possibilité de non-classement (par contre, les classes restent disjointes).

[177] Le logiciel conçu par cette équipe d'IBM autour de Marcotorchino est TEWAT. Très peu d'informations techniques sont données sur ce logiciel, sinon qu'il implémente l'analyse relationnelle. Or dans la pratique, il semble qu'il n'opère pas une double partition, mais simplement une partition des individus, pas des variables.
Une présentation de l'analyse relationnelle de TEWAT, ainsi que d'autres méthodes de classifications usuellement pratiquées sur les données textuelles, peut être consultée dans (Quatrain & Béguinet 1996, Annexe 4).

[178] Richard QUATRAIN (ingénieur chercheur à la DER d'EDF) a expérimenté ce genre d'approche pour l'aide à la reformulation de requête, dans l'interrogation d'une base documentaire.
En revanche, ALCESTE passe à côté de ce degré de liberté. Les mots caractéristiques de chaque classe d'*unités de contexte* sont calculés, mais ensuite chacun de ces mots est attribué à au plus une seule classe, qui détermine *sa clé contextuelle* (unique) (Reinert, Piat 1995). La médiation par la classification des unités de contexte n'a pas été mise à profit pour construire une classification assouplie des mots du corpus.
Quant à Mountaz ZIZI, bien qu'elle reste dans le cadre classique des taxonomies lexicales (thesaurus), elle pressent la valeur sémantique et contextuelle de la manœuvre :
« D'un côté, Salton et McGill suggèrent de faire l'agrégation directement sur les descripteurs. De l'autre Crouch par exemple [Crouch92] agrège les documents plutôt que les descripteurs, cette approche étant justifiée par le fait que pour les descripteurs extraits d'un corpus de documents les fréquences de co-occurrence et donc les similarités avec les autres termes sont souvent trop petites pour que les techniques de classification puissent aboutir à des regroupements réellement pertinents. [(] Mais cette remarque vaut surtout pour des descripteurs à très faible fréquence [...] [et] n'est réellement pertinente que pour des collections de taille relativement petite [)]. [...] Dans le premier cas, chaque terme correspond à une unique entrée dans le thesaurus alors qu'avec l'autre, il existe plusieurs entrées pour le même terme, ce qui va à l'encontre des règles de conception de base des langages d'indexation [...]. D'un autre côté il nous semble qu'il s'agit là d'une manière intéressante de repenser le thesaurus, et peut-être d'apporter une solution élégante aux problèmes d'ambiguïté par la prise en compte automatique et explicite du contexte. » (Zizi 1995, §3.3.2, p. 106)

l'hypothèse est celle d'univocité des documents : un document qui aborde plusieurs points, plusieurs thèmes, n'est jamais classé que dans une seule classe[179].

Le classement réciproque, celui des mots en fonction des documents dans lesquels ils apparaissent, impose quant à lui une conception univoque de chaque mot. On sacrifie les effets de polysémie et d'homonymie, on occulte les variations de contexte et d'usage.

Que l'on imagine encore échapper à la partition initiale en réitérant le processus (par exemple : classification des mots, déduction de groupements de documents ; puis de ces groupements de documents, déduction de groupements de mots), le résultat reste miné, marqué, orienté, par la contrainte initiale.

La partition initiale opérée sur les individus a deux incidences négatives dans notre contexte : d'une part elle impose à l'ensemble des individus une structure qui ne leur correspond pas ; d'autre part, *elle ne forme des groupes de variables que sur leurs propriétés de discriminance*. Illustrons cette deuxième incidence dans le cas d'une classification des documents effectuée dans le but de recueillir des groupements de mots : les groupements de mots obtenus ne comporteront que des mots discriminants pour le corpus considérés et les classes de documents calculées. On ne peut donc pas espérer trouver des groupes de mots généraux qui « fonctionnent » ensemble (par exemple les mots qui forment la toile de fond du domaine, du genre,...), ou des groupes de mots spécifiques mais correspondant à un caractère transverse à la partition calculée (par exemple, des mots exprimant le doute, ou le fait qu'on soit dans une perspective théorique ou à un stade expérimental, alors que les classes regroupent des documents par thématique).

En résumé, le classement des documents en fonction des mots nie le caractère multifacettes de certains documents ; le classement des mots en fonction des documents nie les variations de sens et la polysémie des unités lexicales. Chacune de ces opérations repose donc sur une hypothèse non satisfaisante, qui mine la qualité d'ensemble. Si l'on peut tirer certains groupements significatifs, tout un pan de la description reste néanmoins dans l'ombre.

d) Les axes d'une analyse factorielle

La donnée de départ est un ensemble d'individus, caractérisés chacun par les valeurs qu'ils adoptent sur un ensemble de variables. Dans le cas des Communautés, on peut établir un ensemble de contextes (un ensemble de périodes, ou un ensemble de paragraphes, ou un ensemble de documents), et les variables qui les caractérisent sont les unités qu'ils incluent.

L'analyse factorielle consiste à calculer une suite ordonnée d'axes, orthogonaux les uns aux autres. L'ensemble des axes reconstitue exactement l'espace que décrivait les variables initiales, mais la position des axes a changé. Dans le système d'axes initial, chaque axe correspond à une variable. Dans la nouvelle base, un axe correspond mathématiquement à une *combinaison linéaire* des variables initiales. L'algorithme fournit si on le désire la *contribution* de chaque variable à chaque axe. L'intérêt du nouveau système d'axes est que le premier axe est le meilleur axe possible pour rendre compte, sur une seule dimension, des *différences* entre les individus dans leur ensemble. Si par exemple ces individus sont grosso modo de deux sortes, alors le premier axe se positionne automatiquement en plaçant ces deux catégories à ses deux extrémités. En ajoutant le deuxième axe, on obtient la meilleure représentation, en deux dimensions, des contrastes entre les individus ; etc.

Revenons à l'application qui nous concerne, celle par exemple qui considère un ensemble de paragraphes décrits par les unités qu'ils comportent. Le calcul nous obtient une suite d'axes. Le premier est associé à un groupe d'unités, qui *dans leur ensemble* apparaissent dans une famille de contextes et n'apparaissent pas dans le reste des contextes. Ces unités ont donc, dans leur ensemble, un comportement groupé. L'axe suivant nous révèle de la même façon un groupe d'unités, qui a lui aussi cette propriété de comportement groupé vis-à-vis des contextes. On peut poursuivre ainsi l'investigation d'axe en axe, et recueillir des groupements d'unités.

[179] L'expérience montre que forcer la description du corpus au format d'une partition est décevant et peu efficace. ALCESTE par exemple opérait une classification ascendante hiérarchique de ses *unités de contexte* (segments de l'ordre de la phrase) ; entre 1990 et 1991, son concepteur Max REINERT renonce à la propriété d'exhaustivité : certaines unités de contexte s'avèrent non classables, et sont désormais non classées.

On forme ainsi, par le biais des contributions des variables pour chaque axe, des groupements de variables. Ces groupements ont les propriétés qui nous intéressent. Notamment, on a la possibilité de voir apparaître des recouvrements, lorsqu'une variable contribue à plusieurs axes. Une variable peut aussi n'être affectée à aucun groupe : variable correspondant exactement à un axe, ou variable qui ne contribue significativement qu'aux derniers axes, eux-mêmes peu significatifs.

L'utilisation que nous faisons ici des axes est à notre connaissance inédite. La manière habituelle de recourir à une analyse factorielle pour obtenir des classes consiste à opérer une classification automatique sur les individus, en définissant une distance en fonction des coordonnées dans les *n* premiers axes de la nouvelle base. Le passage par l'analyse factorielle sert à réduire le nombre de variables nécessaires au calcul de la distance, et à augmenter la « netteté » du contour des classes (l'omission des derniers axes gomme les comportements marginaux et les petits écarts). Mais l'étape proprement dite de classification nous ramène au cas de figure précédent, sur les classifications automatiques : nous avons vu que ces algorithmes fournissaient des classes d'équivalences, ce qui ne convient pas.

L'approche proposée a ses propres difficultés.

L'approche étant nouvelle, un certain nombre de questions délicates restent à résoudre, notamment le passage, des valeurs (scalaires) des contributions, à la décision (booléenne) d'intégration ou non au groupement.

En outre, la signification des axes ne correspond pas complètement au critère de formation des communautés. L'analyse factorielle se focalise sur les variables discriminantes ; les variables globalement peu discriminantes sont reléguées dans les derniers axes, voire dispersées sur plusieurs axes. Cela va donc à l'encontre du souhait de ne pas neutraliser les unités participant à plusieurs isotopies. On ne veut pas non plus fuir les isotopies banales : au contraires, ces isotopies, en formant une unité, reçoivent une description synthétique.

Accessoirement, l'analyse factorielle est une procédure complexe, pour laquelle il est raisonnable d'utiliser des logiciels du commerce. Ces logiciels sont souvent coûteux, ce qui peut devenir dissuasif pour le déploiement d'une application comme DECID.

2. Choix actuel

a) Une solution par adaptation

La partie précédente a montré qu'aucun algorithme à notre connaissance ne constitue des groupes d'unités répondant aux propriétés des Communautés. Une réflexion d'équipe a permis de trouver une solution, en prenant comme base un algorithme connu. L'algorithme retenu est celui des *Nuées Dynamiques* de E. Diday (Diday 1971). Il présente un certain nombre de qualités pour notre application, mais est conçu pour fournir une partition en classes d'équivalences. La solution consiste à apporter quelques modifications à l'algorithme, plus conformes au comportement souhaité.

L'algorithme finalement mis au point a été implémenté en Ada par Pascal OBRY (EDF-DER, département SID).

Les paragraphes suivants développent tout ceci : en quoi consiste l'algorithme, quelles sont ses qualités, quelles modifications ont été apportées, quelles difficultés subsistent. On trouvera en annexe tout le détail sur l'algorithme implémenté et les discussions techniques sur les choix réalisés.

b) Les Nuées dynamiques de Diday

La présentation que donne Michel Volle des Nuées dynamiques (Volle 1985, §XVII) convient parfaitement pour expliquer les principes de la méthode et introduire les concepts qui lui sont propres. Nous reprenons donc ses mots dans les paragraphes qui suivent.

Le nom fort bien choisi qu'a reçu cette méthode lui confère d'emblée un certain prestige auprès de ceux qui, ne connaissant pas l'analyse des données, éprouvent devant cette dénomination poétique l'impression agréable (mais trompeuse) d'être au bord de la compréhension intuitive d'un mystère. Il ne s'agit cependant que d'un outil logique comparable aux autres.

Supposons donné un ensemble *E* sur lequel sont définies une distance *d (x,y)* entre éléments et une distance *D (X,Y)* entre sous-ensembles. Prenons au hasard *k* paquets de *p* points chacun dans *E* ; nous appellerons chacun de ces paquets un *noyau*. Ces noyaux nous permettent de définir une partition de *E* en *k* classes, chaque classe comprenant les éléments qui sont plus proches d'un des noyaux que de tous les autres noyaux. A partir de cette partition, on définit une nouvelle famille de noyaux en associant à chaque classe de la partition l'ensemble de *p* points qui en est le plus proche. Puis on recommence : à cette nouvelle famille de noyaux va être associée une nouvelle partition, etc. Il est facile de démontrer que le procédé converge sous certaines conditions : on finit par aboutir à une partition et à une famille de noyaux qui se correspondent.

Ainsi la méthode des nuées dynamiques permet de construire par itération, à partir d'une famille absolument quelconque de noyaux, une partition en *k* classes. Mais cette construction contient une part d'arbitraire : la partition obtenue dépend du choix initial des noyaux. Pour compenser dans une certaine mesure cet arbitraire, on applique la méthode plusieurs fois de suite, en partant à chaque fois d'une famille différente de noyaux tirée au hasard. On obtient ainsi plusieurs partitions en *k* classes. Si l'on considère deux éléments quelconques, ils peuvent avoir été classés ensemble dans certaines de ces partitions, et classés séparément dans d'autres. On appelle *forme forte* –encore une dénomination bien trouvée– un ensemble d'éléments qui auront été classés ensemble dans toutes les classifications. Ces formes fortes déterminent une partition de *E* qui peut fort bien contenir plus de *k* classes. Les formes fortes qui ne comprennent qu'un élément ne présentent pas d'intérêt, car elles concernent des éléments qui semblent « inclassables ». Par contre, les formes fortes qui comprennent beaucoup d'éléments sont intéressantes : pour qu'un groupe d'éléments ait résisté aux aléas dus aux choix des différentes familles de noyaux, il fallait qu'il fût bien homogène.

c) Atouts de cet algorithme

La méthode se distingue par ses performances, en terme de temps de calcul et d'espace mémoire requis :

- Comme on n'utilise pas à chaque étape toutes les distances entre toutes les paires d'individus, mais seulement les distances des individus aux noyaux, l'algorithme est approprié au traitement de grandes populations.
- La convergence est rapide : la pratique montre que la partition est généralement obtenue en moins d'une dizaine d'itérations.

La méthode est robuste :

- Les itérations permettent de reprendre et de rectifier les traitements, de fait un peu grossiers en raison de leur rapidité (toutes les distances ne sont pas examinées).
- Si l'on peut toujours obtenir une classification, on a également des indicateurs pour évaluer sa qualité.

La méthode est souple et s'adapte aux données, outre leur grand nombre éventuel :

- L'information apportée par un indice de dissimilarité peut être directement exploitée, sans avoir besoin de recourir à une distance[180].
- Le cas échéant, on a la possibilité d'exploiter la connaissance que l'on a de la population à classer en fixant les noyaux initiaux.

Les nuées dynamiques s'appuient de plus sur l'idée forte de noyaux :

- Si les individus qui composent le noyau sont bien choisis, ils sont représentatifs, « typiques » de la classe, et en forment un résumé plus riche et plus tangible que peut l'être un centre de gravité.
- Des contraintes peuvent être imposées aux noyaux, dont les éléments par exemple peuvent être choisis parmi des éléments particuliers de l'ensemble initial.
- L'interprétation des résultats peut être facilitée, et leur présentation allégée.

[180] Mathématiquement, une fonction qui mesure l'écart entre deux objets ne peut être appelée *distance* que si elle vérifie un certain nombre de propriétés (notamment l'*inégalité triangulaire*) ; un *indice de dissimilarité* a moins de contraintes à vérifier (il s'agit simplement d'une fonction *symétrique* qui, a tout couple de points associe un réel *positif*, et telle que la dissimilarité d'un point à lui même est nulle). Toute distance est un indice de dissimilarité, la réciproque n'est pas vraie.

- Les noyaux se moulent en quelque sorte sur la forme de la classe ; l'existence de classes non connexes est un cas particulier qui présente plus de difficulté.

La notion de *forme forte* permet de tempérer un certain nombre de choix arbitraires, ou *a priori*, ou trop tranchés :

- La méthode ne sollicite pas l'indication d'un seuil arbitraire pour la formation des classes, ou la mise au point d'une heuristique de coupure d'un arbre hiérarchique ; mais elle demande que soit fixé à l'avance un ordre de grandeur du nombre de classes à obtenir (c'est le nombre de noyaux, qui correspond au nombre de classes obtenues après convergence). Il peut y avoir éventuellement quelques classes vides, c'est pourquoi on recommande généralement de prendre une valeur plutôt légèrement trop grande que trop petite[181] ; mais il faut tout de même partir d'un bon ordre de grandeur. C'est introduire un *a priori* sur la classification à obtenir, alors que l'on pourrait souhaiter que la partition se façonne directement, et en quelque sorte objectivement, à partir des données. L'issue se dessine du côté des formes fortes, dont le nombre peut varier très largement autour du nombre de noyaux[182].
- La partition obtenue à l'issue d'un tirage dépend de l'ensemble de noyaux de départ : chaque solution est un optimum local, on ne trouve pas nécessairement un optimum global. De même, l'affectation des ex æquo a, dans toutes les méthodes (y compris les classifications ascendantes hiérarchiques), une incidence parfois non négligeable sur le résultat obtenu. La notion de *forme forte* est une première réponse face à la dispersion possible des résultats.
- La méthode est une heuristique, elle n'est pas déterministe et ne fournit pas toujours une solution de la meilleure qualité. La pluralité des solutions se résout d'un point de vue global, par la recherche de formes stables, –les formes fortes. Cela réduit, sans l'éliminer toutefois, l'incertitude quant à la validité de la classification obtenue. Mais les données réelles s'organisent-elles toujours de manière univoque ? On peut d'ailleurs montrer qu'il y a une instabilité inhérente à la démarche même de classification (Benzécri & *al.* 1973a, §B.4.3.2). La méthode est peut-être de ce point de vue plus ouverte que des algorithmes qui trouvent par principe une solution et une seule.

d) Réaménagements effectués

- Une détermination automatique du nombre de noyaux est proposée (dans le cas où l'indice de dissimilarité est borné). Ceci permet d'augmenter les chances de partir d'un bon ordre de grandeur du nombre de noyaux, lorsque l'on en a aucune indication par ailleurs. C'est utile, car un trop grand nombre de noyaux disperse les résultats, et les noyaux se gênent mutuellement ; et un trop petit nombre de noyaux ne laisse pas apparaître des différences et des oppositions significatives.
- Le nouvel algorithme prend dès le départ en compte la possibilité qu'il y ait des individus inclassables, et n'oblige jamais à affecter arbitrairement à une classe un individu qui ne présente aucune similarité avec les autres[183].
- Comme le suggère implicitement (Volle 1985) en proposant une méthode d'agrégation des formes fortes, la constitution des formes fortes est assouplie : une tolérance est introduite, permettant de corriger les erreurs introduites par un tirage qui n'a pas convergé, ou qui a convergé vers un optimum local peu satisfaisant[184].
- Enfin bien sûr, le classement final est revu pour obtenir les propriétés recherchées : classification multiclasse et non-exhaustive.

---

[181] Diday présente ce paramètre comme « le nombre maximum de classes désirées ». Il précise que « quand *K* est trop grand par rapport au nombre de classes qui existent effectivement, des classes vides apparaissent. »

[182] Un calcul simple montre que le nombre de formes fortes se situe en théorie entre $E\,(N/K^T)$ et $min\,(N/2,\ K^T)$, avec les notations précisées dans les conventions ou définies pour les variables globales de l'algorithme. Par exemple, si l'on classe 100 individus, que l'on se donne 10 noyaux, et que l'on procède à 5 tirages successifs, on peut trouver entre 0 et 50 formes fortes.

[183] Cet aménagement correspond à l'introduction de la *classe Z*, décrite dans la présentation détaillée de l'algorithme en annexe.

[184] A la méthode des connexités descendantes, de type *single-linkage*, exposée par (Volle 1985), on préfère une stratégie de type *complete-linkage* qui assure une bonne cohérence interne des classes de formes fortes, car ici se joue la qualité du système de noyaux final.

e) Difficultés qui subsistent

- La manière de définir la taille des noyaux est encore problématique : si l'on s'en tient aux propositions de (Diday 1971), on ne sait vraiment mettre en œuvre que des noyaux de taille fixe égale à 1, alors que des noyaux plus grands, et adaptés à la taille effective des classes, seraient mieux représentatifs.
- La gestion de la *classe Z* (que nous avons introduite pour prendre en compte les éléments atypiques) requiert encore quelques mises au point.
- La convergence n'est démontrée que dans certaines conditions ; en dehors de ces cas, il faut prévoir un traitement assez robuste, qui détecte les cas de non convergence, et qui puisse dans tous les cas proposer une solution acceptable (non aberrante). Des heuristiques frustes ont été mises en place pour gérer ces cas de non convergence, mais une meilleure manière de traiter ces cas serait à étudier.
- La fonction d'agrégation-écartement, qui oriente toute la dynamique de construction des classes, présente un certain nombre de défauts ; des corrections sont proposées mais restent à valider.