Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ?

Bénédicte Pincemin

Paru dans la revue Sémiotiques, *n° 17, décembre 1999, pp. 71-120.*

Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ?

Bénédicte PINCEMIN (CNRS, LLI - Paris XIII & MoDyCo - Paris X)

Résumé
Le sème est souvent compris comme une primitive utilisable pour décrire les sens des mots. Or, tel que le présente F. Rastier, il est défini par et dans un contexte, et apparaît comme le résultat de l'interprétation d'un texte. Cette conception du sème permet de comprendre les succès et les limites d'analyses statistiques ou distributionnelles de textes. Certaines propriétés mathématiques des modélisations se révèlent linguistiquement inadéquates ; choisir le cadre de la sémantique interprétative conduit alors à redéfinir les traitements par-delà le seul ajustement des paramètres. Cette perspective oriente également certains choix concernant les outils et les étapes du TAL : dictionnaires, analyseurs morpho-syntaxiques, corpus et étiquetage. Un nouveau mode de classification automatique est présenté, comme moyen de repérage et de représentation des isotopies et donc de la thématique d'un texte. Ces classes sémantiques calculées sont d'une autre nature que celles définies par des experts.

Abstract
Semes are usually understood as semantic primitives, in order to describe meanings. As for F. Rastier, semes are contextually defined, and are the result of an interpretative path. This modelisation accounts for successes and limits of some statistical and distributional approaches. Some of the mathematical properties implied are not linguistically appropriate, so that the processing has to be defined on new bases. Text semantics also sheds new light on NLP tools and processes : dictionaries, taggers, corpus. A new clustering method is proposed as a means to modelize isotopies, and thus texts topics. These computed semantic classes are inherently different from the ones manually defined by experts.

1.Introduction

La réalisation de logiciels, instrumentant une analyse systématique de très grands corpus de textes, a des vertus expérimentales : les résultats concrets obtenus peuvent être spectaculairement révélateurs d'une certaine justesse du modèle sous-jacent à l'analyse — ou au contraire de son inadéquation[1]. Plus encore, les comportements lexicaux et linguistiques observés sont en mesure de manifester avec force des propriétés de la langue, telle qu'elle est en usage dans les textes.

La dimension sémantique n'est pas en reste. Un certain nombre d'applications, comme la recherche d'informations sur le texte intégral (notamment sur Internet) ou le calcul de représentations synthétiques de collections de documents (dans la mouvance des logiciels d'exploration statistique de bases textuelles pour la veille technologique), visent avant tout à rendre compte des thèmes présents dans les textes. Le recours éventuel, dans le traitement, à des connaissances lexicales ou morpho-syntaxiques, est considéré comme un moyen de mieux approcher la langue, au service d'une saisie de ce qui fait sens et devient information pour le lecteur.

Deux perspectives annoncent la pertinence des sèmes, unités de sens décrites par une sémantique différentielle telle que la sémantique interprétative. D'une part, dans le cadre de réalisations informatiques sans prétention linguistique, des heuristiques de classification de mots donnent des résultats étonnamment significatifs, mais pas entièrement satisfaisants. Les rectificatifs tentés par tâtonnement sur les paramètres de l'algorithme de calcul ne contribuent ni au dépassement

[1] Ces observations ont le statut de symptômes, non de diagnostic : autrement dit, elles ne permettent pas à elles seules d'évaluer pleinement une théorie, cf. les propositions de Marandin (1993) pour les analyseurs syntaxiques.

des problèmes entrevus (homogénéité, équilibre, interprétabilité des classes), ni même à la compréhension des propriétés linguistiques qui entrent en jeu dans ces traitements et qui contribuent déjà à leur succès partiel. D'autre part, un certain nombre de difficultés classiques rencontrées par les traitements automatiques des langues sont analysables à la lumière d'une sémantique différentielle textuelle, obligeant à reconsidérer la conception de la langue ancrée dans les algorithmes et les formats de codage.

Pourtant, le potentiel pratique, en terme d'application dans des systèmes informatisés, d'une théorie comme la sémantique interprétative de François Rastier (1987a), est communément perçu comme problématique. Des objections théoriques s'élèvent, en même temps que des réalisations expérimentales s'inspirent de la théorie sans parvenir à l'adopter pleinement. L'ambition de ces lignes est d'indiquer une nouvelle lecture de la sémantique interprétative, fidèle à ses orientations fondatrices, éclairante quant aux observations en sémantique de corpus, et compatible avec une mise en œuvre informatique.

2.Brève introduction au concept de sème

a)Cadre théorique : la sémantique interprétative de François Rastier

Sèmes et classes sémantiques

Nous adoptons ici une sémantique différentielle : le sens y est décrit à l'intérieur de la langue (et non en rapport avec des référents dans le monde physique ou avec des concepts mentaux), par les éléments de signification (*sèmes*) qui sont communs à différents mots[2] (*sèmes génériques*), ou qui les contrastent, les singularisent (*sèmes spécifiques*). Par exemple, 'couteau' et 'fourchette' ont en commun le sème /couvert/, et 'couteau' se distingue de 'fourchette' par le sème /pour couper/.

A travers ces deux rôles (généricité et spécificité), les sèmes organisent donc une structuration du lexique en classes : les sèmes génériques regroupent des mots et délimitent des classes, à l'intérieur desquelles les sèmes spécifiques distinguent les mots comme autant d'éléments de ces classes. En fait, linguistiquement, deux types de structures interviennent (Rastier 1987b : 87). A un premier niveau, les mots se groupent en *taxèmes*, structure paradigmatique, qui se présente mathématiquement comme un *ensemble*, homogène, dans lesquels tous les éléments ont un rôle équivalent. A un second niveau, ces taxèmes se groupent eux-mêmes en *domaines*, selon une structure qui n'est plus ensembliste mais *méréologique*, comme le rapport d'une partie à un tout : les mots de chaque taxème s'articulent en différents rôles dans le domaine. Autrement dit, la structure ensembliste induit des rapports de substituabilité, la structure méréologique induit des rapports de complémentarité.

Un corrélat important de cette organisation en classes, est que tout mot n'est pas interdéfini avec tout autre, mais que les différences ne sont pertinentes que dans les délimitations opérées par les classes. Il ne s'agit pas simplement d'une économie descriptive, c'est aussi le moyen de prévenir des dérives sémasiologiques : différents homonymes n'ont pas de raison d'être définis les uns en fonction des autres, puisqu'ils ne se rencontrent pas dans les mêmes contextes. Les phénomènes d'ambiguïté sont ainsi considérés à leur juste mesure.

Description de la signification des mots : pas de lexique sans contexte

La désignation plus courante de *sémantique componentielle* (Sabah 2000) met l'accent sur la représentation individuelle de chaque mot par un ensemble de sèmes. Il s'agit en fait d'un tout autre courant linguistique, mais le commun recours aux « sèmes » dans la modélisation suscite des confusions, et induit une vision erronée de la sémantique interprétative. En effet, pour la sémantique interprétative, les sèmes ne sont pas des composants sémantiques préexistants, dont l'assemblage servirait à décrire le sens de chaque mot : les sèmes ne sont pas un langage de représentation sémantique, dans lequel chaque signification pourrait être représentée, avec par exemple 'couteau' = /couvert/ + /pour couper/. La perspective est exactement inverse : les sèmes ne se définissent que dans la confrontation de plusieurs unités linguistiques. Si bien que ce sont donc plutôt les sèmes qui sont décrits par un ensemble d'occurrences (/couvert/ = {'fourchette', 'couteau', 'cuillère'}) ; et ce n'est qu'indirectement qu'un ensemble de sèmes peut être attribué à un mot.

[2] Nous choisissons de parler de *mot* pour ne pas alourdir l'exposé, mais il serait plus exact de parler de *lexème* (qui peut être une partie de mot : une racine, un préfixe) ou d'*unité lexicale* (dont des « mots en plusieurs mots »). La définition linguistique du mot reste complexe, sinon problématique.

Dans le même mouvement, on renonce d'emblée à faire des sèmes des primitives, qui seraient universelles, irréductibles, autonomes et exhaustives. Les sèmes sont relatifs aux contextes qui mettent les mots en relations, donc *a fortiori* aux langues. Le niveau de détail de la description en sèmes s'ajuste en fonction du contexte de l'analyse sémantique : il y aurait toujours une description plus fine possible (et donc la quête d'« atomes de sens » serait sans fin), mais en fait une plus grande finesse n'est pertinente que si les distinctions introduites entrent effectivement en jeu au sein des textes considérés (autrement dit, la description est résolument opportuniste). Tel ensemble de sèmes activés dans un texte rend compte d'une lecture de ce texte parmi d'autres possibles, sans que l'on puisse conclure que cette lecture soit unique, complète ou définitive. D'ailleurs, le texte ne contient pas un sens qu'il s'agirait d'extraire ; alors, plutôt qu'une saisie du « contenu », le rôle de la linguistique est de discerner des points d'appui et des contraintes orientant la construction d'un sens, et d'apporter des critères pour évaluer la plausibilité d'une lecture. Avec les sèmes, la description sémantique s'affirme donc relative.

Le texte et le corpus comme terrains d'observation et de construction des sèmes

Le rôle du contexte textuel est par conséquent primordial. Le texte n'est pas une suite arbitraire de mots ou de phrases, on lui reconnaît une propriété de cohésion (le texte est à propos de quelque chose, il a une certaine unité thématique) et une propriété de progression (il ne se répète pas, mais au fil de son déroulement apporte des éléments nouveaux complémentaires). Ces deux propriétés se traduisent respectivement par deux opérations interprétatives, l'*assimilation*, qui établit un lien de similitude, d'homogénéité, d'harmonie sémantique, et la *dissimilation*, qui identifie des contrastes et souligne l'apport sémantique spécifique des occurrences les unes par rapport aux autres. Si chaque mot est en soi (en langue) *a priori* porteur de tel ou tel sème (*sème inhérent*), c'est par abstraction et généralisation des usages en contexte. Les opérations d'assimilation et de dissimilation peuvent remodeler la signification du mot par *inhibition* d'un sème inhérent (par exemple, dans un usage métaphorique), ou par *propagation* / *activation* d'un *sème afférent* (par exemple, une connotation). Cette sémantique est donc par essence dynamique (tout ne peut être prévu et enregistré dans un dictionnaire) et contextuelle, donc relative au texte et à son genre, à l'intertexte (un texte n'est jamais perçu isolément, il se situe par rapport à d'autres textes ; le rassemblement des textes en corpus peut être conçu comme une matérialisation de cette réalité sémantique), à la pratique interprétative ou type de lecture, au lecteur lui-même.

L'*isotopie* est le nom de l'accord sémantique des mots au sein d'un texte : en termes de sèmes, une isotopie correspond à la récurrence d'un sème, c'est-à-dire à son apparition répétée dans un passage du texte[3]. L'étendue du passage n'est pas définitoire, c'est la perception effective d'une répétition qui importe pour l'identification d'une isotopie. L'isotopie donne donc une description unifiée de phénomènes sémantiques à tous les paliers (mot, phrase, texte), ce qui est satisfaisant compte tenu de l'élasticité de l'expression linguistique, à savoir que l'on peut condenser un texte en un résumé d'un paragraphe, voire par un mot, et réciproquement développer un mot en une définition, une paraphrase. Or — et là encore, il faut inverser les idées reçues sur la sémantique interprétative — l'isotopie préexiste au repérage des sèmes[4]. En effet, le lecteur aborde un texte avec une présomption d'isotopie : pour le comprendre, il s'attend à y trouver une certaine cohérence, une certaine unité. Quelques indices (sèmes inhérents à quelques mots, contexte de lecture) aiguillent son attention vers telle isotopie, à partir de laquelle il recherche et établit les sèmes pour la confirmer. En tant que forme privilégiée de manifestation des sèmes dans les textes, l'isotopie est donc un observable essentiel au plan sémantique.

[3] Pour la simplicité de l'exposé, nous parlons d'isotopie, en l'utilisant dans un sens élargi incluant les *paratopies*. « 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, qui pourrait être appelé *paratopie*, est à l'œuvre dans les anaphores associatives » (Rastier 2001 : 202)

[4] « [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 1987a : 11-12)

b)Le point de vue des traitements automatiques des langues

Perspectives prometteuses

Cette théorie sémantique présente au moins trois familles d'atouts pour la conception de traitements linguistiques.

Tout d'abord, la théorie convainc par la justesse de ses observations, et par les phénomènes dont elle rend compte. Par exemple, la perspective descriptive adoptée (rendre compte des usages linguistiques dans les textes, *vs* une perspective normative établissant les bons usages à partir d'un modèle théorique de la langue), perspective descriptive que Rastier radicalise en affirmant la compulsivité de l'interprétation (*i.e*. on ne peut pas s'empêcher de donner du sens), n'est pas sans affinités avec la recherche de traitements robustes (le traitement automatique fournit un résultat malgré les irrégularités et les cas imprévus rencontrés)[5]. La sémantique interprétative prévient aussi des erreurs et difficultés auxquelles se heurtent classiquement les traitements automatiques, comme la résolution des ambiguïtés, puisque la description ne part pas de la forme graphique et isolée des mots mais de leur répartition dans des classes se manifestant en contexte. L'approche textuelle prend également acte de diversités peu à peu reconnues : diversité des langues, diversité des genres textuels, diversité des interprétations d'un même texte, diversité des lexicalisations d'un thème.

L'objet d'étude que se donne la sémantique des textes est doublement pertinent au regard des besoins actuels. C'est bien à des textes réels que les applications automatiques sont utilement confrontées, non à des exemples calibrés ou à des phrases isolées. Et l'ultime visée est le plus souvent sémantique : les efforts se concentrent souvent sur les traitements syntaxiques, avec des insatisfactions (modélisations trop complexes et raffinées par rapport à leur utilisation, implémentations peu performantes), alors que les représentations syntaxiques calculées ne sont généralement pas une fin en soi.

Enfin, la sémantique peut être conçue comme pleinement linguistique, sans dépendre d'une représentation du monde physique ou mental, d'une réalité externe. Certes, elle tourne ainsi ostensiblement le dos aux sémantiques référentielles, massivement dominantes dans les traitements automatiques des langues et en intelligence artificielle, via le recours aux ontologies. Néanmoins, cette nouvelle voie, qui libère d'une lourde tâche de description du monde, est séduisante, pour peu que l'on se donne un recul critique par rapport aux approches coutumières. Cette sémantique linguistique s'affranchit également de fondements cognitifs ou d'une validation par anthropomorphisme : rendre compte des effets sémantiques n'oblige pas à emprunter les mêmes voies que la perception et l'intelligence humaine[6]. Enfin, selon un opportunisme de bon aloi, la description est souple et s'ajuste au contexte. Elle est conçue pour une application et est bien adaptée à elle, ce qui est préférable à une visée trop générale et universelle, jamais complète et donc décevante.

Questions et doutes

Mais les hésitations et les difficultés à mettre en œuvre informatiquement cette théorie n'ont pas toujours trouvé réponse. Le linguiste apporte son expertise sur l'identification et la description des phénomènes de langue ; mais cette description peut être rationnelle sans être formelle, ni surtout limitée à la puissance expressive des formalismes du moment, selon un état de l'art somme toute contingent. Mieux, Rastier explique[7] qu'une théorie linguistique, pour être implémentable, doit se prêter à son propre appauvrissement : c'est un gage de robustesse. D'où une invitation à faire de nécessité vertu, et à reconnaître non seulement que le linguiste ne peut pas tout spécifier, mais aussi qu'il ne le doit pas.

---

[5] Inversement, les astérisques qui épinglent les énoncés rejetés par une linguistique normative (à la suite de Chomsky) sont à mettre en rapport avec les échecs et les rejets d'un analyseur rivé à un lexique préétabli et à une grammaire stricte et rigide.

[6] De même que l'avion vole sans battre des ailes, la modélisation linguistique peut être pertinente sans faire état de la structuration du cerveau et de la mémoire, du fonctionnement des neurones, de la nature et de la disposition des capteurs visuels et auditifs, etc. La linguistique peut être guidée par les résultats de la psycholinguistique ou des neurosciences par exemple, sans être liée à l'avancement et à la maîtrise des connaissances dans ces domaines.

[7] Nous reprenons ici (et dans la suite de cette partie) des éléments apportés par Rastier, en discussion et commentaire de ce paragraphe *Questions et doutes*, et publiés dans le dialogue « Sur les traitements automatiques », liste électronique *Sémantique des Textes*, volume 7, numéro 3, 3 avril 2001.

Les propositions concrètes de mise en œuvre informatique sont ainsi plus indicatives qu'incitatives : les techniques évoquées (calcul de l'écart-réduit[8], classifications hiérarchiques, réseaux neuronaux, reconnaissance des formes) sont mentionnées pour certains comportements intéressants, sans préjuger de leur complète adéquation, ni préciser des choix pourtant déterminants dans leur mise en œuvre (par exemple, quelles entrées et quelles sorties fournir au réseau de neurones). Bien entendu, la manière de recourir à tel ou tel formalisme ne peut se décrire entièrement que dans le cadre d'une application donnée. Il reste que la sémantique interprétative pourrait, d'une manière générale, suggérer certains modes de mise en œuvre, et en exclure d'autres. Quant à la description linguistique elle-même, elle ne s'impose pas d'inventaires systématiques ni d'explicitation complète et définitive, directement retransposable en algorithme : en ce sens, la thèse de Ludovic Tanguy (1997) est un véritable et admirable travail de formalisation de la structure de classes associée aux sèmes.

Il y a finalement des méprises de deux ordres. D'abord, méprise sur les objectifs de la réflexion sémantique : il ne faut pas attendre ici « la théorie du sens des langues », sur laquelle se calquerait une implémentation ; mais plutôt, est réuni tout un faisceau d'éléments de sémantique, cohérents et clarifiants, qui peuvent avoir une incidence dans de multiples implémentations, et qui donnent des points d'appui pour orienter et évaluer les choix de modélisation. De fait, c'est le statut de la théorie qui fait problème : on ne peut demander à une théorie sémantique ce qu'on demande à une grammaire (au sens chomskien). Les « sciences de la culture » et les « sciences de la nature » ont des modes d'objectivation différents, ce qui est encore rarement perçu.

Le deuxième ordre de méprises est plus sournois, il s'agit de toutes ces confusions que nous avons en partie dénoncées ci-avant : l'assimilation des sèmes à des primitives, à des traits sémantiques qui se composent pour décrire le sens, etc. Ces méprises de bonne foi portent sur l'ensemble du courant saussurien, lorsque sont assimilés signifié et concept, sémantique et ontologie, etc.

Pour l'artisan des Traitements Automatiques des Langues, il reste donc bien des étapes et des choix à préciser, tout en restant dans l'esprit de la théorie : contribution des éléments d'analyse textuelle à une application donnée, rôle et place des outils. Par exemple, pour trouver la juste interaction avec un analyseur syntaxique, qui travaille dans les limites de la phrase, il faut sans doute revoir son rôle habituellement central pour pouvoir le concilier avec une sémantique qui s'intéresse à tous les paliers (mots, phrases, texte) de façon unifiée[9].

Enfin, même au niveau de la théorie, bien des chantiers sont ouverts, bien des questions restent à clarifier. Les typologies de textes proposées jusqu'alors sont rejetées, tout en affirmant la nécessité de prendre en compte les genres textuels — genres qui ne se résument pas à de simples types de textes (Rastier 2001). L'utilisation de dictionnaires de langue informatisés est envisagée pour le repérage automatique des sèmes (notamment les indicateurs de domaines, qui font écho à l'éventail des contextes d'usage de la langue), alors même que sont niés l'existence d'une langue générale[10] et le bien-fondé d'une atomisation du sens au niveau des mots.

La détermination du local par le global (c'est l'isotopie qui permet l'identification des sèmes) pose comme caractéristique fondamentale la non compositionnalité du sens ; les différentes composantes sémantiques interagissent en hétérarchie ; et le cercle herméneutique dont la validité est affirmée évoque de façon inquiétante les boucles sans fin d'un programme qui ne parvient pas à converger… Or les ordinateurs et les outils de calculs à notre disposition sont par conception compositionnels, séquentiels, déterministes. Tout l'art du concepteur de programme consiste alors à reporter le plus loin possible ces caractéristiques inhérentes, et à penser l'algorithme du calcul autrement : par exemple, recours aux statistiques (pour une vue synthétique et globale), modularité (avec la conception orientée objet, les modèles hétérarchiques des agents), l'usage de fonctions aléatoires, etc. Et la question de la modélisation reste ouverte : quelle représentation des textes donner pour le calcul ?

De fait, la théorie ne cherche pas à donner accès à une représentation du sens, seulement à repérer les contraintes linguistiques sur la construction du sens : la sémantique relève de la

---

[8] En statistique textuelle, l'écart réduit est utilisé comme indicateur *(i)* de mots caractéristiques à une partie d'un corpus, ou *(ii)* de mots contextuellement associés à un mot donné — voir par exemple (Rastier 1995).

[9] Les travaux en cours de Denise Malrieu (MoDyCo – Paris X) pourront éclairer cette interaction des descriptions lexicale, phrastique et textuelle. Elle utilise un analyseur morpho-syntaxique pour une caractérisation des genres textuels, dans l'esprit de la sémantique interprétative.

[10] Plus exactement, « il y a bien une langue générale, mais pour le lexique elle se réduit à un inventaire de morphèmes et à des principes constructifs. En ce sens, on peut dire que le lexique n'appartient pas à la langue, alors même qu'il la résume pour l'opinion la plus généralement reçue. » (Rastier, *SdT*, 7 (3) – cf. note 7).

linguistique, mais le sens n'appartient pas à la langue et n'est pas immanent aux textes. Alors qu'avec la syntaxe, l'idée (illusoire) était d'être capable de décrire tout texte quel qu'il soit, indépendamment de son domaine ou de son genre, la sémantique interprétative vise la construction de représentations dynamiques, partielles, relatives, non uniques. Quant à la diversité interne des langues et à la relativité de la description, elles interrogent l'intérêt même d'une automatisation, puisqu'il n'y a plus la généralité de la description pour gager la réutilisabilité de l'application informatique. Si réutilisabilité il y a, ce sera au niveau de la méthodologie mise au point pour construire une application, non au niveau du logiciel développé lui-même.

3.Manifestations expérimentales du rôle sémantique du contexte : les sèmes incognito

a)Modèle de l'espace vectoriel : la thématique n'est pas dans les mots

Le modèle de l'espace vectoriel est issu des travaux de Salton dans le domaine de la recherche d'information (Salton & McGill 1983) (Baeza-Yates & Ribeiro-Neto 1999). Tout texte est représenté en fonction des mots qu'il contient (ou en fonction des « mots-clés » issus d'une indexation manuelle ou d'une analyse automatique du texte), par un vecteur dans un espace mathématique dont chaque dimension correspond à un mot. Les dimensions sont orthogonales, si bien que chaque mot concourt de façon indépendante des autres mots à la caractérisation du texte. Pour affiner la représentation, une pondération[11] ajuste les coordonnées pour traduire, de façon significative et équilibrée, l'influence plus ou moins importante des différents mots dans chaque texte.

Cette représentation est évidemment réductrice, mais son but est de donner un moyen opérationnel d'évaluer le degré de similarité entre deux textes, par une mesure de l'écart entre leurs vecteurs représentatifs. Le produit scalaire, opération de base pour confronter deux vecteurs[12], se calcule concrètement comme la somme de contributions ponctuelles, dimension par dimension. Du point de vue des textes, c'est donc assimiler leur ressemblance globale (de texte à texte) à un cumul de ressemblances ponctuelles (mot par mot).

La technique de l'espace vectoriel est actuellement la plus appropriée, par sa simplicité de mise en œuvre et son efficacité, pour la réalisation de systèmes documentaires sur des fonds en texte intégral[13]. C'est le cas par exemple des moteurs de recherche d'Internet, ou encore de l'application de diffusion ciblée d'informations à EDF (Bommier-Pincemin 1999)[14]. Dans ce cadre d'expérience, un point fort et un point faible ressortent, qui retiennent notre attention puisqu'ils sont analysables à travers le concept d'isotopie. Examinons-les tour à tour.

Alors que la représentation vectorielle est très souple et peut être utilisée aussi bien pour une requête (formulée par un ou quelques mots-clés) que pour un texte entier (Salton 1988), les similarités calculées sont d'autant plus justes que les rapprochements peuvent se baser sur un faisceau de mots, plutôt que sur un mot seul (même à forte pondération). En effet, un mot isolé se prête à des interprétations divergentes, notamment par polysémie et homonymie. Lorsque plusieurs mots se retrouvent ensemble d'un texte à un autre, ils ont pour effet de cerner un domaine, donc de motiver la similarité par une thématique précise. Leur récurrence groupée dans les textes est la trace d'une isotopie : c'est alors bien sur une même composante sémantique que les deux textes sont rapprochés.

---

[11] La pondération caractérise la significativité d'un mot dans un texte. Elle est calculée par une fonction généralement sensible à la fois à la forte présence du mot dans le texte (pour valoriser le vocabulaire relatif à la thématique dominante du texte), et à la rareté du mot dans le corpus (pour mettre en valeur l'apport spécifique de chaque texte, et neutraliser le vocabulaire non thématique : mots grammaticaux, liés à la forme, etc.).

[12] Le produit scalaire intervient par exemple dans le calcul du cosinus, mesure de similarité classiquement choisie.

[13] En particulier, pour des fonds dont le volume et l'évolution constante et rapide conduisent à renoncer à une analyse documentaire et un catalogage par des professionnels.

[14] Une base de profils de destinataires, représentative des activités de l'organisme, est construite et mise à jour automatiquement. Dans le cas de l'application de diffusion ciblée d'EDF, les destinataires décrits dans la base sont tous les chercheurs responsables de projet à la Division Recherche et Développement de l'entreprise. Leurs profils sont calculés à partir de l'analyse automatique des textes établissant leurs programmes de recherche annuels. Cette base de profils permet alors la recherche, pour un document donné, des chercheurs les plus concernés et des experts les plus qualifiés sur un sujet : l'application indique les profils les plus proches du document, au sens d'une mesure de similarité entre textes. Ceci permet un envoi sélectif, d'où le nom de diffusion ciblée.

Ce gain en précision se double d'un gain en complétude. En effet, une notion n'est pas toujours mentionnée de la même façon dans les textes, ce qui se traduit par la difficulté de trouver « le bon mot-clé » dans l'interrogation d'un moteur de recherche. En revanche, des regroupements de mots dans le contexte d'un texte introduisent une redondance sém(ant)ique, et l'évocation d'un thème reste dans le cadre global du vocabulaire du domaine : les recouvrements de vocabulaire permettent de trouver les textes en relation, par delà les variantes de leur expression linguistique[15].

D'où le choix du texte comme mode privilégié d'interrogation de l'application d'EDF : l'utilisateur qui cherche à faire la diffusion ciblée d'un document est très vivement invité à soumettre sa demande sous la forme du texte de son document ou d'un passage caractéristique de celui-ci, plutôt que sous la forme d'un ou deux mots-clés représentatifs –le résultat attendu est en effet bien meilleur. Le même genre d'observations a été fait pour les moteurs de recherche d'Internet : la tendance des utilisateurs est de lancer une recherche à partir d'un ou deux mots, or il y a un net saut qualitatif entre les pages sélectionnées à partir d'un seul mot, et celles qui comportent plusieurs mots de la requête et qui s'avèrent globalement plus pertinentes. Ce constat conduit par exemple à faire du nombre de mots communs à la requête et au document un critère décisif pour l'évaluation de la pertinence (Clarke & *al*. 1997).

Côté points faibles, le modèle de l'espace vectoriel est lourdement tributaire de la désarticulation du texte sous forme d'une série de mots, coupés de leur ancrage contextuel et isolés sur des dimensions orthogonales. Seule la pondération introduit un rééquilibrage global, au pouvoir expressif réduit. Le calcul de rapprochement ne fournit qu'un score cumulatif, qui dans le meilleur des cas réunit *a posteriori* des mots en relation d'isotopie, mais qui peut tout aussi bien juxtaposer des mots sans relation significative entre eux, issus de contextes différents. Du caractère numérique et additif de la mesure dérivent deux cas pathologiques opposés : un rapprochement effectué sur un seul mot à forte pondération (mais dépourvu de contexte), et un rapprochement résultant d'une accumulation de mots de faible importance et sans lien sémantique consistant.

Quelques exemples observés dans le contexte de l'application de diffusion ciblée d'EDF illustrent ces différents cas de figure. Les données ont été recueillies lors de l'envoi de l'annonce de soutenance de (Bommier-Pincemin 1999), comprenant le résumé de la thèse et des indications de date et de lieu. Pour chaque destinataire proposé, l'application indique les six mots (au plus) qui ont le plus contribué au rapprochement du profil avec le document soumis (l'annonce de soutenance). On retrouve les cas positifs et négatifs indiqués :

(i) contextuatisation mutuelle des mots et isotopie sous-jacente :

rapprochement avec profil 1 : `linguistique, interprétation, informations, automatique, partir, texte…`

rapprochement avec profil 2 : `SGML, corpus, documents, DTD, électronique, intégral…`

Les deux destinataires « expliqués » par l'une et l'autre de ces séries, sont clairement concernés par une thématique commune avec la thèse : linguistique computationnelle pour le premier, édition électronique et codage des corpus pour le second.

(ii) décontextualisation d'un mot isolé, suscitant un rapprochement à cause de sa forte pondération :

rapprochement avec profil 3 : `diffusion, département, contexte.`

Le rapprochement est basé uniquement sur ces trois mots, et les pondérations de 'département' et 'contexte' sont négligeables, si bien que 'diffusion' suscite à lui seul la sélection du profil — et à EDF, il peut s'agir d'un tout autre domaine, d'*'équation de diffusion de la chaleur'* par exemple. Lorsque le modèle vectoriel est appliqué à des requêtes textuelles, comme ici, plutôt qu'à des requêtes par un ou quelques mots-clés, les rapprochements sur un seul mot isolé sont rares. C'est en revanche un des principaux facteurs d'erreurs dans l'usage des moteurs de recherche internet.

(iii) cumul de mots sans lien contextuel significatif :

rapprochement avec profil 4 : `salle, système, mesure, prises, moyen, documentation,…`

---

[15] « En principe, on pourrait avoir différents textes qui traitent de sujets tout à fait en rapport en termes complètement distincts. Si tel est le cas, une méthode par comparaison de vocabulaire ne convient pas pour la caractérisation des textes. En pratique, on ne décrit pas facilement des faits identiques sans utiliser une bonne partie du vocabulaire en commun. Un tel recouvrement de vocabulaire est alors repérable par des méthodes bien conçues de mise en correspondance d'éléments textuels. » (Salton, Allan, Buckley 1994). Les auteurs précisent que, certes, il existe des cas où des domaines différents partagent un même vocabulaire. Ils donnent comme exemple rencontré le cas d'un rapprochement entre une phrase sur le football américain, et une sur la théorie des jeux (mathématiques probabilistes), en raison de mots comme 'games', 'play' et 'team(s)'. Il faut admettre que de tels recouvrements de vocabulaire existent, mais restent relativement rares ; mieux, ils ne sont pas non plus totalement dénués de signification — le déplacement métaphorique garde une portée sémantique.

rapprochement avec profil 5 : `Paris, thèse, partir, conception, porte, système,…`

La lecture de ces listes permet de diagnostiquer immédiatement le peu de pertinence de ces destinataires, alors que le score de similarité (qui a sélectionné ces propositions) est trompeur.

La pratique d'applications mettant en œuvre le modèle de l'espace vectoriel fait donc ressortir concrètement le dénuement sémantique du mot isolé, par contraste avec le renforcement et la contextualisation mutuels d'un ensemble de mots en contexte. La pratique de l'utilisation de mots isolés est issue de l'interrogation des bases documentaires et bibliographiques. Elle réussit dans ce cadre d'origine, puisque les mots-clés utilisés par les documentalistes sont choisis dans un référentiel (thesaurus, liste d'autorité) et sont intrinsèquement dotés d'un contexte par leur positionnement dans ce référentiel. Elle échoue pour les moteurs de recherche basés sur le modèle de l'espace vectoriel, dès lors que le texte est représenté par ses mots extraits de leur contexte et isolés les uns des autres.

b)L'indispensable recours au contexte pour l'interprétation

Dans le domaine de la consultation de grandes bases de textes électroniques, le succès des outils de calcul de concordances est remarquable. Soit c'est la fonctionnalité centrale de parcours et d'analyse des textes ; soit elle cohabite avec d'autres fonctionnalités beaucoup plus évoluées (calculs statistiques opérant des sélections, des contrastes, génération de visualisations graphiques), mais son rôle et son utilisation restent majeurs. Or le principe même des concordances, c'est de présenter un mot dans ses contextes. Une des principales exploitations des concordances consiste à repérer les cooccurrents significativement associés au mot pris comme point de départ. Là encore, la pratique montre bien ce besoin de retour au texte et au contexte : même sélectionné sur un critère statistique raisonné, le mot isolé ne se suffit pas à lui seul, son interprétation fait appel à ses contextes d'occurrences pour restituer les isotopies dans lesquelles il s'inscrit et percevoir sa valeur sémantique.

4.Construction automatique de contextes : explication des limites des heuristiques à la lumière la théorie sémantique des isotopies

a)La reformulation de requêtes en recherche d'informations

Les travaux sur la recherche d'information dans des bases de documents en texte intégral a donc évidemment révélé l'indigence des recherches sur quelques mots, dépourvus de tout contexte. Un axe d'amélioration très couramment envisagé est de compléter la requête en ajoutant à chaque mot l'ensemble de ses synonymes, pris dans un dictionnaire ou une terminologie structurée (thesaurus, réseau sémantique) (Fluhr 2000). L'idée est de corriger ainsi une « imperfection de la langue »[16], à savoir la non correspondance entre les mots et les concepts : en particulier, une même idée peut être exprimée de multiples façons, et par plusieurs mots (synonymie). Compléter chaque mot de la requête par l'ensemble de ses synonymes, éventuellement aussi par des mots voisins plus génériques (hyperonymes) ou plus spécifiques (hyponymes), assurerait donc le repérage du concept associé dans tout texte, quelle que soit l'expression choisie pour le formuler.

Les résultats de ce correctif ne sont pas à la hauteur des espérances. La sémantique interprétative apporte plusieurs éléments d'explication. Il y a tout d'abord erreur sur le fonctionnement sémantique de la langue, qui d'une part ne repose pas sur la composition de la signification des mots (le sens se construit en contexte, dans l'interaction des mots), et d'autre part ne correspond pas à un appariement de la langue avec des concepts extra-linguistiques. Ensuite, la manière même de procéder (ajout, pour chaque mot, d'un paquet de synonymes) est doublement décontextualisée : *(i)* elle suppose qu'un mot[17] est doté d'un seul ensemble de synonymes ; *(ii)* l'apport se fait en considérant chaque mot de la requête indépendamment, sans même exploiter le contexte minimal que celle-ci fournit déjà.

[16] Mais est-ce la langue qui est inadéquate, ou au contraire le modèle que l'on s'en fait ? Ce numéro prend le parti (fructueux) de la seconde alternative.

[17] Il n'y a pas d'ensemble de synonymes associé, hors contexte, à une unité *lexicale*, et donc à plus forte raison encore à une unité *graphique* (chaîne de caractères) susceptible d'homonymie.

Une autre voie est beaucoup plus porteuse : c'est celle explorée par exemple par les *Live Topics* du moteur de recherche *AltaVista* (Bourdoncle 1997) (Fig. 1). A partir d'un premier ensemble de documents sélectionnés par la requête, une classification automatique de l'ensemble des mots qui apparaissent dans ces documents produit quelques groupements de mots, associés entre eux au vu de leurs cooccurrences. Cette approche est plus compatible avec la sémantique interprétative : l'exploitation des cooccurrences permet de capter des associations sémantiquement motivées (mots indexés sur une même isotopie), l'association des mots est pleinement relative au contexte (les classes sont recalculées pour chaque requête, et dépendent aussi des documents présents dans l'espace de recherche), enfin les mots proposés ne sont pas uniquement en relation d'équivalence sémantique (synonymie) mais décrivent des rapports variés.

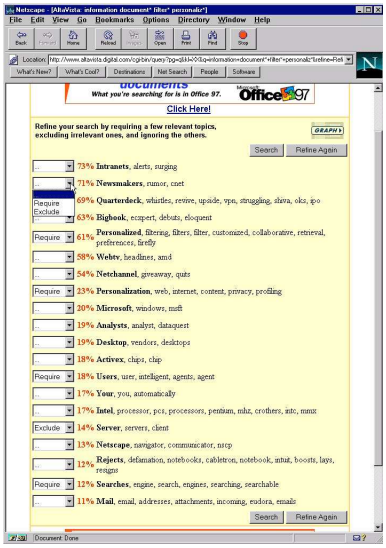

Fig. 1 : LiveTopics

b)Classifications automatiques : le moule inaperçu des algorithmes standards

Les classifications automatiques de mots en fonction des contextes (documents, syntagmes) dans lesquels ils apparaissent font l'objet de multiples travaux : en effet, les expérimentations produisent des résultats intéressants, pas entièrement satisfaisants mais prometteurs.

Les classes regroupent des mots qui ont une affinité sémantique contextuelle, relativement au corpus sur lequel s'est effectué le calcul. Les associations peuvent tenir compte de l'ordre de succession des cooccurrents, et être calculées au premier ordre (mots qui cooccurrent — associations à dominante syntagmatique : « associates ») ou au second ordre (mots qui ont des ensembles de cooccurrents similaires — associations à orientation paradigmatique : « parallels ») (Schütze & Pedersen 1993). Par exemple, ces auteurs obtiennent, pour le mot 'wrote' dans le *Oxford Hector Pilot Corpus* :

relation	mots sélectionnés avec le plus fort score
left associates	he, yeats, lewis, who, edward, jack, richard, owen, tom, morrissey
right associates	book, poem, letter, poetry, letters, stories, poems, books, novel, article
left parallels	wrote, writes, describes, spoke, recalled, remarked, answered, liked, remembered, studied
right parallels	wrote, write, writing, read, written, copy, reading, author, published, poems

L'interprétation des regroupements n'est pas toujours aisée : le calcul d'étiquettes nommant chaque classe peut aider, mais surtout les fonctionnalités de navigation facilitant le retour aux textes sont des plus utiles. Même avec cet effort nécessaire d'interprétation en corpus, il reste des classes peu significatives et des intrus, que l'on s'efforce d'élaguer et de corriger. L'élagage se traduit par un filtrage très sélectif, en particulier la non prise en compte des faibles fréquences. L'attention se centre sur l'obtention de « bonnes » classes, plutôt que sur la construction d'une représentation (fidèle, complète) du corpus. Les ajustements et corrections jouent sur les nombreux paramètres des calculs : définition des contextes, normalisation des variantes linguistiques, choix des fonctions de distance, choix de l'algorithme de classification, seuils. Ces ajustements se font au coup par coup, corpus par corpus, heuristiquement, en fonction du caractère satisfaisant des classes obtenues plutôt que par une interprétation linguistique des calculs.

Ce faisant, la prégnance des algorithmes existants (partitionnement, classification hiérarchique ascendante ou descendante) fait oublier leurs caractéristiques très fortes et restrictives : les classes sont constituées en partition (un élément se retrouve dans une et une seule classe), et sont considérées sur le modèle des classes d'équivalence, avec les propriétés de réflexivité, symétrie, transitivité (Warnesson 1985). Or ces propriétés structurelles ne sont pas celles de la langue : pas d'exacte synonymie, homonymies, glissements de sens,… et plus généralement, non superposition des plans du contenu et de l'expression, comme le souligne Hjelmslev. Dans notre cadre d'analyse, si les classes visent à traduire des isotopies, il faut qu'un mot puisse être affecté à plusieurs classes (il peut porter plusieurs sèmes l'indexant sur des isotopies différentes), il faut aussi qu'il puisse n'être affecté à aucune (pas de participation significative à la formation d'une isotopie dans le cadre du corpus considéré) (cf. § 6.b). Les algorithmes produisent donc naturellement une représentation déformée des associations

sémantiques : les contraintes mathématiques, qui forcent la répartition complète des mots dans des classes disjointes, induisent des distorsions linguistiques.

Quelques tentatives intéressantes sont expérimentées pour contourner cette limitation des classifications automatiques de mots. La première consiste à exploiter la dualité de la modélisation, et à classer non pas les mots en fonction des contextes, mais les contextes en fonction des mots (Meunier & *al*. 1997). On obtient donc une répartition des contextes en classes, et de chacun de ces groupes de contextes on dérive un ensemble des mots les plus représentatifs. Rien n'empêche un mot d'être caractéristique de plusieurs classes, et un mot peut également n'être associé à aucune classe. L'inconvénient de cette tactique est de reverser sur les contextes les contraintes dont on a libéré le mot. La classification de chaque contexte dans une et une seule classe n'est pas moins inadaptée à la nature linguistique des textes, et là encore induit des distorsions.

Une tactique voisine considère les associations de mots que l'on peut tirer d'une analyse factorielle des contextes. L'analyse factorielle calcule un système d'axes, optimal pour la représentation du corpus, en rendant compte des interactions des mots en contexte mieux que le système d'axes initial qui fait de chaque mot une dimension indépendante dans l'espace de représentation. Mathématiquement, la transformation est une combinaison linéaire : les axes calculés se traduisent comme une somme de mots pondérés. Un axe associe plusieurs mots, et un mot peut répartir sa contribution sur plusieurs axes[18]. La difficulté vient ici des limites expressives de la combinaison linéaire : la somme additive opère un cumul, qui ne préjuge pas des interactions effectives entre les mots ; la pondération établit un continuum entre la forte contribution et l'absence de contribution, alors qu'il y a des seuils qualitatifs (par exemple, entre inhérence et afférence), et d'autre part elle introduit une dépendance entre des contextes qui ne sont pas interdéfinis (un mot polysémique voit son influence réduite, car répartie entre ses diverses significations).

Ces observations sur l'inadéquation des pondérations s'étendent aux classifications floues. Celles-ci assouplissent les classifications traditionnelles, en donnant ici pour chaque mot son degré d'appartenance à une classe (Fig. 2). Comme l'influence totale d'un mot doit se redistribuer dans ses différentes contributions, une contrainte sémasiologique s'instaure, qui exacerbe la polysémie (émiettement du mot) au lieu de la résoudre en contexte[19]. La remarque sur le continuum (au paragraphe précédent) vaut également pour le degré d'appartenance, qui ne rend pas compte des sauts qualitatifs et des seuils linguistiques.

Des paramétrages des algorithmes de classification nous avons donc repoussé la question à la conception des algorithmes ; et il faudrait encore remonter en amont et considérer la forme des opérations élémentaires (calculs de distance) mobilisées par les classifications.

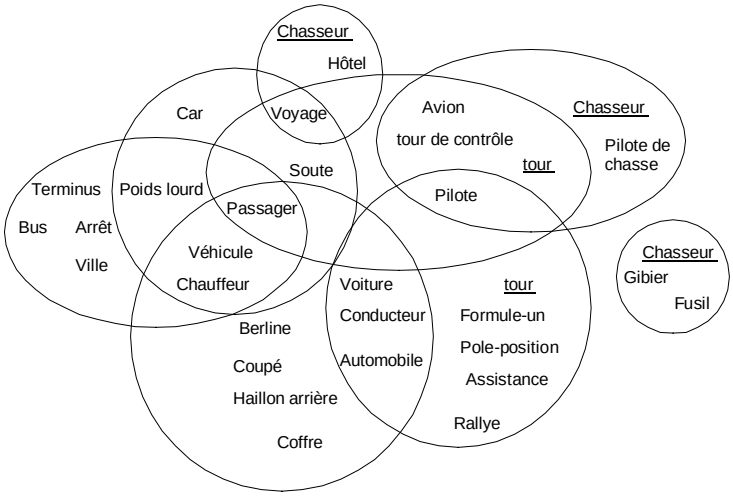

Fig. 2 : Regroupement de termes par classification floue (Gonzalez Rubio & Guizol 1997)

Dans ce diagramme, « les associations trouvées dans les textes ont permis une séparation de divers engins motorisés selon les caractéristiques qui leur sont propres. Par exemple, le terme de 'coffre' est spécifique d'une voiture de tourisme […]. Par contre les termes tels que 'chasse' et a fortiori 'chasseur' ne sont pas pertinents pour la séparation des documents. Corrélativement, leur pouvoir discriminant sera reconnu faible. »

[18] Cette transformation et réduction du système d'axes en fonction des associations des mots est le cœur de la technique LSA, *Latent Semantic Analysis* (Deerwester & *al*. 1990), ou *Latent Semantic Indexing* pour les applications à la recherche d'information (Baeza-Yates & Ribeiro-Neto 1999) (Manning & Schütze 1999).

[19] L'ampleur des phénomènes d'ambiguïté observée avec les traitements automatiques des langues est artificielle : pour le lecteur ou l'interlocuteur humain, les situations d'ambiguïtés sont rares et pour la plupart délibérément produites.

c)Une modélisation fortement binaire

Les classifications procèdent par agrégations ou divisions successives pour modeler plusieurs classes à partir d'un ensemble initial d'items. La procédure s'effectue en de multiples étapes élémentaires, qui à chaque fois considèrent *deux* entités : deux éléments, un élément et une classe, ou deux classes (deux sous-ensembles à agréger pour une classification ascendante, ou deux parties résultant d'une scission pour une classification descendante). Or, même si « plusieurs » commence à « deux », et que deux occurrences de sèmes soient le minimum pour étayer une isotopie, rien ne permet de conclure que les constructions et les interactions linguistiques, notamment sémantiques[20], se laissent analyser en interactions binaires. Considérons les classifications de mots en fonction de leur cooccurrence dans un voisinage de l'ordre de la phrase : l'idée est de grouper peu à peu des mots à partir de dépendances syntagmatiques d'un mot avec un autre. Or les structures actancielles mobilisent une constellation de rôles, ou encore les prédicats peuvent avoir un nombre variable d'arguments, deux mais aussi bien un ou trois. De même, deux occurrences d'un sème ne sont pas toujours suffisantes pour ancrer une isotopie, c'est plutôt une convergence sémantique d'un ensemble d'occurrences qui confirme la présomption d'isotopie. La validité d'une décomposition binaire des interactions linguistiques suppose la possibilité d'une modélisation compositionnelle de la langue, ce que théorie et observations tendent à infirmer (Nazarenko 1998). Pour être plus proche des réalités linguistiques, il faudrait donc repenser les fonctions de similarité, de distance, d'association, pour construire et introduire dans les traitements des fonctions d'évaluation globale de cohésion, de concentration.

Les limitations d'une modélisation binaire ont été perçues, mais non résolues, dans les modèles probabilistes (notamment pour les applications en recherche d'information). En première approche, l'hypothèse d'indépendance entre les mots est admise comme une approximation. Ensuite, l'introduction graduelle de dépendances d'ordre 1, puis d'ordre 2, etc., se heurte d'un côté à la limitation brutale des interactions linguistiques (s'arrêter à l'ordre *n* c'est ne pas pouvoir décrire l'ordre *n+1*), de l'autre à la complexification combinatoire de la modélisation, dont les rouages échappent à une compréhension d'ensemble fine et défient les capacités de calcul.

Sur un autre terrain encore, les formalismes de représentation visuelle des relations lexicales et des structures sémantiques (graphes, réseaux) développés avec l'intelligence artificielle et la linguistique computationnelle, induisent fortement une conception binaire des interactions entre unités linguistiques. Les relations sont notées par des segments ou des flèches, qui n'ont jamais que deux extrémités. Des notations permettent d'indiquer des variations dans la multiplicité de la réalisation d'une extrémité, mais la relation reste fondamentalement binaire, la multiplication des exemplaires ne modifiant pas la nature de la relation.

d)Cooccurrence : à la recherche du bon voisinage

Par définition, le calcul de cooccurrences suppose la définition de zones dans lesquelles les unités (mots) apparaissent (« occurrent ») ensemble (« co- »). Curieusement, chaque réalisation 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). Examinons tour à tour les arguments avancés pour la détermination du voisinage de cooccurrence.

La phrase, l'énoncé

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. Ce voisinage a également une pertinence cognitive, en ce que son empan correspond aux capacités de la mémoire à court terme.

Ces atouts linguistiques se doublent de vertus statistiques : la phrase courante a une taille suffisamment petite pour être sélective. D'une part cela évite le foisonnement de cooccurrences

[20] Mais la sémantique n'est guère dissociable des « autres niveaux » de description linguistique (syntaxe, morphologie, lexique, prosodie, pragmatique…) (Rastier 1987b).

examinées ; et d'autre part, la significativité statistique des écarts de répartition ainsi mesurés est accrue.

Cependant, le découpage d'un texte en phrases est problématique. La syntaxe reconnaît plutôt la proposition grammaticale. Le concept de phrase, et son marquage typographique, sont d'ailleurs apparus tardivement. Sémantiquement, les frontières de la phrase sont perméables, et les coupures que l'on opère en isolant ces contextes ont une part d'arbitraire et d'exclusion brutale.

Le paragraphe

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.

Le paragraphe a aussi pour lui une dimension cognitive, au plan de la mémorisation comme de la perception. S'en tenir au paragraphe évite l'excès qu'il y aurait à ne faire aucune différence entre des mentions que séparent plusieurs pages et des mentions qui voisinent dans les mêmes lignes : l'appréhension 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.

Le repérage des paragraphes n'est pas moins problématique que celui des phrases. Le retour à la ligne est interprété différemment selon le contexte : longueur des paragraphes, présentation sous forme de liste, etc.

Le texte

Le texte se présente comme une unité sémantiquement autonome. 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 » (tels que des dépêches, ou des annonces sur les forums électroniques), le texte est l'unité considérée pour décrire les relations entre les mots. Pour les documents plus longs, un découpage selon les parties logiques (sections, chapitres) donne des contextes plus développés que le paragraphe, tout en ayant une certaine autonomie, une longueur *a priori* plus régulière et une cohérence interne forte. Une partie logique est en effet du même ordre qu'un texte. Comme lui, elle a 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*.

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.

La difficulté vient principalement de la diversité des genres textuels, avec *(i)* des variations de taille, pouvant susciter des déséquilibres au niveau des calculs statistiques, *(ii)* des variations lexicales, qui mêlent au vocabulaire thématique le vocabulaire rédactionnel conventionnel pour le type de texte, et *(iii)* la nature plus ou moins homogène ou composite des documents.

e)Contexte syntaxique et sémantique distributionnelle

Dans le courant des calculs d'associations lexicales à partir de corpus, une branche de travaux se détache, en revendiquant son recours à une description linguistique tant pour la définition des unités lexicales (prise en compte notamment d'unités composées) que pour l'identification des contextes (interrelations plus précises que la proximité syntagmatique). Ces travaux se réclament de l'approche distributionnelle initiée par Harris (Habert & Zweigenbaum 2001). L'analyse syntaxique permet de repérer les mots qui se substituent les uns aux autres à une position précise (c'est-à-dire dans une relation de rection ou de dépendance identifiée à l'égard d'autres mots ou classes de mots fixés). Ce « test de commutation » en corpus fournit des paradigmes[21]. Les résultats les plus spectaculaires sont obtenus sur l'analyse des groupes nominaux dans des corpus correspondant à des pratiques

[21] Les classes ainsi construites s'apparentent à des *taxèmes* de la sémantique interprétative : elles décrivent un certain type d'isotopies (microgénérique). Elles peuvent être réutilisées comme éléments pour la construction d'autres isotopies.

professionnelles dans lesquelles une terminologie précise a cours (Habert & Fabre 1999) : les classes reconstituent des familles de synonymes, des paires d'antonymes, des variations scalaires ou des énumération de valeurs sur une même dimension sémantique.

Par exemple, dans le contexte du *Guide de planification des réseaux électriques*, Assadi & *al*. (1995) obtiennent des classes d'adjectifs comme :

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

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

Pour générer ce genre de classes, les calculs se focalisent sur la structure interne des termes composés, selon une description stable et unifiée présentant les trois caractéristiques *(i)* binaire (décomposition en une Tête et une Expansion), *(ii)* asymétrique (l'Expansion dépend syntaxiquement de la Tête et n'est pas interchangeable avec elle), et *(iii)* récursive (la structure des termes plus complexes se laisse décomposer comme une série de duos Tête – Expansion imbriqués). Toute unité linguistique composante de termes est donc contextuellement et linguistiquement associée à l'ensemble des unités Expansions dont elle est la Tête, et à l'ensemble des unités Têtes dont elle est Expansion.

Habert & *al*. (1997) généralisent en quelque sorte ce type d'approche, en limitant l'ensemble de la description syntaxique à un ensemble de dépendances binaires élémentaires, incluant la reconnaissance de dépendances à distance et de variantes par dérivation, élision ou insertion[22]. Par exemple, un syntagme nominal complexe extrait par l'extracteur terminologique LEXTER, décrit par l'arbre d'analyse (Fig. 3a), est transformé en arbre normalisé binaire (Fig. 3b), puis ramené à quatre dépendances élémentaires (Fig. 3c).

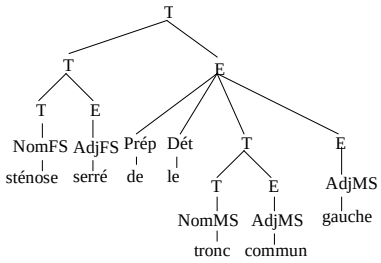

Fig. 3a : Arbre LEXTER

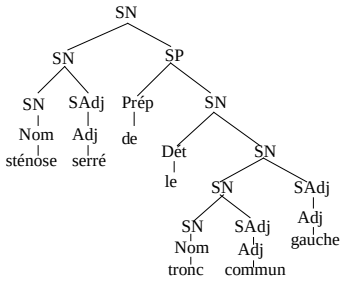

Fig. 3b : Arbre normalisé

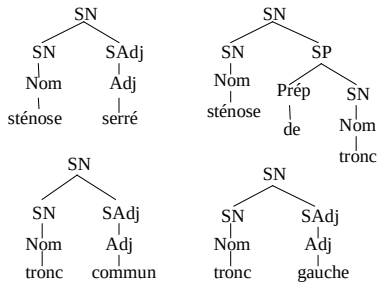

Fig. 3c : Dépendances élémentaires

Cette simplification syntaxique a de nombreuses vertus : au plan de la faisabilité, elle peut être obtenue entièrement automatiquement par des analyseurs robustes, et permet ainsi de sortir des cas d'école en travaillant sur des corpus vastes de textes réels. En outre, elle met en valeur les régularités dans les associations de mots, par delà les variantes d'usage[23]. En revanche, cette décomposition extrême des structures syntaxiques se paie par un morcellement des contextes, et par une représentation fortement compositionnelle (les structures *n*-aires sont décomposées en structures binaires).

L'intuition de Rajman (1995, §III.5), comme celle des travaux précédents, est que l'utilisation exclusive des relations syntaxiques est doublement avantageuse par rapport à la simple utilisation de la

[22] L'identification des variantes de termes s'inscrit dans la suite des travaux de Christian Jacquemin. Sont mises en évidence des équivalences comme : *'sténose sévère'* / *'sévérité de sténose'* ; *'angine de poitrine instable'* / *'angine instable'* ; *'tronc commun gauche'* / *'tronc gauche'* ; *'obstruction sévère d'une artère coronaire'* / *'obstruction coronarienne sévère'*.

[23] Il y a moins d'associations binaires différentes que de combinaisons *n*-aires différentes : la description est donc plus redondante, moins dispersée. Elle permet ainsi le repérage de recouvrements partiels, la mise en évidence de points communs même s'il n'y a pas stricte identité des syntagmes complets.

phrase comme contexte de mise en relation des unités. Au plan de la théorie, on s'en tient à des relations identifiées linguistiquement. Au plan pratique, la combinatoire des liens à examiner est fortement réduite, ce qui soulage les calculs, voire conditionne la faisabilité de l'expérimentation (Besançon & *al*. 1999) (Rajman & *al*. 2000). Notre discussion s'établira sur le fait que cette réduction ne s'opère pas sans pertes. Concrètement, pour l'énoncé :

Le chat de la voisine miaule sur la palissade blanche

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 Rajman pour éviter l'association « pathologique » '(miauler, palissade)'), il nous semble dangereusement réducteur de limiter les interactions sémantiques à la syntaxe. Une paire comme '(miauler, palissade)' ou '(miauler, 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 déborde 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.

Ne serait-ce que par cette contribution directe à la délimitation et à la contextualisation des unités linguistiques, la syntaxe est un support important de la construction du sens. Cependant, elle n'enferme pas celui-ci. La typographie, la morphologie, la phonétique, sont des exemples d'autres dimensions qui interviennent au plan sémantique, avec l'établissement d'autres voisinages, d'autres relations. Il ne faut donc pas vouloir en réclamer plus à la syntaxe que ce qu'elle peut nous dire de la langue : elle n'en condense pas tout le potentiel sémantique (Habert & Nazarenko 1996).

5.La modélisation linguistique : le poids des présupposés structurels

Les modèles traditionnels des traitements automatiques des langues masquent certains aspects linguistiques soulignés par la sémantique interprétative. Plusieurs limites et difficultés sont ainsi identifiables et analysables.

a)Le rôle des ressources lexicales : dictionnaires, terminologies

Techniquement, l'analyse se base sur la reconnaissance des mots, à partir de laquelle s'opère une construction du sens. Cette architecture est appropriée à l'ordinateur, à ses calculs et manipulations de réécritures ; mais elle est en flagrante contradiction avec la non compositionnalité de la langue. Dans les textes en effet, c'est le global qui détermine le local : le lecteur aborde le texte avec une vue d'ensemble (c'est un document de telle longueur / épaisseur, indiqué par telle personne, édité dans telle collection, etc.) ; puis au fil de sa lecture, c'est le contexte qui lui permet d'identifier les unités linguistiques, leur délimitation et leur interprétation ne s'imposent pas d'emblée.

L'incidence déterminante des différents paliers de contexte s'est d'ailleurs montrée de façon criante à l'occasion des tentatives de traduction automatique basées sur le mot à mot. Quant à une analyse purement ascendante, — du mot à la phrase, de la phrase au texte —, elle conduit rapidement au foisonnement d'ambiguïtés artefactuelles (pour le locuteur humain, dans les pratiques courantes, bien peu fréquents sont les cas d'hésitation entre plusieurs interprétations), avec une démultiplication combinatoire problématique.

Pourtant, les dictionnaires ont légitimement leur place dans l'analyse : ils enregistrent des connaissances sur la langue[24] (sur la parenté des mots, sur leurs significations d'usage) qui sont en partie (et inconsciemment) à l'esprit du lecteur. Pour comprendre leur rôle, on peut s'appuyer sur la distinction entre sèmes inhérents et afférents (cf. §2.a). A une unité lexicale sont associés des sèmes inhérents, à savoir des traits sémantiques qui lui sont habituellement associés (du moins, dans la pratique de lecture considérée : les mots sont perçus différemment dans la lecture d'un roman ou celle d'un règlement). Lors de l'interprétation du texte par le lecteur, l'unité lexicale peut recevoir de nouvelles déterminations, de nouvelles connotations, notamment par propagation dans certaines

[24] Bien que décrivant « la langue générale », les dictionnaires supposent implicitement une maîtrise des genres textuels et des variations d'usage qu'ils induisent : car la signification des mots est aussi réglée par les types de textes. Les dictionnaires ne comportent tout au plus que des marques de « domaines », qui situent quelques secteurs ou pratiques sociales.

constructions (par exemple énumérations, attribution définitoire) : c'est l'afférence de sèmes, en contexte. Et inversement, les opérations interprétatives peuvent inhiber, virtualiser, un sème inhérent.

Autrement dit, les dictionnaires sont utiles, dans la mesure où ils rendent compte de sèmes inhérents, effectivement à l'œuvre dans l'interprétation ; mais leur intervention est une contribution et non un passage obligé et définitif dans le traitement, puisque ce qu'ils enregistrent n'est ni complet (il y a d'innombrables possibilités d'afférence en contexte), ni toujours valide (le contexte prime, et peut écarter une composante sémantique). Le dictionnaire est consulté, sans pour autant jamais détenir le sens d'un texte. L'activité de lecture a une dynamique, que la sémantique interprétative décrit en terme d'actualisation et de virtualisation de sèmes, et qu'il serait vain de vouloir reporter dans un dictionnaire, par nature statique.

A titre d'exemple, les travaux sur les chaînes lexicales à la suite de Morris et Hirst (1991)[25] sont intéressants, mais ne rendent compte que très partiellement des phénomènes sémantiques. La démarche consiste à observer les cooccurrences en contexte de mots mis en relation dans le thesaurus. Ainsi, sur un article de magazine, et en utilisant le Roget's Thesaurus, Morris et Hirst (1991) relèvent au fil du texte les séries du mots suivantes[26], issues chacune d'un secteur du thesaurus :

	mots relevés au fil du texte	étendue[27]
chaîne 1	surbubs, driving, Volkswagen, car's, lights, commuters, traffic, Volks, apartment,…	1-44
chaîne 2	afflicted, darkness, panicky, mournful, exciting, deadly, hating, aversion, cruel,…	2–12, 16, 24
chaîne 3	married, wife, wife, wife.	13-15, 27
chaîne 4	conceded, tolerance.	19-20
chaîne 5	virgin, pine, bush, trees, trunks, trees.	31-33

L'apparition récurrente, en contexte, de mots issus d'un même secteur du thesaurus, est une manifestation de la formation d'isotopies génériques, et donc une confirmation expérimentale partielle de la pertinence descriptive du concept d'isotopie. Cependant, ce procédé est pauvre par rapport à ce que décrit la sémantique interprétative, car il est statique : seules les chaînes lexicales prévues par le thesaurus peuvent être repérées, et le thesaurus fixe les interrelations entre les mots indépendamment des usages textuels. Plusieurs points d'introduction d'une dynamique interprétative seraient donc à prévoir : la construction d'isotopies non enregistrées préalablement, la diffusion ou l'inhibition des sèmes en contexte. Pour cela, le repérage statistique d'associations de mots en corpus est en mesure de fournir les éléments d'une description complémentaire des chaînes lexicales, relative à un contexte intertextuel. La possibilité de rétroaction de la description globale (chaînes lexicales déjà repérées, leur consistance, leur entrelacement) sur la description locale (la reconnaissance et la récurrence de sèmes) ajouterait la dynamique de la formation des isotopies au sein de chaque texte.

b)Langue générale ou genres textuels

L'idée que l'on puisse décrire une *langue générale* (par opposition à des *langues de spécialité*, ou *sous-langages* selon Harris), une évidente *langue naturelle*, est démentie par une approche textuelle. En fait, tout texte réel s'inscrit dans une pratique sociale et par là même relève d'un discours et d'un genre (Rastier & *al*. 1994). Expérimentalement, des analyses statistiques multidimensionnelles sur corpus confirment que l'usage de la langue est contrasté entre les genres, et ce à tous les plans (lexique, syntaxe, etc.). Comme il n'y a pas de langue moyenne, il n'y a donc pas d'outil de traitement automatique des langues parfaitement « tout terrain », sinon par un paramétrage qui l'ajuste après un « profilage » du corpus (Illouz & *al*. 1999). L'analyse sémantique n'est évidemment pas en reste, et les isotopies thématiques dépendent des genres textuels[28].

---

[25] C'est le premier article qui a fait référence sur ce sujet ; Hirst et St-Onge ont signé plus récemment, dans (Fellbaum 1998), un chapitre dans la suite de ces travaux.

[26] Au total, neuf chaînes sont construites.

[27] Passages du texte où se réalisent les chaînes, repérés par les numéros de phrases (il y a au plus trois phrases entre un mot d'un chaîne et le mot suivant de la même chaîne).

[28] « Dans l'analyse thématique, [le corpus] doit être restreint à bon escient pour pouvoir caractériser la spécificité des discours et des genres : les thèmes du roman ne sont pas ceux de l'essai ni du poème. Par exemple, en dépouillant un corpus trop étendu qui mêlait des romans et des essais dans la période 1830-1870, nous nous sommes aperçu que les sentiments du roman n'étaient pas ceux de l'essai. Par exemple, le sentiment de 'fraternité', récurrent dans les ouvrages de Leroux, et celui d''équité' chez Proudhon n'ont pas été relevés dans les romans, à l'exception confirmatrice des *Misérables*, œuvre qui, à dire vrai, alterne des chapitres romanesques et d'autres qui relèvent du genre de l'essai.

Avec l'automatisation, dont un intérêt est la réutilisation et la réitération multiple d'un même traitement, la tentation est de vouloir n'envisager la langue que sous l'angle de ce qui est commun, général, universel. Or les outils sont confrontés non seulement à la diversité des langues (les applications multilingues ou dans le domaine de la traduction rencontrent les limites d'un transcodage par correspondance directe ou via un langage pivot), mais donc aussi à la diversité interne de chaque langue. La réutilisabilité se reporte au plan des méthodologies d'analyse.

c)Graphes lexicaux, graphes conceptuels, réseaux sémantiques

Nous avons indiqué comment les représentations graphiques de la langue et du lexique contribuent à une conception binaire des interactions entre unités linguistiques, par la manière dont sont notés les liens (§ 4.c). Les nœuds des réseaux ne sont pas non plus une représentation neutre : ils délimitent et isolent des unités sémantiques. Mais cette autarcie du nœud est tout illusoire : le graphe est à nouveau un objet sémiotique, peut-être même une forme particulière de texte, et le nœud reçoit une part essentielle de son sens de son entour structurel : avec quels autres nœuds est-il mis en relation, en concurrence ? Des phénomènes de propagation sémantique s'établissent, mais ils ne sont ni unidirectionnels, ni uniformes : les relations d'héritages et de transitivité sont contrariées par des cas particuliers, des ruptures.

La disposition même du graphe sur le plan n'est pas sans effets sémantiques. Une organisation tabulaire (Fig. 4a), qui aligne les nœuds selon des lignes et des colonnes, avec une majorité de liens horizontaux et orientés de gauche à droite, favorise une lecture syntagmatique qui enchaîne les nœuds en succession, et suscite la construction par le lecteur de formes composées, de séquences[29]. Tournier (1986) parle même de « mirage lexicométrique », pour ces longs enchaînements suscités par une lecture (trop) spontanément transitive des graphes, et qui simulent des constructions linguistiques en fait non réalisées en corpus. Une autre disposition, étoilée (Fig. 4b), met en valeur un ou plusieurs centres, pris comme points d'entrée et comme contexte d'ensemble. Les effets de proximité ou d'éloignement, de positionnement intermédiaire ou en périphérie, ajoutent un supplément de sémantique aux liens explicitement représentés par un segment ou une flèche.

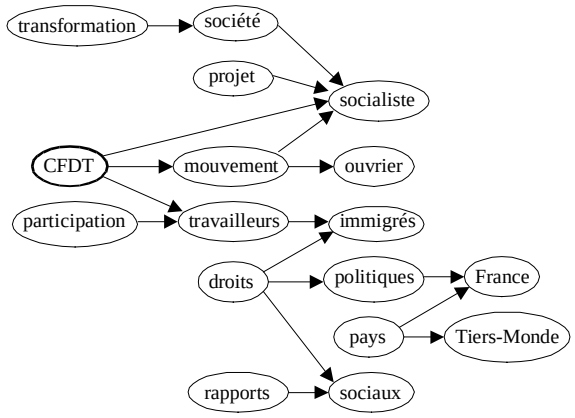

Fig. 4a : Weblex - Lexicogramme récursif (CFDT) : placement hiérarchique gauche-droite

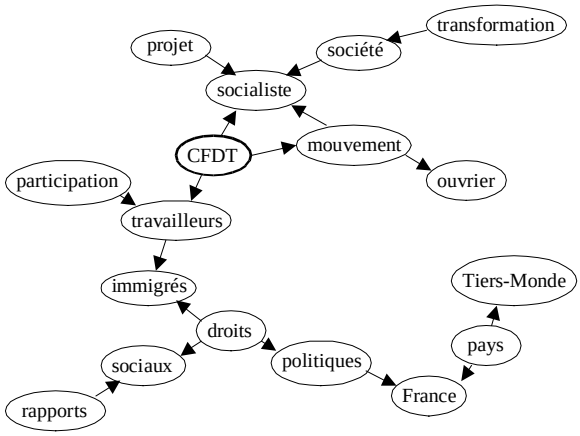

Fig. 4b : Weblex - Lexicogramme récursif (CFDT) : disposition non hiérarchique (en étoile)

L'étiquetage des liens est problématique, car il apparaît contingent et restrictif. La tradition des référentiels documentaires et des thesaurus, et l'intérêt opératoire de l'héritage (élégance et économie descriptive), ont focalisé l'attention sur les relations d'*hyperonymie* / *hyponymie* (généricité / spécificité), auxquelles s'est ajoutée la *méréonymie* (partie / tout). La description lexicale avait aussi

---

A supposer même que le même mot se rencontre dans des genres différents, rien n'assure qu'il se rapporte aux mêmes thèmes : 'amour' se rencontrera certes en poésie et dans le roman, mais le thème de l'Amour diffère pourtant avec ces genres : il n'a pas la même molécule sémique, ni les mêmes lexicalisations, ni les mêmes antonymes ». (Rastier 2001, 205-206)

[29] Nous avons à l'esprit les observations faites sur les lexicogrammes récursifs, générés par l'outil *Weblex* (anciennement *Lexploreur*), `http://lexico.ens-lsh.fr/local/weblex.html`, cf. le manuel rédigé par Serge Heiden.

familiarisé avec la *synonymie* (ou mots associés, variantes) et l'*antonymie*[30]. Mais pour l'analyse thématique en corpus, la question de la pertinence de ces relations se pose : les facilités opératoires escomptées (héritage, équivalence) sont déçues par les usages linguistiques (non transitivité, exceptions). D'autres graphes (par exemple les graphes conceptuels proposés par Sowa) font usage d'étiquettes casuelles. Or l'inventaire des cas sémantiques reste une question ouverte ; l'affinité facilement établie entre d'une part les nœuds et les mots ou les syntagmes, d'autre part la structure casuelle et la structure de la proposition, donne comme point de départ le palier de la phrase, pour une description décevante du texte (simple extension cumulative)[31].

d)L'analyse morpho-syntaxique comme outil d'extraction d'unités de sens

Dans le contexte de l'analyse automatique de textes intégraux, l'objectif de déduire du texte un ensemble de mots-clés s'est traduit par une transposition, trop littérale, des techniques d'indexation mises au point par les documentalistes pour les bases bibliographiques. Le traitement consiste en une série de transformations orientées par la forme conventionnelle des mots-clés, et opère de multiples réductions :

- la lemmatisation gomme les variations flexionnelles (singulier / pluriel, masculin / féminin, temps verbal, etc. [32]) : or l'identité de sens entre ces diverses formes n'est pas toujours assurée. Les relations morphologiques, flexionnelles ou dérivationnelles (mots de la même famille), établissent des conditions favorables à une relation sémantique forte, mais l'usage ou le contexte peuvent également contraster et spécifier les occurrences morphologiquement apparentées[3334].
- le primat est donné au nom, au syntagme nominal, à l'exclusion des autres catégories morphosyntaxiques. Lorsque le verbe est considéré, c'est souvent dans la mesure où il peut prendre l'allure d'un mot-clé par nominalisation[35]. D'autres catégories sont purement et simplement ignorées, alors qu'elles ne contribuent pas moins à la sémantique du texte, par exemple les ponctuations (Bourion 1998).
- les formes syntagmatiques complexes sont recherchées sous la forme de groupes nominaux. Or des calculs statistiques comme les segments répétés (Lafon & Salem 1983) montrent que les unités forgées par les figements sont souvent très significatives et ne sont pas là où on les attend, elles s'affranchissent des délimitations syntaxiques.[36]

---

[30] Les contraires sont effectivement fortement liés, des observations en corpus le confirment (Justeson & Katz 1991).

[31] Pourtant, un graphe casuel pourrait décrire de façon unifiée l'articulation interne d'un thème, pouvant se manifester à tous les paliers du mot, de la phrase et du texte (Rastier & *al.* 1994).

[32] Dans les langues à cas, la neutralisation du cas pourrait aussi être une perte d'information dommageable. Certains mots ou certaines acceptions sont significativement associé(e)s à tel ou tel cas (cf. par ex. Castot 1981) pour eux-mêmes ou pour leur(s) argument(s) syntaxiques, ce que la sémantique interprétative identifie comme des valences casuelles internes ou externes (Rastier 1987b) et décrit par des sèmes casuels (Rastier & *al.* 1994). Pour le français ou l'anglais, langues sans marquage morphologique des cas, ce type d'information serait fourni par un analyseur syntaxique, en termes de fonction grammaticale. La sémantique n'est pas une partie de la linguistique aux côtés de la syntaxe : les structures syntaxiques participent à la construction du sens. Les insuffisances et les échecs des chaînes de traitement automatique par couches (procédant successivement à l'analyse morphologique, puis syntaxique, puis sémantique, puis pragmatique), de même que la remarquable stabilité des résultats d'analyses statistiques de corpus réalisées selon de multiples représentations des textes (segmentation en chaînes de caractères de longueur fixe, formes graphiques des mots, lemmes, catégories morpho-syntaxiques) (Brunet, *à paraître*), illustrent la non validité d'une telle division trop stricte de la linguistique.

[33] Sans compter que même la notion de « type », d'« unité lexicale », n'a qu'une cohésion abstraite, puisque les occurrences ne sont pas porteuses d'une identité de sens à travers les contextes.

[34] Geffroy & *al.* (1973) montrent que la lemmatisation groupe des mots qui ont une fréquence très différente (et donc un comportement distributionnel différent, au sens de la statistique lexicale). Ils observent aussi, sur des corpus politiques, que pour toute une collection de noms, le singulier est susceptible de renvoyer à une notion abstraite, alors que le pluriel est essentiellement concret : 'honneur' (16 singuliers, 1 pluriel), 'exploitation' (5 sg., 2 pl.), 'propriété' (9 sg., 3 pl)., 'travail' (79 sg., 20 pl.), 'société' (44 sg., 7 pl.), etc.

[35] Cependant, les travaux récents de Bourigault et Fabre (2001) témoignent d'une remise en question de ce primat implicite du syntagme nominal. Alors que l'outil LEXTER a été conçu pour l'extraction de groupes nominaux (en vue du repérage de termes complexes dans des documents techniques), l'outil SYNTEX, son successeur, s'intéresse également aux syntagmes verbaux, dans un environnement d'analyse distributionnelle.

[36] Les segments répétés prennent linguistiquement des formes très diverses, au point que Fiala (1986) est amené à distinguer des « segments saturés syntaxiquement » et ceux « non saturés syntaxiquement (incomplets, ne

- en procédant mot par mot, au mieux phrase par phrase, l'analyseur effectue la cueillette des mots-clés au fil du texte, sans vision d'ensemble : or les unités sémantiques ne sont pas des données à reconnaître, mais sont construites dans la dynamique de la lecture et avec une vision globale du texte[37].
- le résultat du traitement, ce sont des mots extraits du texte et élevés au rang de mots clés. Mais à la différence des mots-clés des documentalistes, les mots extraits souffrent d'une décontextualisation brutale et complète : ils sont coupés de leur entour et de leur positionnement textuel, sans pour autant se définir par un rattachement précis au sein d'un langage documentaire ou d'un thesaurus. Cette décontextualisation est évidemment pénalisante au plan sémantique.

e)L'étiquetage sémantique

L'étiquetage sémantique d'un corpus consiste à attribuer à chaque unité lexicale une étiquette qui indique son sens, parmi tous les sens recensés pour ce mot dans un dictionnaire donné. L'attention se focalise d'abord sur la désambiguïsation (Ide & Véronis 1998) : lorsqu'un mot est polysémique ou a des homographes, lequel des différents sens recensés lui attribuer ? C'est en fait la motivation première de cette tâche d'étiquetage : résoudre une fois pour toutes, au niveau d'un corpus, les ambiguïtés lexicales. Avec un peu de recul, vient la discussion sur le choix du référentiel lexical, et de sa granularité : comment dénombrer et caractériser les sens qui décriraient les variations sémantiques observées ? Enfin, c'est le procédé même d'association sens – mot qui pose question.

C'est une conception trop naïve qui résumerait le fonctionnement des dictionnaires à un inventaire de mots mis en correspondance avec des sens ou « concepts ». Le dictionnaire est un objet hautement linguistique (ou sémiotique), dans lequel la définition ne s'identifie pas avec un contenu sémantique isolable, mais construit un contexte donnant des point d'appuis pour interpréter et donner sens à l'entrée lexicale. Ainsi se comprennent la diversité des dictionnaires (non seulement sur la formulation des définitions, mais aussi sur les entrées recensées et sur les divisions et la structuration des significations associées à un mot) et la pertinence des citations et exemples illustratifs des usages. Le dictionnaire fonctionne sur le mode de la suggestion et non sur celui de la représentation.

L'étiquetage qui fixe une signification à chaque mot considéré opère une triple approximation. La première est la localisation du sens au niveau du mot : le sens se construit par isotopies, souvent diffuses, et n'est pas strictement isolable dans un ou quelques mots, si ce n'est par simplification descriptive. Une approximation moins brutale ajouterait la possibilité d'étiqueter des zones de contexte (par exemple des paragraphes). Seconde approximation, les étiquettes (qui recensent les significations possibles pour chaque mot) sont définies hors contexte, elles ne sont pas relatives aux corpus soumis à l'étiquetage : l'affectation de l'étiquette (la désambiguïsation éventuelle) se fait sur une décision locale (à l'échelle de la phrase, par exemple). Une fois l'étiquetage fait, une recomposition du corpus ne modifie pas l'étiquetage : en particulier, la description sémantique d'un texte ne dépend pas ici de l'intertexte dans lequel il s'insère. De fait, l'enjeu même de l'étiquetage est bien de faire ce travail d'analyse sémantique « une fois pour toutes », et qu'il soit réutilisable dans divers contextes : sinon, à quoi bon conserver, par l'intermédiaire d'un codage à même le texte, la valeur sémantique calculée pour chaque mot ? Ceci nous conduit à la troisième approximation : le caractère figé[38] de l'étiquetage sémantique. La sémantique interprétative indique la dynamique de la construction de l'interprétation, or l'étiquetage explicite une lecture statique du texte. Il ne rend pas compte par exemple d'un parcours dans le texte, au cours duquel la sémantique globale, et par ricochet

---

formant pas une unité grammaticale) ». Les segments non saturés peuvent être la trace de parallélismes rhétoriques (*'nieras-tu que tu', 's'il est vrai que vous la'*, Salem 1984). Ce sont encore des « conjonctions de formes grammaticales *'à la', 'et de la'*, etc., auxquelles nous ne sommes pas en mesure de donner un statut grammatical précis » (Lafon & Salem 1983). Stylistiquement marquées aussi, et non sans portée sémantique, on relève des tournures et des phraséologies : *'tous les', 'tous ces'* dans les appels vindicatifs du révolutionnaire *Père Duchesne* (Salem 1984) ; *'de l'ensemble des', 'des intérêts de classe des'* dans les résolutions de confédérations syndicales (Lafon & Salem 1983). De plus, l'enchevêtrement des segments répétés dissipe l'illusion d'une segmentation du texte simple, nette, stable, sans recouvrements ni restes ; on décompte par exemple dans un corpus syndical 'nouveau type' (fréq. 6), 'nouveau type de développement' (4), *'type de développement'* (11), 'développement économique' (5) (Lafon 1986).

[37] Par exemple la délimitation des unités dépend du contexte : [Bayard] [monte] [au créneau] *vs* [Rocard] [monte au créneau]. Voir (Rastier 1997).

[38] La constitution d'un lexique recensant les sèmes repérables ensuite dans les textes fait conclure à Cavazza (1996) : « la description n'est que de l'interprétation figée. »

locale, se trouve modelées. L'approximation se conçoit si l'on se situe dans une application donnée, qui adopte un point de vue déterminé en fonction de ses finalités. Le jeu d'étiquettes est alors une formalisation d'attentes de lecture[39] dans ce contexte.

6.Pistes et repères pour une mise en œuvre informatique

a)Préalables méthodologiques : ressources, outils

La constitution d'un corpus

Pour le linguiste, le corpus est un outil de travail essentiel, puisque les textes sont les observables de la langue. Pourtant, il n'est pas d'ensemble de textes qui fournisse une image complète et universelle de la langue. L'effort n'a donc pas à porter sur l'acquisition du corpus parfait (parfaitement équilibré, parfaitement délimité, parfaitement représentatif à défaut d'exhaustivité…), dont la « solidité » serait le gage de la valeur des observations tirées. Ce qui fait la validité du corpus, c'est son interprétabilité —à savoir l'explicitation des critères sur lesquels il est rassemblé, la pratique langagière dont il se fait écho. Et ce qui fait la pertinence du corpus, c'est l'adéquation du sens qu'on lui accorde avec les objectifs des analyses qui y sont menées.

Dans l'optique de la sémantique interprétative, l'étude sémantique requiert à l'évidence des corpus de textes, et non des corpus d'échantillons (« du » texte « au kilomètre »)[40] (Péry-Woodley 1995). Première raison, le texte est une articulation centrale dans le jeu des différents paliers de contexte (lexie, période, paragraphe, texte, intertexte), qui interviennent directement dans la construction du sens. L'échantillon est au contraire, par définition, une partie, un extrait –et non une totalité et unité contextualisante. Deuxième raison pour rejeter les échantillons, la construction du texte lui confère une certaine autonomie à l'égard de réalités extra-linguistiques (le texte les représente comme en creux), ce qui rend concevable la conduite d'une analyse essentiellement déployée à partir d'un corpus, sans recours à d'autres données.

Le codage des structures textuelles

Paradoxalement, les conceptions et les pratiques d'un codage adapté à la nature des textes et de la langue, telles que les reflète la *Text Encoding Initiative* (TEI)[41] principalement, focalise la description sur ce qui est méta-textuel (l'en-tête) et extralinguistique (identification de référents : dates, lieux, personnes). Les formes d'articulations textuelles élémentaires et fondamentales, qui construisent des paliers de contexte et typent leurs interrelations, apparaissent de façon dispersée[42] et non systématique à travers des éléments standards (par ex. le paragraphe, les listes). Le caractère familier de ces éléments standards en fait un repérage de surface évident, sans que soit perçue leur portée sémantique et leur structuration globale des paliers du texte.

La définition de l'élément ' ' de la TEI[43] est un premier pas dans cette direction. Il unifie les différentes divisions d'un texte en parties (chapitres, sections, etc.) comme un même procédé, de découpage selon un niveau. Cette modélisation rend bien compte du type d'interaction sémantique qui

---

[39] Ces attentes de lecture imperturbables seraient plutôt un conditionnement, auquel est soumis sans états d'âme l'automate-lecteur.

[40] Bien que rejetant ici globalement ces approches, on peut néanmoins distinguer d'une part le recours à une technique d'échantillonnage rigoureuse visant à rendre compte de la diversité interne et intertextuelle des textes (dans la tradition du corpus de Brown), et d'autre part l'accumulation de « données linguistiques » ou « textuelles » dont la masse énorme serait censée assurer la représentativité (« more data is better data »). Voir (Péry-Woodley 1995) pour une discussion documentée.

[41] Ce groupe de travail international se penche sur la question du codage SGML des textes. Il a mis au point des conventions de codage des corpus sous forme de recommandations (Sperberg-McQueen & Burnard 1994, Burnard & Sperberg-McQueen 1996), qui se veulent suffisamment détaillées et complètes pour être utilisables pour tout corpus dans une multiplicité de contextes d'utilisations. L'objectif est de favoriser ainsi les échanges scientifiques et la réutilisabilité des corpus codés (Ide & Véronis 1995).

[42] Ce qui est en cause dans cette « dispersion », ce n'est pas l'effort de la TEI en vue d'une description modulaire pour s'adapter à la diversité des genres textuel ; c'est la multiplicité des éléments répertoriés. Le détail des éléments est certes utile pour un encodage précis en vue de l'archivage du corpus et de sa réutilisabilité dans de multiples contextes. Mais l'inventaire des éléments encodables se déploie sans vraiment de considération pour leurs parentés structurales, en termes d'organisation du texte et d'effets sémantiques produits.

[43] On trouve une structure analogue déjà dans LaTeX, une vingtaine d'années auparavant.

se joue dans cette structuration, et qui sous-tend de multiples réalisations « de surface » (chapitres etc.). D'autres types d'interaction sémantique[44] sont ainsi à mettre en valeur, avec à la clé une saine économie descriptive.

Les unités linguistiques

La détermination du local par le global éclaire les difficultés de l'indexation automatique des textes, qui extrait les « mots-clés » au fur et à mesure de l'analyse, mot par mot, au mieux phrase par phrase. Or l'analyse linguistique ne consiste pas à reconnaître des unités préexistantes et serties par la syntaxe, mais elle construit l'unité à partir du contexte global. D'où l'efficacité d'heuristiques comme l'apprentissage endogène du logiciel LEXTER (Bourigault 1994)[45] : l'articulation interne d'un syntagme nominal complexe peut se déduire de l'observation des occurrences de ses composants dans le corpus. Une généralisation de cette approche consiste à ne faire de l'analyse locale (morpho-syntaxique) qu'une préanalyse, les unités linguistiques n'étant construites qu'avec une vision d'ensemble de cette préanalyse du corpus, ou / et d'autres contextes globaux (unités déjà construites par ailleurs : archive, description d'un genre textuel, etc.)[46].

La persistance d'antagonismes comme la querelle sur la lemmatisation (Brunet 2001), ou encore l'alternative indexation libre *vs* contrôlée[47], montre que chacun des choix a de bonnes justifications tant théoriques qu'expérimentales. La résolution de ces oppositions ne sera pas obtenue par ce qu'une option l'emporterait sur l'autre, mais par le dépassement de ce qui les sépare. Ainsi, la détermination du local par le global invite à pratiquer une lemmatisation partielle et réglée par le contexte : la réduction n'est opérée que si les formes qu'elle condense ont le même comportement dans le corpus, à savoir participent à la construction d'isotopies analogues. La lemmatisation n'est donc ni toujours bonne, ni toujours mauvaise – sa pertinence est contextuelle. En ce qui concerne l'opposition indexation libre / contrôlée, il s'agit là encore d'affirmer la prédominance du contexte : un lexique prédéfini (enregistrant par exemple des connaissances générales sur la langue) aide et complète la description, dans la mesure où il n'exclut pas la formation d'autres unités en contexte (cf. indexation libre) et où ses apports ne sont retenus que quand ils ne sont pas infirmés par le contexte.

b)La classification automatique pour le repérage et la modélisation de thématiques

Interprétation des classes comme la modélisation d'isotopies : un exemple

L'expérience de Pichon et Sébillot (1999), sur un corpus d'articles du *Monde Diplomatique*, illustre comment des regroupements calculés de mots contribuent à une analyse sémantique.

Une classification automatique sur les noms les plus fréquents du corpus produit des ensembles qui s'interprètent comme des thématiques : par exemple, la thématique des /négociations/, pour l'ensemble {'négociation', 'accord', 'création', 'position'}, ou encore /territoire/, pour {'autorité', 'région', 'territoire'}. La variation sémantique d'un mot donné (par exemple 'militaire') peut alors être décrite selon les mots (adjectifs, noms) qui apparaissent le plus souvent dans son voisinage (à +/- 5 mots d'écart), dans les paragraphes relevant de telle ou telle thématique (*i.e.* comportant au moins deux mots de l'ensemble définissant celle-ci). Voici le relevé obtenu, au voisinage de '*militaire*' :

- dans le contexte de /territoire/ et /négociations/ : `force`, `Etats-Unis`, `américain`, `grand`, `puissance`, `économique`, `intervention`, `politique`, `présence`, `aide`.

---

[44] Exemples de structures sémantiques élémentaires : la condensation d'une attente globale (titre / partie, résumé / texte, chapeau / article), l'ouverture (notes, bibliographie), le parallélisme induisant une composition d'ensemble et un jeu de contrastes (listes, tableaux), la mise en relief d'une expression, celle d'un passage, les voisinages et les lignées intertextuels (Pincemin, *à paraître*).

[45] La « *composante sémantique* de *raisonnement par analogie*, incluse par F. Debili [thèse, 1982] dans son analyseur [syntaxique] [est] très similaire, dans son principe, à notre procédure de *désambiguïsation endogène* […]. Notre contribution se situe au niveau de la validation empirique de cette idée. » (Bourigault 1994, § IV.2.1.)

[46] Le choix de ce contexte d'ensemble, contexte de référence pour la construction des unités, peut être varié, pour refléter différents points de vue. Il doit être guidé par le critère d'*interprétabilité du corpus* (Pincemin 1999) et par le *profilage*, contribuant à l'homogénéité linguistique (Illouz & *al*. 1999).

[47] Ces antagonismes se résolvent d'eux-mêmes dans certains contextes ; mais ils sont manifestes en analyse des données textuelles, en lexicométrie, en recherche d'informations sur le texte intégral, terrains qui nous intéressent ici dans la perspective de traitements automatiques à visée sémantique et textuelle.

- dans le contexte de /territoire/ (à l'exclusion de /négociation/) : `moyen`, `opération`, `régime`, `russe`, `victoire`, `base`, `massif`, `occupation`.
- dans le contexte de /négociations/ (à l'exclusion de /territoire/) : `action`, `ordre`, `effort`, `responsable`, `pays`, `dépense`, `atlantique`, `Europe`, `OTAN`, `organisation`.

On observe ainsi que, dans le contexte de la thématique /territoire/, 'militaire' prend une connotation /guerrière/[48]. Avec la thématique /négociations/, la sémantique de 'militaire' et de ses voisins converge sur le caractère /organisé/ et /structuré/ des intervenants militaires. Autrement dit, les regroupements de mots qui figurent dans les mêmes contextes peuvent être la trace d'isotopies, au sens de la sémantique interprétative. En effet, s'ils évoquent une thématique, c'est grâce à la récurrence d'un élément sémantique — un sème — qu'ils partagent.

Des propriétés linguistiques à la modélisation informatique

Une classification automatique basée sur les cooccurrences est donc en mesure de contribuer à une modélisation et à un repérage automatique des isotopies[49]. Nous avons vu combien la sémantique interprétative fait de l'isotopie un point d'appui central pour la description sémantique. Il s'agit alors de *concevoir un algorithme* de classification *en adéquation avec les propriétés linguistiques* des isotopies, plutôt que de pallier tant bien que mal les distorsions induites par les algorithmes standard (cf. § 4). La sémantique interprétative donne les orientations suivantes :

- La classification est non exhaustive : tout mot n'a pas à entrer dans une classe (cette contrainte engendrant d'ailleurs classiquement une classe « divers », sans cohérence interne et sans lien avec le reste de la classification). En effet, selon le corpus, un mot peut ne pas contribuer à la formation d'une isotopie significative.
- La classification est multi-classe : un mot peut être indexé sur plusieurs isotopies. Lorsque l'algorithme force le classement d'un mot dans une seule classe, il résout brutalement et de façon inadéquate les nuances de sens, les polysémies ; cela induit d'ailleurs parfois la construction de classes artificiellement mitigées.
- La classification n'est pas dominée par les mots au bon pouvoir discriminant (ni trop fréquents, ni trop rares, et univoques[50]), mais elle regroupe les mots en fonction de leur répartition dans le corpus. Il peut ainsi se former des classes de mots très généraux, par exemple.
- Plusieurs paliers de contexte structurent les textes, et ont chacun une pertinence sémantique propre : le voisinage au sein d'une même période est modelé par différents types de relations de dépendances ; le voisinage dans un paragraphe ancre les mots sur un même fond sémantique ; des relations à longue distance se jouent également à l'échelle du texte, notamment sous l'effet des genres textuels. Les cooccurrences sont donc à considérer pour chacun de ces différents types de contextes, pour former des isotopies de natures différentes, complémentaires[51]. La question du choix du meilleur voisinage pour les cooccurrences était indûment limitative (ce qui explique d'ailleurs la variété des options prises) (§ 4.d). Si l'on s'accorde maintenant à considérer qu'il n'y a pas de voisinage supérieur aux autres, mais que leur valeur varie (selon l'application, le genre du texte, etc.)[52], le pas suivant consiste à considérer conjointement et complémentairement tous les paliers de contexte, sans se limiter à un seul à chaque fois. Les paliers sont différents mais ne sont pas indépendants : les isotopies établies à un palier (comme toute unité déjà construite : mot

[48] Nous notons entre barres obliques les éléments sémantiques mis en évidence, selon la convention d'écriture courante pour les sèmes.

[49] Le résultat d'une telle classification ne se limite pas à des isotopies (réitération d'un même sème) mais décrit aussi des paratopies (apparition groupée de sèmes différents mais associés dans la construction d'un thème), cf. note 3.

[50] L'analyse factorielle moyenne les diversités d'usage, si bien que les mots polysémiques sont versés au centre de l'espace, zone confuse et peu informative. De même, les coefficients d'association ou d'information mutuelle valorisent les dépendances rares et exclusives.

[51] Concrètement, la définition de ces contextes allie des critères linguistiques, typographiques, et des critères de taille. Il s'agit d'inclure des effets cognitifs de perception visuelle du texte et de mode de mémorisation. La *période* évite une définition trop syntaxique de la proposition, ou typographique de la phrase. Le *paragraphe* forme une unité sémantique, et ne se superpose pas nécessairement avec l'alinéa délimité par un retour à la ligne. (Pincemin, *à paraître*).

[52] Voir par exemple l'article de Grefenstette, « Evaluation techniques for Automatic Semantic Extraction : Comparing Syntactic and Window Based Approaches », *in* (Boguraev & Pustejovsky 1996).

composé, paradigme…) peuvent être réutilisées comme composantes d'isotopies aux paliers supérieurs.

- La distinction entre sèmes inhérent et afférent peut être reprise dans la structuration interne de la classe en un noyau (sous-classe de mots pour lesquels le sème est inhérent) entouré de satellites (sous-classe de mots pour lesquels le sème est afférent).
- Les interactions linguistiques ne se laissent pas décomposer en relations binaires. Plutôt que de classer des mots sur la base de fonctions qui évaluent la ressemblance du comportement de mots pris deux à deux, il faudrait faire intervenir des mesures de cohérence d'ensemble de groupements de mots, considérés simultanément[53] [54]. Une telle mesure valoriserait d'emblée des caractéristiques stables (prise en compte de l'ensemble des mots concernés, avec leurs interactions) et contrastives (prise en compte du corpus), récurrentes modulo des variations mineures au regard de l'ensemble[55].

Cette modélisation par classes répond à des exigences fondamentales de la sémantique interprétative, tout en résolvant certaines difficultés rencontrées par d'autres implémentations :

- Ces classes introduisent une détermination du local par le global, et évitent une description trop compositionnelle. D'une part, elles sont construites à partir de la considération d'ensemble d'un corpus. D'autre part, le repérage des sèmes en contexte ne s'établit pas occurrence par occurrence, mais l'isotopie n'est établie que sur la présence dans un même voisinage de plusieurs éléments différents de la classe (dont sans doute au moins un du « noyau », *i.e.* pour lequel le sème est inhérent).
- La description est dynamique : chaque corpus permet de construire un nouveau jeu de classes, relatif au corpus[56].
- La granularité de la description est opportune, relative au corpus et efficace dans son contexte. De même que les sèmes ne sont pas des primitives irréductibles et exhaustives, les classes construites par la classification automatique trouvent les repères d'une description complète et pertinente dans l'ensemble de textes que circonscrit le corpus.
- Les éléments de la description sémantique sont les classes elles-mêmes, non les sèmes qui serviraient à les nommer et qu'elles représenteraient. Ainsi se trouve écarté le problème de l'énumération des sèmes par de désignations mnémoniques, précises, univoques (Cavazza 1996). Le sème est entièrement défini par la composition de la classe[57]. Il représente le « dénominateur commun » sémantique que l'on perçoit à sa lecture. Bien entendu, rien n'empêche de donner une étiquette à la classe comme un raccourci interprétatif, une sorte de commentaire, mais cette étiquette n'est ni exclusive, ni tenue d'être fixée une fois pour toutes : ce n'est pas elle qui définit la classe, c'est la classe qui lui donne sens.
- A la compulsivité de l'interprétation répond la robustesse de la description par classes : la cooccurrence de quelques mots d'une même classe suffit à donner une indication thématique, même si d'autres éléments du contexte relèvent de la même thématique et ne sont pas recensés dans la classe. Autrement dit, il n'y a pas d'impératif de complétude pour les classes ; la qualité ne s'évalue pas par la composition de chaque classe, mais par leur équilibre d'ensemble et leur capacité descriptive pour le corpus considéré.

---

[53] Dias & *al.* (2000) ont des préoccupations analogues, pour l'extraction d'unités lexicales complexes ; ils définissent une mesure de cohésion qui ne procède pas par amorçage, *i.e.* qui ne décrit pas les structures syntagmatiques *n*-aires comme une composition de structures binaires.

[54] La mesure mathématique servirait donc à déterminer les groupes de mots représentatifs d'isotopies, en retenant les plus « cohérents ». Des critères linguistiques pourraient être introduits pour composer préalablement des groupes « bons candidats » et réduire ainsi la combinatoire du calcul.

[55] Cette notion d'approximation globale, de tolérance, permettrait de trouver un équilibre entre le trop strict et le trop lâche. Par exemple, lorsque la description s'appuie sur la théorie des graphes (les relations binaires calculées étant transcrites comme des arrêtes, les mots étant les nœuds), les *composantes connexes* sont trop sensibles à la transitivité, alors que les *cliques* sont trop restrictives. Lorsque l'on fait appel à une description ensembliste, l'homogénéité d'un ensemble pourrait être grande même si aucune propriété n'est commune à *tous* les éléments.

[56] En revanche, la dynamique s'inscrivant dans le temps de la lecture est mal représentée dans cette modélisation par classes sémantiques. Les formalismes mettant en jeu l'apprentissage, comme les réseaux neuronaux, semblent mieux placés pour rendre compte de cette temporalité : Rastier (1987b) cite l'exemple d'une interprétation de « The astronomer married a star », où, pour le mot 'star', le réseau de neurones commence par sélectionner 'corps céleste', pour ensuite préférer 'étoile du cinéma'.

[57] Dans sa modélisation informatique de la sémantique interprétative, Tanguy (1997) définit lui aussi les sèmes non par leur nom, mais par l'isotopie qui les manifeste, et qu'il représente comme un ensemble structuré d'occurrences.

c)Classes sémantiques manuelles ou calculées

Deux objets différents

L'entreprise de constitution de classes sémantiques n'est pas l'apanage des approches automatisées. Par exemple, un concept comme celui des classes d'objets conduit à constituer des groupements de mots sur des critères linguistiques, minutieusement observés par un linguiste (Le Pesant & Mathieu-Colas 1998). Les thesaurus et les réseaux sémantiques patiemment constitués par des équipes d'experts sont encore d'autres propositions d'organisation sémantique du lexique. Certaines de ces ressources se réclament explicitement de la sémantique interprétative : voir par exemple les réalisations présentées dans (Rastier & *al*. 1994), (Cavazza 1996), (Antoine 1994), (Vaillant 1997).

Par l'économie de moyens et de temps qui le caractérise, le traitement automatique est-il appelé tôt ou tard à remplacer le travail des experts, — à les soulager de leur tâche fastidieuse ? Ou inversement, la qualité des réalisations mobilisant l'intelligence et l'expertise humaines, auxquelles l'ordinateur ne saurait se substituer, condamne-t-elle à la médiocrité les efforts des traitements automatiques ? Le débat ne se situe pas là. Ces approches s'inscrivent dans des démarches complémentaires, et ne sont donc pas en soi substituables[58]. Donnons quelques points d'un parallèle contrastif :

	classes sémantiques construites par une expertise humaine	classes sémantiques calculées
valeur théorique	Des critères servent de guides pour constituer et délimiter les classes : celles-ci sont l'expression d'une théorie linguistique au plan de la sémantique du lexique.	La méthode de calcul des classes est mise au point en fonction de considérations linguistiques, mais les classes elles-mêmes sont une description contingente d'un corpus dans le cadre d'une application, elles préjugent pas des structures en langue[59].
part de détermination, part d'implicite	Lorsqu'une analyse sémantique utilise ces classes, définies *a priori*, elle reconnaît dans un texte ce qui a été recensé et explicité dans les classes. Il y a en quelque sorte prédétermination : on ne trouve jamais que ce qui a été préalablement enregistré.	L'ensemble des classes fournit une vue d'ensemble. Elle fait ressortir des régularités pas toujours perçues. Elle ne s'inscrit pas dans des repères préétablis et stables.
nature de l'apport sémantique	Les classes reflètent une connaissance partagée ; elles sont une expression de la *doxa*, telle que l'opinion générale de la civilisation ambiante, ou le savoir commun d'une communauté de personnes dans une pratique professionnelle.	Les « déformations » des classes par rapport aux attentes communes sont significatives d'aspects sémantiques des textes du corpus.
délimitation et complétude	L'objectif est d'obtenir une description stable et complète, tout au moins pour un contexte applicatif fixé. La stratégie de constitution de la ressource peut donc procéder par pans, en visant déjà cette stabilité et cette complétude pour des parties, des modules, peu à peu rassemblés.	Chaque calcul produit une série de classes déterminées. Mais la multiplicité des calculs possibles et des corpus donne à chaque résultat une valeur d'inachèvement et d'ouverture.
contexte sémantique	La structure sémantique ainsi définie s'interprète indépendamment d'un corpus de textes : elle est déjà elle-même son propre contexte, en ce sens que chaque classe s'inscrit dans l'organisation et l'équilibre dessinés par	Les classes s'interprètent par rapport à un corpus, et en connaissant les principes du calcul. Une très bonne connaissance du corpus est communément requise pour comprendre ce qui sous-tend chaque regroupement.

[58] Une troisième voie conjugue les approches manuelle et automatisée, en utilisant les calculs comme aide à l'expert pour sa tâche de construction de classes (dégrossissement initial, suggestions, indications, etc.) : « L'étude des graphes [calculés] et des contextes partagés entre formes aide effectivement à construire des classes sémantiques. Elle permet d'étudier les emplois et les proximités entre formes dans une langue de spécialité, même mal connue, en s'affranchissant des préjugés induits par la langue générale. » (Habert 1998, § 4.4.1) (Habert & Nazarenko 1996).

[59] « En analysant le champ lexical des sentiments, nous ne postulons pas que ce champ soit uniforme, ni qu'il soit une unité de langue ; il contient sans doute plusieurs taxèmes, mais ne constitue pas un domaine délimité par l'incidence d'une pratique sociale : il s'agit donc d'un regroupement *ad hoc*, convoqué par la pratique descriptive. » (Rastier 2001 : 207)

	l'ensemble.	
mode d'utilisation	Les classes construites manuellement sont « présentables », elles sont adaptées à une lecture humaine.	Les classes calculées (de par leur nombre, leur volume, leur composition) rendent souvent leur lecture difficile, voire irréaliste. En revanche, elles peuvent servir pour une description sémantique exploitée par un traitement automatique.
malléabilité, dynamique	Les remaniements sont limités, car ils ont des incidences sur l'ensemble de la description. De plus, du fait de l'usage dans la durée, les modifications apportées peuvent périmer les utilisations antérieures.	Le travail se situe sur la définition des principes d'organisation de la langue sur lesquels baser le calcul. Des réajustements sont toujours possibles, et facilement envisageables : il n'y a pas de limites à relancer le calcul. La variation des paramètres peut même être un moyen d'observation du comportement linguistique.
portée (réutilisabilité)	visée générale : la ressource lexicale ainsi constituée n'est pas destinée à un usage ponctuel, mais doit servir dans la durée (moyennant son entretien, sa mise à jour). Sa généralité tient à sa possibilité d'être réutilisée, sa couverture pouvant très bien être celle d'un domaine de spécialité, dans le cadre d'une application donnée.	visée singulière : les classes sont calculées pour le corpus considéré. Un changement de corpus motive un nouveau calcul. La description est relative et volatile. Ce ne sont pas les classes qui sont réutilisées, mais la méthode et les outils pour les construire. Si le calcul est utilisé en vue de construire un lexique sémantique de référence, il intervient comme une aide (point de départ, suggestions) dans un processus qui mobilise une élaboration humaine : il appelle des corrections et des ajustements.

Questionnement des attentes

Cette distribution des rôles (les classes sémantiques manuelles pour une formalisation explicite et stable d'une connaissance sur la langue, les classes calculées comme étape de représentation sémantique intégrée à un traitement automatique) amène à reconsidérer un certain nombre d'attentes vis-à-vis des classes sémantiques : les propriétés attendues des classes obtenues par classification automatique n'ont pas nécessairement à se conformer aux modèles que l'on se donne pour la description lexicale sémantique. Bouaud & *al*. (1997) observent des écarts entre graphes lexicaux calculés et ontologies établies par des experts, par delà certains accords généraux entre les deux structures. En fait, l'évaluation des regroupements de mots calculés ne devrait pas se mesurer à l'aune des référentiels terminologiques établis[60] ; elle supposerait plutôt l'appréciation de leur utilité dans des analyseurs sémantiques d'un nouveau type, pour des applications de mise en relation de passages textuels par exemple (recherche d'information, navigation hypertexte). La comparaison des structures manuelle et calculée garde sa pertinence sur un autre terrain, celui de la compréhension et de la caractérisation théorique de la nature profonde et singulière de chacune des deux approches.

L'esthétique des classes n'est pas primordiale pour les classes calculées, si elles sont destinées à être reprises par la machine. A savoir, la taille ou le nombre des classes peut être réglé par des considérations linguistiques, sans que cela corresponde à des effectifs cognitivement bien appréhendables par la mémoire humaine. Un ordre logique des classes, et des mots dans les classes, est utile à la lecture humaine pour embrasser le contenu de la classification, il n'est pas nécessairement exploité par une machine. Une classification manuelle recherche un inventaire systématique, alors que dans la classification automatique peuvent s'inscrire des « trous », qui peuvent être révélateurs de singularités ou des limites du corpus, et qui peuvent par ailleurs ne pas gêner un traitement robuste.

Les classes calculées ont généralement une allure plus hétérogène. Une part de cette hétérogénéité est artificielle, celle introduite par les algorithmes de classification. En raison de la

[60] Habert et Zweigenbaum (2001) conviennent que la transposition des bancs d'évaluation est problématique pour les approches non supervisées, à savoir pour les classes construites par classification automatique, sans introduction de connaissances *a priori* (sinon les propriétés structurales des classes déterminées par l'algorithme de traitement). Le cas est différent pour les approches supervisées (où l'on indique par exemple des mots pour amorcer les classes, ou une terminologie structurée à enrichir, ou des patrons morpho-syntaxiques propres au repérage de certains éléments) : il y a alors une idée *a priori* du résultat, un modèle stable et déterminé, unique, par rapport auquel on peut mesure la conformité des regroupements calculés.

répartition stricte et complète qu'ils imposent, ils génèrent notamment une classe « divers ». Ceci doit être corrigé par la conception de nouveaux algorithmes (§ 6.b).

Il y a également une hétérogénéité linguistique naturelle des classes. Si le traitement examine les cooccurrences sans restriction sur la nature morphosyntaxique des mots (par exemple, ne considérer que les noms) ou sur leur construction (lorsque le but est de rassembler les mots en relation de substituabilité), les classes obtenues n'ont pas l'allure de paradigmes. Et même si l'on introduit des contraintes pour obtenir une homogénéité morphosyntaxique, les classes mêlent des mots qui, dans une même thématique, entretiennent des relations diverses (Fig. 5). La recherche d'homogénéité tient sans doute en partie à l'objectif de trouver des procédures automatiques qui approchent les réalisations manuelles. Pour ces dernières, l'homogénéité est traditionnelle : pour les thesaurus, elle résulte explicitement des règles de construction, un concept s'exprimant sous la forme normalisée d'un descripteur substantif singulier. Pour la construction de réseaux sémantiques, la réalisation la plus complète et la plus populaire, *WordNet*, a volontairement organisé la description partie du discours par partie du discours (un réseau pour les noms, un pour les verbes, etc.), arguant que les relations sémantiques diffèrent d'une catégorie à l'autre (Fellbaum 1998).

Or il n'est pas du tout évident que des classes sémantiques, représentatives d'isotopies, et utilisables pour une analyse sémantique automatique des textes, doivent se conformer à ce principe d'homogénéité paradigmatique. Les sèmes ne suivent pas le cloisonnement des parties du discours, par exemple il semble pertinent que 'coureur', 'course', 'courir' puissent être rassemblés dans une même classe. Plus largement encore, des classes à vocation sémantique peuvent s'ouvrir à la diversité non seulement des parties du discours, mais aussi d'informations linguistiques de tous ordres. Par exemple, une classe pourrait mettre en évidence la convergence d'usage entre un temps, une ponctuation, telle construction syntaxique et tels items lexicaux[61]. Les relations de transformation syntaxique et de dérivation morphologique instaurent des équivalences, correspondant à des situations favorables de relation sémantique. Si le but n'est pas la construction de paradigmes mais de classes de mots en relation sémantique (d'isotopie), alors passer par une étape de normalisation syntaxique et dérivationnelle n'apporte pas un gain évident : il s'agit d'un traitement complexe, et il ne délimite pas l'ensemble des interrelations pertinentes (d'autres relations indirectes et de voisinages sont potentiellement pertinentes).

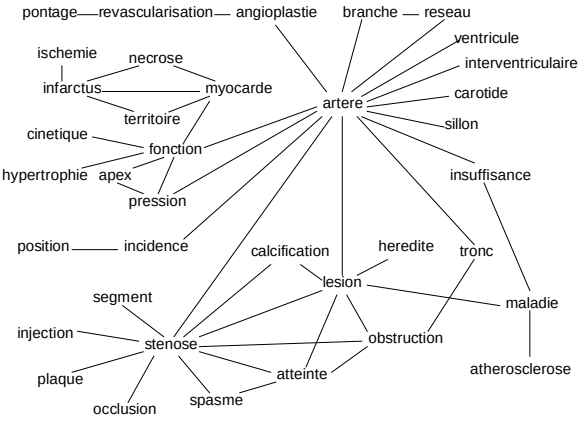

Fig. 5 : Hétérogénéité d'un graphe calculé

« Cette première approche du principal sous-graphe connexe met en évidence une hétérogénéité certaine. Ce sous-graphe mêle en effet des organes et des sites corporels ('*artère*', '*branche*', '*réseau*', '*ventricule*', '*interventriculaire*', '*carotide*', '*sillon*'), des affections corporelles localisées ('*sténose*', '*occlusion*', '*calcification*', '*lésion*', '*atteinte*', '*obstruction*'), des actes médicaux ('*pontage*', '*revascularisation*', '*angioplastie*'), une affection particulière ('*ischémie*', '*infarctus*', '*myocarde*', '*nécrose*', '*territoire*'), et enfin des groupes au fonctionnement peu clair : ('*cinétique*', '*fonction*', '*hypertrophie*', '*apex*', '*pression*') et ('*position*', '*incidence*'). »

(Habert & *al*. 1997)

[61] Sur le théâtre classique en vers, Beaudouin (2000) montre statistiquement l'association entre certaines thématiques ('amour', 'mort') et certaines formes rythmiques des alexandrins. La symétrie des deux hémistiches se retrouve sous de multiples angles d'analyse — phonétique (positions de chaque voyelle, accentuation), morphosyntaxique (délimitation des mots, position des catégories grammaticales), etc. —, si bien que Beaudouin conclut à la figure « feuilletée » du rythme.

7.Conclusion

A la suite de Prié (1995), ce parcours rapide des pratiques actuelles d'analyse automatique de textes est orienté par la mise en évidence d'une « trame conceptuelle résistante » de la sémantique interprétative : quels sont les principes essentiels qui fondent cette sémantique, et comment se retrouvent-ils (ou non) dans les modélisations informatisées ? Prié centre son étude autour des quatre points qu'il retient de la sémantique interprétative, pour les confronter ensuite tour à tour à différents formalismes : la détermination du local par le global, qui se manifeste en particulier dans la non compositionnalité du sens ; la dynamique de la construction de l'interprétation, en contexte (par exemple par des opérations d'assimilation ou de dissimilation) ; le non-déterminisme de l'interprétation, qui admet pour un texte une pluralité de sens (multiplicité qui n'est ni unicité, ni infinité) ; le rôle central joué par le concept d'isotopie, à la base de toute construction d'unités sémantiques d'ordre supérieur, et présidant à la plupart des opérations interprétatives.

Dans cet article, nous rejoignons Prié en cheminant en sens inverse, en accordant toute notre attention à ce que les traitements automatiques des textes révèlent, par leurs succès et leurs limites, du fonctionnement sémantique de la langue, tel que le décrit la sémantique interprétative. Si le concept de *sème* est central, il est trop souvent mal compris : nos observations mettent l'accent sur la pertinence des *isotopies* pour la modélisation et l'analyse des corpus, et d'ailleurs une relecture de la théorie ne dément guère cette primauté de l'isotopie sur les sèmes (cf. note 4).

Chemin faisant, plusieurs seuils qui limitent les applications actuelles se révèlent, ainsi que des possibilités de les dépasser. Au plan des techniques, les classifications automatiques disponibles imposent des contraintes de structuration des classes inadaptées à la description linguistique : l'explicitation des propriétés mathématiques souhaitables[62] définit des spécifications pour la mise au point de nouveaux algorithmes. Beaucoup d'oppositions persistantes devraient se résoudre, la dynamique de la description donnant les moyens de mutualiser des options finalement non exclusives : définition des contextes de cooccurrence, lemmatisation ou non, indexation libre et indexation contrôlée. La clarification des objectifs de construction de classes sémantiques conduit à discerner et reconnaître la valeur de deux conceptions différentes. D'une part, des classes sémantiques homogènes et stables organisent une description de la langue, et impliquent une expertise humaine (construction manuelle, ou semi-automatique). D'autre part, des regroupements sémantiques calculés dans le contexte d'un corpus sont un moyen essentiel de sa représentation thématique, rendant compte de la construction dynamique et globale des isotopies. Là encore, l'automatisation ne remplace pas le maniement humain du langage, mais elle contribue à une autre mode de perception de la langue et de son fonctionnement en corpus.

Les pistes proposées à l'issue de cette réflexion sont prospectives[63]. Plutôt que d'indiquer une solution, — une modélisation conforme à la sémantique interprétative —, ces lignes ouvrent surtout des voies de recherche difficiles mais importantes : construction de l'analyse sémantique guidée par le contexte d'une vue d'ensemble, malgré la compositionnalité des calculs ; conception et intégration de mesures de similarité non binaires ; dynamique de l'interprétation d'un texte dans le fonctionnement et les objectifs des applications analysant des données textuelles.

Je tiens à remercier Benoît Habert et François Rastier pour leurs nombreuses suggestions et précisions constructives, qui ont eu une incidence décisive pour le mûrissement de ce texte.

[62] A savoir non exhaustivité de la classification, possibilité de multiclassement, mesures intrinsèquement ensemblistes plutôt que binaires, etc. cf. § 6.b).

[63] Un premier algorithme de classification multiclasse non exhaustive a été proposé et implémenté (Bommier-Pincemin 1999), mais n'a pas encore été utilisé pour le repérage d'isotopies en corpus.

8.Bibliographie

ANTOINE Jean-Yves (1994) - *Coopération syntaxe-sémantique pour la compréhension automatique de la parole spontanée*, Thèse de Doctorat en Signal-Image-Parole, Institut National Polytechnique de Grenoble, 12 décembre 1994, 319 pages.

ASSADI Houssem, BOURIGAULT Didier, GROS Cécile (1995) - « Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances », *3èmes Journées internationales d'Analyse Statistique de Données Textuelles*, Rome, décembre 1995.

BAEZA-YATES Ricardo, RIBEIRO-NETO Berthier (1999) – *Modern Information Retrieval*, Reading (Massachusetts) : Addison-Wesley.

BEAUDOUIN Valérie (2000) – *Rythme et rime de l'alexandrin classique : étude empirique des 80 000 vers du théâtre de Corneille et Racine*, Thèse de Doctorat, EHESS, Paris, 2 volumes —à paraître chez Paris : Champion, coll. « Lettres numériques ».

BEDECARRAX C., WARNESSON I. (1989) - « Relational analysis and dictionaries », *Applied stochastic models and data analysis*, 5, pp.131-151.

BENZECRI Jean-Paul & *al.* (1973) - *L'Analyse des Données*, tome I : *La taxinomie*, Dunod ; rééd. 1984, 643 pages.

BERNI CANANI Ugo (1986) - « Information retrieval et analyse de textes », Actes du Colloque *Méthodes quantitatives et informatiques dans l'étude des texte –Hommage à Charles Muller*, Nice, 5-8 juin 1985, Genève - Paris : Slatkine - Champion, pp. 79-88.

BESANÇON Romaric, RAJMAN Martin, CHAPPELIER Jean-Cédric (1999) – « Textual Similarities based on a Distributional Approach », *International Workshop on Similarity Search* (IWOSS'99), Firenze (Italy), septembre 1999.

BIBER Douglas (1988) - *Variation across speech and writing*, Cambridge University Press, 315 pages.

BOGURAEV Branimir, PUSTEJOVSKY James (1996) – *Corpus Processing for Lexical Acquisition*, Cambridge (Massachusetts) : The MIT Press, « Language, Speech and Communication » series, 245 pages.

BOMMIER Bénédicte (1993) - *Recherche d'une typologie des commandes de la DER par analyse statistique des données textuelles*, Rapport de stage de fin d'études, Ecole Centrale Paris, 2 tomes, 73 + 235 pages.

BOMMIER-PINCEMIN Bénédicte (1999) – *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Linguistique, Université de Paris IV – Sorbonne, 6 avril 1999, 806 pages.

BOUAUD Jacques, HABERT Benoît, NAZARENKO Adeline, ZWEIGENBAUM Pierre (1997) – « Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles », *Ingénierie de la connaissance*, Roskoff, pp. 207-223.

BOURDONCLE François (1997) - « LiveTopics : recherche visuelle d'information sur l'Internet », *Proceedings of RIAO'97 « Computer-Assisted Information Searching on Internet »*, 25-27 juin 1997, Montréal, pp. 651-654.

BOURIGAULT Didier (1994) - *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Thèse de Doctorat, Mathématiques et informatique appliquées aux sciences de l'homme, Ecole des Hautes Etudes en Sciences Sociales, Paris, 351 pages.

BOURIGAULT Didier, FABRE Cécile (2001) – « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de grammaire* « Sémantique et Corpus », 25, Université du Mirail.

BOURION Evelyne (1998) – « Ponctuation et accès sémantique aux banques textuelles », Actes du colloque *A qui appartient la ponctuation ?*, Liège, 12-14 mars 1997, Paris, Bruxelles, Duculot, pp. 409-435.

BRUNET Etienne (1999) – « Le triple double V », *Littérature Informatique Lecture : De la lecture assistée par ordinateur à la lecture interactive*, Alain VUILLEMIN et Michel LENOBLE (dir.), Presses universitaires de Limoges, pp. 61-85.

BRUNET Etienne (2001) – « Qui lemmatise dilemme attise », *Lexicometrica*, 2, 19 pages.

BRUNET Etienne (*à paraître*) – « Formalisation et quantification des textes. Le domaine français », *Actes du 2e séminaire de l'Ecole interlatine de hautes études de linguistique appliquée « Mathématiques et traitement de corpus »*, San Millán de la Cogolla (Espagne), 19-23 septembre 2000, 19 pages.

BURNARD Lou, SPERBERG-MCQUEEN C. M. (1996) - « La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange », traduction française par François ROLE du document TEI U5, *Cahiers GUTenberg*, 24, pp.23-151.

CAILLEZ F., PAGES J.-P. (1976) - *Introduction à l'analyse des données*, S.M.A.S.H., Paris.

CASTOT J.-J. (1981) – « Analyse statistique des emplois de mots en langue russe : cas des noms et rection des verbes », in Jean-Paul BENZECRI & *al*., *Pratique de l'Analyse des données*, tome 3 : *Linguistique et lexicologie*, pp. 230-240.

CAVAZZA Marc (1996) - « Sémiotique textuelle et contenu linguistique », *Intellectica*, 1996/2, 23, pp. 53-78.

CLARKE Charles L.A., COORMACK Gordon V., TUDHOPE Elizabeth A. (1997) - « Relevance Ranking for One to Three Term Queries », *Proceedings of RIAO'97 « Computer-Assisted Information Searching on Internet »*, 25-27 juin 1997, Montréal, pp. 388-400.

DEERWESTER Scott, DUMAIS Susan T., FURNAS Georges W., LANDAUER Thomas K., HARSHMAN Richard (1990) – « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, 41, pp. 391-407.

DIAS Gaël, GUILLORE Sylvie, BASSANO Jean-Claude, PEREIRA LOPES José-Gabriel (2000) – « Extraction automatique d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire », *Traitement automatique des langues*, 41 (2), pp. 447-472.

FABRE Cécile, BOURIGAULT Didier (2001) – « Linguistic clues for corpus-based acquisition of lexical dependencies », *Proceedings of the Corpus Linguistics 2001 Conference*, UCREL Technical Papers, 13, Lancaster University, pp 176-184.

FELLBAUM Christiane (ed.) (1998) – *WordNet : an electronic lexical database*, Cambridge (Massachusetts) : The MIT Press, « Language, Speech and Communication » series, 423 pages ; la première section de l'ouvrage reprend en partie *Five Papers on WordNet*, `http://www.cogsci.princeton.edu/~wn/`.

FIALA Pierre (1986) - « Inventaires distributionnels et opérateurs textuels dans le *Rivage des Syrtes* de Julien Gracq. Structures syntaxiques et faits stylistiques » , Actes du Colloque *Méthodes quantitatives et informatiques dans l'étude des texte –Hommage à Charles Muller*, Nice, 5-8 juin 1985, Genève - Paris : Slatkine - Champion, pp. 381-391.

FLUHR Christian (2000) – « Indexation et recherche d'information textuelle », J.-M. PIERREL (dir.), *Ingénierie des langues*, Paris : Hermès science publications, pp. 235-251.

GEFFROY Annie, LAFON Pierre, TOURNIER Maurice (1974) - « L'indexation minimale - Plaidoyer pour une non-lemmatisation », E.N.S. de Saint-Cloud, 30 pages - Communication au *Colloque sur l'Analyse des corpus linguistiques* : « Problèmes et méthodes de l'indexation maximale », Strasbourg, 21-23 mai 1973.

GONZALEZ-RUBIO Ruben, GUIZOL Jacques (1997) - « Un système de recherche et de filtrage d'information multilingue », *Proceedings of RIAO'97 « Computer-Assisted Information Searching on Internet »*, Montréal, 25-27 juin 1997, pp. 773-782.

HABERT Benoît (1998) – *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*, Habilitation à diriger des recherches en linguistique, Université Lille III – Charles de Gaulle, 16 décembre 1998, Villeneuve d'Ascq.

HABERT Benoît (2000) – « Création de dictionnaires sémantiques et typologie des textes », Actes des Journées Scientifiques 1999 « Philologie électronique et assistance à l'interprétation des textes », Jean-Emmanuel TYVAERT (dir.), *Recherches en linguistique et Psychologie cognitive*, 15, Presses Universitaires de Reims, pp. 171-188.

HABERT Benoît, BARBAUD Philippe, DUPUIS Fernande, JACQUEMIN Christian (1997) – « Simplifier des arbres d'analyse pour dégager les comportements syntactico-sémantiques des formes d'un corpus », *Lexicometrica*, 0, 29 pages.

HABERT Benoît, FABRE Cécile (1999) – « Elementary Dependancy Trees for Identifying Corpus-Specific Semantic Classes », *Computers and the Humanities*, 33 (3), pp. 207-219.

HABERT Benoît, NAZARENKO Adeline (1996) – « La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience », *Journées sur l'acquisition des connaissances*, Sète : AFIA, pp. 137-142.

HABERT Benoît, ZWEIGENBAUM Pierre (2001) – « Contextual Acquisition of Information categories : what has been done and what can be done automatically ? — version 2 ».

HARRIS Zellig (1990) – « La genèse de l'analyse des transformations et de la métalangue », *Langages* « Les grammaires de Harris et leurs questions », 99, pp. 9-20.

HJELMSLEV Louis (1968) - *Prolégomènes à une théorie du langage*, suivi de *La structure fondamentale du langage*, traduction de Una CANGER, Annick WEWER et Anne-Marie LEONARD, Minuit, coll. Arguments, 1984, 231 pages.

IDE Nancy, VERONIS Jean, éd. (1995) – *Text Encoding Initiative — Background and Context*, Dordrecht (The Netherlands) : Kluwer Academic Publishers, 246 pages, réédition de *Computers and the humanities*, 29 (1-3), 1995.

IDE Nancy, VERONIS Jean (1998) – « Introduction to the Special Issue on Word Sense Disambiguation : The State of the Art », *Computational Linguistics*, 24 (1), pp. 1-40.

ILLOUZ Gabriel, HABERT Benoît, FLEURY Serge, FOLCH Helka, HEIDEN Serge, LAFON Pierre (1999) – « Maîtriser les déluges de données hétérogènes », Actes de l'Atelier *Corpus et TAL : pour une réflexion méthodologique*, Anne CONDAMINES, Marie-Paule PERY-WOODLEY et Cécile FABRE (éd.), Conférence TALN'99, Cargèse, 12-17 juillet 1999, pp. 37-46.

JUSTESON John S., KATZ Slava M. (1991) - « Co-occurrences of Antonymous Adjectives and Their Contexts », *Computational Linguistics*, 17 (1), pp. 1-19.

LAFON Pierre (1986) - « Pour une nouvelle unité de segmentation des textes à partir de l'Inventaire des Segments Répétés. Application à la résolution générale du congrès de la CFDT en 1976 » , Actes du Colloque *Méthodes quantitatives et informatiques dans l'étude des texte –Hommage à Charles Muller*, Nice, 5-8 juin 1985, Genève - Paris : Slatkine - Champion, pp. 531-540.

LAFON Pierre, SALEM André (1983) - « L'inventaire des segments répétés d'un texte », *M.O.T.S.*, 6, pp. 161-177.

LE PESANT Denis, MATHIEU-COLAS Michel, éd. (1998) – *Les classes d'objets*, *Langages*, 131.

LEBART Ludovic, SALEM André (1994) – *Statistique textuelle*, Paris : Dunod, 350 pages.

MANNING Christopher D., SCHÜTZE Hinrich (1999) – *Foundations of Statistical Natural Language Processing*, Cambridge (Massachsetts) : The MIT Press, 680 pages.

MARANDIN Jean-Marie (1993) – « Analyseurs syntaxiques. Equivoques et problèmes », *Traitement automatique des langues*, 34 (1), pp. 5-34.

MEUNIER Jean-Guy, BISKRI Ismaïl, NAULT Georges, NYONGWA Moses (1997) – « ALADIN et le traitement connexionniste de l'analyse terminologique », *Proceedings of RIAO'97 « Computer-Assisted Information Searching on Internet »*, Montréal, 25-27 juin 1997, pp. 661-664.

MICHELET Bertrand (1988) - *L'analyse des Associations*, Thèse de Doctorat, Information Scientifique et Technique, Université de Paris VII, 26 octobre 1988, 407 pages.

MORRIS Jane, HIRST Graeme (1991) - « Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text », *Computational Linguistics*, 17 (1), pp. 21-48.

NAZARENKO Adeline (1996) - « Une méthode d'étiquetage sémantique pour l'analyse des données textuelles », *Actes de TALN'96*, Marseille, 22-24 mai 1996, pp. 228-237.

NAZARENKO Adeline (1998) - *Compositionnalité, Traitement automatique des langues*, 39 (1) –direction du numéro et Présentation, pp. 3-7.

PERY-WOODLEY Marie-Paule (1995) - « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues*, 36 (1-2), pp. 213-232.

PICHON Ronan, SEBILLOT Pascale (1999) – « Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience », *Actes de la conférence TALN'99*, Cargèse, 12-17 juillet 1999, pp. 279-288.

PIERREL Jean-Marie (1994) – « Représentations conceptuelles et intelligence artificielle », *Traitements informatisés de corpus textuels*, Eveline MARTIN (dir.), Paris : Didier érudition, coll. « Etudes de sémantique lexicale », pp. 81-106.

PIERREL Jean-Marie (2000) – *Ingénierie des langues*, (direction du volume), Paris : Hermès science publications, coll. Informatique et systèmes d'information, 354 pages.

PINCEMIN Bénédicte (1999) - « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », Actes de l'Atelier *Corpus et TAL : pour une réflexion méthodologique*, Anne CONDAMINES,

Marie-Paule PERY-WOODLEY et Cécile FABRE (éd.), Conférence TALN'99, Cargèse, 12-17 juillet 1999, pp. 26-36.

PINCEMIN Bénédicte (*à paraître*) – *Traitement automatique de la textualité*, Paris : Champion, coll. « Lettres numériques ».

PLOUX Sabine, VICTORRI Bernard (1998) – « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement Automatique des Langues*, 39 (1), pp. 161-182.

PRIE Yannick (1995) - *Contribution à une clarification des rapports entre Sémantique Interprétative et Informatique*, Mémoire de DEA, Informatique, Université de Rennes I & Laboratoire d'Intelligence Artificielle et de Sciences Cognitives de l'Ecole Nationale Supérieure des Télécommunications de Brest, août 1995, 100 pages.

RAJMAN Martin (1995) - *Apports d'une approche à base de corpus aux techniques de traitement automatique du langage naturel*, Thèse de Doctorat, Informatique et Réseaux, Ecole Nationale Supérieure des Télécommunications, Paris, 18 décembre 1995, 267 pages.

RAJMAN Martin, BESANÇON Romaric, CHAPPELIER Jean-Cédric (2000) – « Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire », *Traitement automatique des langues*, 41 (2), pp. 549-578.

RASTIER François (1987a) - *Sémantique interprétative*, Presses Universitaires de France, 277 pages.

RASTIER François (1987b) – « Représentation du contenu lexical et formalismes de l'intelligence artificielle », *Langages* « Sémantique et intelligence artificielle », 87, pp. 79-102.

RASTIER François (dir.) (1995) - *L'analyse thématique des données textuelles*, Paris : Didier, 282 pages.

RASTIER François (1997) - « Défigements sémantiques en contexte », in Martins-Baltar M., éd., *La locution entre langue et usages*, coll. Signes, E.N.S. Fontenay-St Cloud Editions, diff. Ophrys, pp. 305-329.

RASTIER François (2001) – *Arts et sciences du texte*, Presses Universitaires de France, 311 pages.

RASTIER François, CAVAZZA Marc, ABEILLE Anne (1994) - *Sémantique pour l'analyse –De la linguistique à l'informatique*, Masson, coll. Sciences cognitives, 252 pages.

RASTIER François, PINCEMIN Bénédicte (1999) – « Des genres à l'intertexte », *Cahiers de praxématique* « Sémantique de l'intertexte », 33, pp. 83-111.

REINERT Max (1991) - « Système ALCESTE : une méthodologie d'analyse des données textuelles présentée à l'aide d'une application », Actes des *Jornades Internacionals d'Analisi de Dades Textuals*, 10-12 décembre 1990, pp. 144-161.

SABAH Gérard (2000) – « Sens et traitements automatiques des langues », J.-M. PIERREL (dir.), *Ingénierie des langues*, Paris : Hermès science publications, pp. 77-108.

SALEM André (1984) - « La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes », *Les Cahiers d'Analyse des Données*, IX (4), pp. 489-500.

SALTON Gerard (1988) - « On the relationship between theoretical retrieval models », *Informetrics*, 87/88, Select Proceedings of the *1st International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*, Diepenbeck, Belgium, 25-28 August 1987, pp. 263-270.

SALTON Gerard, ALLAN James, BUCKLEY Chris (1994) - « Automatic structuring and retrieval of large text files », *Communications of the ACM*, 37 (2), février 1994, pp. 97-108.

SALTON Gerard, MCGILL Michael J. (1983) - *Introduction to Modern Information Retrieval*, McGraw-Hill.

SCHÜTZE Hinrich, PEDERSEN Jan (1993) - « A Vector Model for Syntagmatic and Paradigmatic Relatedness », *Proceedings of the 9th Annual Conference of the UW Center for the New OED and Text Research*, Oxford (England), pp. 104-113.

SPARCK JONES Karen, WILLETT Peter (1997) – *Readings in Information Retrieval*, San Francisco (California) : Morgan Kaufmann.

SPERBERG-MCQUEEN C.M., BURNARD Lou (eds) (1994) – *TEI P3 - Guidelines for Electronic Text Encoding and Interchange*, Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), Association for Literary and Linguistic Computing (ALLC), Chicago / Oxford : Text Encoding Initiative. Edition hypertexte sur CD-ROM : coll. Electronic book library, vol. 2, Providence (USA, RI) : Electronic Book Technologies, Inc.

TANGUY Ludovic (1997) - *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration informatique d'un modèle de la sémantique interprétative*, Thèse de Doctorat, Informatique, Université de Rennes I, 7 mai 1997, 207 pages.

TEIL Geneviève (1991) - *CANDIDE : un outil de sociologie assistée par ordinateur pour l'analyse quali-quantitative de gros corpus de textes*, Thèse de Doctorat, Information Scientifique et Technique, Ecole des Mines de Paris, 24 septembre 1991, 492 pages.

THLIVITIS Théodore (1998) - *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension de textes*, Thèse de Doctorat, Informatique, Université de Rennes 1, 29 juin 1998, 218 pages.

TOURNIER Maurice (1986) - « Dans l'ombre portée des sigles confédéraux : un mirage lexicométrique (C.G.T. et C.F.D.T. en 1972) », Actes du Colloque *Méthodes quantitatives et informatiques dans l'étude des texte – Hommage à Charles Muller*, Nice, 5-8 juin 1985, Genève - Paris : Slatkine - Champion, pp. 841-853.

VAILLANT Pascal (1997) - *Interaction entre modalités sémiotiques : De l'icône à la langue*, Thèse de Doctorat, Sciences Cognitives, Université de Paris-Sud (Orsay), 16 septembre 1997, 293 pages.

VOLLE Michel (1985) - *Analyse des données*, Economica, coll. Economie et statistiques avancées, 324 pages.

WARNESSON Isabelle (1985) - « Applied linguistics : optimization of semantic relations by data aggregation techniques », *Applied stochastic models and data analysis*, 1, pp.121-141.