La polysémie dans la langue générale et les discours spécialisés*

Cécile Fabre[*], *Benoît Habert*[°], *Dominique Labbé*[+]

Nous remercions D. Bourigault et E. Naulleau, grâce auxquels nous avons pu utiliser ZELLIG sur les résultats de Lexter et d'AlethIPGN, dans le cadre de deux conventions entre l'ENS de Fontenay-St Cloud et la DER-EDF. Notre analyse doit beaucoup au travail commun avec eux ainsi qu'avec J. Bouaud, A. Nazarenko et P. Zweigenbaum. Un grand merci également à F. Issac pour son travail sur les sorties graphiques.

1. Observation en corpus de la polysémie

L'idée d'une polysémie plus restreinte en langage spécialisé n'est pas nouvelle [Lerat, 1995]. Nous nous donnons ici les moyens de vérifier empiriquement cette hypothèse et de caractériser finement le fonctionnement de la polysémie nominale et adjectivale dans deux corpus, l'un relevant de la langue spécialisée (discours médical entre spécialistes), l'autre étant proche de la langue générale (discours politique à l'adresse du grand public). Nous ne partons pas de représentations sémantiques *a priori* des noms. Au rebours, notre analyse s'inscrit dans un courant de travaux qui visent à induire le fonctionnement sémantique des mots à partir de leurs contextes d'emploi dans le prolongement des travaux de Harris [1988, 1989]. De nombreuses recherches, depuis une dizaine d'années, ont été consacrées à l'acquisition automatique de classes sémantiques à partir de corpus (cf. Grefenstette [1994a] pour une présentation générale de ces travaux). Les regroupements ainsi opérés et leurs points de contact renseignent sur le degré de polysémie des mots examinés, ainsi que, plus généralement, sur l'organisation sémantique du domaine étudié. L'approche dominante peut être divisée en trois étapes selon Grefenstette [1994b] : 1) extraction des cooccurrents d'un mot, 2) association à chaque mot de l'ensemble de ses cooccurrents et mise en évidence de la proximité/distance des mots deux à deux en fonction des cooccurrents qu'ils partagent, 3) découpage en classes en fonction des proximités entre mots. A chaque étape, les techniques mises en œuvre varient. Dans la chaîne de traitement ZELLIG [Habert *et al.*, 1997b], nous utilisons un contexte syntaxique, et non une fenêtre graphique (x mots avant et après le mot considéré), pour choisir les cooccurrents. Les mots sont ensuite rassemblés à partir d'un seuil fixé de contextes partagés.

Nous proposons ici d'adopter une approche contrastive en confrontant ZELLIG à deux corpus très différents. Nous souhaitons ainsi vérifier si les fonctionnements syntaxiques des mots sont un indice fiable de leur

fonctionnement sémantique, et s'ils le sont au même titre dans les deux corpus. [Bouaud *et al.*, 1997] ont décrit les résultats du traitement par ZELLIG du corpus médical. Nous résumons leurs analyses et mettons donc ici plus nettement l'accent sur les résultats obtenus à partir du corpus de langue générale et sur leur apport en matière d'observation et d'analyse de la polysémie.

Nous présentons tout d'abord dans la partie suivante les deux corpus que nous analysons. La méthode de regroupement des dépendances syntaxiques que nous utilisons pour dégager des classes sémantiques est rappelée en **3.** Nous donnons en **4.** une vue globale des résultats obtenus sur les deux corpus. Enfin, nous les évaluons plus finement en **5.**, en insistant sur la possibilité de visualiser la dimension polysémique des mots du corpus.

2. Deux corpus : langue générale et langue spécialisée

Le recours à deux corpus a également pour rôle de limiter les jugements impressionnistes. Pour éviter les dérives interprétatives, les propositions de regroupements ou de dégroupements sémantiques que nous opérons sur la base des comportements observés en corpus ont été confrontées, pour chaque corpus, à des spécialistes du domaine et du discours en cause : à l'expérience du troisième auteur [Labbé, 1990], pour le corpus politique ; pour le corpus médical, à celle des personnes qui ont travaillé à la compréhension automatique de ce type de langage médical [Zweigenbaum, 1994].

2. 1. *MENELAS* : le langage médical des maladies coronariennes

L'expérience utilise en premier lieu le corpus rassemblé dans le cadre du projet MENELAS [Zweigenbaum, 1994] de compréhension de comptes-rendus d'hospitalisation dans le domaine des maladies coronariennes. Ce corpus (désormais MENELAS) se compose d'un extrait de manuel mais surtout de comptes rendus d'hospitalisation, ainsi que de lettres de médecins hospitaliers aux médecins traitants. Il contient 84 839 mots (occurrences) pour 6 191 formes différentes.

2. 2. MITTERRAND1 : les interventions radio-télévisées du premier septennat.

Le second corpus a été rassemblé dans le cadre d'une étude du discours politique français contemporain [Labbé, 1990]. Il comporte l'ensemble des prestations radio-télévisées de F. Mitterrand au cours de son premier septennat : allocutions, entretiens, conférences de presse, soit

68 textes. Les transcriptions diffusées par le service de presse de l'Elysée ont été systématiquement contrôlées à l'aide des enregistrements audio et vidéo. Le corpus (désormais MITTERRAND1) est entièrement lemmatisé : à chaque forme graphique, on associe un lemme (son entrée dans les dictionnaires), et sa catégorie grammaticale. Le corpus comporte 305 124 occurrences, 14 362 formes graphiques et 7 700 lemmes différents.

2. 3. Convergences et écarts

Les deux corpus sont, en totalité pour MITTERRAND1, en partie pour MENELAS (les lettres et les comptes rendus) de l'oral transcrit, énoncé au cours d'un discours ou d'un interview dans le premier cas, dicté dans le second. Les phénomènes propres à l'oral (pauses, interruptions, reprises) en sont par contre exclus. Signalons le décalage de taille : MENELAS représente à peu près le quart de MITTERRAND1. Le poids du temps —un septennat entier — se fait sentir pour MITTERRAND1. Cette dimension diachronique est absente de MENELAS. En outre, MENELAS associe des documents de trois "genres" spécifiques (énoncé didactique, compte rendu, lettre aux collègues), MITTERRAND1 mêle des documents plus variés, aux situations de communication diverses.

MENELAS constitue sans conteste possible un corpus de langue spécialisée, à la fois par sa thématique et par les genres spécifiques qui le structurent : ils correspondent à des situations de communication où le récepteur, le plus souvent unique, partage la culture du locuteur (même partiellement comme dans le cas de l'extrait de manuel). Du fait de sa variété thématique et de sa portée (le message présidentiel a pour destinataire l'ensemble des Français), le corpus MITTERRAND1 appartient plutôt à la langue générale même s'il est caractéristique du discours politique français contemporain. Certes, les notions de "langue spécialisée" et de "langue générale" sont deux archétypes qu'on ne rencontre jamais à l'état pur. MENELAS contient aussi de la langue générale et le corpus MITTERRAND1 de la langue spécialisée. Mais, pour l'essentiel, la spécialisation du vocabulaire, dans le discours présidentiel, ne vient pas d'un savoir technique partagé par des spécialistes, comme dans le cas de MENELAS, mais des questions du moment, et de ce que les interlocuteurs pensent être les attentes des Français envers leurs dirigeants.

3. Regroupements à partir de dépendances syntaxiques

Le logiciel ZELLIG permet de visualiser des regroupements de mots obtenus à partir du calcul des contextes qu'ils partagent. Ce traitement

s'effectue en deux temps, par extraction de dépendances syntaxiques élémentaires puis par regroupement des mots sur la base des distributions communes. Cette chaîne de traitement est décrite précisément dans [Habert et *al.*, 1997b]. Nous en résumons ici les principales étapes à travers un exemple.

Pour rapprocher les entrées lexicales sur la base des contextes dans lesquels elles apparaissent et dégager certaines des relations sémantiques qu'elles entretiennent, le logiciel s'appuie sur les arbres d'analyse produits par des extracteurs de groupes nominaux[1], pour en extraire les arbres élémentaires sous-jacents[2]. Cette normalisation permet de mettre au jour les régularités que masquent les groupes nominaux complexes extraits. Par exemple, pour la séquence *le prochain sommet des grands pays industriels,* ZELLIG fournit quatre arbres élémentaires à partir de l'arbre d'analyse complet. Les voici, accompagnés de leur transcription syntaxique :

a. *prochain sommet*	[SN [SADJ [ADJ *prochain*]] [SN [N *sommet*]]]
b. *sommet des pays*	[SN [SN [N *sommet*]] [SP [Prép des] [SN [N *pays*]]]]
c. *grand pays*	[SN [SADJ [ADJ *grand*]] [SN [N *pays*]]]
d. *pays industriels*	[SN [SN [N *pays*]] [SADJ [ADJ industriels]]]

On obtient donc à partir de ce groupe nominal complexe une série de dépendances élémentaires binaires, c'est-à-dire entre deux mots pleins (nom ou adjectif), correspondant chacune à une relation entre un gouverneur (ou tête) et un mot régi (argument ou modifieur). Elles ne correspondent donc pas à de simples proximités textuelles, mais à des relations de dépendance vérifiées dans les arbres d'analyse.

Chaque mot est alors assorti de la liste de ses distributions. On obtient ainsi des classes distributionnelles dans une position donnée d'une dépendance élémentaire. Par exemple, dans MITTERRAND1, les noms *peuple* et *monde* partagent avec *pays* le contexte c. (noms qualifiés par l'adjectif antéposé *grand)* ; l'adjectif *dernier* partage avec *prochain* le contexte a. (adjectifs antéposés au nom *sommet).*

A partir des contextes élémentaires, ZELLIG construit un graphe dont les nœuds sont les lemmes et les arêtes sont les contextes partagés par deux lemmes. Pour limiter les proximités artificielles (liées par exemple à l'emploi de mots sans valeur sémantique discriminante, comme certains adjectifs très courants et constituant de ce fait des cooccurrents banals pour un vaste ensemble de noms : *certain, autre,* etc.), un nombre minimal de contextes partagés est requis pour placer une arête entre deux lemmes : dans cette expérience, soit 5, soit 10, pour MENELAS et soit 10, soit 15, pour MITTERRAND1. La taille plus importante de MITTERRAND1 et le plus grand nombre de dépendances élémentaires générées obligent à écrémer davantage.

[1] *Ici Lexter [Bourigault, 1994] et AlethIPGN (développé par GSI-ERLI dans le cadre du projet Eureka GRAAL). Cf. [Habert et al., 1997a] pour une comparaison de ces deux extracteurs.*

[2] *Soulignons une disparité dans le traitement des deux corpus. MENELAS a été catégorisé par AlethIPCAT, étiqueteur développé par GSI-ERLI : de nombreux mots restent non reconnus, les erreurs de catégorisation foisonnent. Les programmes d'extraction de groupes nominaux intervenant en aval en pâtissent forcément. A l'inverse, pour MITTERRAND1, l'étiquetage et la lemmatisation ont donné lieu à une vérification et une correction soigneuses par le troisième auteur. La transformation des étiquettes en étiquettes utilisables par Lexter a été effectuée par les deux premiers auteurs.*

Le graphe fournit deux formes de regroupement de lemmes à travers ses composantes connexes (CC) et ses cliques (KC). Une composante connexe est une partie de graphe telle qu'il y ait un chemin entre deux nœuds quelconques. Une clique est un graphe où chaque nœud est relié à tous les autres par une arête. La figure 1 montre la quarante-cinquième clique obtenue au seuil 10 sur MITTERRAND1. Chaque arc est étiqueté par les contextes que partagent les deux nœuds qu'il relie. La mention des lemmes est remplacée par un tilde (~) dans les contextes ; en outre, le nombre d'occurrences de chaque contexte est indiqué. Par exemple, entre les nœuds PAYS et NATION, le contexte :

9 *ensemble de ~*

doit être lu ainsi : on trouve 9 fois le contexte *ensemble de pays* (nœud gauche, numéro **<0>**) et 1 fois (c'est la valeur par défaut) le contexte *ensemble de nations* (nœud droit, numéro **<2>**). ZELLIG permet également de visualiser les contextes propres à chaque nœud (qu'il ne partage avec aucun nœud du graphe). Nous ne les avons pas reproduits dans la figure faute de place.

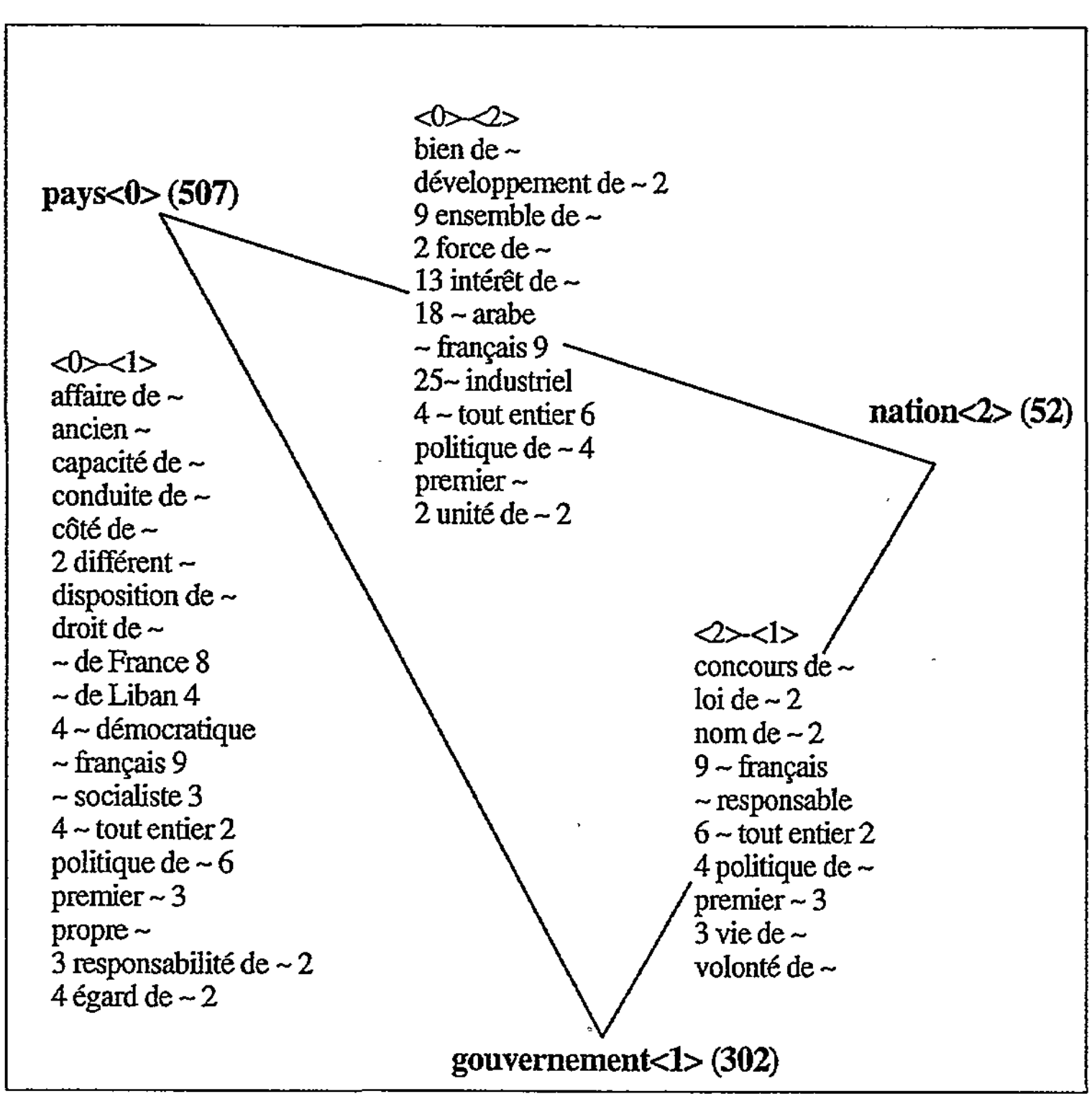

figure 1
MITTERRAND1 KC 45, seuil 10

4. Présentation globale des regroupements obtenus

4. 1. MENELAS

Sur les données produites par ZELLIG, quels que soient l'extracteur et le seuil utilisés, on observe une répartition similaire des résultats : une première composante pléthorique qui semble regrouper plusieurs catégories sémantiques en intersection, suivie d'une série de composantes beaucoup plus petites, mais également plus homogènes.

La première composante de seuil 10 met en évidence des lemmes "attracteurs", qui partagent des contextes avec de nombreux autres lemmes. C'est le cas des mots LÉSION, STÉNOSE, ARTÈRE, IVA[3]. Autour de ces attracteurs, on remarque des zones denses, enchevêtrées, où les cliques foisonnent : cette même composante fait apparaître les groupes {LÉSION STÉNOSE ATTEINTE RESTÉNOSE MALADIE PONTAGE ANGIOPLASTIE ARTÈRE RÉSEAU DROITE} et {ARTÈRE RÉSEAU TRONC IVA BRANCHE SEGMENT INTERVENTRICULAIRE DIAGONALE}. Enfin, on note des groupes de nœuds dont le rattachement au reste de la composante est ténu. L'arête {ANGIOPLASTIE CORONAROGRAPHIE} est le seul lien qui rattache au reste de la composante l'étoile autour du mot BILAN, constituée des nœuds {CONTRÔLE ÉPREUVE EXPLORATION PLAN EXAMEN}, et qui est homogène sémantiquement : il s'agit des examens, qui sont parfois des interventions, comme le manifeste EXPLORATION ou le lien de CORONAROGRAPHIE avec ANGIOPLASTIE.

[3]*Les nœuds du graphe sont indiqués en petites majuscules, les contextes en italiques. Nous entourons d'accolades les listes de nœuds que nous décrivons comme des regroupements sémantiques pertinents.*

L'interprétation de cette première composante déclenche deux types de réactions. La volonté de faire "éclater" certains regroupements qui paraissent hétérogènes, en fonction de connaissances générales sur la langue ou sur le monde : ARTÈRE, STÉNOSE, ANGIOPLASTIE semblent relever de trois catégories distinctes, les organes, les affections au sens large (cf. le suffixe *-ose*), les interventions médicales (cf. le suffixe *-plastie* et la base *angio-* identifiant le site corporel concerné). Seconde réaction : la volonté de voir si certains regroupements, *a priori* surprenants, ne manifestent pas toutefois des comportements sémantiques propres au domaine. C'est le cas d'ÉPREUVE qui ici rentre avant tout dans le syntagme *épreuve d'effort* : une série normée d'efforts demandés à un patient cardiaque à fins d'examen.

Les autres composantes sont souvent réduites : entre 3 et 5 nœuds. Certaines dégagent des fragments de catégories sémantiques comme :

— la gravité d'un dysfonctionnement : {MINIME MODÉRÉ NET SIGNIFICATIF IMPORTANT}. Une partie de cette catégorie se retrouve dans une autre composante, liée également à d'autres formes : {SÉVÈRE DIFFUS}.

— la localisation au sein du myocarde ou par rapport aux affections qui le touchent : {LIMITÉ ANTÉRIEUR POSTÉRIEUR LATÉRAL INFÉRIEUR} où LIMITÉ joue le rôle d'intrus. Le seuil 5 rattache ce groupe d'adjectifs aux

autres adjectifs essentiels du corpus. Mais il y constitue une zone dense et homogène, d'ailleurs enrichie : {ANTÉRO-APICAL POSTÉRO-INFÉRIEUR ANTÉRO-LATÉRAL}.

D'autres manifestent des comportements syntaxiques particuliers. C'est le cas des adjectifs ou déterminants (considérés comme des adjectifs par l'étiqueteur employé) et qui peuvent se placer avant le nom : {AUTRE PREMIER SEUL DEUX DOUBLE TROIS MÊME DERNIER PROCHAIN}. Il est intéressant de voir par exemple que DOULEUR est quantifié, au même titre que des mots qui le sont plus habituellement (JOUR, MOIS, SEMAINE), probablement en tant que symptôme dont la fréquence et la datation comptent.

Une configuration particulière est celle des composantes à trois nœuds où un nœud joue le rôle d'intermédiaire. C'est le cas, par exemple, de {MAUVAIS BON BEAU}. On y trouve des liens d'hyponymie, le rapprochement de synonymes et d'antonymes, ou encore l'articulation entre deux fonctionnements d'un même mot. Ces structures renseignent aussi sur des comportements spécifiques de mots courants. Dans ce type de texte, BEAU fonctionne comme un diagnostic positif sur les possibilités d'évolution d'un site corporel. Et, hors connaissance du domaine, on aurait du mal à prédire les noms qu'il peut effectivement modifier : *calibre branche lit.* Ce qui semble d'ailleurs qualifié ici, ce n'est pas la partie du corps concernée, mais sa capacité à remplir sa fonction typique : une *belle branche d'artère* est une branche qui permet au sang de circuler au mieux.

4. 2. MITTERRAND1

La première composante, obtenue au seuil 15 (v. **figure 2**, page suivante) regroupe à la fois des noms et des adjectifs, le mot POLITIQUE par son double fonctionnement assurant la transition entre les deux groupes.

Trois ensembles se dégagent :

— un vocabulaire géopolitique des pays et des institutions : {GOUVERNEMENT MINISTRE MONDE FRANCE PAYS EUROPE PEUPLE RÉGION FRANÇAIS}

— des adjectifs "étiquetant" des niveaux, des strates de la réalité : {ÉCONOMIQUE SOCIAL POLITIQUE INDUSTRIEL NATIONAL FRANÇAIS EUROPÉEN AMÉRICAIN}

— une série de noms très généraux centrés autour du mot POLITIQUE et caractérisant ses modes d'application (conditions, effets) : {PLAN DÉCISION MOYEN SITUATION PROBLÈME POLITIQUE AFFAIRE}.

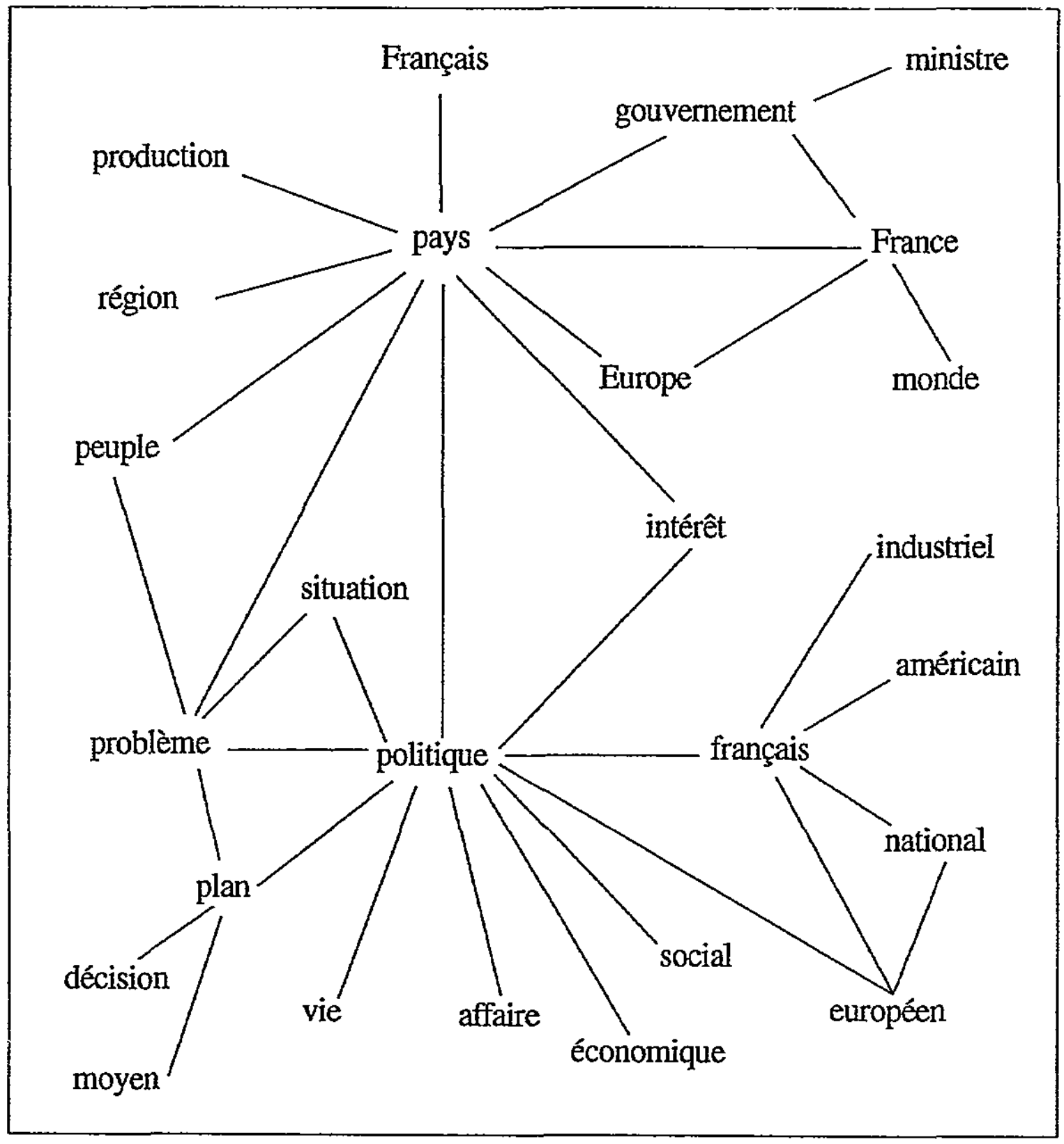

figure 2
MITTERRAND1 CC 1, seuil 15

Cette configuration se retrouve, sous une forme plus complexe, dans la première composante de seuil 10. Elle est formée de 107 nœuds, ce qui rend difficile une lecture globale. Les éléments attracteurs sont encore GOUVERNEMENT, autour duquel gravitent les noms désignant des instances politiques ou des communautés humaines : {RESPONSABLE PARTI MAJORITÉ CHEF RÉPUBLIQUE PRÉSIDENT ÉTAT PEUPLE FRANCE EUROPE PAYS HOMME FRANÇAIS MINISTRE POLITIQUE NATION}, FRANCE, PAYS (et les entités géopolitiques), ainsi que POLITIQUE. Une portion importante du graphe est consacrée aux adjectifs : on retrouve le vocabulaire des "strates", principalement organisé autour de l'adjectif FRANÇAIS. Il est composé à la fois d'autres adjectifs de nationalité, ou plus généralement de localisation spatiale : {AMÉRICAIN ALLEMAND AFRICAIN SOVIÉTIQUE INTERNATIONAL ÉTRANGER EUROPÉEN}, d'adjectifs désignant

des localisations temporelles : {ACTUEL NOUVEAU}, et des champs d'activité : {INDUSTRIEL ÉCONOMIQUE AGRICOLE PUBLIC PRIVÉ}. Au-delà de ces groupements très denses autour de quelques pôles attracteurs, sur lesquels nous reviendrons, on observe des associations plus spécifiques, comme par exemple le vocabulaire de la force : {FORCE MOYEN ARME ARMEMENT ÉQUILIBRE DÉCISION} ou le vocabulaire de l'entreprise : {ENTREPRISE SECTEUR INDUSTRIE}.

Les composantes de taille plus modeste présentent plus d'homogénéité, qu'il s'agisse d'homogénéité sémantique (vocabulaire du temps : {TEMPS ANNÉE MOIS}, du nombre : {MILLION MILLIER}) ou d'homogénéité linguistique : les éléments de déterminants complexes sont par exemple regroupés dans une composante ({ENSEMBLE NOMBRE} : *un ensemble* ou *un grand nombre de mesures, de pays, de travailleurs,* etc.) tout comme les adjectifs antéposés. Cette importance de l'antéposition distingue nettement MITTERRAND1 de MENELAS. Cette particularité tient principalement au poids des adjectifs évaluatifs : {BON FORMIDABLE GRAND PETIT TRÈS-GRAND}. Y contribue également un ensemble d'adjectifs ordinaux ou indiquant le rang dans une série : {PROCHAIN DERNIER DEUXIÈME PREMIER}.

5. Évaluation

L'interprétation des regroupements produits par ZELLIG aboutit pour les deux corpus à délimiter des ensembles plus ou moins facilement "dénommables". Pour l'étude de la polysémie éventuelle, les noms qui figurent à l'articulation de plusieurs ensembles fournissent un point d'observation privilégié.

5. 1. Univocité conceptuelle pour MENELAS

Prenons pour MENELAS STÉNOSE, probablement la forme qui, dans la première composante, possède le plus de "voisins" immédiats et qui donc, vu la variété de ces associations, semble la plus susceptible de polysémie. Une sténose se caractérise par son aspect — la longueur et le diamètre, notamment — *sténose à 70%* (P1), par son degré de gravité : *sténose serrée* (P2) et par sa localisation, en particulier sur une artère ou une partie d'artère : *sténose circonflexe, sténose de tronc* (P3). On constate également qu'une sténose fait l'objet d'un acte thérapeutique (P4) et qu'elle peut être considérée comme un processus : *apparition de sténose* (P5).

Si l'on considère les contextes que STÉNOSE partage avec chacun de ses voisins, on peut établir le classement suivant :

— seul LÉSION partage avec STÉNOSE tous ses types de contextes : la forte proximité entre ces deux lemmes se trouve confirmée à la fois par le nombre de contextes partagés, par le nombre des occurrences de contextes concernés et par le nombre de types de contextes partagés, *i. e.* par la diversité sémantique de ces contextes.

De nombreux lemmes partagent avec STÉNOSE plusieurs types de contextes (3 ou 4) mais pas tous. Parmi cet ensemble, certaines sous-catégories se dégagent :

— 5 lemmes (INSUFFISANCE ATTEINTE MALADIE PLAQUE DOULEUR) ont en commun de partager toutes les propriétés de STÉNOSE, sauf P4. À l'exception de PLAQUE, ils fonctionnent tous dans le corpus comme des hyperonymes de STÉNOSE en désignant des pathologies de manière plus générique, à partir d'un diagnostic moins précis.

— ISCHÉMIE et INFARCTUS se distinguent de l'ensemble : ils ne peuvent être localisés que par rapport au cœur. Ils s'écartent de ce fait de la famille de STÉNOSE pour laquelle la localisation artérielle est centrale (P3).

— 4 lemmes (IVA ARTÈRE DROITE BRANCHE) se rapprochent essentiellement de STÉNOSE du fait de P4 et P1, mais surtout de P3 (nombre important à la fois de contextes et d'occurrences de contextes).

— 2 lemmes (OCCLUSION RESTÉNOSE) partagent les propriétés P1, P2, P3 et P4. Ils se distinguent de ARTÈRE et DROITE mentionnés ci-dessus par le fait que, cette fois, P1 et P2 jouent un rôle important.

— 5 lemmes (ANOMALIE ATHÉROME RÉTRÉCISSEMENT DILATATION TRÈS) se rapprochent de STÉNOSE sur la base de ses trois attributs principaux : P1, P2 et P3.

— 3 lemmes (PONTAGE ANGIOPLASTIE SEGMENT) ne partagent que deux attributs avec STÉNOSE. Dans ces cas-là, c'est la variété des contextes introduisant un attribut donné[4] qui fait apparaître une proximité sur les graphes, mais cette proximité reste très partielle.

[4] *Sur les 14 contextes que PONTAGE et ANGIOPLASTIE partagent avec STÉNOSE, 9 ou 10 introduisent une localisation sur une artère ; 10/16 contextes partagés entre SEGMENT et STÉNOSE introduisent une localisation.*

Dans le cas du corpus MENELAS, on observe donc un comportement monosémique des mots, même si un même mot peut partager des facettes différentes avec ses voisins : le graphe met alors en évidence les différentes composantes du sens, c'est-à-dire les différents éléments qui entrent dans la compréhension du concept.

5. 2. Polysémie massive pour MITTERRAND1

Les regroupements formés à partir du corpus MENELAS peuvent être mis en relation avec les classes de concepts du domaine [Bouaud et *al.*,

1997]. Dans le cas de MITTERRAND1, l'accès à la catégorisation semble plus problématique, dans la mesure où les classes qui se dessinent constituent des ensembles thématiques vastes, souvent hétérogènes, ou très généraux. Si l'objectif d'induction de classes conceptuelles, qui s'est avéré fructueux dans le cas du corpus technique, semble moins adapté au traitement du corpus MITTERRAND1, que peut nous apporter le logiciel pour l'étude d'un corpus plus proche de la langue générale, et caractérisé par une grande polysémie des nœuds du graphe ?

5. 2. 1. — Des possibilités limitées de catégorisation sémantique

L'observation des sorties produites sur MITTERRAND1, pour les noms et les adjectifs, fait apparaître quelques classes sémantiques que l'on peut sans difficulté étiqueter :

— {GOUVERNEMENT PARTI MAJORITÉ OPPOSITION} : instances politiques
— {HOMME GENS PAYS} : acteurs
— {FRANCE PAYS EUROPE MONDE} : géographie
— {EUROPÉEN NATIONAL FRANÇAIS} : géographie
— {INDUSTRIEL ÉCONOMIQUE SOCIAL MILITAIRE} : champ d'activité

En outre, les contextes partagés et les contextes propres à chacun des mots permettent de préciser les caractéristiques de chaque nœud. Par exemple, EUROPÉEN et FRANÇAIS partagent : *crise ~, défense ~ , marché ~, production ~, société ~,* alors que FRANÇAIS possède en propre : *département ~, délégation ~, jeunesse ~, contingent ~.* Cependant, cette tentative de catégorisation, qui révèle des thématiques banales du discours politique, ne résiste pas toujours à un examen plus attentif des contextes. Par exemple, le rapport entre FRANCE et MONDE (rapport d'opposition) est différent du rapport entre MONDE et PAYS à cause de l'homographie entre le singulier et le pluriel. Ainsi, l'expression *les pays arabes* est synonyme de : *le monde arabe,* comme *le pays* l'est de *la France,* alors que cette dernière est en opposition à MONDE (notamment dans l'expression *la France et le reste du monde).* Cet exemple montre que les techniques actuelles d'extraction enregistrent des relations spatiales deux à deux qui, en langue générale, ne définissent pas toujours des classes sémantiques homogènes.

5. 2. 2. — Une plus grande complexité des relations attestées

L'observation des cliques met au jour, dans le vocabulaire de F. Mitterrand, des rapports sémantiques traditionnels comme la synonymie (FRANÇAIS et NATIONAL, PAYS et NATION), l'antonymie (PRIVÉ et PUBLIC, RICHE et PAUVRE, PAYS et MONDE), l'hyperonymie (POLITIQUE et DÉFENSE, PAYS et MONDE). Ces relations sont toujours partielles et sont

rendues complexes par de nombreux glissements de sens. Par exemple, la synonymie n'est jamais complète : CATÉGORIE n'est synonyme de COUCHE que dans certains contextes précis *(catégorie sociale).* Ou encore, le lien entre FORCE et MOYEN provient de quelques contextes communs — *dissuasion, nucléaire, tactique, stratégique* — alors que ces deux mots apparaissent dans beaucoup d'autres contextes qui leur sont propres. Par exemple, les *forces françaises* ou les *forces américaines* sont synonymes d'armée et interviennent dans des contextes différents de : *moyens de production, moyens de communication,* etc. De même, MAJORITÉ et OPPOSITION ne sont antonymes que dans le contexte parlementaire et sont indépendants dans d'autres expressions fréquentes : *la grande majorité de ~, être en opposition à ~.* Enfin, l'hyperonymie est encore plus complexe. Par exemple POLITIQUE englobe : ACTION, DÉCISION, DÉCLARATION, DÉFENSE, MESURE, MOYEN, NUCLÉAIRE, PLAN et peut se substituer à eux, dans certains contextes plus ou moins limités (synonymie partielle). Ou encore PAYS, employé au pluriel englobe un emploi particulier de monde, partie de l'univers : *~ arabe, ~ riche, ~ développé, ~ occidental.* Généralement, les relations se combinent. On peut en trouver une illustration assez claire dans les deux types de contextes reliant POLITIQUE et DÉFENSE : *~ de l'emploi, ~ de la paix vs ~ française, ~ commune, ~ du pays.* La première série indique une synonymie entre les deux mots, l'un pouvant se substituer à l'autre sans modifier le sens : *la défense de l'emploi* est une certaine politique de l'emploi alors que *la défense du pays* est incluse dans la politique du pays comme un de ses aspects parmi d'autres.

Tout cela conduit à un certain éparpillement sémantique et à une complexité plus grande que dans la description du vocabulaire technique.

5. 2. 3. — D'autres modes de caractérisation du sens

Les résultats obtenus sur le corpus MITTERRAND1 ont donc mis au jour une polysémie massive des principaux termes. Cela rend nécessaire l'invention de nouveaux modes d'analyse pour le vocabulaire général. Dans cette perspective, deux points nous semblent fondamentaux : la structuration sémantique des mots polysémiques et les associations sémantiques non réductibles à une relation lexicale traditionnelle.

En premier lieu, il s'agit de repérer les sphères d'influence des mots en s'appuyant sur les cliques, mais aussi en effectuant des retours au texte. Ainsi, dans ce qui apparaît au premier abord comme des relations de synonymie partielle, un examen plus approfondi fait souvent apparaître une spécialisation des termes. Par exemple, GRAND et BON remplissent tous deux apparemment les mêmes fonctions d'amplification et de mélioration mais celles de GRAND sont beaucoup plus puissantes que celles de BON : chez F. Mitterrand, une *grande majorité* est toujours confortable alors qu'une *bonne majorité* a toutes les chances de n'être que

relative. GRAND exprime ainsi une amplification forte *(césure, danger, angoisse, coupure, injustice, péril, deuil)* quand BON exprime une amplification faible *(majorité, part, partie)* ; en outre, on trouve des contextes de mélioration forte avec GRAND *(enthousiasme, satisfaction, sagacité, prestige)* et de mélioration faible avec BON *(observation, occasion, question, remarque).*

Plus largement, l'examen attentif du contexte permet de résoudre bien des synonymies ou les hyperonymies apparentes, dissipant la sensation de flou et de foisonnement qui se dégage d'un premier examen des graphes. Prenons pour exemple le mot POLITIQUE qui est certainement le plus "foisonnant". Chez Mitterrand, la politique, employée sans épithète, est toujours l'affaire des Français alors que, lorsqu'il traite de la politique économique, il n'emploie pas le substantif *Français* mais : *salariés, chefs d'entreprise, commerçants, agriculteurs.* D'où la nécessité de considérer que POLITIQUE ne constitue pas un seul mot mais bien deux voire plus. La polysémie des principaux mots du lexique obligerait donc à éclater les nœuds des graphes en plusieurs champs. Si l'on examine le mot POLITIQUE déjà cité, quatre cliques définissent apparemment trois champs sémantiques (les lieux, les thèmes, les objectifs) :

— POLITIQUE, FRANÇAIS, EUROPÉEN : champ géographique

— POLITIQUE, SOCIAL, ÉCONOMIQUE : champ thématique

— POLITIQUE, DÉVELOPPEMENT : champ thématique

— POLITIQUE, FORCE, ÉQUILIBRE : champ télique

Cependant, à l'intérieur de ces vastes champs sémantiques, les variations de sens peuvent être importantes, ce qui oblige à opérer à nouveau d'autres partitions. Par exemple le vocabulaire de la politique économique *(bataille, crise, décision, force, guerre, inégalité, indépendance, pouvoir, succès)* est nettement plus tendu et combatif que le vocabulaire de la politique sociale *(choc, inégalité, harmonie, problème, réalité, vie),* ce dernier étant apparemment plus de l'ordre de la description que de l'action. De même, la géographie se confond souvent avec la thématique. Par exemple, chez Mitterrand *politique européenne* ne désigne pas l'action conduite au niveau de la Communauté européenne, voire du continent entier, mais toujours *la politique européenne de la France.* Lorsque F. Mitterrand parle des politiques conduites au niveau européen — par exemple dans le domaine agricole ou douanier — il utilise l'adjectif *commun (aux pays membres)* et non pas l'adjectif *européen.* Sans doute est-ce parce que, à ses yeux, la seule *politique européenne* est celle que décident les chefs d'Etat et de gouvernement.

Comme on le voit à l'aide de ces quelques exemples, sous l'apparence "molle" du vocabulaire général, certaines associations de mots peuvent être très bien spécialisées, mais c'est le retour aux contextes qui permet de

lever les ambiguïtés et d'établir cette spécialisation. Comme dans le cas de MENELAS, l'inspection des contextes permet de caractériser un lien apparemment surprenant. Au-delà des relations lexicales traditionnelles, on repère alors des sèmes communs, qui permettent de rapprocher des mots dans l'univers discursif selon des modalités extrêmement diverses. Par exemple, si les mots PAYS et NATION sont reliés par une grande diversité de contextes (v. **figure 1,** p. **19**), les mots DÉCISION et QUESTION ont en commun exclusivement des termes qui recensent *(dernier ~, deuxième ~, premier ~, quatrième ~, nombre de ~),* ou des mots qui désignent des individus *(~ de monsieur, ~ de ministre).* Ainsi, alors que dans le premier cas on a affaire sans conteste à une synonymie, dans le deuxième le lien repère des mots dont le point commun est de désigner des événements ponctuels ordonnés dans le temps et constituant des actes d'énonciation. De même, dans une des cliques au seuil 10, la cohérence du groupement des quatre mots PAYS, POLITIQUE, INTÉRÊT et DROIT se manifeste dans l'apparition constante de certains contextes sémantiquement proches au niveau de chaque arête : *conduit ~, développe ~, responsable de ~, défense de ~, service de ~, respect de ~.* Cet ensemble signale donc des entités marquées positivement, dont il s'agit de se porter garant.

Si la catégorisation conceptuelle semble décevante pour caractériser finement le discours général, les groupements permettent par contre de mettre au jour des correspondances plus subtiles entre les mots, en signalant des sèmes partagés.

6. Conclusion

L'étude comparée des sorties obtenues avec ZELLIG à partir de MÉNÉLAS et de MITTERRAND1 semble indiquer qu'un corpus de langue générale n'offre pas la même organisation sémantique qu'un corpus de langue spécialisée dans la mesure où les délimitations entre classes sémantiques en fonction des contextes partagés s'avèrent moins nettes et les possibilités de catégorisation sémantiques plus limitées. A l'inverse, d'autres phénomènes — comme la polysémie, l'interpénétration de plusieurs champs lexicaux, les glissements de sens — se manifestent ici et constituent probablement des entrées plus pertinentes pour l'analyse d'un corpus de cette nature. Cette expérience tend à montrer qu'en langue spécialisée, il est possible de mettre en œuvre une sémantique "conceptuelle" (classement des mots par concepts) alors que l'analyse d'un corpus en langue générale appelle également une sémantique "interprétative" permettant de rendre compte des glissements sémantiques en fonction des thèmes ainsi que des liens plus ténus entre les mots autour de sèmes communs. Elle vérifie également le fait qu'un langage spécialisé

vise à décrire une classe limitée de phénomènes avec des mots relativement univoques, alors que le vocabulaire général vise à communiquer un nombre illimité d'expériences dans des situations imprévisibles et faiblement codifiées [Sager, 1986].

Cette première description des résultats sur MITTERRAND1 nous convainc donc de poursuivre l'analyse de ce type de corpus en adoptant d'autres objectifs que ceux qui avaient guidé les premières expérimentations, et en favorisant, au détriment de la catégorisation conceptuelle, l'étude des facettes nominales et adjectivales, manifestées par des liens extrêmement divers (conceptuels, sémantiques, formels) qui permettent de tracer les différents modes d'apparition des mots dans le discours.

D'un point de vue plus technique, plusieurs observations tracent des perspectives afin d'améliorer les critères de regroupement et la lisibilité des liens exhibés. Tout d'abord, le groupe nominal semble un moins bon observatoire des fonctionnements sémantiques pour MITTERRAND1 que pour MENELAS. L'analyse des verbes et des pronoms dans [Labbé, 1990] montre d'ailleurs bien qu'une partie du fonctionnement du discours mitterrandien tient à des jeux subtils sur l'énonciation (entre *je, nous, la France, le Président de la République)* et aux modalisations des verbes. Dans une phase ultérieure, il s'agira donc d'affranchir ZELLIG de la contrainte consistant à intervenir en aval d'analyseurs spécialisés dans l'analyse des groupes nominaux. Enfin, il s'avère que la lecture des contextes permet seule de caractériser un lien et d'éviter des jugements approximatifs (nous avons vu que le spectre des relations est extrêmement large, entre par exemple une synonymie et un rapprochement fondé sur des critères essentiellement syntaxiques). Notre objectif est donc d'essayer de caractériser des types de contextes. D'après nos premières observations, il semble par exemple que le rôle des contextes soit sensiblement différent selon qu'il s'agit de contextes gauche ou droite, adjectivaux ou nominaux : les synonymes sont reliés par un plus grand nombre de structures de type N de N, qui semblent plus discriminantes que des contextes adjectivaux. Il s'agit donc de mettre en œuvre des filtres plus fins pour faciliter la compréhension des regroupements, l'objectif étant à terme de déterminer les critères permettant d'accorder des poids différents aux liens constatés.

*(*ERSS — Université de Toulouse-le-Mirail)*

(°Equipe Linguistique et Informatique — Ecole Normale Supérieure de Fontenay St Cloud)

(+CERAT — IEP Université Pierre Mendès-France de Grenoble)

Bibliographie

BOUAUD (J.), HABERT (B.), NAZARENKO (A.), ZWEIGENBAUM (P.)
1997, "Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation de deux modélisations contextuelles", *in Ingénierie de la connaissance,* Roscoff.

GREFENSTETTE (G.)
1994a, "Corpus-Derived First, Second and Third Order Affinities", *in EURALEX,* Amsterdam.

1994b, *Explorations in Automatic Thesaurus Discovery,* Dordrecht, Kluwer Academic Publisher.

HABERT (B.), HERVIOU-PICARD (M.-L.), BOURIGAULT (D.), QUATRAIN (R.), ROUMENS (M.)
1997a, "Un Outil et une méthode pour comparer deux extracteurs de groupes nominaux", *in 1ères Journées Scientifiques et Techniques FRANCIL.*

HABERT (B.), BERTRAND-GASTALDY (S.), NAZARENKO (A.), DUPUIS (F.), NAULLEAU (E.), LEMIEUX (M.), DELISLE (C.)
1997b, "Recyclage d'analyses syntaxiques automatiques pour le repérage de variantes de termes", Montréal, *Actes de Coopération franco-québécoise.*

HARRIS (Z.)
1988, *Language and Information,* New York, Columbia University Press.

HARRIS (Z.), GOTTFRIED (M.), RYCKMAN (T.), MATTICK (P.), DALADIER (A.), HARRIS (T. N.), HARRIS (S.)
1989,"The Form of Information in Science, Analysis of Immunology Sublanguage", *Boston Studies in the Philosophy of Science,* vol. 104, Kluwer Academic Publisher.

LABBÉ (D.)
1990, *Le Vocabulaire de François Mitterrand,* Paris, Presses de la Fondation nationale des Sciences Politiques.

LERAT (P.)
1995, *Les Langues spécialisées,* Paris, PUF (Linguistique nouvelle).

SAGER (N.)
1986, "Sublanguage : Linguistic Phenomenon, Computational Tool", p. 1-18, *in Analyzing Language in Restricted Domains : Sublanguage Description and Processing,* R. Grishman, R. Kitteredge, eds., New Jersey, Lawrence Erlbaum Associates, Hillsdale.

ZWEIGENBAUM (P.), consortium MENELAS
1994, "MENELAS : An Access System for Medical Records Using Natural Language", p. 117-120, *in Computer Methods and Programs in Biomedicine,* 45.