Numéro d'ordre | | | | | | | | |

UNIVERSITÉ PARIS X-NANTERRE

UFR LLPHI

THÈSE

PRÉSENTÉE PAR

ANTONIO BALVET

POUR OBTENIR LE GRADE DE

DOCTEUR EN SCIENCES DU LANGAGE

APPROCHES CATÉGORIQUES ET NON CATÉGORIQUES EN LINGUISTIQUE DES CORPUS SPÉCIALISÉS

APPLICATION À UN SYSTÈME DE FILTRAGE D'INFORMATION

Soutenue publiquement le 11 Décembre 2002, devant le jury

M.	Christian FLUHR	Rapporteur
M.	Benoît HABERT	Examinateur
M.	Bernard LAKS	Directeur
M.	Éric LAPORTE	Examinateur
M.	Célestin SEDOGBO	Examinateur
Mme	Antoinette RENOUF	Rapporteur

Remerciements

Je remercie tout d'abord Antoinette Renouf et Christian Fluhr d'avoir accepté la lourde charge de rapporteur. Leur regard sur ce travail m'a permis d'aborder des perspectives nouvelles, les questions qu'ils m'ont adressées m'ont incitées à clarifier certains points.

Je souhaite également adresser mes remerciements à l'ensemble des membres de la convention CIFRE régissant le présent travail de thèse : Bernard Laks, Célestin Sedogbo et Éric Laporte pour la qualité de leur encadrement au cours de cette thèse. Par ses remarques, tout au long de ce travail, Bernard Laks m'a permis d'apporter un éclairage épistémologique aux études sur corpus exposées ici. De son côté, Célestin Sedogbo, en m'accueillant au sein du laboratoire de recherche du groupe Thales, m'a permis de bénéficier d'un environnement humain et matériel propice à la recherche, tout en me confiant des responsabilités dans le cadre du projet CORAIL, qui m'ont permis de valider certaines des hypothèses centrales de cette thèse. Enfin, les troisième et quatrième chapitres de cette thèse doivent beaucoup à la rigueur avec laquelle Éric Laporte a relu et commenté mon travail.

Mes remerciements vont également à Max Silberztein et Dominique Dutoit, les auteurs des principaux outils mis en œuvre pour ce travail : respectivement Intex et le Dictionnaire Intégral. En effet, sans les conseils et l'aide que m'ont apportée Max Silberztein et Dominique Dutoit, cette thèse, dans ses aspects techniques liés au système CORAIL, n'aurait pas pu être menée à bien.

Les membres de l'UMR MoDyCo, notamment Marcel Cori, Benoît Habert, Sophie David, Ali Tifrit, et René Lavie ont toute ma gratitude, pour leur relecture attentive de mes travaux, leurs conseils et les discussions informelles qui m'ont permis d'affiner certains points développés ici.

Je souhaite également remercier Maurice Gross, Blandine Courtois et Christian Leclère, pour leur accueil au sein du LADL, la promptitude et la patience avec laquelle ils ont toujours répondu à mes questions, même les plus naïves.

Merci également à l'ensemble des membres du groupe DAS/HIT de Thales Research & Technologies pour leur aide au quotidien, ainsi que leurs critiques constructives tout au long de ma thèse, en particulier Thierry Poibeau, Frédéric Meunier et Nathalie Richardet. Merci également à Olivier Grisvard, Rodrigo Reyes et Pascal Bisson d'avoir bien voulu

partager leurs compétences, sans oublier Claire Laudy, Bénédicte Goujon, David Faure et Camal Tazine.

Une thèse est faite de chemins détournés, de voies qu'on abandonne en se jurant de repasser par là, plus tard ... Merci à Alain Polguère et Sylvain Kahane d'avoir bien voulu éclairer ces chemins de traverse.

L'ensemble du point de vue adopté dans cette thèse doit beaucoup aux discussions informelles avec Danièle Dubois et Sophie David, qui m'ont fait découvrir une vision non catégorique des problèmes linguistiques ; qu'elles en soient remerciées. Je souhaite également exprimer ma gratitude envers Karine Baschung, qui m'a toujours encouragé au cours de mon parcours universitaire et professionnel.

Enfin, cette partie ne serait pas complète sans la mention des personnes avec qui j'ai partagé interrogations, astuces et savoir-faire au sujet de Intex : principalement Cédrick Fairon, Anne Dister, Nathalie Friburger, Sébastien Paumier, Matthieu Constant, Jean Senellart et Elisabeth Ranchod.

Table des Matières

REMERCIEMENTS 2

TABLE DES MATIÈRES 4

TABLE DES FIGURES 10

TABLE DES EXEMPLES 11

TABLE DES FORMULES 12

GLOSSAIRE 13

LISTE DES SIGLES ET ABRÉVIATIONS 17

INTRODUCTION 19

CHAPITRE 1 POUR UNE LINGUISTIQUE DES CORPUS 23

1.1. LINGUISTIQUE STRUCTURALE ET DISTRIBUTIONNALISME 24

1.1.1. La recherche d'une démarche scientifique 24

1.1.1.1. Le monde à connaître 25

1.1.1.2. L'apport saussurien 26

1.1.1.3. Bloomfield, la science du langage 28

1.1.2. Classification et linguistique structurale 31

1.1.2.1. La linguistique comme entreprise catégorisante 31

1.1.2.2. Le modèle classique de la catégorisation 32

1.1.2.3. Influences du modèle classique sur une science du langage 36

1.1.3. Quelques notions fondamentales 39

1.1.3.1. Unité 39

1.1.3.2. Système 40

1.1.3.3. Signe 41

1.2. DU DISCONTINU DANS LE DISTRIBUTIONNALISME 42

1.2.1. Le distributionnalisme de Harris, un processus de découverte 42

1.2.1.1. La primauté des observables 42

1.2.1.2. Notion de distribution 43

1.2.1.3. Notion d'unité linguistique 44

1.2.2. Le distributionnalisme catégorique comme théorie linguistique 44

1.2.2.1. Un modèle de la Langue ... 44
1.2.2.2. L'objection chomskyenne au processus de substitution ... 46
1.2.2.3. Adéquation descriptive ... 49
1.2.2.4. Adéquation prédictive ... 51
1.2.2.5. Adéquation explicative ... 54
1.3. DISTRIBUTIONNALISME ET PROBABILITÉS ... 56
1.3.1. Herdan, le glissement vers un distributionnalisme probabiliste ... 57
1.3.1.1. Motivations linguistiques pour une approche probabiliste ... 57
1.3.1.2. Une vision quantitative de l'opposition Langue/Parole ... 58
1.3.1.3. Une théorie linguistique non grammaticale ... 60
1.3.2. Un changement de paradigme ... 61
1.3.2.1. Du catégorique au probable ... 62
1.3.2.2. Vers une théorie non catégorique et non logique ... 62
1.3.3. Le distributionnalisme probabiliste comme théorie linguistique ... 64
1.3.3.1. Théorème de Gold et apprentissage à partir d'exemples positifs ... 64
1.3.3.2. L'argument de la Pauvreté du Stimulus ... 66
1.3.3.3. Grammaticalité et probabilités ... 67
1.3.4. Critères d'adéquation d'un modèle probabiliste des faits langagiers ... 68
1.3.4.1. Adéquation descriptive ... 68
1.3.4.2. Adéquation prédictive ... 69
1.3.4.3. Adéquation explicative ... 70
1.4. CONCLUSION ... 71
CHAPITRE 2 DÉTECTION D'UNITÉS LINGUISTIQUES ET THÉMATIQUES POUR LA RECHERCHE D'INFORMATION ... 77
2.1. LA RECHERCHE D'INFORMATION ... 80
2.1.1. Notion d'information ... 81
2.1.1.1. Définition quantitative ... 82
2.1.1.2. Définition fonctionnelle ... 84
2.1.2. Les marqueurs thématiques en Recherche d'Information ... 87
2.1.2.1. Indexation manuelle et marqueurs thématiques ... 88
2.1.2.2. La variation dans l'indexation humaine ... 90
2.1.2.3. Indexation automatique et sélection de descripteurs de documents ... 92

2.1.3. Limites des approches basées sur des descripteurs en Recherche d'Information..... 95

2.1.3.1. L'approche « sac de mots » 96

2.1.3.2. Pertinence d'une base de descripteurs figés 98

2.1.3.3. Prise en compte du point de vue des utilisateurs 98

2.1.4. Recherche d'information basée sur des unités lexicales complexes 100

2.1.4.1. Analyses linguistiques automatisées et Recherche d'Information, une difficile intégration .. 101

2.1.4.2. Un retour à l' « empirisme » ? 105

2.2. EXTRACTION DE MARQUEURS THÉMATIQUES LINGUISTIQUES PAR ANALYSE DISTRIBUTIONNELLE 107

2.2.1. Analyse distributionnelle discontinue des corpus spécialisés 108

2.2.1.1. Élaboration d'une grammaire d'un domaine de spécialité 108

2.2.1.2. Extraction terminologique 110

2.2.1.3. Extraction d'information à partir de schémas conceptuels 111

2.2.1.4. Analyse thématique automatique fondée sur une ontologie sémantique 116

2.2.1.5. LIZARD, un assistant linguistique pour l'extraction de signatures thématiques 118

2.2.2. Ressources linguistiques issues d'une analyse classique 123

2.2.2.1. Thesauri et ontologie(s) 123

2.2.2.2. Une base de signatures thématiques sous la forme d'une table du lexique-grammaire 124

2.2.3. Distributionnalisme probabiliste pour la découverte de signatures thématiques : détection de collocations 128

2.2.3.1. Définition 129

2.2.3.2. Quelques techniques d'extraction de collocations 129

2.2.3.3. Transformation d'un corpus en n-grammes 131

2.2.3.4. Quelques résultats d'une fouille de corpus spécialisé 136

2.2.4. Ressources linguistiques issues d'une analyse probabiliste 139

2.2.4.1. Des bases de collocations pour la recherche d'information 140

2.2.4.2. Des collocations aux grammaires locales probabilistes 140

2.3. CONCLUSION 141

CHAPITRE 3 LE FILTRAGE D'INFORMATION 145

3.1. APERÇU HISTORIQUE DE LA NOTION DE FILTRAGE D'INFORMATION 146

3.1.1. Naissance d'un concept : la veille économique 146

3.1.1.1. Les *Business Intelligence Systems* 147

3.1.1.2. De la *SDNI* à la *SDI* 147

3.1.2. TREC et le filtrage d'information ... 148

3.1.2.1. Une conférence d'évaluation internationale ... 148

3.1.2.2. Des débuts hésitants ... 149

3.1.2.3. Une stabilisation tardive ... 150

3.2. APPROCHES POUR LE FILTRAGE D'INFORMATION ... 151

3.2.1. « Filtrage d'information » basé sur un moteur de recherche et d'indexation ... 151

3.2.1.1. Principes d'indexation automatique ... 151

3.2.1.2. PRISE, SMART et dérivés ... 153

3.2.2. Filtrage d'information par reconnaissance de mots-clés ... 154

3.2.2.1. Principe des expressions rationnelles ... 154

3.2.2.2. SIFT et Infoscope, deux systèmes fondateurs ... 155

3.2.3. Filtrage d'information par reconnaissance d'expressions typiques d'un domaine ... 157

3.2.3.1. Notion de signature thématique ... 157

3.2.3.2. Des unités lexicales complexes comme descripteurs ... 158

3.3. PROBLÈMES D'ÉVALUATION DES SYSTÈMES DE FILTRAGE D'INFORMATION ... 159

3.3.1. Quelques métriques de la recherche d'information ... 159

3.3.1.1. Précision et Rappel ... 160

3.3.1.2. F-mesure, P&R ... 160

3.3.2. Les métriques TREC pour le filtrage d'information ... 161

3.3.2.1. Utilité ... 161

3.3.2.2. TREC-5, une remise en cause du protocole d'évaluation ... 163

3.3.2.3. Association de l'utilité et d'autres mesures ... 164

3.3.2.4. Fonctions linéaires / non linéaires d'utilité et métriques associées ... 167

3.3.2.5. Métriques orientées vers la précision ... 168

3.4. PROBLÈMES DE MODÉLISATION D'UNE TÂCHE COMPLEXE : LE FILTRAGE D'INFORMATION ... 171

3.4.1. Problèmes de constitution d'une référence ... 172

3.4.1.1. Représentativité quantitative/qualitative des corpus ... 172

3.4.1.2. Des données observables : le vocabulaire spécialisé ... 174

3.4.2. Le filtrage d'information, une tâche complexe ... 175

3.4.2.1. Subjectivité ou expérience ? ... 175

3.4.2.2. Filtrage d'information et catégorisation ... 177

3.4.2.3. Décision de sélection binaire et satisfaction de contraintes ... 178

3.5. CONCLUSION 180

CHAPITRE 4 FILTRAGE D'INFORMATION PAR SIGNATURES THÉMATIQUES, MISE EN ŒUVRE EN MILIEU INDUSTRIEL 183

4.1. LE SYSTÈME CORAIL 183

4.1.1. Une plate forme industrielle de gestion des documents électroniques : PRIAM 184

4.1.1.1. Architecture fonctionnelle 184

4.1.1.2. Phases de veille, phases de crise 186

4.1.1.3. L'alliance filtrage/extraction d'information 187

4.1.2. TALN et recherche d'information par analyse locale 188

4.1.2.1. La recherche de la qualité en recherche d'information 188

4.1.2.2. Principes d'une analyse locale 190

4.1.2.3. La technique des cascades de transducteurs 191

4.1.3. CORAIL, un système de FI par cascades de transducteurs 192

4.1.3.1. Intex pour le filtrage d'information 192

4.1.3.2. Prétraitements 193

4.1.3.3. Décision de sélection 195

4.2. LIZARD, UN ASSISTANT LINGUISTIQUE POUR LA DÉCOUVERTE DE SIGNATURES THÉMATIQUES 198

4.2.1. Motivation 199

4.2.1.1. Automatiser l'analyse distributionnelle des corpus 199

4.2.1.2. Harmoniser et centraliser les ressources lexicales 199

4.2.2. Fonctionnalités principales 200

4.2.2.1. Une plate forme multi-agents distribuée 200

4.2.2.2. Extraction de formes schématiques 203

4.2.2.3. Passage de formes schématiques à des schémas de sous-catégorisation 206

4.2.2.4. Génération de bases de données lexicales 207

4.2.3. Une base de données lexicales pour la recherche d'information 209

4.3. MESURE DES PERFORMANCES DU SYSTÈME CORAIL 210

4.3.1. Un corpus professionnel 211

4.3.1.1. Un corpus financier 211

4.3.1.2. Quelques éléments stylistiques 211

4.3.1.3. Structuration en thèmes 213

4.3.2. Mesure des performances ... *215*

4.3.2.1. Protocole d'évaluation quantitative ... 215

4.3.2.2. Indicateurs de performance ... 216

4.3.2.3. Discussion des résultats ... 218

4.3.3. Questions d'utilisabilité ... *221*

4.3.3.1. Ébauche d'une évaluation ergonomique ... 221

4.3.3.2. Quelques résultats ... 223

4.4. CONCLUSION ... 225

CHAPITRE 5 CONCLUSION ET PERSPECTIVES ... 228

5.1. UN CADRE POUR UNE LINGUISTIQUE DES CORPUS ... 228

5.2. LINGUISTIQUE DE CORPUS ET RECHERCHE D'INFORMATION ... 230

5.3. LINGUISTIQUE ET CATÉGORIES ... 232

RÉFÉRENCES BIBLIOGRAPHIQUES ... 235

ANNEXE I : LE SYSTÈME CORAIL ... 253

INTERFACE D'ÉDITION DE FILTRES EN MODE CLIENT-SERVEUR (APPLET JAVA) ... 254

TABLE DES CAPTURES D'ÉCRAN DU SYSTÈME CORAIL ... 254

MANUEL UTILISATEUR DU MOTEUR DE FILTRAGE EXPÉRIMENTAL CORAIL ... 269

INTRODUCTION ... 269

TABLE DES FIGURES DU MANUEL D'UTILISATEUR ... 291

ÉVALUATION ERGONOMIQUE ... 292

GRAMMAIRES LOCALES UTILISÉES POUR L'ÉVALUATION ERGONOMIQUE ... 293

TABLE DES GRAMMAIRES LOCALES UTILISÉES POUR L'ÉVALUATION ERGONOMIQUE ... 293

ANNEXE II : GRAMMAIRES LOCALES POUR LE FILTRAGE D'INFORMATION ... 312

TABLE DES GRAMMAIRES LOCALES UTILISÉES PAR LE SYSTÈME CORAIL ... 312

TABLE DES AUTOMATES-PATRONS UTILISÉS PAR LE SYSTÈME CORAIL ... 314

TABLE DU LEXIQUE-GRAMMAIRE POUR LE THÈME 19 DU CORPUS FIRSTINVEST ... 314

Table des Figures

Figure 1 : un schéma conceptuel pour l'extraction d'information par le système Autoslog.. 113

Figure 2 : un extrait d'une base de données lexico-grammaticales du domaine financier..... 126

Figure 3 : automate-patron, générant les grammaires locales correspondant aux constructions figées acceptant la forme active 127

Figure 4 : mesures d'utilité pour trois scénarios d'évaluation 162

Figure 5 : décisions de sélection d'un système de filtrage d'information et mesures d'utilité correspondantes 165

Figure 6 : architecture fonctionnelle de la plate forme PRIAM 185

Figure 7 : conception classique des rapports entre activités de *push* et de *pull* 187

Figure 8 : PRIAM, une interdépendance entre *push* et *pull* 188

Figure 9 : interface utilisateur du système CORAIL, édition de grammaires locales pour le filtrage d'information 196

Figure 10 : visualisation des filtrats, acheminés par courrier électronique 198

Figure 11 : architecture de l'assistant linguistique LIZARD 201

Figure 12 : LIZARD, extraction de formes schématiques 205

Figure 13 : LIZARD, deuxième phase de généralisation 207

Figure 14 : LIZARD, génération de noyaux de bases de données lexicales 208

Figure 15 : base de signatures thématiques extraites d'un corpus financier 210

Figure 16 : tableau synthétique de la répartition en thèmes du corpus Firstinvest 214

Figure 17 : scores de rappel et de précision pour deux versions du système CORAIL, comparés à un système aléatoire 217

Table des exemples

Exemple 1 : extraction d'information sur une phrase décrivant les conséquences d'un attentat 111

Exemple 2 : étapes principales du prétraitement d'un corpus en vue d'en extraire des collocations..... 134

Exemple 3 : expansions associées à la tête « AOL » 136

Exemple 4 : scores d'entropie conditionnelle des expansions de la tête « AOL » 136

Exemple 5 : quelques 2grammes fortement cohésifs 138

Exemple 6 : les noms propres construits sur la tête « Jean » (extrait) 138

Exemple 7 : phases d'analyse d'un moteur de filtrage d'information générique..... 192

Table des formules

Formule 1 : t-test 130

Formule 2 : score d'information mutuelle..... 131

Formule 3 : cohésion lexicale..... 137

Formule 4 : information maximale..... 137

Formule 5 : test du Khi-2 218

Glossaire[1]

Amorces (*triggers*). Éléments lexicaux associés à des séquences (suites de caractères, mots) de façon régulière. Ainsi, par exemple, dans le domaine financier, la mention d'un montant peut être associé à une opération de vente d'une société.

Analyse de surface (*shallow parsing*). Analyse syntaxique minimale fondée sur des séquences d'étiquettes morpho-syntaxiques. À ce niveau, le système d'étiquetage n'a généralement pas accès aux informations de sous-catégorisation.

Analyse locale. Analyse syntaxique minimale, fondée sur la description de séquences inférieures à la phrase. Ce type d'analyse est souvent réservé aux domaines spécialisés, dans lesquels la phraséologie est plus fixe que dans la langue générale. Ainsi, par exemple, l'expression des dates, ou d'un montant pour une transaction, peuvent être décrits par une grammaire dite locale.

Apprentissage automatique. Paramétrage d'un système automatique par des données, à partir desquelles le système induit des règles. Dans le cas d'un apprentissage supervisé, les données à traiter sont accompagnées de la réponse désirée, au cours de la phase de paramétrage. Dans le cas d'un apprentissage non supervisé, les règles induites le sont à partir des seules données fournies au système.

Bruit. Indicateur de performance utilisé dans l'évaluation de systèmes de recherche d'information, proportion de documents non pertinents parmi les réponses des systèmes évalués.

Cascade d'automates ou de transducteurs à états finis (*finite state automata/transducers cascades*). Processus itératif d'analyse d'un texte, au cours duquel les éléments reconnus au cours d'une première analyse sont marqués, et utilisées par les analyses ultérieures.

---

[1] Les définitions du glossaire sont inspirées, pour partie, de celles données dans (Poibeau, 2002). Les termes anglais correspondants sont mentionnés entre parenthèses dans les cas où ils font partie des termes utilisés en français.

Corpus. Ensemble de productions linguistiques (ex. : discours transcrit, textes) formant un échantillon d'une langue donnée. Les corpus peuvent être construits de façon à être le plus représentatifs de la langue étudiée, ils peuvent être considérés sous deux points de vue : en tant qu'échantillons, ou bien comme extraits d'une langue. Dans les expérimentations, on distingue généralement entre corpus d'entraînement et corpus d'apprentissage. Le corpus d'entraînement sert au paramétrage des systèmes, le corpus d'apprentissage sert à tester la validité des règles induites au cours de l'apprentissage ; il est constitué de données inconnues du système évalué.

Désambiguïsation/levée d'ambiguïtés syntaxiques (*disambiguation*). Procédure visant à limiter le nombre d'hypothèses élaborées au cours d'une analyse syntaxique automatique.

Entité nommée (*named entity*). Ensemble des noms de personnes, d'entreprises, et de lieux présents dans un texte donné.

Étiquetage (*tagging*). Opération visant à assigner à chaque mot d'un texte une étiquette (ex. : une partie du discours).

Extraction d'information (*information extraction*). Activité de recherche d'information visant à la mise à jour automatique de bases de données relationnelles à partir de textes en langue naturelle. Ainsi, un système d'extraction d'information traitant des descriptions d'attentats (MUC-3, MUC-4), viserait à renseigner les champs « nombre de blessés », « localisation géographique », ou encore « type d'arme utilisé », d'un formulaire (*template*) fixe.

Filtrage d'information. Sélection et acheminement de documents tirés d'un flux d'information textuelle (ex. : fil de dépêches journalistiques), sur la base d'une comparaison binaire (correspondance/non correspondance) entre le profil informatif de chaque document et celui du besoin en information exprimé par un ensemble d'utilisateurs. En filtrage d'information, seuls les documents pertinents sont acheminés vers les utilisateurs.

Filtre. Dans le cadre d'un système de filtrage d'information, sous-éléments d'un profil d'utilisateur. Un filtre peut être constitué par une séquence d'expressions à rechercher dans les documents, ou une conjonction/disjonction/négation de ces expressions (opérateurs booléens).

Grammaire locale (*local grammar*). Grammaire généralement limitée à l'analyse d'éléments dont la productivité syntaxique est limitée. Ainsi, l'expression des dates, en français, peut être analysée par une grammaire locale. Il est possible d'imbriquer ou d'associer des grammaires locales afin d'étendre le degré de localité.

MUC. Conférence internationale d'évaluation de systèmes de compréhension automatique de messages en langue naturelle, organisée principalement par le DARPA et le NIST. Cette conférence est essentiellement consacrée aux systèmes d'extraction d'information, elle a donné lieu à la validation des approches basées sur des cascades de transducteurs à états finis pour les applications en Recherche d'Information.

Opérateurs booléens. Opérateurs de la logique booléenne : disjonction (OU), conjonction (ET) et négation (NON) sont les opérateurs de base, permettant de générer l'ensemble des fonctions d'évaluation logique (implication etc…). OU et ET sont des opérateurs binaires, NON est un opérateur un-aire.

Précision (*precision*). Taux de documents pertinents retrouvés par un système de recherche d'information, par rapport à l'effectif des réponses du système.

Profil d'utilisateur. Modélisation des besoins en information d'un utilisateur donné. Le profil peut être basé sur une explicitation des besoins, ou représenté par l'ensemble des documents consultés.

Rappel (*recall*). Taux de documents pertinents retrouvés par un système de recherche d'information par rapport à l'effectif de référence.

Recherche d'information (*information retrieval*). Activité visant à (re)trouver et présenter l'information pertinente à chaque utilisateur des systèmes de recherche d'information. La recherche d'information peut être mise en œuvre de façon manuelle, semi-automatique (interactive), ou complètement automatique.

Routage d'information (*routing*). Sélection et acheminement de documents tirés d'un flux d'information textuelle (ex. : fil de dépêches journalistiques). L'ensemble des documents sont évalués, en termes de pertinence, par rapport à un besoin en information donné. En routage d'information, l'ensemble des documents traités sont ordonnés en fonction de leur score de pertinence et acheminés vers les utilisateurs.

Silence. Indicateur de performance utilisé dans l'évaluation de systèmes de recherche d'information, proportion de documents pertinents non trouvés parmi les réponses des systèmes évalués.

Transducteur à états finis (*finite states transducer*). Graphe représentant un ensemble de séquences (ex. : caractères, mots) en entrée, et qui leur associe des séquences produites en sortie. Les transducteurs peuvent être utilisés pour associer aux séquences reconnues des informations structurées : balises (HTML, XML), mots-clés.

TREC. Conférence internationale d'évaluation de systèmes de fouille de textes (*text retrieval*). Cette conférence reprend le fonctionnement de MUC, elle est consacrée à différentes activités de RI, de l'indexation des documents à l'interrogation vocale de bases de données, en passant par le filtrage d'information. Elle a donné lieu à la diffusion de variantes des moteurs d'indexation et de recherche PRISE et SMART pour l'ensemble des tâches de fouille de textes.

Liste des sigles et abréviations

ARPA. Advanced Research Projects Agency, autre appellation du DARPA.

AP. Associated Press, agence de presse diffusant des dépêches journalistiques en langue anglaise.

AFP. Agence France Presse, diffusant des dépêches journalistiques en langue française.

CORAIL. Composition de Requêtes assistée par Agents Intelligents Linguistiques, système de filtrage d'information à base d'analyse locale par application de cascades de transducteurs à états finis. Ce système a été mis en œuvre au laboratoire Thales Research & Technologies, par le département DAS-HIT (Department of Advance Software, Human Interaction Technologies), il a permis d'évaluer la faisabilité industrielle d'une approche linguistique pour le filtrage d'information.

DARPA. Department of Advanced Research Projects Agency, dependant du gouvernement fédéral américain.

FI. Filtrage d'Information, voir TREC.

LIZARD. LInguistic wiZARD, assistant linguistique pour l'élaboration de grammaires locales mises en œuvres dans le cadre de systèmes de recherche d'information basés sur des cascades de transducteurs à états finis.

MUC. Message Understanding Conference, conférence d'évaluation des systèmes de compréhension automatique de messages en langue naturelle, organisée principalement par le DARPA et le NIST.

NIST. National Institute for Standards and Technologies, institut national nord-américain des standards et technologies.

OT. Optimality Theory, théorie linguistique développée dans (Prince & McCarthy, 1993), définissant un cadre formel basé sur la notion de hiérarchie de contraintes universelles.

P/R. Précision/Rappel.

RI. Recherche d'Information.

SDI. Selective Dissemination of Information, diffusion sélective d'information.

SDNI. Selective Dissemination of New Information, diffusion sélective de la nouvelle information.

SIG. Special Interest Group, groupement d'intérêts ; dans le domaine de la recherche appliquée, les SIG ont une influence particulière aux États-unis.

TALN. Traitement Automatique des Langues Naturelles.

TREC. Text REtrieval Conference, conférence d'évaluation des systèmes de RI, organisée principalement par le DARPA et le NIST.

INTRODUCTION

Les études sur des données linguistiques observables et attestées, centrées sur la Parole[1], longtemps cantonnées au rang de simples outils descriptifs par les tenants d'une linguistique abstraite centrée sur la Langue[2], connaissent un regain d'intérêt depuis une dizaine d'années, tant au niveau national[3] qu'international.

Le domaine de l'ingénierie linguistique, de son côté, devant répondre de façon pragmatique à des besoins opérationnels par l'élaboration de systèmes (logiciels) automatiques d'analyse linguistique, a toujours favorisé les études sur corpus, considérées comme des échantillons, si possible représentatifs, des données linguistiques à traiter. Autrement dit, l'ingénierie linguistique, passant outre l'anathème chomskyen de l'étude de la Parole comme « chasse aux papillons », s'est toujours appuyée sur des données linguistiques attestées. Ce mouvement n'a fait que s'accentuer sous la pression d'une demande toujours plus importante de la part des utilisateurs finaux, allant dans le sens d'une meilleure couverture, d'une plus grande fiabilité des systèmes fournis, conjuguée à l'intérêt grandissant, en Europe notamment, pour des corpus multilingues, issus de pratiques effectives[4], en quantité suffisante pour le paramétrage desdits systèmes.

Ce constat amène les questions suivantes.

Quel statut ont, aujourd'hui, les études sur corpus ?

Sont-elles capables de répondre aux objections chomskyennes ?

---

[1] Au sens saussurien, c'est-à-dire des productions linguistiques présentant les marques d'une individualité.

[2] Autrement dit des productions présentant les marques d'un fonctionnement collectif.

[3] Voir (Habert B. *et al.*, 1997), consacré aux *linguistiques* de corpus.

[4] Par exemple, corpus alignés pour la traduction automatique, corpus oraux pour la reconnaissance de la parole, ou encore corpus professionnels pour la recherche d'information.

Autrement dit, réunissent-elles les conditions pour dépasser le domaine de l'empirie dans lequel elles étaient jusqu'ici confinées ?

Les études sur corpus sont-elles capables d'être autre chose que des sources de modèles opérationnels utilisables en ingénierie linguistique, autrement dit peuvent-elles fournir la base d'une théorie linguistique ?

Ces questions posent de façon détournée celle de la place d'une étude scientifique de la Parole, en tant qu'acte individuel, opposée de façon classique à la Langue et au Langage, dans la conception structuraliste saussurienne.

Pour répondre à ces interrogations, nous tenterons de voir dans quelle mesure les développements récents dans le domaine des approches linguistiques guidées par des données observables et attestées fournissent les conditions d'une approche scientifique des phénomènes relevant de la Parole.

Deux problématiques traverseront l'ensemble de notre exposé.

La première a trait aux études sur corpus, passage obligé dans la conception d'un système d'analyse linguistique automatisé, envisagées sous l'angle de leur statut, c'est-à-dire en tant que théorie linguistique. Nous tenterons, dans l'ensemble de notre exposé, de justifier une prise de position pour une linguistique des corpus, spécialisés notamment, dépassant le cadre purement descriptif et empirique pour atteindre le niveau d'une réelle réflexion théorique.

La deuxième problématique sous-tendant l'ensemble de notre présentation a trait à la question de la variation linguistique, tant en production qu'en compréhension. Cette variation, observable à tous les niveaux (lexical, stylistique, pragmatique, sémantique) dans les productions linguistiques, tant spontanées (conversation) que codifiées (rédaction d'une dépêche journalistique), représente un défi pour tout concepteur de système linguistique automatisé. Or, manifestement, elle est loin de constituer un frein à la communication humaine, il semblerait au contraire qu'elle en soit un prérequis.

Nous serons amené, par le biais du phénomène de la variation, à aborder des problèmes liés à la reconnaissance d'un invariant, une unité (linguistique), dans un flux mouvant. En d'autres termes nous serons amenés à aborder des problèmes d'analyse et de catégorisation. Ces problèmes seront étudiés tant dans le domaine linguistique que dans celui de l'activité de filtrage d'information. Nous tenterons essentiellement de montrer la nécessité de prendre en compte le caractère non catégorique et non logique des catégories construites par des locuteurs humains dans l'optique d'une tâche de classification de textes, dont le filtrage d'information constitue une spécialisation.

Nous nous pencherons, dans une première partie, sur le domaine le plus formalisé des études sur corpus : le distributionnalisme. Nous tenterons de définir le cadre épistémologique du courant initié par Harris dans les années 1950, ainsi que les limites, en termes de théorie linguistique, imposées par ce cadre originel. Nous verrons comment, par des approches refusant le postulat catégorique et logique communément admis, ces limites peuvent être dépassées. Cette première partie sera ainsi l'occasion d'examiner deux approches complémentaires des données linguistiques observables, dans une perspective tant théorique que pratique (la mise au point d'un système linguistique automatisé) :

- une approche classique, restant dans un cadre catégorique où les éléments linguistiques délimités en corpus entretiennent des relations régies par des principes logiques ;
- une approche non classique, abordant de front l'ensemble des phénomènes rejetés par les linguistiques abstraites (le générativisme notamment), à savoir des phénomènes ayant trait principalement à la variation dans les productions langagières.

Dans une deuxième partie, nous tenterons de mettre en œuvre les deux approches complémentaires ci-dessus sur un corpus spécialisé, issu d'une pratique effective de diffusion sélective d'information par des experts humains du domaine de la finance. Ce corpus spécialisé nous servira à paramétrer un système de filtrage d'information, présenté plus loin. Nous tenterons de montrer comment, tant par une approche catégorique classique que par une approche probabiliste linguistiquement motivée, il est possible d'extraire de ce corpus un

ensemble d'unités lexicales complexes, associées à des thèmes et des sous-thèmes du domaine financier : des signatures thématiques.

La troisième partie, consacrée à un sous-domaine de la recherche d'information, le filtrage d'information, nous permettra d'aborder la question de l'élaboration d'un système automatique de filtrage d'information, reposant sur des analyses partielles faisant appel aux signatures thématiques extraites des corpus spécialisés. Nous examinerons ainsi les résultats des conférences d'évaluation nord-américaines TREC (Text REtrieval Conference), notamment en ce qui concerne les procédures d'évaluation de systèmes de filtrage d'information. Cette partie sera, notamment, l'occasion d'une réflexion sur les problèmes rencontrés au cours des différentes éditions de TREC pour l'évaluation des systèmes de filtrage d'information.

La quatrième partie de notre exposé est consacrée à la présentation d'un prototype de système de filtrage d'information en temps contrôlé, CORAIL, développé dans un cadre industriel. Cette présentation sera l'occasion de statuer sur la faisabilité d'un système de filtrage d'information reposant sur des procédures d'analyse linguistique automatisées, visant à répondre à une demande de diffusion ciblée d'information. Au cours de cette présentation, nous tenterons d'évaluer l'apport d'une telle analyse automatisée, en nous basant aussi bien sur des données chiffrées que sur une réflexion plus qualitative.

Enfin, nous tenterons, dans une dernière partie, de mettre en perspective les problèmes abordés au cours de notre exposé, notamment au sujet du statut des études sur corpus, des rapports entre linguistique et recherche d'information, ainsi que des relations entre la linguistique et les catégories. Cette dernière partie sera également l'occasion d'esquisser des pistes de recherche sur les sujets abordés au cours de cette thèse.

CHAPITRE 1

Pour une linguistique des corpus

Quels sont les fondements, méthodologiques et conceptuels, d'une linguistique attachée aux observables linguistiques ?

D'un point de vue épistémologique, quel statut conférer aux données linguistiques attestées, produites dans un contexte particulier, à destination d'un public particulier, par rapport à des énoncés construits, produits par les chercheurs ?

La linguistique structurale, européenne dans un premier temps, américaine dans un deuxième temps, semble se caractériser par une orientation générale en faveur des productions linguistiques effectives. Cette centration sur les observables est partagée par le courant distributionnaliste, incarné par Harris. De leur côté, les recherches appliquées, dans le domaine de l'ingénierie linguistique, ont massivement recours à des données linguistiques effectives dans le but de paramétrer les systèmes élaborés, souvent grâce à des approches statistiques.

Quel lien peut-on tirer entre le domaine de la linguistique théorique et celui de la recherche appliquée ? Plus précisément, quel lien peut-on tirer entre les trois domaines suivants, ayant tous pour objet d'étude les productions linguistiques effectives : Recherche d'Information (RI), linguistique de corpus et ingénierie linguistique ?

Dans le cadre d'une application des principes de la linguistique de corpus au domaine de la RI, qui sera l'objet des chapitres suivants, quel est le statut des analyses visant à révéler, au sein de corpus spécialisés, des unités linguistiques particulières, associées de façon relativement stables à des thèmes : les signatures thématiques ? Nous avons évoqué le statut de ces analyses dans le domaine applicatif, comme tenant essentiellement du paramétrage. Du point de vue d'une théorie linguistique, à quelles conditions ces observations peuvent-elles acquérir un statut scientifique ?

Dans ce premier chapitre, consacré aux fondements méthodologiques et conceptuels d'une linguistique des corpus, nous tentons essentiellement de montrer que l'étude des distributions des éléments linguistiques, telle qu'exposée dans (Harris, 1951), constitue plus qu'une simple méthode de travail en vue de découvrir les unités d'une langue étudiée, ainsi que leurs propriétés. Le distributionnalisme est issu d'une vision fortement empreinte de pragmatisme dans l'étude du langage. De ce fait, cette approche des faits langagiers est, le plus souvent, perçue comme un ensemble de procédures, de méthodes plutôt que comme une véritable théorie scientifique sur le langage. Toutefois, à condition de dépasser le cadre structuraliste classique inspiré d'une conception logiciste des relations entre éléments linguistiques, ainsi que les travaux de Harris nous y invitent implicitement, il est possible de voir dans l'étude des distributions un véritable cadre méthodologique et conceptuel, dans l'optique de l'émergence d'une linguistique des corpus, constituée comme un domaine de recherches à part entière, non plus comme une simple méthode d'exploration.

Ce premier chapitre est l'occasion de préciser les notions sur lesquelles nous basons l'ensemble de notre travail ; dans un premier temps, nous évoquons les sources multiples à l'origine du courant distributionnaliste, puis nous revenons sur quelques notions fondamentales héritées du structuralisme classique saussurien. Dans un deuxième temps, nous nous centrons sur la méthode distributionnelle telle qu'exposée par Harris dans son ouvrage paru en 1951, ainsi qu'à ses évolutions, notamment le recours à des algorithmes statistiques tels qu'exposés dans les travaux précurseurs de Herdan. Ce faisant, nous tentons de voir comment le changement de paradigme que représente le passage d'une conception catégorique et logiciste du distributionnalisme à une conception probabiliste est susceptible de fournir les bases d'une théorie linguistique à part entière, guidée par les observables.

1.1. Linguistique structurale et distributionnalisme

1.1.1. La recherche d'une démarche scientifique

La linguistique structurale, dans laquelle nous voyons les germes du distributionnalisme, tente de concilier deux impératifs antagonistes, face aux phénomènes qu'elle cherche à organiser en une théorie scientifique : un mouvement d'abstraction,

nécessaire afin de se détacher de la contingence empirique (les faits dans leur seule matérialité), et un mouvement discriminant, s'attachant aux particularités les plus fines des phénomènes observés.

1.1.1.1.Le monde à connaître

Historiquement, nous voyons dans les principes exprimés dans la métaphysique aristotélicienne les conditions de l'émergence d'une démarche expérimentale, dans le cadre de la construction d'un appareil scientifique basé sur des phénomènes observables. Dans la conception aristotélicienne, en effet, le monde et sa structuration sont pensés comme accessibles à travers les propriétés des objets de ce monde, voire de leurs propriétés en tant que médiatisées par le langage. De plus, la science des « êtres en tant qu'êtres » est vue comme une « science recherchée[1] », donc en construction. Pour cette raison, nous qualifions la position aristotélicienne comme celle d'un monde à connaître, plutôt qu'à reconnaître[2], motivant l'observation des objets de ce monde, effectuée de la façon la plus méthodique possible, de manière à aboutir à une caractérisation de ces objets dans les termes de leur essence, c'est-à-dire les propriétés qui leur sont à la fois habituelles et nécessaires. La caractérisation de la position aristotélicienne sur le monde, comme origine de la démarche expérimentale en tant qu'instrument de la connaissance est l'objet du passage ci-dessous.

> Aristote fait une part à l'expérience, soit en tant que la sensation est pour la raison une manière d'exercer son pouvoir d'intuition, soit même en tant que la sensation a pour fonction de saisir le contingent. L'esprit expérimental est même si développé chez Aristote qu'il faut voir en lui le plus puissant des promoteurs de la science expérimentale chez les Anciens. C'est grâce à lui et à son école, qu'il y a eu dans l'Antiquité, en dehors

[1] « Nous négligeons d'ordinaire le fait que la description la plus communément donnée de la nouvelle discipline [la métaphysique aristotélicienne] est « la science recherchée ». À la différence de toutes les autres sciences, elle ne part pas d'un objet donné mais de la question de savoir si son objet existe », (Bourdeau, 2000, p. 3).

[2] Dans la conception Platonicienne, le monde et les objets qu'il contient ne sont qu'apparence. Platon recherche des Principes essentiels, non par l'observation du monde mais par l'exercice de la philosophie. Par ailleurs, cette philosophie est imprégnée d'une mythologie postulant l'accès au savoir comme une réminiscence d'un savoir perdu. Dans cette conception, le monde et les observables qu'il contient ne peuvent fournir la base d'aucune connaissance véritable.

> de l'astronomie, une certaine somme des connaissances sur les phénomènes naturels et quelque soupçon de la méthode propre aux sciences de la nature.
>
> (Hamelin, 1985, p.79)

Nous cherchons à montrer ici que le choix d'une visée abstrayante opposée à une visée discrétisante donne nécessairement deux positions antagonistes sur le monde, comme lieu d'observation des phénomènes, donc deux approches dans la construction d'une théorie scientifique de ces phénomènes, à savoir une approche rationnelle (abstraction) opposée à une approche qualifiée dans la tradition anglo-saxonne d'empirique[3] (discrétisation). La filiation que nous tentons d'établir ici entre linguistique structurale, distributionnalisme et métaphysique aristotélicienne est motivée par le fait que le cadre fourni par cette métaphysique est porteur de limitations intrinsèques en ce qui concerne les théories scientifiques qu'il permet de construire. Ces limitations sont explorées plus bas, notamment par le biais du postulat catégorique et logique.

1.1.1.2.L'apport saussurien

L'œuvre de Saussure, fondatrice de la linguistique comme étude des structures, apparaît essentiellement comme un mouvement vers les phénomènes langagiers, donc un mouvement vers les observables, pour l'étude desquels l'auteur définit un cadre conceptuel et méthodologique.

L'une des avancées théoriques de l'œuvre saussurienne a trait à la nécessaire abstraction par rapport aux données empiriques, évoquée plus haut : les phénomènes langagiers sont pris comme résultant essentiellement d'un compromis social. Cette notion de compromis est essentielle, en ce qu'elle fonde deux domaines d'observation linguistique :

- le domaine de la Parole, lieu des particularismes, domaine le plus descriptif, le plus proche des données observables ;

---

[3] Nous modulerons plus bas cet antagonisme : le terme « empirique » étant épistémologiquement marqué, comme synonyme de non science, par des auteurs tels que Comte, notamment : « 'Une stérile accumulation de faits incohérents' : c'est ainsi que Comte caractérise l'empirisme. La formulation d'hypothèses est donc préalable, l'observation et l'expérimentation, tout aussi nécessaires, venant cependant en conséquence. Il ne s'agit de rien de moins que de « réconcilier » les deux modes d'établissement de la vérité (rationnel, expérimental) », (Comte, 1996, p. 9).

- le domaine de la Langue, lieu des régularités, visant un maximum de cohérence dans les observations.

En posant ces deux domaines d'observation, la linguistique structurale saussurienne pose les conditions d'une réflexion dépassant le niveau empirique, descriptif. Elle pose également deux cadres méthodologiques : le premier ayant trait au recueil des données, le deuxième à leur interprétation.

La deuxième avancée que nous souhaitons souligner est celle ayant trait à la caractérisation des unités linguistiques, les signes, comme essentiellement et nécessairement arbitraires. On peut rattacher cette caractérisation à la prise en compte de la dimension sociale du langage humain, elle a pour conséquence de nier toute relation naturelle entre la face signifiante (la forme) et la face signifiée (le contenu conceptuel) des signes.

Le postulat de l'arbitraire des signes linguistiques implique que les éléments d'une langue donnée ne sont conçus qu'en ce qu'ils s'opposent à d'autres éléments, autrement dit ils ne possèdent pas de valeur intrinsèque[4] mais bien une valeur qui ne peut être que relative, résultant des relations d'opposition avec les autres éléments (ou système).

Outre les avancées méthodologiques et conceptuelles consignées dans le *Cours de Linguistique Générale* (*CLG*), fondant la linguistique comme une approche scientifique des phénomènes langagiers en synchronie et effectifs, l'œuvre saussurienne marque une étape primordiale en ce qu'elle prend position et fait acte de science en postulant des unités, abstraites par rapport à une réalité (ex. : un signal acoustique). Cette abstraction première, exprimée dans les termes saussuriens comme la discrétisation dynamique de la matière (phonique, par ex.) et de la pensée l'un par l'autre, constitue, à nos yeux, une évocation du recours nécessaire à une démarche catégorisante – en termes saussuriens, l'adoption d'un point de vue – dès l'étape de description.

Autrement dit, toute étude linguistique repose sur un effort visant à organiser le réel continu en un ensemble d'éléments discontinus, contenus dans des classes : « la langue ne se présente pas comme un ensemble de signes délimités d'avance, dont il suffirait d'étudier les

[4] La position saussurienne sur le langage est celle d'un objet complètement conventionnel, où aucun déterminisme naturel ne joue, puisque même les onomatopées sont culturelles. Cette position est à la base de celle d'arbitraire du signe, dont nous montrerons plus bas l'importance, en termes de scientificité de l'étude des observables linguistiques.

significations et l'agencement ; c'est une masse indistincte où l'attention et l'habitude peuvent seules nous faire trouver des éléments particuliers », (de Saussure, 1972, p.146). En cela, on peut voir une avancée primordiale en termes scientifiques : toute description linguistique dépasse le niveau empirique ; toute description est déjà une analyse.

Nous verrons plus bas que le processus d'élaboration de ces classes est fortement influencé par l'héritage aristotélicien, notamment le recours à la logique formelle comme instrument (*organon*) de science, qui a pour résultat des catégories (ex. : phonèmes, morphèmes, parties du discours) mutuellement exclusives et discontinues.

En tant que démarche scientifique basée sur des observables, le *CLG* présente une tension inévitable entre abstraction, nécessaire à l'établissement de classes d'éléments, et discrétisation, nécessaire à l'édification de relations d'opposition déterminant la valeur d'éléments particuliers, et du même coup l'ensemble du système d'une langue. Nous basant sur le postulat de cette tension, nous interprétons l'ensemble des analyses linguistiques saussuriennes comme issues d'un équilibre, d'une harmonisation de ces deux contraintes fondamentales. Le distributionnalisme, influencé par les principes structuralistes, nous apparaît, lui aussi, parcouru de cette tension entre abstraction et discrétisation, se répercutant jusque dans les analyses proposées par Harris.

1.1.1.3.Bloomfield, la science du langage

Le distributionnalisme de Harris, toute comme le structuralisme saussurien[5], se caractérise par une position fondant ouvertement l'étude des observables linguistiques comme démarche scientifique. Notamment, la réaffirmation par Harris du caractère fondamentalement arbitraire de la relation entre la face matérielle et la face immatérielle des signes, donne lieu à une démarche centrée uniquement sur les observables (la face signifiante) des éléments linguistiques, dans laquelle le sens de ces éléments n'intervient qu'en tant que critère distinctif. En effet, le distributionnalisme est marqué par l'héritage bloomfieldien, nourri de l'expérience acquise au cours des campagnes d'étude des langues indiennes nord-américaines, ainsi que des principes du behaviorisme, prédominant à cette époque aux États-unis.

[5] Structuralisme européen et américain doivent cependant être distingués, en ce que l'ouvrage de Saussure n'était que peu diffusé outre-Atlantique.

L'arbitraire, chez Harris, rejoint l'arbitraire saussurien en tant que fondement d'une étude scientifique des faits langagiers.

> The only body of data required for the whole analysis of language is the indication that certain sound sequences, out of some large sample, are utterances of the language (with normal acceptance, or less) while others are not, and that certain ones are repetitions of each other. Structural linguistics shows how these utterances can be characterized as a set of constructions on certain discrete elements. Mathematical linguistics shows that the characterization can be made in terms of other sets, defined by certain relations among these linguistic elements, and that *entities in the new set are arbitrary and are defined only by the relations among the new sets*[6].
>
> (Harris, 1968, p.1)

Cette exclusion des phénomènes sémantiques a été justifiée par le rejet du mentalisme, qui peut être associé à la diffusion du courant behavioriste outre-Atlantique, se superposant aux principes scientifiques existants : héritage aristotélicien et structuralisme bloomfieldien. La recherche d'une scientificité pour une linguistique pensée à cette époque comme essentiellement descriptive, est palpable dans le passage ci-dessous.

> It is widely recognized that forbidding complexities would attend any attempt to construct in one science a detailed description and investigation of all the regularities of a language. Cf. Rudolf Carnap, *Logical Syntax of Language* 8: "Direct analysis of (languages) must fail just as a physicist would be frustrated were he from the outset to attempt to relate his laws to natural things – trees, etc.. (He) relates his laws to the simplest of constructed forms – thin straight levers, punctiform mass etc." Linguists meet this problem differently than do Carnap and his school. Whereas the logicians have avoided the analysis of existing languages, linguists study them; but instead of taking parts of the actual speech occurrences as their elements, they set up very simple elements which are merely associated with features of speech occurrences.
>
> (Harris, 1951, p.16)

[6] Italiques ajoutés.

On voit à l'oeuvre dans ce passage, la même tension, évoquée plus haut pour le structuralisme saussurien, entre objets du monde étudiés dans une approche scientifique, et nécessaire abstraction par rapport aux « choses naturelles » (*natural things*). On voit également quel rôle jouent les corpus pour Harris, en tant qu'échantillons de Langue et non comme simple accumulation de faits de Parole : les lois, ou règles, que le linguiste cherche à établir doivent être mis en rapport avec des objets, non plus naturels, mais bien construits, théorisés. Ces objets représentent la même fonction que celle des modèles en physique, par exemple : ils constituent une version simplifiée d'un objet du monde, dont les paramètres sont contrôlés. En quelque sorte, ils entretiennent un rapport d'analogie avec l'objet du monde étudié, i.e. la Langue. Malgré les limites qu'il reconnaît aux études sur corpus[7], la position harrissienne n'est pas limitée à la simple description des langues étudiées : « when the linguist offers his results as a system representing the language as a whole, *he is predicting that the elements set up for his corpus will satisfy all other bits of talking in that language*[8] », (Harris, 1951, p. 17). Il est possible d'interpréter cette remarque de deux façons différentes : d'un point de vue limité à l'étude des phonèmes d'une langue, ou bien d'un point de vue plus large, adopté ici, étendant les principes décrits dans (Harris, 1951) à l'ensemble des domaines d'étude du langage.

Nous avons posé, dans la partie précédente, que les analyses de Harris étaient issues d'une équilibration entre deux contraintes opposées : l'abstraction par rapport aux données linguistiques visant à une cohérence maximale par la généralisation de régularités constatées sur des exemples particuliers, et la discrétisation par la prise en compte des propriétés les plus particulières des éléments étudiés, dans l'optique d'une recherche de complétude maximale, garante d'une adéquation descriptive du modèle en construction.

Nous insistons ici sur la notion d'équilibre, qui nous semble contenir en germe les deux approches distributionnelles possibles des faits de langue, développées plus bas : l'approche classique (catégorique et logique) et l'approche probabiliste. En effet, en matière d'équilibre, deux conceptions sont possibles : celle d'un équilibre statique, opposée à celle

[7] Voir plus bas.

[8] Italiques ajoutés.

d'un équilibre dynamique. Le premier suppose des objets du monde intrinsèquement équilibrés et stables, à tout le moins à l'échelle temporelle humaine. Le second voit dans l'état d'équilibre le résultat d'un processus dynamique, passant par la neutralisation de contraintes opposées (ex. : la force de gravitation opposée à la force de frottement). Ces deux visions de la notion d'équilibre, qui conditionnent deux conceptions de la notion de règle linguistique, ont souvent été illustrées par la métaphore du cristal opposé à celle de la flamme[9], qui a pour mérite de réconcilier deux positions opposées dans l'étude des faits langagiers : la première cherchant des loi et des règles (cristal), la seconde pensant le monde en probabilités, en termes de régularités (flamme).

1.1.2. Classification et linguistique structurale

Nous nous attachons ici aux liens étroits entre l'étude du langage, perçue comme une activité scientifique, et le modèle classique du processus de construction de catégories à des fins scientifiques, hérité du modèle aristotélicien. Nous tentons, dans un premier temps, de caractériser la linguistique comme étant essentiellement une démarche structurant le réel, autrement dit une démarche catégorisante, puis nous exposons, dans un second temps, les principes de cette démarche structurante, passant par la constitution d'un système de catégories, dont les frontières sont conçues comme étanches (position catégorique), déterminées par des principes logiques. Enfin, nous examinons l'influence du modèle classique de la catégorisation sur les théories linguistiques, élaborées dans ce cadre catégorique et logique.

1.1.2.1.La linguistique comme entreprise catégorisante

La position exprimée par Labov : « If linguistics can be said to be any one thing it is the study of categories: that is, the study of how language translates meaning into sounds through the categorization of reality into discrete units and sets of units », (Labov, 1973, p. 342), nous apparaît partagée par l'ensemble des disciplines scientifiques s'appuyant sur des observables (des objets du monde). La linguistique structurale partage avec ces sciences la préoccupation de classer les observables, de décomposer des phénomènes complexes en unités plus simples, et de rendre compte des relations entre les observables par une théorie, devant posséder nécessairement un pouvoir descriptif, explicatif et prédictif adéquat.

[9] Voir (Piattrelli-Palmerini, 1975).

Ainsi, l'objet premier de la phonologie, domaine d'application privilégié des principes structuraux, qui fonde cette activité comme démarche scientifique, est l'étude des phonèmes, considérés comme unités abstraites, caractérisées par des faisceaux de traits distinctifs (des propriétés) mutuellement exclusifs (ex. : sourde/sonore, avant/arrière), organisées en un système (i.e. le système phonologique d'une langue donnée). Suivant la démarche aristotélicienne, l'essentiel des questions phonologiques se résume à la question de la nature de l'objet étudié, par exemple : tel phénomène observé est il une instance d'une unité phonologique (phonème, syllabe) ou non ?

De même, en morphologie, en syntaxe, ainsi que dans l'ensemble des champs de recherche de la linguistique dite structurale, la question véritablement scientifique intervient à partir du moment où les observables sont abstraits de leurs caractéristiques les moins générales, autrement dit leurs accidents, regroupés en un réseau de relations d'opposition (un système). De la même façon qu'en phonologie, la question fondamentale de l'ensemble de la linguistique structurale est d'ordre métaphysique au sens aristotélicien : tel mot est-il décomposable en morphèmes (préfixe, suffixe) ou non, tel groupe de mot fonctionne-t-il comme un seul mot ou non, telle phrase est elle bien formée ou non, tel énoncé fait il sens ou non[10] ?

En tant qu'entreprise intéressée fondamentalement par l'établissement de classes d'éléments, nous qualifions l'étude des faits langagiers, menée dans le cadre structural, comme étant essentiellement une entreprise catégorisante, fondée sur la métaphysique aristotélicienne, structurée par les deux contraintes de ce modèle classique : loi de contradiction et loi du tiers exclu, d'où découlent l'ensemble des contraintes additionnelles, qui ont donné naissance aux différents courants issus de la souche structurale (voir *infra*).

1.1.2.2.Le modèle classique de la catégorisation

Le modèle classique de la catégorisation, c'est-à-dire, le processus qui permet d'établir des classes, dans la perspective d'une activité scientifique, d'y inclure ou d'en exclure des objets du monde (pris au sens large : objets matériels, conceptuels) afin d'aboutir à une vision cohérente de celui-ci, peut être synthétisé comme suit. Il repose principalement sur des

[10] De même, l'ensemble des questions scientifiques dans d'autres domaines, tels que la physique (classique), ou l'astronomie, est d'ordre métaphysique : par exemple, tel corps céleste est il une étoile ou non, tel élément (ex. : un électron) est il une onde ou une particule ?

contraintes de structuration de type logique, ainsi que sur la mise en œuvre de raisonnements logiques, pensés par Aristote en tant qu'instruments de science.

Ce modèle, exposé dans *La Métaphysique* d'Aristote, repose sur une distinction fondamentale opérée entre l'essence des objets du monde et leurs accidents. L'essence est considérée comme l'élément définitoire des choses, les accidents étant des propriétés incidentes (ni nécessaires ni habituelles). Le modèle classique repose donc sur la prise en compte de propriétés des objets du monde, ainsi que de leur caractère nécessaire ou non, suffisant ou non, en tant que définition de ces objets. La catégorisation, qui est souvent définie comme la faculté de percevoir du Même dans la diversité, peut donc être reformulée comme suit : la faculté de percevoir l'essence des choses plutôt que leurs accidents. Dans ce modèle, les catégories sont définies par des conjonctions (au sens logique) de conditions (ou propriétés) nécessaires et suffisantes (ou CNS).

Ce modèle est principalement structuré par deux contraintes :

- la loi de non contradiction, qui stipule qu'une chose ne peut pas à la fois être et ne pas être ;
- la loi du tiers exclu, qui stipule qu'une chose *doit* être ou ne pas être.

Ces deux contraintes peuvent servir de base à une description des propriétés des choses, intégrées à un modèle scientifique en construction, selon un principe binaire (+/-, vrai/faux, 0/1)[11]. Dans ce modèle, les catégories possèdent des frontières bien définies (loi de contradiction), par ailleurs, tous les membres d'une catégorie donnée sont perçus comme ayant le même statut : par exemple, dans la catégorie des mammifères, aucune gradation n'est envisageable dans ce modèle entre deux membres de la catégorie, tels qu'un chien et un ornithorynque[12].

[11] Le binarisme en linguistique peut aussi être vu comme une représentation optimale des propriétés des éléments décrits, voir à ce sujet (Herdan, 1962, p. 132).

[12] Des précautions semblent devoir être prises quant à cette affirmation, notamment dans le cadre de l'induction, chez Aristote, autrement dit la généralisation d'une loi à partir de quelques individus d'une classe, jugés les plus saillants : « l'induction est une condensation de l'expérience, analogue à celle qui s'opère machinalement quand des sensations se groupent autour de l'une d'entre elles *qui est plus intense* » (Robin, 1973, p. 291), italiques ajoutés.

Dans la vision classique, le monde est nécessairement structuré de façon taxinomique, les classes d'objets hiérarchisées selon que leur essence est plus ou moins générique (les traits sont plus ou moins partagés par l'ensemble des objets du monde). Ainsi, au bas de l'arbre taxinomique, se trouvent les individus, ou éléments les moins génériques (dont l'essence est la moins partagée). Le haut de la hiérarchie est dominé par les genres les plus génériques, auxquels Aristote impose la contrainte de disposer d'un contenu, ce qui évite l'inclusion à la hiérarchie des classes les principes platoniciens tels que l'Un et l'Être, tellement génériques qu'ils peuvent s'appliquer à tout.

On peut voir dans la construction d'un modèle scientifique guidé par les observables la reprise de la démarche aristotélicienne, conférant à la logique formelle le rôle d'instrument, de méthode. Il convient, toutefois, de distinguer le processus de construction de classes à partir des propriétés des objets du monde, des classes proposées par Aristote proprement dites. En effet, à l'instar des modèles scientifiques forgés au cours de l'Antiquité, le modèle en dix classes, tel que l'expose Aristote, a fait l'objet d'une remise en cause justifiée, au cours du développement des sciences de la nature. Dans le domaine de la linguistique structurale, Benveniste, notamment, a dénoncé l'influence du système conceptuel de la langue grecque sur les classes proposées par Aristote[13] (voir ci-dessous), à l'occasion d'une réflexion sur l'interdépendance entre langue et pensée.

> Aristote pose (…) la totalité des prédicats que l'on peut affirmer de l'être, et il vise à définir le statut logique de chacun d'eux. Or, il nous semble (…) que ces distinctions sont d'abord des catégories de langue, et qu'en fait Aristote, raisonnant d'une manière absolue, retrouve simplement certaines des catégories fondamentales de la langue dans laquelle il pense.
>
> (Benveniste, 1966, p. 66)

[13] Pour Aristote, une catégorie représente tout ce qu'il est possible d'attribuer à un objet du monde, autrement dit toutes les prédications médiatisées par le langage. Ainsi, Aristote est amené à proposer un système à dix catégories (ex. : substance, quantité, qualité, lieu, temps, possession etc…), qui constituent de toute évidence un inventaire des prédications possibles dans sa langue.

Cette influence de la langue grecque sur le système de catégories proposé par Aristote remet en cause la validité de ce système en-dehors du contexte dans lequel il a été élaboré. Ceci étant, il n'en reste pas moins que le mode de constitution d'une hiérarchie conceptuelle des objets du monde, obtenue par l'observation de leurs propriétés matérielles, structurée par les deux contraintes : loi de contradiction et loi du tiers exclu, nous apparaît fondamentalement inchangé. Ce mode de constitution de classes d'objets semble avoir été repris sans discussion dans le cadre de l'étude des faits langagiers, en dépit de l'évolution historique de la démarche expérimentale en science, essentiellement suite à la transition cartésienne. Bien que nous souscrivions à la remise en cause des modèles aristotéliciens, ne sauvant les phénomènes qu'imparfaitement[14], nous soulignons le fait que la logique aristotélicienne ne semble pas remise en cause en tant qu'instrument (*organon*) de science. La structuration du monde, à laquelle tendent toutes les sciences expérimentales, reste sous-tendue par les deux contraintes fondamentales de la logique et des catégories aristotéliciennes : le principe de non contradiction et celui du tiers exclu.

Ainsi, Auroux attribue à cette fondation logique l'émergence d'une position catégorique (reposant sur des catégories fondées sur les lois citées plus haut) et logiciste sur les phénomènes langagiers, reprise de façon plus ou moins explicite par l'ensemble de la linguistique structurale : « C'est dans l'œuvre logique d'Aristote que trouve son point de départ la théorie des parties du discours qui formera le cœur de la tradition grammaticale occidentale », (Auroux 1994, p. 34).

Ce fondement catégorique et logique des théories linguistiques s'étend au domaine de l'étude des distributions des éléments linguistiques.

(Auroux, 1994, p. 175)

> Derrière la théorie des parties du discours, il faut reconnaître quelque chose qui est la propriété essentielle du langage humain et qu'on peut énoncer comme étant sa nature catégorielle : une expression linguistique ne correspond pas simplement à la concaténation d'unités indifférenciées, c'est-à-dire que le langage humain n'est pas simplement un monoïde libre (…). Les mots doivent être catégorisés et leurs possibilités

[14] Notamment dans le domaine de l'astronomie, dans lequel les observations de Galillée ont permis de remettre en cause les modèles aristotéliciens.

> d'association dépendent de leur appartenance aux diverses catégories. Il s'agit là d'une découverte essentielle (on peut l'attribuer à Platon et voir en Aristote son premier théoricien) pour l'histoire scientifique de l'humanité.

Le conditionnement souligné par Auroux entre « possibilités d'association » (axe syntagmatique) et « appartenance aux diverses catégories » (axe paradigmatique) constitue, en effet, la base des études distributionnelles.

1.1.2.3.Influences du modèle classique sur une science du langage

Nous avons posé la catégorisation comme question centrale de la linguistique. Il en découle naturellement que le modèle de la catégorisation sous-jacent aux recherches linguistiques revêt une importance capitale. L'adoption de l'approche classique de la catégorisation est considérée par certains auteurs, notamment les linguistes cognitivistes, comme la condition du développement d'une linguistique théorique. Pour ces auteurs, sous l'impulsion de la phonologie, l'approche aristotélicienne aurait été étendue à l'ensemble des domaines de la linguistique, et se serait également enrichie de contraintes supplémentaires.

Taylor, dans son ouvrage de linguistique cognitive (Taylor, 1989), attribue ainsi aux phonologues fonctionnalistes (Troubetzkoy, Jakobson, Martinet) l'introduction de la notion de primitive, caractérisant les traits phonologiques appelés à être formalisés en un système de traits binaires. Taylor voit donc dans la notion de primitive le fondement d'une linguistique abstraite, autonome (indépendante des phénomènes cognitifs) et modulaire, par l'extension de la notion de primitive à l'ensemble des éléments linguistiques (ex. : sèmes). Cette extension est vue comme la condition de l'émergence d'une linguistique théorique et mentaliste reposant sur le postulat de l'innéité de la faculté de langage, c'est-à-dire son fondement génétique, telle que formalisée par Chomsky, Fodor et autres auteurs du courant générativiste-transformationnel.

La linguistique d'essence structurale s'est ainsi constituée comme une discipline cherchant les fondements essentiels des objets linguistiques, et a repris les postulats du modèle aristotélicien :

1. les catégories ont des frontières bien définies ;
2. les éléments linguistiques sont régis par les contraintes de la loi de contradiction et de la loi du tiers exclu ;

3. les propriétés distinctives des éléments peuvent être exprimées sous une forme binaire, dans une optique de principe optimal de description.

L'effet de cette adhésion au modèle classique nécessaire dans une perspective objectivante a principalement été, selon des auteurs tels que Taylor, de constituer la linguistique comme science (ou à tout le moins de démarche scientifique), en sortant, notamment, du rôle quasi-exclusivement descriptif qui lui était dévolu jusque là : en posant la question de l'acquisition de la faculté de langage, le générationalisme impose à tout modèle linguistique de disposer de moyens de prédiction et d'explication des phénomènes étudiés. Toutefois, l'adhésion sans condition au modèle aristotélicien des catégories s'est également traduite par l'introduction d'une visée logiciste[15], discrétisante, en contradiction avec des phénomènes dont le caractère holiste et flou a été souligné par des auteurs tels que Wittgenstein. Des observations et des expériences menées par des linguistes tels que Labov, Lakoff et Langacker, des psycholinguistes tels que Rosch, notamment, ou encore des anthropologues tels que Levi-Strauss, ont par ailleurs révélé l'existence de processus de catégorisation déviants, par rapport au modèle aristotélicien classique fondé sur le principe du tiers exclu, tant dans des cultures non occidentales[16] que dans les cultures nourries de l'héritage philosophique classique[17].

Ainsi, l'adoption d'une démarche logiciste dans la construction de catégories d'éléments linguistiques, mises en œuvre dans le cadre de la construction d'une théorie scientifique, bien qu'historiquement nécessaire pour la constitution d'une science du langage,

---

[15] En opposition d'ailleurs, avec le positivisme de Comte.

[16] Par exemple, les Dyirbal, un groupe d'aborigènes australiens, et leur système de catégorisation traditionnel tripartite décrit dans (Lakoff, 1987), structuré par un principe de ressemblance d'aire de famille plutôt que par les lois de non contradiction et du tiers exclu.

[17] De son côté, (Labov, 1973) relate des expériences de dénomination d'items familiers, dans des cultures occidentales (i.e. nord-américaine), tels que des tasses (*cup*) et des bols (*bowl*), visant à mettre en évidence les principes de catégorisation de ces objets, en faisant varier certaines de leurs propriétés (taille, circonférence, présence d'anse). Ces expériences ont essentiellement révélé des principes de catégorisation non logiques et non catégoriques, en fonction d'une distance par rapport à un exemplaire jugé le plus représentatif (prototype). Le constat d'un tel gradient d'appartenance catégorielle remet en cause la validité de la conception aristotélicienne des catégories.

semble être également source de difficultés. Ces difficultés, soulignées par les linguistes cognitivistes américains ainsi que par l'ensemble des linguistes « de terrain » nord-américains (sociolinguistes notamment)[18], ont également été remarquées par des auteurs tels que Fuchs, dans le domaine de l'ingénierie linguistique, et Auroux, que nous citons ci-après et se traduisent essentiellement par le problème de la construction d'une théorie monocatégoriale *versus* polycatégoriale[19].

> Selon les auteurs, les [unités segmentant la chaîne parlée] ne doivent appartenir qu'à une seule catégorie (monocatégorisation), soit peuvent relever de plusieurs catégories (polycatégorisation). comme il arrive que, dans des contextes différents, une même forme manifeste des propriétés catégorielles différentes, pour sauver la monocatégorisation, ceux qui la soutiennent, ont développé deux stratégies théoriques: i) l' ellipse qui permet de conserver l'unicité catégorielle (/un savant à un [homme] savant/) ; ii) l' homonymie qui assure que deux formes appartenant à des catégories différentes ne sont pas la même entité linguistique (fr. /que/ pronom relatif et fr. /que/ conjonction de coordination).
>
> (Auroux, 1994, p. 154)

Le recours à l'ellipse, autrement dit l'introduction d'opérations invisibles, ainsi que l'homonymie complexifient les théories linguistiques construites dans un cadre monocatégorial, en ce que ces deux opérations impliquent nécessairement des choix de catégorisation, puisqu'il s'agit de trancher, l'appartenance d'un élément à une classe donnée. Confrontés aux phénomènes mentionnés par Auroux, les théories monocatégoriales doivent supposer, en plus du domaine des observables, un domaine non directement observable[20].

---

[18] Ainsi, (Manning, 2002), dans sa revue de la tradition catégorique et logiciste en linguistique formelle, attribue à (Sapir, 1921, p. 38) le constat que « all [categorical] grammars leak ».

[19] Où les éléments peuvent appartenir à une *versus* plusieurs catégories différentes.

[20] La distinction entre le plan des formes de surface et celui des formes profondes du générationalisme, par exemple, peut être compris comme la manifestation du processus décrit par Auroux. Le générativisme se caractérise par une position monocatégoriale, à tous les niveaux, notamment en ce qui concerne la grammaticalité.

> Le recours à la traditionnelle classification en parties du discours (noms, adjectifs, verbes, adverbes, articles, pronoms, conjonctions, prépositions, interjections) pose un certain nombre de problèmes dont nous allons donner quelques exemples. (…)
>
> Polycatégorie par « dérivation impropre » : rouge, juste, informatique, linguistique sont à la fois noms et adjectifs ; rire, pouvoir sont à la fois noms et infinitif. ; clair, fort, juste sont à la fois adjectifs et adverbes. Faut-il créer les catégories « adjectif-nom », « adjectif-adverbe » ? (…)
>
> Aucune classification réellement satisfaisante ne s'est imposée : on se heurte au problème d'un continuum rebelle à toute classification rigide.
>
> (Fuchs 1993, p. 91)

L'extrait cité ci-dessus montre une autre conséquence de l'adoption d'un point de vue logiciste et catégorique (et monocatégorial) dans la construction des classes linguistiques (i.e. les parties du discours) pour des applications en ingénierie linguistique, qui poussent l'auteur à poser la question de la création de nouvelles classes pour sauver les phénomènes[21], sachant que toute nouvelle classe remet nécessairement en cause l'équilibre de l'ensemble du système construit jusque là.

1.1.3. Quelques notions fondamentales

1.1.3.1. Unité

Historiquement, l'émergence d'une linguistique des structures, ou systèmes, est liée à la prise en compte du matériau sonore du langage. Ce matériau pose un défi aux études linguistiques : là où l'écrit, qui est déjà une formalisation et une normalisation de l'oral, propose des unités évidentes[22] (ex. : des mots typographiques), l'oral n'est que variation. La naissance de la phonologie peut être vue comme celle d'une approche scientifique du langage, qui amène à se poser la question véritablement linguistique concernant l'étude de la langue orale : comment distinguer dans un flux ininterrompu (des accidents), un signal continu, les

[21] Dans le passage cité, il s'agit de la polycatégorialité par dérivation impropre.

[22] Du moins pour les langues possédant une tradition ancienne d'édition et de diffusion d'écrits fondamentaux : législatifs, religieux, ou encore philosophiques.

unités (l'essence) de la Langue ? Cette question revient à poser le problème de la reconnaissance du même - une unité donnée, telle qu'un phonème - dans l'autre : un signal acoustique variable et difficile à segmenter, c'est-à-dire, le problème de la discrétisation/catégorisation du réel.

En posant la notion d'unité, la linguistique pose nécessairement celle de la représentation mentale de ces unités, relativement stable et qui permet de guider la reconnaissance des phonèmes, par exemple. La phonologie pose, dans le même mouvement, la nécessaire abstraction par rapport au matériau linguistique premier, qui, ménagée en degrés, fournira l'ensemble des unités linguistiques : morphèmes, phrases, sèmes, et des angles d'approche de la Langue correspondants. Ce faisant, elle pose nécessairement les deux plans introduits par Saussure : celui de la Parole et celui de la Langue.

La notion d'unité, dans le cadre de la construction d'un modèle de la Langue, est donc fondamentale en ce qu'elle est nécessairement une construction, un objet mental abstrait. Ces objets sont nécessairement plus ou moins découplés du matériau linguistique qu'ils visent à organiser. Le recours à l'abstraction semble partagé par l'ensemble des auteurs que nous avons évoqués jusqu'ici : d'Aristote à Saussure, en passant par Comte, la position dominante est celle d'une discrétisation nécessaire d'un réel continu. Cette discrétisation n'est possible que dans l'optique où la démarche scientifique adopte un point de vue par rapport au réel.

Nous nous inscrivons dans une telle démarche d'abstraction en posant, pour le domaine qui nous intéresse : la linguistique de corpus appliquée à la recherche d'information, des unités dépassant les bornes traditionnelles du mot typographique. Ces unités forment la base sans laquelle aucune approche raisonnée du problème n'est possible, nous les nommons **signatures thématiques**. Nous donnerons une définition plus complète de cet objet dans le deuxième chapitre, retenons simplement, à ce stade de l'exposé, le statut d'unité que nous lui conférons.

1.1.3.2.Système

La notion de système, ou de structure, base de la linguistique structurale, est généralement définie comme un réseau de relations entre éléments, ou unités d'un ensemble d'objets, en l'occurrence des faits linguistiques. Cette notion est centrale pour notre problème en ce qu'elle fait le postulat raisonnable que les objets du monde ne sont pas connaissables directement, mais bien plutôt par les relations qu'ils entretiennent entre eux. De la notion de système découle d'ailleurs la notion de valeur saussurienne, qui a l'élégance de ne faire appel

qu'à des critères linguistiques, assurant par là même un degré de cohérence interne de l'objet modélisé, sans faire appel à un appareil formel externe (ex. : mathématiques, logique). Nous reviendrons sur cette notion de valeur dans la suite de notre développement, notamment au sujet de la méthode distributionnelle.

Par ailleurs, la formalisation de la notion de système permet d'aborder les problèmes de construction de grammaire, ainsi que celle, plus large, de modèle d'un ensemble de données linguistiques observables.

1.1.3.3.Signe

La notion de signe, ou association d'un ensemble de propriétés linguistiques observables (formes) et de propriétés non observables directement (sens, dans une acception large), est centrale pour notre propos. En effet, le problème qui nous occupe revient à (re)trouver les signes associés à un domaine de spécialité, au moyen d'une analyse linguistique, afin d'en faire un recensement, de les organiser en un système (ou plusieurs sous-systèmes) traduits en un format interprétable par une machine.

L'ensemble de propriétés observables auquel nous nous intéresserons sera constitué des mots contenus dans des textes de spécialité à vocation informative. L'ensemble des propriétés non observables directement sera constitué par l'expertise d'opérateurs humains, explicitée dans la mesure du possible, qui servira de base à la construction des signes. Cette expertise permet d'associer un ensemble de formes à un ensemble de thèmes (ex. : finance, terrorisme, vache folle) d'un domaine de spécialité.

Le distributionnalisme, comme l'ensemble de la linguistique structurale, s'est fondé sur les acquis du modèle aristotélicien, notamment dans son versant classique, tel qu'initié par Harris et perpétué par ses héritiers. On peut donc raisonnablement s'attendre à ce que le distributionnalisme classique rencontre les mêmes difficultés que ceux évoqués plus haut, à savoir le recours à des opérations invisibles (une structure cachée) et la nécessité de devoir décider de l'appartenance catégorielle de certains éléments. Cependant, ainsi que nous avons tenté de le montrer, la conception classique des catégories ne constitue pas le seul modèle disponible des catégories, ce qui nous servira à instaurer, dans la suite de notre exposé, une distinction entre un distributionnalisme fondé sur une vision catégorique et logiciste, que nous qualifions comme s'inscrivant dans un cadre discontinu, d'un distributionnalisme fondé sur une prise en compte de la dimension probabiliste, donc continue, dans les phénomènes langagiers.

1.2. Du discontinu dans le distributionnalisme

Dans cette partie, nous traitons du courant distributionnaliste, tel que défini et mis en œuvre par Harris. Ainsi que nous l'avons fait plus haut pour la linguistique structurale, nous insistons sur quelques notions essentielles qui nous seront utiles dans l'ensemble de notre développement.

1.2.1. Le distributionnalisme de Harris, un processus de découverte

Les travaux fondateurs de Harris, centrés sur des procédures de découverte des unités d'une langue donnée et de leurs propriétés (le système de la langue), sont marqués d'une vision catégorique. Par catégorique, nous entendons une conception basée sur les principes logiques : loi du tiers exclu et loi de non contradiction, qui amène à poser, pour une unité donnée, une appartenance catégorielle (une fonction) unique. Dans cette conception, les propriétés des unités linguistiques sont destinées à former la base d'une hiérarchie (taxinomie), suivant les principes classiques de la catégorisation, tels qu'exposés plus haut.

1.2.1.1.La primauté des observables

Nous avons tenté de préciser le cadre épistémologique (voir *supra*) et méthodologique du distributionnalisme, nous insistons ici sur le caractère systématique qui ressort des études de (Harris, 1951). Cette systématicité est l'expression d'une primauté accordée aux observables linguistiques, à l'exclusion des aspects non directement observables[23]. Nous voyons plusieurs conséquences à une telle démarche systématique. La première est l'impossibilité d'une étude réellement systématique, en tant que réalisée par un opérateur humain. En effet, volontairement ou non, l'humain catégorise, a des attentes, oublie, bref il ne prend en compte qu'une partie du réel. L'étude des observables est donc nécessairement une étude imparfaite, incomplète, résultant d'un compromis double : celui du dialogue entre l'ordre réel et la pensée humaine, ainsi que celui de l'incomplétude des données, le langage possédant un caractère infini (bien que dénombrable). On comprend dès lors qu'il faille viser l'exhaustivité afin de limiter les effets de ce compromis.

[23] Ainsi, dans son ouvrage de 1951, Harris n'aborde les aspects sémantiques, pragmatiques, ou sociaux du langage qu'en termes de bornes, au-delà desquelles le distributionnalisme ne s'aventure pas.

Nous avons placé l'œuvre de Harris à la croisée des chemins du structuralisme et du behaviorisme, et nous l'avons caractérisée comme une systématique. Il en découle nécessairement une prise de position en faveur des études sur le terrain.

> In both the phonologic and the morphologic analyses the linguist first faces the problem of setting up relevant elements. To be relevant, these elements must be set up on a distributional basis: *x* and *y* are included in the same element *A* if the distribution of *x* relative to the other elements *B*, *C* etc., is in some sense the same as the distribution of *y*.
>
> Since this assumes that the other elements *B*, *C*, etc., are recognized at the time when the definition of *A* is being determined, *this operation can be carried out without some arbitrary point of departure only if it is carried out for all the elements simultaneously*. The elements are thus determined relatively to each other, and on the basis of the distributional relations among them.
>
> (Harris, 1951, Methodological preliminaries, p.7)

Dans cet extrait, le passage que nous soulignons vise à mettre en évidence l'aspect systématique de l'analyse harrissienne, condition d'une absence d'arbitraire dans les observations menées sur corpus.

1.2.1.2.Notion de distribution

L'un des postulats essentiels de l'approche distributionnaliste est que le matériau linguistique présente des régularités. Ce postulat est la base de toute approche raisonnée d'un problème linguistique, qui reste le même quelle que soit la nature des unités cherchées : (re)trouver les unités d'une langue donnée par une étude portant sur un échantillon de cette langue. Harris donne la définition suivante de la notion de distribution.

> The distribution of an element is the total of all environments in which it occurs, i.e. the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements.
>
> (Harris, 1951, p. 15)

La notion de distribution respecte le caractère arbitraire entre signifiant et signifié : elle constitue une propriété observable, voire quantifiable des unités linguistiques étudiées. Par cette notion de distribution sont introduits les deux axes d'analyse, qui jouent le même rôle qu'en linguistique structurale classique : les axes syntagmatique et paradigmatique. En effet, on peut traduire « la somme de tous les environnements » d'un élément donné comme les contraintes observées sur l'axe syntagmatique ; chaque profil distributionnel particulier définit un paradigme (une classe d'éléments) particulier.

1.2.1.3.Notion d'unité linguistique

Comme le signe saussurien, une unité, ou un élément, dans le sens de Harris, est donc constituée d'un ensemble de propriétés observables (i.e. une forme, et un profil distributionnel) et d'un ensemble de propriétés non directement observables (un sens). Cependant, Harris délaisse quelque peu le sens au profit de la forme, en étoffant la notion même de forme : là où elle semblait parfois être évidente chez Saussure, Harris se caractérise par une approche prudente du problème du relevé des unités d'une langue. Ainsi, les formes ne sont telles qu'en ce qu'elles s'opposent à d'autres formes, et non par leurs propriétés absolues (ex. : propriétés acoustiques). Par là même qu'elles s'opposent, elles construisent un réseau de relations entre elles : un système. Cette approche purement fonctionnaliste amène Harris à considérer des constituants dépassant le cadre du mot typographique : constituants discontinus, constituants longs, tant aux niveaux phonologique, morphologique que syntaxique.

1.2.2. Le distributionnalisme catégorique comme théorie linguistique

Nous examinons ici le distributionnalisme catégorique du point de vue de sa capacité à jouer le rôle d'un modèle pour une science des faits langagiers. Nous abordons donc les aspects liés à l'adéquation descriptive, explicative et prédictive de l'approche harrissienne.

1.2.2.1.Un modèle de la Langue

(Harris, 1968, p. 20) révèle l'objet du distributionnalisme : « Given the properties of language (…), it follows that we should be able to define discrete elements, and should then be able to describe language as certain well-formed sequences of classes of them ». Tout le programme distributionnel est contenu dans ce passage. (Harris, 1968) constitue, en effet, un

exposé théorique et formel de la démarche distributionnelle, là où (Harris, 1951) constituait essentiellement un recueil méthodologique. Dans un mouvement comparable à celui entrepris par Saussure, Harris pose, comme axiome, la nature discrète, non des observables linguistiques eux-mêmes, mais bien de ces observables, en tant qu'ils sont amenés à jouer le rôle d'unités pour une théorie linguistique.

L'objet premier d'une démarche scientifique passe donc, pour Harris, par la délimitation des unités d'une langue, aux niveaux phonologique, morphologique et syntaxique. Cette découverte est assurée par ce que Harris nomme des procédures, pour lesquelles il voit une traduction possible, sous la forme d'un langage formel. Bien que l'automatisation des procédures d'analyse ne soit pas, à notre connaissance, mentionnée explicitement par Harris, (Harris, 1968) contient en germe les principes d'une linguistique des corpus formelle et automatique.

Harris pose la grammaire d'une langue comme objet à modéliser, autrement dit il pose la reconnaissance des seules séquences bien formées comme problème à résoudre pour une théorie linguistique.

> We begin with an experimental method for establishing the ultimate discrete elements, the phonemic distinctions, for each language separately (…). A recurrent stochastic process on these elements then distinguishes words (…), and another and different recurrent stochastic process on words distinguishes sentences (…). The latter process can also be stated in the form of an axiomatic theory which, given the word list of a language and a set of axiomatic sequences, obtains the sentences (more precisely, the sentence structures) of the language.
>
> (Harris, 1968, p. 20)

Dans cette conception théorique, la détermination (découverte) des éléments est aussi importante que celle des opérations (*stochastic process*, *axiomatic theory*) qui permettent d'aboutir aux (structures de) phrases : « the determination of the elements is as important as the operations upon these elements » (Harris, 1968, p. 20). Cette identification se base sur la détection, par des locuteurs, d'une « répétition » entre deux séquences données : « the elements are determined by speakers' identical recognition of a relation of 'repetition' between utterances » (*idem*), autrement dit la détermination des unités repose

fondamentalement sur un processus de discrétisation du réel, de catégorisation, qui regroupe les éléments assurant la même fonction dans une même classe.

1.2.2.2.L'objection chomskyenne au processus de substitution

L'ensemble de la démarche distributionnelle repose sur l'hypothèse structuraliste des axes syntagmatique et paradigmatique. Le premier, en effet, permet d'aborder le matériau linguistique dans sa linéarité, et ainsi d'aboutir à la description des relations de contiguïté entre unités. Le second permet de construire des classes d'éléments, en fonction d'un comportement similaire, sur l'axe syntagmatique. La construction de ces classes d'éléments repose sur les opérations classiques de segmentation et de substitution, qui découlent des axes syntagmatique et paradigmatique. Or, bien que ces opérations soient justifiées dans le cadre d'une « simple » description (ex. : description d'un système phonologique, ou morphologique), Chomsky rejette la validité du principe de substitution, comme moyen d'accéder aux propriétés, notamment sémantiques, des unités.

> In any example of linguistic material, no two words can be expected to have exactly the same set of contexts. On the other hand, many words which should be in different contexts will have some context in common. (…) Thus substitution is either too narrow, if we require complete mutual substitutability for co-membership in a syntactic category (…), or too broad, if we require only that some context be shared.
>
> (Chomsky, 1955, pp. 129-145)

En rejetant le principe de substitution, Chomsky rejette toute approche partant des observables et cherchant à identifier la fonction (phonologique, morphologique, syntaxique) des éléments étudiés. Or, ce principe est à la base de l'ensemble des approches automatiques les plus récentes dans le domaine de l'apprentissage de contraintes de sélection et de sous-catégorisation des éléments lexicaux, servant de base à l'induction de grammaires à partir de données observables[24]. Quelle est, réellement, la portée de cet argument ?

[24] Voir, par exemple : (Abney, 1996 b.), (Goldsmith, 2001), (van Zaanen, 2001), (Osborne, 1999), (McMahon, 1994), (Finch, 1993), (Hutchens, 1995), (Lee, 1997), ou encore (Schulte im Walde, 1998).

Afin de préciser la portée de l'argument chomskyen au principe de substitution, examinons les deux énoncés ci-dessous.

1. John is eager to please
2. John is easy to please

Les opposants au principe de substitution avancent que la différence essentielle entre ces deux énoncés, à savoir le renversement de rôle pour *John* (agent dans 1, patient dans 2), ne peut pas être corrélée avec une différence observable dans la répartition des formes. En effet, les deux énoncés sont construits, sur le même schéma : **N0 is Adj to V**. Autrement dit, on se trouverait dans les deux cas face à une même forme de surface, alors que les formes profondes (l'interprétation) de ces énoncés sont différentes[25]. Cet exemple a servi à remettre en cause de façon catégorique la validité des approches guidées par les observables en tant qu'approches scientifiques : le structuralisme européen, le distributionnalisme de Harris et les approches statistiques.

Sans remettre en cause complètement l'objection soulevée ci-dessus, il est possible d'en atténuer la portée. Signalons une première réfutation, pratique, de cet argument : la différence fine entre les deux énoncés a trait au domaine sémantique, qui reste dans une large part difficilement traitable par les approches automatiques. Cette limite n'empêche pas les réalisations pratiques opérationnelles : les cas où un système automatique aurait à opérer une distinction de l'ordre de celle existant entre 1) et 2) sont marginaux, en termes d'application. De plus, si on considère la transformation en **It is Adj to V N0**, seule 2') paraît attestable :

1'. * It is eager to please John
2'. It is easy to please John.

[25] Signalons que la discussion de cet exemple est bien un problème de catégorisation, bien qu'il ne s'agisse plus de reconnaître du même dans l'autre, mais bien de l'autre (deux structures profondes) dans du même (une même forme de surface). Autrement dit, on se trouve dans le cas évoqué plus haut par Auroux, du recours à un principe d'homonymie, dans un cadre monocatégorial : 1 et 2 ont la même forme de surface, mais leur forme profonde est différente.

Autrement dit, il paraît beaucoup plus probable de trouver 2') que 1'), et *eager* et *easy* n'ont pas la même valeur, puisqu'ils se distinguent par au moins un contexte (fabriqué) distributionnel. Nous considérons donc que la cible première de l'objection formulée plus haut est la faiblesse des principes de regroupement d'éléments en fonction d'une similarité de profil distributionnel, qui ne remet pas, pour autant, en cause la validité de l'approche partant des observables. Sur ce point, l'argument est justifié, c'est d'ailleurs la principale critique qu'il est possible d'adresser à l'encontre de l'ensemble de la méthode exposée dans (Harris, 1951). La conséquence que doit en tirer une linguistique partant des corpus est de tenter de formuler des principes systématiques et explicites de regroupement d'éléments en classes à partir de leur profil distributionnel.

Signalons également que l'argument s'applique à d'autres cas, qui ne font pas appel à une représentation fine des rôles casuels. Ainsi, dans les deux énoncés ci-dessous, construits sur le même schéma **N0 V N1**, les indices formels ne permettent de dire rien de plus que : les deux verbes considérés, *voir* et *manger*, partagent au moins une partie de leur profil distributionnel.

3. Le chat voit la souris
4. Le chat mange la souris

Quelle conclusion tirer de cet exemple, beaucoup plus simple que le premier, et qui ne permet pas, non plus, de distinguer la différence essentielle existant entre le signifié attaché à *voir* et celui attaché à *manger*[26] ? Une première conclusion pourrait être la réaffirmation de l'objection chomskyenne contre la procédure de substitution, et le rejet de toute approche guidée par les observables seuls. Cependant, qu'a-t-on vraiment dit au sujet de ce type d'approche avec ces exemples ? Principalement qu'un locuteur natif n'a aucun mal à distinguer 1) de 2) et 3) de 4), sans plus de contexte que celui fourni par ces quatre énoncés, pour la bonne et simple raison qu'il connaît déjà le sens de *to be eager to* versus, *to be easy to*, et de *voir* versus *manger*.

Pour mieux comprendre dans quel piège l'objection posée par ces énoncés fait tomber les approches guidées par les observables, faisons un détour par la langue vernaculaire, et

[26] Dans le premier cas, la souris est toujours vivante, dans le second elle ne l'est plus.

considérons des éléments tels que *machin*. En français, *machin* peut remplacer n'importe quel substantif, il peut également être utilisé comme verbe. Dans chaque cas, *machin* respecte les contraintes morphologiques de l'élément remplacé[27]. En remplaçant le verbe de 3) et 4) par *machin(er)*, on obtient :

3'. Le chat machine la souris.

Là encore, en-dehors d'informations apportées par le contexte situationnel, rien ne permet d'interpréter 3') comme un événement au cours duquel une souris est vue ou bien mangée. En quelque sorte, *voir* et *manger* deviennent des homonymes par la transformation subie.

Le détour pris par 3') nous a surtout permis de comprendre l'importance du contexte (situationnel) pour l'interprétation de ces énoncés, dans le cas où leur sens n'est pas déjà connu. Or, justement, les approches guidées par les données se placent dans une telle position de découverte de la valeur des éléments linguistiques uniquement à partir de leur comportement distributionnel, c'est-à-dire uniquement à partir des places dans lesquelles on les trouve ; le sens de ces éléments n'est utilisé que d'un point de vue distinctif[28]. De ce fait, l'objection soulevée par 1) et 2) est d'autant plus amoindrie qu'elle se place à l'extérieur du domaine que les approches distributionnelles entendent explorer[29].

1.2.2.3.Adéquation descriptive

Les corpus collectés et transcrits se trouvent, de fait, au centre de l'approche distributionnelle, envisagée principalement comme méthodologie descriptive systématique. L'introduction à l'ouvrage de 1951 est éloquente.

---

[27] Ainsi, *machin* respecte l'accord en genre et en nombre s'il remplace un substantif : *un machin*, *deux machins, Machine est venue*. Dans les cas où il remplace un verbe, *machin* est employé comme le radical d'un verbe du premier groupe : *je machine*, *tu machines*, *ils machinent*…

[28] Autrement dit, 4) n'est pas perçu comme une répétition de 3) par un locuteur natif.

[29] La réfutation de l'objection chomskyenne au principe de substitution est développée dans (van Zaanen, 2001), dans le cadre d'un apprentissage de régularités structurelles guidé uniquement par des exemples positifs, grâce à un algorithme non supervisé, nommé ABL (*Alignment Based Learning*).

> This volume presents methods of research used in descriptive, or, more exactly, structural, linguistics.
>
> Starting with the utterances which occur in a single language community at a single time, these procedures determine *what may be regarded as identical in various parts of various utterances*[30].
>
> (Harris, 1951)

On le voit, pour Harris linguistique structurale et descriptive semblent quasiment synonymes ; le programme que se donne le distributionnalisme est la détermination des éléments qui peuvent être considérés comme identiques. L'objet de la méthode harrissienne, rappelé à plusieurs reprises, est la détection des régularités de comportement distributionnel, que nous interprétons comme la détection d'un même face à la variation, c'est-à-dire une tâche de catégorisation.

Par sa centration sur les observables, le distributionnalisme vise une bonne adéquation descriptive. Cependant, l'ouvrage fondateur de 1951 est imprégné d'une prudence vis-à-vis des notions les plus fondamentales telles que mots, ou morphèmes. En effet, la particularité du distributionnalisme de 1951, qui se retrouve d'ailleurs dans d'autres écrits, tels que ceux de Herdan[31], est de considérer les différents éléments que sont les phonèmes, les morphèmes et les éléments phrastiques comme autant d'unités, pourvues de propriétés distributionnelles. Cette unité de traitement permet à Harris d'aborder, avec les mêmes méthodes et la même simplicité, aussi bien des phénomènes phonologiques, que morphologiques ou syntaxiques, dans des langues appartenant à des groupes différents, comme, par exemple : l'anglais, le français, les langues bantoues ou encore les langues sémitiques.

Une lecture superficielle du programme distributionnel pourrait faire croire que le but poursuivi par Harris est la constitution d'une hiérarchie d'éléments linguistiques, sur le modèle taxinomique binaire de la phonologie de l'époque. Or, dès l'introduction à son ouvrage, Harris insiste sur la relativité de l'identité de comportement distributionnel recherchée : ainsi, il dit « what may be regarded as identical » et non pas « what *is* identical »

---

[30] Italiques ajoutés.

[31] Voir *infra*.

au sujet de ce comportement. L'ensemble de l'ouvrage est écrit sur le même ton, par ailleurs un certain nombre des procédures destinées à regrouper des éléments linguistiques sur la base de leur comportement distributionnel sont autant de contournements de l'impératif catégorique sous-tendant le processus de découverte. Il mentionne ainsi, dans l'ensemble des étapes d'analyse d'une langue, visant à en isoler les différents éléments, une phase préliminaire, jugée indispensable, dénommée explicitement « approximation ». L'approximation se décline en procédures basées sur une « similarité grossière » (*rough similarity of environment*), sur des simplifications, des généralisations, ou encore sur une identité distributionnelle partielle (*partial distributional identity*). Il donne l'exemple de l'élément *root*, en anglais, dans les contextes suivants : *watch it grub for –s*, *those –s look withered to me*, *the eleventh – of 2048 is 2*, *that's the – of the trouble*. Harris souligne que la mise en rapport de ces différents contextes d'occurrence est une prise de décision, qui ignore délibérément les différences de dépendance de *root* par rapport au reste de l'énoncé : les éléments suivants *grub for roots*, *the root of the problem* et *the nth root of x*, sont des expressions figées en anglais, alors que dans les deux autres énoncés, *root* pourrait être remplacé par n'importe quel mot appartenant à la classe des « parties de végétaux ».

La forte adéquation descriptive recherchée par Harris ne se fait cependant pas dans le sens d'un empirisme, qui hypothèquerait toute construction scientifique à partir des observables : Harris rejoint en effet Saussure sur ce point, en affirmant que « [t]he elements of linguistics are not direct descriptions of portions of the flow of speech » (Harris, 1951, The status of linguistic elements, p. 18). Cette position n'est pas réservée au domaine de la phonologie : « speech is a set of complex continuous events (…) and the ability to set up discrete elements lies at the base of the present development of descriptive linguistics » (*idem*). On le voit, tant pour Saussure que pour Harris, toute description est déjà une analyse, et non pas une simple accumulation de faits. On peut voir dans le recours à des procédures d'approximation la mise en œuvre d'un principe visant à assurer un maximum de cohérence aux observations, allié à la recherche d'une complétude maximale donnée par la description envisagée comme systématique.

1.2.2.4. Adéquation prédictive

Nous avons présenté le distributionnalisme comme une démarche centrée sur les observables, donc visant une complétude maximale dans les observations. Nous avons également avancé que ce principe de complétude était associé à un principe de cohérence. Ce principe, visant à induire des règles générales à partir des exemples observés, peut être vu

comme une ambition de dépasser le niveau purement descriptif. Les approches rationalistes, telles que le générativisme, ont essentiellement fondé leurs critiques du distributionnalisme sur les limites d'une théorie linguistique fondée sur des observables incomplets : les corpus. Le recours aux procédures d'approximation, évoquées plus haut, peut être vu comme la réponse de Harris à l'argument de l'incomplétude fondamentale des corpus.

> A major reason for the use of approximation techniques here is the inadequacy of the usual linguistic corpus as a sample in respect to the distribution of morphemes (…) even a corpus large enough to yield almost all the morphemes of the language will (…) fail to give us anything like all the environments of each morpheme. The number of mathematically statable sequential permutations of the morphemes in a language is very great. Some of these sequences will practically never occur.
>
> (Harris, 1951, p. 253)

Par ailleurs, Harris est conscient du comportement idiomatique de certains éléments, qui constitue une deuxième justification pour la mise en oeuvre de procédures d'approximation.

> The impracticability of obtaining an adequate corpus is increased by the fact that some utterances are rare not merely because of the great number of possible morphemically different utterances, but also because of a special rarity which we may call a culturally determined limitation.
>
> In view of all this, it would be desirable, in grouping the morphemes into classes, to devise such an approximation as would disregard at least these culturally determined limitations.
>
> (*idem*)

Harris prend donc position sur un problème soulevé plus tard en linguistique quantitative, au sujet duquel nous citons (Muller, 1973), concernant la validité des tests statistiques réalisés sur des corpus, considérés soit comme des échantillons de Langue, soit comme des extraits de Langue elle-même.

> On distinguera donc deux types de raisonnement fondamentalement différents.
>
> Ou bien on raisonne sur un **texte fini**, par exemple *L'illusion*. Sachant que dans ce texte le substantif (…) a une fréquence relative de 0,18, on est parfaitement fondé à tirer de cette fréquence une probabilité (…). Mais cette probabilité ne s'applique qu'à un tirage non exhaustif, ou à la rigueur à un tirage exhaustif de très faible amplitude par rapport à l'étendue du texte.
>
> Ou bien on raisonne sur la **langue** de ce texte, sur la population parente dont ce texte est un échantillon, et dont on ne connaît les caractères qu'à travers ce texte.
>
> (Muller, 1973, p.112)

La position de Harris par rapport aux corpus, donc aux observables, est celle de données empiriques tirées du domaine de la Parole, prises comme reflétant des contraintes générales relevant du domaine de la Langue. Cette position est justifiée par la centration, dans les travaux ultérieurs, sur les sous-langages[32] : ceux-ci sont vus comme reflétant des contraintes plus fortes que celles de langue générale[33].

Ainsi, les conditions d'adéquation prédictive du distributionnalisme harrissien concernent essentiellement les « degrés de liberté d'occurrence » des unités linguistiques. Elles reposent essentiellement sur l'induction de règles à partir des exemples étudiés en corpus, qui permettent de prédire les jugements de grammaticalité de séquences construites à partir d'éléments dont les comportements distributionnels sont extrapolés. En d'autres termes, l'adéquation prédictive du distributionnalisme doit être vue sous l'angle de la capacité d'abstraction par rapport aux données empiriques connues. Cette abstraction passe essentiellement par deux types de procédures : les procédures dites d'approximation et la promotion (*setting up*) d'éléments linguistiques au rang d'unités d'un système.

En ce qui concerne les procédures d'approximation, qui visent à permettre la construction de classes d'éléments de comportement distributionnel *similaire*, nous avons vu

---

[32] Voir les études menées dans le domaine des langues de spécialité, telles que l'immunologie, (Harris, 1989).

[33] La prolifération des études portant sur les sous-langages, en linguistique de corpus, peut être vue comme le reflet de la position harrissienne sur les corpus. Voir, par exemple (Morin, 1999), (Daille, 2002), (Faure, 2000), (Hamon, 2000), ou encore (Bourigault, 1994).

plus haut, à propos de l'objection au principe de substitution, que la lacune principale du distributionnalisme, du moins dans sa forme parue dès 1951, réside dans l'absence de systématicité et de formalisation de ces procédures.

Les procédures de promotion, de leur côté, sont autant d'hypothèses faites sur l'appartenance catégorielle des éléments considérés, confirmées ou infirmées par l'observation de nouveaux corpus.

Tant les procédures d'approximation que celles de promotion sont prisonnières du cadre catégorique adopté par Harris, qui rend d'autant plus difficile toute extrapolation que tout contre-exemple, y compris unique, permet de remettre en cause l'ensemble du système en construction.

1.2.2.5.Adéquation explicative

Nous nous sommes penchés sur les conditions d'adéquation descriptive et prédictive de l'approche distributionnaliste. Quelle peut-être l'adéquation explicative d'une démarche centrée sur le comportement distributionnel des éléments auxquels elle s'intéresse ? La question du sens vient rapidement, ainsi que le souligne Martinet.

> Fonder les classes d'unités significatives sur les compatibilités, c'est-à-dire sur un comportement strictement matériel, se heurte à la conviction que ce qui fait l'unité d'une telle classe est ce qu'il y a de sémantiquement commun à toutes les unités qui y figurent.
>
> (Martinet 1985, p. 109)

Qu'en est-il de la conviction dont parle Martinet, chez Harris ? En effet, une théorie linguistique tiendrait là une explication au moins partielle des effets de sens, ainsi que des contraintes de distribution relevées : le sens d'un énoncé pourrait être envisagé comme une représentation reposant au moins pour partie sur la représentation des contraintes distributionnelles des éléments, inversement ces contraintes pourraient être vues comme des effets de contraintes de sens. Cependant, Harris ne parle du sens qu'en tant qu'élément distinctif, quelque soit le niveau d'analyse (phonétique, morphologie, syntaxe), la position harrissienne semble être la même : ne considérer le sens des énoncés qu'en tant qu'il est le

même ou non[34], le détail des différences restant hors de portée. On reconnaît l'influence behavioriste de l'héritage bloomfieldien dans cette position. Le sens des énoncés n'est donc pas ce dont Harris cherche à rendre compte.

À première vue, l'ouvrage fondateur de 1951 paraît vide, quant à la portée explicative de l'approche distributionnelle. Il est nécessaire de prendre une voie détournée pour appréhender ce que le distributionnalisme permet d'expliquer, qui est contenu dans l'objet-même de l'ouvrage de 1951 : promouvoir des objets du monde (linguistique) au rang d'éléments, construire un système des éléments d'une langue.

Ainsi, ce que permet d'expliquer le distributionnalisme, c'est l'émergence d'un système d'éléments linguistiques. Pour cette raison, la démarche distributionnelle a pu être reprise dans le domaine de l'acquisition des langues, de l'enseignement, mais également dans celui de l'apprentissage automatique. La portée explicative du distributionnalisme a donc trait aux notions saussuriennes de valeur, de système, d'unité, d'axes paradigmatique et syntagmatique, c'est l'objet du passage ci-dessous.

> Harris suggested how the structural and distributional regularities could work together to support language acquisition and use: "when only a small percentage of all possible sound-sequences actually occurs in utterances, one can identify the boundaries of words, and their relative likelihoods, from their sentential government […]"
>
> (Pereira, 2000, p. 1241)

La position exprimée par Pereira est intéressante à plus d'un titre : elle s'inscrit dans le cadre d'un renouveau du programme distributionnel, par l'abandon d'une vision catégorique, dont nous avons vu les difficultés qu'elle comportait pour une linguistique partant des observables. Pereira voit, dans le programme distributionnel, les conditions de l'émergence de systèmes linguistiques, tant dans la phase d'acquisition que dans l'ensemble de l'utilisation de la compétence linguistique. Cependant, Pereira note l'insuffisante formalisation et systématisation des principes distributionnels débouchant sur de tels systèmes, qui rejoint les remarques faites ci-dessus au sujet des conditions d'adéquation prédictive du programme de Harris.

[34] C'est l'idée sous-tendant le recours à la notion de « répétition ».

> While Harris discussed the functional role of distributional regularities in language, he proposed no specific mechanisms by which language users could take advantage of those regularities in language acquisition and use. In particular, it is not obvious that language users can acquire stable distributional information (…) from the limited evidence that is available to them from their linguistic environment. This question created a great opening for Chomsky's rationalist critique of empiricist and structuralist linguistics (…).
>
> (Pereira, 2000, p. 1242)

Les conséquences de cette formalisation insuffisante sont une remise en cause de toute approche guidée par les observables[35] par une approche rationaliste, i.e. le générationvisme.

Face aux lacunes du programme distributionnel, envisagé dans un cadre catégorique trop restrictif, quelle valeur accorder au renouveau des approches probabilistes guidées par les observables, sous la pression, principalement, du domaine de l'ingénierie linguistique ?

1.3. Distributionnalisme et probabilités

L'alliance entre une approche non catégorique, fondée sur un appareil formel en contradiction avec le cadre catégorique classique, et une approche « empirique » des phénomènes langagiers, n'est pas nouvelle. Nous abordons dans cette partie le glissement vers une approche probabiliste (*stochastic*) - initié par des auteurs tels que Herdan - des faits langagiers et de la construction d'un système linguistique, autrement dit une théorie, à partir des observables. L'œuvre de Herdan nous paraît centrale en ce qu'elle fonde, avec ses ouvrages *The calculus of linguistic observations* (1962) et *Quantitative linguistics* (1964), une approche probabiliste de problèmes phonologiques, morphologiques, syntaxiques et stylistiques, dans le prolongement de la linguistique structurale européenne, notamment dans la lignée de Saussure.

[35] Empirique, au sens anglo-saxon.

Dans cette partie, nous nous pencherons donc tout d'abord sur les ouvrages de Herdan cités plus haut, puis nous examinerons les conséquences d'une approche non catégorique et non logique des faits langagiers, telle que proposée récemment par Manning. Enfin, nous tenterons de déterminer les conditions à même de constituer une telle approche en tant que théorie linguistique, notamment grâce au regain d'intérêt pour les approches probabilistes sous l'impulsion de l'ingénierie linguistique.

1.3.1. Herdan, le glissement vers un distributionnalisme probabiliste

1.3.1.1. Motivations linguistiques pour une approche probabiliste

Le programme que se donne Herdan est ambitieux : de la phonologie à la stylistique, en passant par la syntaxe, la morphologie, mais également la linguistique comparative, l'auteur affirme la nécessité de recourir à des outils mathématiques en linguistique structurale, seuls à même de dépasser le niveau empirique. Il entend combler les lacunes des approches connues jusque là, en fondant une linguistique formelle, axée autour d'axiomes et de démonstrations. Le tout premier de ces axiomes motive le recours à des outils mathématiques particuliers, en l'occurrence des outils statistiques. Cet axiome donne toute la vision herdanienne des notions saussuriennes fondamentales, y compris la distinction entre Langue et Parole, fondée sur le caractère arbitraire du signe.

> [W]e derive the definition of a random sample as being obtained by a method of sampling in which the criterion we sample by is uncorrelated with the variable characteristic we are sampling for.
>
> In the area of language, we have a positive hint where to look for such a random variable in de Saussure's axiom of independence of sound and meaning. This is the tenet generally accepted today by linguists that the sounds of which a word consists are independent of its meaning (…). If this were not so, the same concept could not be expressed in different languages by different words. If that axiom is true, then the undoubtedly non-random sequence of words in a literary text (…) should yield a random sample of sounds, phonemes, and also letters, since the criterion we are *sampling by*, i.e. the words arranged according to their meaning, is uncorrelated with what we are *sampling for*, i.e. the individual sounds of the language or the letters of the alphabet.

(Herdan, 1964, p. 6)

Herdan tire la conséquence, sur le plan statistique, du lien arbitraire entre la face signifiante et la face signifiée des signes : il découle de ce lien arbitraire que les productions linguistiques (textes, discours), constituées de séquences non aléatoire d'unités (mots) faisant sens, doivent fournir un échantillon aléatoire (*random sample*) de ces unités. Herdan exprime en termes statistiques l'idée suivante : si le lien entre signifiant et signifié n'était pas arbitraire, un même concept devrait toujours être exprimé de la même façon.

Ce premier axiome est fondateur à plus d'un titre : il appelle nécessairement à considérer les ensembles de productions linguistiques étudiés, autrement dit les corpus, comme autant d'échantillons, au sens statistique, de la Langue. D'autre part, la conséquence logique de ce premier axiome est la prise en compte et la quantification de la variation (i.e. stylistique) dans les productions linguistiques. En d'autres termes, Herdan prend position contre la grammaire générative, déjà féconde d'objections à toute approche statistique des phénomènes langagiers. Il pose les bases d'une linguistique centrée sur la Parole, mais visant la Langue, à travers les observables, considérés comme des événements suivant une loi de distribution donnée[36].

Toutefois, de l'aveu de l'auteur, l'application de méthodes statistiques à l'ensemble des phénomènes langagiers semble impossible, en l'état des moyens informatiques disponibles au milieu des années 1960. De façon générale, les limitations d'ordre technique constitueront un frein à l'approche probabiliste des phénomènes langagiers ; elle justifiera d'ailleurs, comme nous le verrons plus bas, les principales objections formulées par le courant génératiiviste à l'encontre de ces approches non catégoriques et non logiques.

1.3.1.2.Une vision quantitative de l'opposition Langue/Parole

Herdan se donne comme objectif principal de traduire en termes statistiques et quantitatifs les concepts-clés du *CLG* de Saussure, assurant à toute étude linguistique menée dans un cadre structuraliste le statut de démarche scientifique. Ainsi, tant dans son ouvrage de 1962 que dans celui de 1964, la conception d'une linguistique scientifique quantitative

[36] Herdan propose, dans les faits, plusieurs lois de distribution, correspondant à autant de sous-domaines de la Langue : la loi de distribution normale pour les éléments grammaticaux les plus fréquents, la loi de Poisson composée et celle de Waring-Herdan pour les éléments lexicaux, et la loi de Poisson pour les *hapax legomena*.

s'appuie sur la constatation de l'étonnante stabilité des fréquences d'occurrence relatives des unités linguistiques (ex. : phonèmes, morphèmes).

> The phenomenon of the stability of relative frequencies of linguistic forms leads to the statistical view of de Saussure's fundamental distinction between 'la langue' and 'la parole'. According to de Saussure, 'la langue' is the total of linguistic habits which make communication between the members of the speech community possible. It is a social reality, existing for the mass of the people. Roughly, it represents the lexicon of the language in question. 'La parole', on the other hand, is the individual utterance. Whereas 'la langue' is independent of the individual, 'la parole' as the realisation of parts of 'la langue' through speech is dependent upon the individual. So far, it was thought that the former comprised the engrams of the language in the sense of 'lexical forms' (including here, of course, also grammar forms listed in the lexicon), and the latter the words of actual speech. However, the stability of the relative frequencies which we find attached to the various items of a given series of linguistic forms leads inevitably to the conclusion that what 'la langue' comprises are not only engrams as lexical forms, but *these engrams plus their respective probabilities of occurrence*. This is what I have called the statistical view of de Saussure's dichotomy. The basic law of linguistic communication as stated above is then tantamount to the statement that language is the collective term for linguistic engrams (phonemes, word engrams) together with their particular probabilities of occurrence. The engrams concept is thus inseparably connected with that of frequency of occurrence, and if by linguistic normative laws we understand something which regulates the relative frequency of linguistic forms belonging to a certain class, then our statistical conception of 'la langue' implies such normative laws, as whose realisation we must regard the empirically determined frequencies of 'la parole'.
>
> (Herdan, 1962, pp.18-19)

Ce passage est éclairant à plus d'un titre : il donne les clés des conditions d'adéquation explicative de la théorie linguistique proposée par Herdan. Il propose une conception quantitative de l'opposition saussurienne Langue/Parole, tout en s'inscrivant dans un cadre structuraliste classique. En effet, ce passage permet de comprendre l'objectif poursuivi par l'auteur : aboutir, à partir d'une base empirique, à la détermination des « lois normatives » régissant la Parole, c'est-à-dire, les règles d'une grammaire catégorique. La conception herdanienne des rapports entre langue et Parole est celle de la distinction entre population

statistique et échantillon tiré de cette population. En tant que la Parole est un échantillon de la Langue, Herdan considère qu'elle permet d'estimer de façon suffisamment précise la valeur des éléments linguistiques considérés, ce qui motive donc, tant sur le plan descriptif, prédictif qu'explicatif, une démarche scientifique fondée sur des observables linguistiques.

La position adoptée par Herdan sur les rapports entre les sujets parlants et leur langue est centrée sur la fréquence d'usage, la fixation et la propagation des faits langagiers par leur répétition. Autrement dit, pour Herdan, la Langue est un processus dynamique, intégrant, en plus de la dimension sociale, une dimension temporelle. L'ensemble de l'ouvrage de 1964 est imprégné de la conviction que les phénomènes langagiers peuvent et doivent être étudiés avec la même rigueur que les phénomènes naturels (ex. : astronomie, biologie, sociologie) : « all laws of language except those which are basic laws of logic are statistical in nature, since they are behavioural conventions through frequency of use » (Herdan, 1964, p. 18).

1.3.1.3.Une théorie linguistique non grammaticale

Herdan pose l'objet d'une théorie linguistique scientifique comme la reconnaissance des séquences bien formées. En cela, son approche statistique s'inscrit dans la continuité tant du structuralisme saussurien que dans celle du distributionnalisme. Toutefois, Herdan rejette l'ensemble de la tradition grammaticale. Ce rejet de la grammaire, qu'elle soit traditionnelle ou raisonnée, telle que mise en œuvre dans le cadre générativiste, tient à la part de sémantisme que l'auteur attribue aux étiquettes de parties du discours, ainsi qu'à l'ensemble des unités traditionnellement distinguées (ex. : les substantifs désignent préférentiellement des choses, les verbes des actions). Pour l'auteur, ce reliquat sémantique viole la contrainte saussurienne posant la nécessité d'un lien arbitraire entre signifiant et signifié, violation évitée par une approche probabiliste de la distribution des unités linguistiques. Autrement dit, Herdan voit dans cette approche les conditions d'une étude scientifique du langage respectant les principes du structuralisme, passant par la possibilité de décrire les signifiants de toute langue indépendamment de leur signifié.

La démarche initiée dans son ouvrage de linguistique quantitative reprend les acquis du structuralisme, notamment l'œuvre de Saussure en tant que fondatrice d'une méthodologie scientifique dans l'étude du langage. Cette démarche reprend également les acquis du courant distributionnaliste classique, par la centration sur les observables et l'accent mis sur l'étude des énoncés effectivement produits. Cependant, Herdan prend position contre Harris en déplorant la formalisation insuffisante entreprise dans le cadre distributionnel. Naturellement,

Herdan prend également position contre le générativisme, formulant déjà à cette époque ses premières objections à une approche probabiliste en linguistique. Sur ce point précis, la position fondamentalement antagoniste entre les approches chomskyenne et herdanienne est visible dans le postulat, fait par la première, du caractère non essentiellement linguistique des outils statistiques[37]. De son côté, Herdan voit, justement, dans le caractère universel des contraintes traduites par les lois statistiques, les conditions d'une pensée véritablement scientifique[38]. On voit à l'œuvre, d'une part, une conception reposant sur des principes premiers, une nature essentiellement linguistique des faits langagiers, que seule une théorie linguistique peut expliquer, et d'autre part une conception posant l'universalité de certaines contraintes, applicables à tous les phénomènes, y compris linguistiques.

Herdan voit dans le recours à une formalisation linguistique sous tendue par une approche probabiliste les conditions de l'extension réussie des principes structuraux. Autrement dit, il se donne comme modèle la révolution de la phonologie structurale, sous l'impulsion de Troubetzkoy, à l'ensemble des domaines d'étude linguistiques. En termes épistémologiques, l'ambition herdanienne est comparable à la révolution relativiste dans le domaine de la physique : Herdan compare la prise en compte de la dimension probabiliste du langage au changement de paradigme induit en physique par la prise en compte de la dimension temporelle, se traduisant par le passage d'un référentiel de coordonnées cartésien à 3 dimensions vers un référentiel à 4 dimensions.

1.3.2. Un changement de paradigme

Nous voyons dans l'œuvre de Herdan l'émergence d'une réflexion, alternative par rapport au courant génératiste, sur le statut d'une théorie linguistique guidée par une approche probabiliste des observables langagiers. Herdan accorde aux observations sur le terrain un statut d'expérience, au sens scientifique, par là même, on peut voir dans la parution de son ouvrage de 1964 l'acte fondateur d'une linguistique de corpus se constituant comme démarche scientifique, détachée de la contingence empirique.

---

[37] Comme, par exemple, la vérification de l'application de la loi de Gauss à certains phénomènes linguistiques.

[38] Herdan fait d'ailleurs remarquer que l'argument génératiste contre la pertinence de l'application la loi de Gauss en linguistique oblitère le caractère universel de cette loi de distribution, au départ réservée au domaine de l'astronomie et diffusée, entre autres, dans l'étude des populations humaines.

La réflexion initiée par Herdan nous semble fondamentale en ce qu'elle prépare et annonce la diffusion des approches probabilistes en ingénierie linguistique, qui nous paraît avoir fourni les bases d'une réflexion plus générale, tentant de dépasser le seul cadre applicatif. En amorçant une réflexion sur le statut des unités traditionnelles de la linguistique structurale dans un cadre probabiliste, l'œuvre de Herdan nous semble poser une question épistémologique fondamentale : dans quelles conditions une théorie linguistique non catégorique et non logique peut-elle exister ? Par extension, quelles sont les conditions d'adéquation d'une théorie linguistique guidée par les observables langagiers, c'est-à-dire l'émergence d'un Système (supposant une abstraction nécessaire par rapport aux observables) à partir de stimuli langagiers dont le courant générativiste s'évertuera à démontrer la pauvreté.

1.3.2.1.Du catégorique au probable

La linguistique structurale classique, et par la suite le courant générativiste, se caractérise par le postulat catégorique et logique des modèles construits, ainsi que des unités linguistiques intégrées à ces modèles. Nous l'avons vu, on peut attribuer ce postulat catégorique et logique à la reprise d'une conception de la logique formelle comme instrument de science, position héritée d'Aristote et confortée par la position cartésienne revendiquée par Chomsky.

Or, l'émergence d'approches probabilistes, car guidées par les observables dont la variabilité a tour à tour été revendiquée par les tenants des études sur le terrain (sociolinguistes et acquisitionnistes notamment) et dénoncée par les tenants des approches antiempiriques, induit une nécessaire remise en cause de ce postulat catégorique. Cette remise en cause touche l'ensemble des domaines d'une théorie linguistique : des unités (ex. : les phonèmes, les mots), donc des catégories (ex. : parties du discours), aux règles postulées, et donc à l'ensemble du système ainsi construit.

Dans une telle approche, rien n'est jamais démontrable, au sens où on l'entend habituellement, puisque les régularités observées dépendent des données utilisées. On comprend aisément le refus d'une linguistique non démontrable mais seulement probable par les tenants de la linguistique cartésienne de Chomsky, amalgamant positivisme et logicisme.

1.3.2.2.Vers une théorie non catégorique et non logique

L'ambition d'une linguistique scientifique parce que fondée sur la reconnaissance de contraintes universelles visibles dans le comportement distributionnel des unités linguistiques

se pose clairement contre la position cartésienne du générativisme. Cette dernière reste une linguistique catégorique, construite grâce à la logique formelle utilisée comme instrument scientifique, conformément aux principes aristotéliciens.

Dans ce contexte, la parution des travaux de Herdan pose les questions suivantes. Une théorie scientifique doit-elle être catégorique ? La logique doit-elle être au cœur d'une théorie linguistique ?

Les unités d'une telle théorie, que Herdan pose comme discrètes en reprenant les acquis saussuriens et distributionnalistes, n'en auraient pas moins des propriétés relevant du continu, exprimées, par exemple, par des probabilités dans les contraintes de sélection entre unités, plus que comme règles. On voit d'emblée la difficulté pour une théorie, visant une démarche objective, de se baser sur des probabilités, donc un déterminisme empirique, plutôt que sur des règles.

De même, comment envisager l'élaboration d'un système linguistique, reposant sur des unités seulement probables ? La question posée au sujet du système tient en fait à celle de l'équilibre : comment concevoir une stabilité d'ensemble à partir d'éléments potentiellement non stables ? La réponse tient en l'adoption d'une conception dynamique, plutôt que statique, de l'état d'équilibre. En ce sens, la réflexion apportée par Herdan constitue un véritable changement de paradigme, comparable à celui induit par le passage d'une conception classique de la physique à une conception relativiste. Cependant, l'auteur ne tire pas, à notre connaissance, toutes les conséquences, tant méthodologiques que conceptuelles, du changement de paradigme qu'il appelle. En effet, l'objet d'étude privilégié de Herdan est la stylistique, partageant avec la Parole une forte variabilité, pour laquelle il conçoit un modèle non catégorique et non logique permettant, par exemple, de comparer deux auteurs, ou encore d'attribuer la paternité d'une œuvre à un auteur donné.

Le renouveau d'un programme distributionnel non catégorique et non logique, sous l'impulsion d'auteurs tels que Abney, Manning, ou encore Pereira nous paraît poser une assise plus complète que l'œuvre de Herdan, en ce qu'elle définit un cadre théorique et méthodologique, dans la perspective d'une science du langage refusant l'autonomie de la syntaxe et intégrant la part essentielle de variation des phénomènes de Parole. Signalons toutefois que ce programme, dont nous examinons les conditions d'adéquation ci-dessous, reste essentiellement prospectif.

1.3.3. Le distributionnalisme probabiliste comme théorie linguistique

Nous examinons ici la valeur du courant de recherches que nous nommons distributionnalisme probabiliste, en opposition au distributionnalisme catégorique, en tant que théorie linguistique à part entière et non pas seulement de modèle opérationnel dans le cadre de l'ingénierie linguistique. Cet examen se fonde sur les réponses apportées par (Abney, 1996 b.), (Pereira, 2000), (Manning, 2002), ou encore (Finch, 1993) aux principales objections formulées par Chomsky, principalement, à une approche dont l'essence est non linguistique. Nous examinons notamment l'argument de l'impossibilité d'un apprentissage basé sur des exemples positifs, tiré d'une interprétation du théorème de Gold[39], l'argument de la pauvreté du stimulus et l'incompatibilité entre probabilité et grammaticalité.

1.3.3.1. Théorème de Gold et apprentissage à partir d'exemples positifs

Le distributionnalisme, qu'il soit catégorique ou non, est fondamentalement une procédure de découverte du fonctionnement linguistique des unités observées au sein d'échantillons de Langue. En tant que tel, il sert de fondement théorique et méthodologique à l'ensemble des approches automatiques dans le domaine de l'apprentissage des régularités linguistiques à partir de corpus. Or, le générationalisme, de son côté, se caractérise par une position du développement d'une compétence linguistique ne faisant pas appel à des mécanismes d'apprentissage.

L'un des arguments les plus décisifs en défaveur d'un apprentissage à partir des données observables est celui développé dans (Chomsky, 1957 ; 1965), prenant appui sur le théorème de Gold. Dans la conception chomskyenne, l'objectif d'une théorie linguistique étant de rendre compte d'un ensemble de phrases grammaticales, décrites par un langage formel, il est amené à examiner les différents langages existants, afin d'évaluer leur adéquation (descriptive et prédictive) au regard de la tâche à accomplir. Dans le cadre chomskyen, la question du développement d'une compétence linguistique est liée à celle du paramétrage d'un langage formel, afin de n'engendrer et de ne reconnaître que des phrases

[39] (Gold, 1967).

grammaticales. Chomsky montre, dans les ouvrages cités, que seul un type de langage formel est en adéquation avec la nature du problème linguistique : les grammaires dites décidables[40].

Gold, de son côté, définit les conditions générales du paramétrage de différents langages formels, en fonction de la nature des exemples présentés au modèle, dans une situation d' « apprentissage » particulière : un oracle fournit à l' « apprenant » un ensemble d'exemples (des énoncés), auquel il associe un jugement de grammaticalité. Deux situations de paramétrage sont envisagée : la première ne fournit que des exemples dits positifs (grammaticaux), la seconde fournit aussi bien des exemples positifs que négatifs (agrammaticaux). Dans le cadre défini par Gold, l'apprentissage à partir des exemples peut être vu comme l'élaboration d'un algorithme de décision (grammatical/agrammatical) sur un ensemble de phrases conformes à la grammaire qui les a produites[41]. Gold montre que les conditions de la constitution d'un algorithme (réussite ou échec) sont liées au type de la grammaire à apprendre[42] et au paradigme d'apprentissage. Il montre, notamment, que les grammaires décidables, qui constitueraient le niveau nécessaire à la modélisation de la grammaticalité, ne peuvent pas être apprises à la limite, à partir des seuls observables, quel que soit le paradigme d'apprentissage. Les seules grammaires apprenables à la limite sont les grammaires dites hors-contexte (*context free*), les grammaires sensibles au contexte (*context sensitive*) et les automates à états finis.

(Finch, 1993), ainsi que (Pereira, 2000) et (Manning, 2002), remettent en cause l'argument chomskyen en défaveur de l'apprentissage, qui plus est, en ce qui concerne Finch, l'auteur se positionne en faveur de l'apprentissage à partir des seuls exemples positifs, dans un cadre non supervisé. En d'autres termes, Finch remet en cause le paradigme d'apprentissage décrit par Gold et repris par Chomsky : dans un cadre non supervisé, aucun oracle n'est nécessaire.

---

[40] Pour une présentation plus détaillée de l'argument goldo-chomskyen, voir (Finch, 1993), qui pose la constitution modèles linguistiques guidés par les observables dans les termes de la construction d'une théorie scientifique.

[41] Cette conception de l'apprentissage, à partir d'exemples positifs et/ou négatifs, est dénommée « identification de langue à la limite » (*language identification in the limit*).

[42] La position de la grammaire à apprendre au sein de la hiérarchie de Chomsky.

> To put it succinctly, although we know from formal learning theory that we can't learn *all* transformational languages, this is irrelevant because natural language is a particular transformational language. Moreover, what makes it special is the regularity which is evident over nearly all large *finite* sets of sentences, and the Chomsky hierarchy does not classify these at all.
>
> (Finch, 1993, p. 73)

Finch met l'accent sur une lacune de l'argument chomskyen, et remet en cause la classification des langages formels établie par Chomsky. Par ailleurs, Finch caractérise la conception goldo-chomskyenne de l'apprentissage comme trop contraignante, car elle vise à induire les règles d'un ensemble infini de phrases. En restreignant l'apprentissage à un ensemble fini, et en prenant en compte les régularités locales, observables dans ce domaine, Finch pose les conditions d'un réel apprentissage à partir des observables.

1.3.3.2.L'argument de la Pauvreté du Stimulus

Le théorème de Gold est également à la base d'une autre objection chomskyenne, en défaveur de l'apprentissage de la faculté de langage, qui pose que, non seulement l'identification d'un langage formel adéquat à la limite (dans les conditions d'apprentissage définies par Gold) est impossible, mais de plus, l'apprenant est soumis à un ensemble de stimuli trop limité pour mener à bien tout paramétrage. Cet argument pose le stimulus langagier auquel est soumis l'apprenant comme intrinsèquement pauvre, et amène à supposer un principe grammatical inné, universel, génétiquement déterminé : une Grammaire Universelle (GU). Le développement d'une compétence linguistique, pour Chomsky, passe non pas par un apprentissage, mais bien plutôt par un paramétrage de cette GU, c'est-à-dire une sélection parmi un ensemble de primitives. On voit à quel point cette conception *top-down* de l'émergence d'un système linguistique est incompatible avec l'ensemble des approches guidées par les observables (*bottom-up*) : distributionnalisme, catégorique ou non, linguistique de corpus, pédagogie, acquisition des langues, ou encore ingénierie linguistique.

> How poor is the stimulus that the language learner exploits to acquire its native language? (…) [L]inguistic experience is not just a string of words, but it is *grounded* in a rich perceptual and motor environment that is likely to provide crucial clues to the acquisition,

> interpretation and production processes, if for no other reason than for the functional one that much of the linguistic experience is *about* that non-linguistic environment. However, this points to a fundamental weakness in much of the work discussed so far: both in formal grammar and in most computational models of language, language is taken as a completely autonomous process that can be independently analysed.
>
> (Pereira, 2000, pp. 1246-1247)

Cette pauvreté supposée du stimulus fait aujourd'hui l'objet d'une remise en cause par l'ensemble des linguistes cognitivistes d'une part : Lakoff, Langacker et Taylor, notamment. D'autre part, l'ensemble des défenseurs des approches probabilistes en TALN, tant dans leurs applications en ingénierie que dans le domaine de la recherche théorique, militent pour l'abandon d'une conception pauvre du stimulus linguistique, envisagé uniquement sous la forme de suites de caractères, par exemple. Tant les linguistes cognitivistes que des auteurs tels que Finch, Manning ou Pereira, voient dans l'ensemble des paramètres des situations de communication (ex. : contexte situationnel, social, émotionnel) des stimuli riches, rendant possible l'apprentissage à proprement parler de la faculté de langage. En d'autres termes, la pauvreté du stimulus linguistique viendrait de la conception de stimulus linguistique elle-même plus que des informations utilisables dans le cadre de l'apprentissage d'une langue.

1.3.3.3.Grammaticalité et probabilités

Le problème des rapports entre grammaticalité et probabilités peut être résumé par les deux énoncés improbables suivants[43], opposés par Chomsky aux tenants d'approches non catégoriques en linguistique.

5. Colorless green ideas sleep furiously
6. Furiously sleep ideas green colorless

L'objection chomskyenne vis-à-vis des approches probabilistes tient au fait que ces deux énoncés n'ont probablement jamais été prononcés, par conséquent un modèle statistique basé sur des énoncés effectifs attribuerait à 1) et 2) la même probabilité d'occurrence (i.e. 0),

[43] Repris de (Pereira, 2000) et de (Manning, 2002).

alors que 1) est attestable et pas 2). Cette objection tient essentiellement à la reconnaissance par Chomsky, du caractère fondamental de l'abstraction pour la construction d'une théorie linguistique, dont il accuse les approches probabilistes de ne pas pouvoir disposer. Il s'ensuit que, si les approches guidées par les observables sont incapables d'une telle abstraction, elles se trouvent invalidées en tant que fondement d'une théorie linguistique.

L'objection discutée ici tient également à une position, implicite dans l'ensemble de la linguistique structurale, et revendiquée par Chomsky de l'impossibilité de la construction d'une théorie linguistique non catégorique, ne reposant pas sur des contraintes logiques. Cependant, des auteurs tels que Manning, voient dans le langage (sa compréhension comme sa production) un fonctionnement essentiellement continu et quantitatif, commun à l'ensemble des processus cognitifs. Ainsi, d'après (Manning, 2002), l'approche probabiliste de la compréhension du langage naturel revient à voir cette tâche complexe comme l'apprentissage de la probabilité de distribution P(sens | énoncé, contexte). Autrement dit, la tâche linguistique consiste en l'apprentissage de la probabilité conditionnelle associant un sens à un énoncé *et un contexte*. La faculté de langage consiste donc à induire, à partir d'un contexte et d'un énoncé donné une fonction de projection (*mapping*) vers un espace sémantique.

1.3.4. Critères d'adéquation d'un modèle probabiliste des faits langagiers

Dans le cadre de réflexion défini par les travaux de Pereira, Manning et Abney, les conditions d'adéquation d'une théorie probabiliste des faits langagiers sont les suivantes[44].

1.3.4.1. Adéquation descriptive

L'adéquation descriptive de tels modèles doit être assurée par l'adoption de modèles probabilistes capables de couvrir suffisamment les données observées. D'après les auteurs cités, cette adéquation (*fitting*) doit s'appuyer, autant que possible, sur les aspects cognitifs, situationnels, ou encore pragmatiques du langage. En effet, la première objection chomskyenne à l'émergence d'une vision probabiliste d'une théorie linguistique tient à l'argument de la pauvreté du stimulus langagier, exposé plus haut. Chomsky tire de cet

[44] Nous prenons ici le problème de la constitution d'une théorie linguistique comme étant, essentiellement celui de la description, la prédiction et l'explication de phénomènes liés à l'acquisition de la faculté de langage.

argument la nécessité d'une théorie fondée sur le paradigme Principes et Paramètres, il est amené à postuler des structures abstraites, innées et universelles[45], ainsi qu'un mécanisme organique spécialisé d'acquisition du langage (LAD, *Language Acquisition Device*).

Les tenants des approches probabilistes, ainsi que l'ensemble des linguistes cognitivistes, remettent en cause la notion de stimulus pauvre : ils ne nient pas que le langage, considéré sous l'angle d'une chaîne de caractères ou de sons constitue un stimulus insuffisamment riche pour permettre un réel apprentissage, toutefois ils remettent en cause cette vision étriquée du langage, en plaidant pour l'intégration de l'ensemble des stimuli associés, ainsi que pour la réintégration des processus cognitifs dans la construction d'une théorie linguistique[46]. Pereira insiste sur l'information apportée par l'ensemble du contexte dans lequel s'inscrit une production linguistique, ainsi que sur le déterminisme (*grounding*) cognitif et perceptif de cette production : « linguistic experience is not just a string of words, but it is *grounded* in a rich perceptual and motor environment that is likely to provide crucial clues to the acquisition, interpretation and production processes ».

Manning et Pereira posent que l'insuffisance constatée des mécanismes probabilistes d'apprentissage, à l'époque où Chomsky formulait les objections mentionnées plus haut, était de nature technique : ils affirment que le champ des approches probabilistes s'est doté, depuis, de nouveaux algorithmes permettant de dépasser les problèmes posés par la variabilité des observables langagiers. Pour ces auteurs, les conditions d'adéquation descriptive d'une théorie linguistique non catégorique et non logique passent donc par la réfutation de l'argument de la pauvreté du stimulus, ainsi que par le dépassement de limites techniques inhérentes aux premiers formalismes mis en œuvre.

1.3.4.2.Adéquation prédictive

Une fois l'adéquation aux données réalisée, l'adéquation prédictive d'un modèle probabiliste du langage doit se traduire par la capacité d'un tel modèle à généraliser les régularités constatées à de nouvelles données, c'est-à-dire à faire preuve d'une capacité d'abstraction par rapport aux données brutes. Autrement dit, un modèle probabiliste doit pouvoir être capable de concilier les deux impératifs contraires, que nous avons mentionné au

[45] Des Principes, au sens platonicien.

[46] Autrement dit, ils prennent position contre le dogme d'une linguistique autonome, logiciste, déconnectée des autres capacités cognitives majeures.

sujet de la construction d'une théorie scientifique : la complétude, par une bonne adéquation aux données, et la cohérence du modèle construit, permettant de dépasser la contingence empirique[47]. La réfutation des objections chomskyennes dans ce domaine passe, à nouveau, par l'affirmation du caractère technique de l'insuffisance, constatée par les tenants d'une linguistique rationaliste, des modèles probabilistes. Pereira cite, notamment, des procédures de lissage (*smoothing*) des données, susceptibles de fournir la base d'une capacité d'abstraction pour des approches probabilistes.

1.3.4.3.Adéquation explicative

Nous avons exposé les conditions d'adéquation de modèles linguistiques probabilistes, telles que les conçoivent Pereira et Manning. Les ouvrages cités contiennent des réfutations plus développées des objections chomskyennes que ce que nous livrons ici, cependant ces réfutations tiennent, dans l'ensemble, aux capacités descriptive et prédictive de tels modèles. En ce qui concerne l'aspect explicatif des modèles probabilistes, on ne trouvera que peu d'indices dans les ouvrages cités. Nous considérons, pour notre part, ainsi que nous l'avons évoqué pour le distributionnalisme classique, que la capacité explicative d'une théorie linguistique non catégorique doit se focaliser sur le processus d'élaboration d'un système linguistique (Langue) à partir d'observables langagiers (Parole), non restreints à des suites de signes (ex. : phonèmes, graphèmes) prises dans leur dimension linéaire.

En effet, le paradigme chomskyen s'attache essentiellement à expliquer l'acquisition d'un ensemble de comportements langagiers adéquats par le postulat de l'existence de structures abstraites innées et universelles, paramétrées par les stimuli langagiers. Par conséquent, nous considérons que les approches probabilistes doivent, pour accéder au statut de théorie linguistique et pour dépasser le statut de modèle opérationnel dans le cadre de l'ingénierie linguistique, se prononcer sur les conditions de la constitution d'un tel système linguistique à partir des observables. Manning voit dans une variante probabiliste de la théorie de l'optimalité[48] (OT) un cadre pour le développement de modèles linguistiques non catégoriques et non logiques. Il propose un modèle syntaxique reposant sur un principe de satisfaction de contraintes hiérarchisées, rendant mieux compte, d'après ses observations, des

[47] Cette analogie entre construction d'une théorie et induction de règles linguistiques à partir des observables fournit la base de (Finch, 1993).

[48] Voir (Prince & Smolensky, 1993).

pratiques réelles, notamment de la variation dans la production d'énoncés en langue générale. Manning fait reposer l'ensemble de sa conception d'une syntaxe probabiliste, guidée par les observables, centrée sur la variation, sur le modèle probabiliste proposé par (Boersma & Hayes, 2001).

L'enjeu lié à la constitution d'une théorie linguistique non catégorique et non logique est celui de la possibilité de l'existence d'une théorie scientifique non catégorique et non logique. Cette question, examinée par le positivisme comtien dans le domaine de l'épistémologie des sciences, trouve, avec les travaux mentionnés, un début de réponse dans le domaine linguistique.

1.4. Conclusion

Le distributionnalisme classique, outil plus que théorie

Nous avons développé, dans cette première partie, deux approches des faits langagiers partant des observables, que nous avons choisi de qualifier de distributionnalisme classique, d'une part, probabiliste d'autre part. Nous avons tenté de montrer quelle vision des faits linguistiques constituait le fondement de ces deux approches, et quel intérêt présentait l'étude des observables linguistiques, tant dans le cadre d'une construction théorique que dans un cadre applicatif. Nous avons exposé la question sous-tendant l'ensemble des études sur corpus en linguistique, qui est celle de la scientificité : un modèle construit à partir des observables peut-il aspirer au statut de théorie linguistique ?

Pour tenter de répondre à cette question centrale, nous avons exposé les objections du courant générativiste à une science du langage constituée à partir des observables, ainsi que les contre-arguments à ces objections, émanant essentiellement des tenants d'une nouvelle linguistique probabiliste, non catégorique et non logique. Au cours de ce premier chapitre, nous avons tenté de mettre en lumière les motivations tant techniques qu'épistémologiques des tenants de chaque approche. Nous souhaitons ici considérer le débat, opposant

essentiellement les tenants (générativistes) d'une linguistique rationnelle à ceux d'une linguistique guidée par les observables[49], sous un angle plus essentiellement épistémologique.

Les deux positions : harrissienne *versus* chomskyenne, peuvent être vues comme deux démarches scientifiques à part entière, plutôt qu'une démarche empirique opposée à une démarche scientifique. Ces deux positions peuvent être conçues comme un équilibre dynamique, résultant de l'interaction entre les contraintes d'adéquation descriptive, explicative et prédictive, qui reflètent les contraintes premières de complétude *versus* de cohérence. En d'autres termes, il est possible d'adopter une position médiane, concédant aux deux approches le statut de théorie scientifique, en considérant que l'opposition généralement affirmée au sujet de ces deux approches tient à une pondération différente de ces deux contraintes fondamentales.

Ainsi, le générativisme, approche rationaliste, logiciste et principielle, apparaît comme une démarche essentiellement guidée par la contrainte de cohérence, alors que l'approche de Harris, elle, apparaît essentiellement guidée par celle de complétude. Ce qui n'implique pas que chez Harris, la contrainte de cohérence soit absente. En effet, cette contrainte est visible à tous les niveaux d'analyse : rephonémisation, prise en compte de composants longs, postulat (*setting up*) de classes distributionnelles, voire resegmentation morphologique. Jusqu'à la notion de distribution elle-même, qui contient en germe la pondération des deux contraintes complétude/cohérence : elle est définie comme « la somme (totale) des environnements dans lesquels les segments apparaissent ». On a bien là, d'un côté la prégnance des données (les environnements) et, de l'autre, la nécessité de s'en abstraire marquée par l'accent mis sur le caractère cumulatif des distributions. Ce caractère cumulatif appelle d'ailleurs les approches distributionnelles automatiques (ex. : statistiques, réseaux de neurones artificiels), utilisées avec succès dans le domaine de l'ingénierie linguistique.

Au-delà des antagonismes concernant le statut du matériau linguistique, le distributionnalisme peut être vu comme une approche visant à construire un système linguistique avec un minimum de connaissances, alors que le générativisme présuppose un

[49] Nous préférons ce terme à l'adjectif « empirique », souvent employé pour qualifier la démarche consistant à partir des données attestées. Nous considérons, en effet, que la notion d'empirie est trop marquée, dans le domaine épistémologique, comme synonyme d'approche non scientifique.

ensemble maximal de connaissances préalable : des principes universels, génétiquement déterminés, parmi lesquels les stimuli langagier vont sélectionner les plus adaptés.

Points de vue objectif et subjectif pour une science du langage

Au-delà de l'équilibration des deux contraintes de complétude *versus* de cohérence, la question fondamentale que pose l'émergence d'une approche raisonnée des faits langagiers, basée en partie sur la prise en compte de phénomènes de Parole, est celle de l'accommodation d'une visée objective *versus* subjective. Cette question trouve une réponse dans la position exprimée par Saussure.

> L'analyse des unités de la langue, faite à tous les instants par les sujets parlants, peut être appelée *analyse subjective* ; il faut se garder de la confondre avec l'*analyse objective*, fondée sur l'histoire.
>
> (…) Le grammairien est souvent tenté de voir des erreurs dans les analyses spontanées de la langue ; en fait l'analyse subjective n'est pas plus fausse que la « fausse » analogie. La langue ne se trompe pas ; son point de vue est différent, voilà tout. Il n'y a pas de commune mesure entre l'analyse des individus parlants et celle de l'historien, bien que toutes les deux usent du même procédé : la confrontation des séries qui présentent un même élément. *Elles se justifient l'une et l'autre, et chacune conserve sa valeur propre ; mais en dernier ressort celle des sujets importe seule, car elle est fondée directement sur les faits de langue*[50].
>
> L'analyse historique n'en est qu'une forme dérivée. Elle consiste au fond à projeter sur un plan unique les constructions des différentes époques. (…) Le mot est comme une maison dont on aurait changé à plusieurs reprises la disposition intérieure et la destination. L'analyse objective totalise et superpose ces distributions successives ; mais pour ceux qui occupent la maison, il n'y en a jamais eu qu'une.
>
> (Saussure, 1972, pp. 251-253)

Ainsi, la position exprimée par Saussure est celle d'une conciliation des deux points de vue, dans l'optique d'une étude scientifique des faits langagiers. Nous voyons, avec les

[50] Italiques ajoutés.

partisans des approches non catégoriques et non logiques (essentiellement, les tenants de la linguistique cognitive et ceux d'une approche probabiliste des faits langagiers), les conditions de l'émergence d'un nouveau point de vue sur l'étude scientifique des faits langagiers, apte à concilier les deux visées identifiées par Saussure : objective et subjective. Nous voyons également dans une linguistique non catégorique et probabiliste la résolution des difficultés introduites par la fidélité au paradigme catégorique, qui apparaissent dans l'ensemble de la linguistique d'inspiration structurale. La position de Saussure, militant pour un compromis jugé nécessaire entre la démarche objective et la démarche subjective, nous paraît être le reflet d'une telle tension entre l'insaisissable essence des observables et la nécessité d'en poser une. Harris, de son côté, dans son entreprise classifiante, ne cesse d'introduire des moyens de contourner la rigidité, non linguistiquement opératoire, des principes de non contradiction et de tiers exclu, par les procédures d'approximation.

La conciliation des deux points de vue évoqués ci-dessus nous paraît fondamentale, en ce que l'adhésion trop stricte, dans le domaine linguistique, au principe catégorique, a eu pour conséquence une vision normative sur la Langue : la position catégorique sur les énoncés naturels, violant certaines contraintes considérées comme des règles, ne peut être que celle d'un rejet, d'une négation de l'évidence d'un phénomène contredisant le modèle. On comprend facilement la raison d'un tel rejet : intégrer de tels énoncés non canoniques à un modèle catégorique implique une modification de l'ensemble du système construit. Or, la pratique réelle de la langue, ainsi que les applications concrètes (ex. : ingénierie linguistique, pédagogie) semblent bien éloignées de la vision idéale d'une langue constituée d'énoncés dont les condition de bonne formation, les intentions pragmatiques et la charge sémantique sont clairement identifiables. Nous ajoutons qu'on ne peut comprendre autrement la désaffection, de la part de l'ingénierie linguistique, des modèles et de l'approche chomskyenne des faits langagiers, visible dans le recours aujourd'hui massif aux approches statistiques, partielles et locales[51] (*chunking*, cascades de transducteurs) pour la construction de systèmes de traitement automatique des langues (ex. : traduction automatique, recherche d'information, systèmes de question-réponse, systèmes de reconnaissance de la parole).

La faillite du générationalisme dans le champ des applications nous semble être attribuable à un point de vue objectiviste implicite sur les faits langagiers, alors que les performances de ces applications sont tributaires des « détails » linguistiques que sont les

[51] Voir (Abney, 1996 a.), (Vergne, 2002) et (Roche & Schabes, 1997).

hésitations, la violation de certaines contraintes (syntaxiques, pragmatiques, sémantiques) dans la formulation des énoncés, le recours à l'implicite, voire à la communication non-verbale, autrement dit un point de vue subjectif.

Vers une linguistique continue

Nous avons vu plus haut les difficultés posées, tant dans le champ strictement linguistique que dans celui des applications pratiques d'une adhésion trop stricte au paradigme catégorique et logique, autrement dit à une vision discontinue des phénomènes langagiers. Nous voyons dans l'émergence d'une syntaxe probabiliste, alliée aux acquis de la linguistique cognitive, l'acte fondateur d'une science du langage non catégorique, non logique, qualifiée par ses défenseurs de linguistique continue. Le développement d'une nouvelle approche des observables linguistiques permet, non seulement de dépasser les limites pratiques du paradigme catégorique, mais également de faire évoluer l'ensemble du champ des recherches linguistiques.

En effet, la question épistémologique que pose la constitution d'outils théoriques non catégoriques est la suivante : la science est-elle nécessairement logique ? Pour être scientifique, une science doit-elle être nécessairement catégorique ? De façon plus large, le modèle classique des catégories, régies par les lois de non contradiction et du tiers exclu, est-il le seul viable en tant que support d'une science ? La question des rapports entre observables et abstraction, résolue au XIXème siècle par le positivisme comtien sur le plan philosophique, trouve ainsi des échos dans l'émergence d'une linguistique scientifique continue.

Soulignons, cependant, que les objectifs fixés par Maning, Pereira, Abney et autres partisans d'une linguistique continue, restent du domaine du programme, ainsi que la conclusion de (Manning, 2002) le montre.

> There are many phenomena in syntax that cry out for non-categorical and probabilistic modeling and explanation. The opportunity to leave behind ill-fitting categorical assumptions and to better model probabilities of use in syntax is exciting. (…) The frequency evidence needed for parameter estimation in probabilistic models requires a lot more data collection, and a lot more careful evaluation and model building than traditional syntax, where one example can be the basis of a new theory, but the results can

> enrich linguistic theory by revealing the soft constraints at work in language use. This is an area ripe for exploration by the next generation of syntacticians.
>
> (Manning, 2002)

Manning insiste sur le lourd investissement nécessaire à une approche probabiliste des phénomènes langagiers, notamment dans la collecte de corpus équilibrés, représentatifs d'une pratique effective des langues naturelles, seule à même de fournir les données nécessaires à l'élaboration d'une syntaxe probabiliste. En effet, dans les domaines spécialisés, les réalisations menées dans le cadre de l'induction de grammaire à partir de données textuelles[52] ont montré leurs limites en ce qu'elles sont difficilement généralisables à des corpus non spécialisés.

[52] Voir, par exemple (Klein & Manning, 2001), (Soderland, 1997), ou (van Zaanen, 2001).

CHAPITRE 2

Détection d'unités linguistiques et thématiques

pour la recherche d'information

> [I]t is evident that too little is known about either linguistics or information science to justify dogmatic assertions about the relation between them. This conclusion immediately leads to one recommendation: go and find out more about them.
>
> (Spärck Jones & Kay, 1973, p. 200)

Nous avons vu, dans le chapitre précédent, quel pouvait être le statut scientifique d'une étude des phénomènes langagiers, centrée sur leur face observable. À ce sujet, nous avons évoqué un cadre théorique et méthodologique émergent, intégrant des phénomènes tels que la variation intra et interindividuelle dans la production d'énoncés relevant d'un domaine de spécialité : la théorie de l'optimalité, dans sa variante probabiliste[1]. Nous avons également vu comment ce cadre théorique et méthodologique permettait de concilier les deux points de vue identifiés par Saussure : le point de vue objectif (collectif), c'est-à-dire le domaine de la Langue, et le point de vue subjectif (individuel), c'est-à-dire le domaine de la Parole. La Recherche d'Information (désormais RI) partage avec les études linguistiques sur corpus un même objet d'études : les observables linguistiques. Ces deux domaines d'étude partagent également la nécessité de réconcilier les deux points de vue sur ces observables : ainsi, en RI, la détermination de la valeur (fonction) informative d'un document peut être vue comme relevant d'une conciliation entre point de vue objectif et individuel, de façon analogue à la détermination de la valeur (fonction) d'un élément dans le domaine linguistique.

[1] (Boersma & Hayes, 2001).

La RI, comprenant différentes sous-tâches spécialisées, est centrée sur les documents, un terme englobant aussi bien les documents textuels que multimédias (ex. : archives sonores). Elle s'est constituée en tant que science de l'information (*information science*), à partir des systèmes d'information traditionnels, tels que bibliothèques ou centres de documentation, par l'adoption de normes et de procédures standardisées pour l'archivage et la recherche de documents pertinents (ex. : une liste d'ouvrages correspondant à des critères définis par un utilisateur du système). Ces procédures standardisées visent essentiellement à :

- obtenir une description abrégée du contenu des documents lors de leur archivage ;
- apparier une requête d'utilisateur du système d'information avec les descriptions de contenu des documents archivés, afin de fournir une liste la plus exhaustive possible des documents susceptibles de combler le besoin en information de cet utilisateur.

On comprend, dès lors, que si la langue dans laquelle sont élaborés les documents est envisagée comme un vecteur du contenu de ces documents, alors l'étude de ce vecteur constitue une priorité, notamment dans une optique d'automatisation des processus d'archivage et de recherche des documents pertinents. Ce lien entre linguistique de corpus et recherche d'information a donné naissance, dès les années 1960, à de nombreux programmes de recherche et de développement, dans l'optique d'un apport mutuel entre les deux disciplines citées. On retrouve ainsi la trace de l'émergence d'une recherche d'information basée sur des études linguistiques aussi bien dans (Bar-Hillel, 1964), qui constitue un examen critique des pratiques dans le domaine de la recherche d'information, que dans (Coyaud, 1972) et (Spärck Jones & Kay, 1973), consacrés aux relations entre linguistique et recherche d'information[2]. Ces trois ouvrages serviront de base au présent chapitre, consacré à l'application de procédures de découvertes d'unités linguistiques, suivant les principes du distributionnalisme, dans le cadre de la recherche d'information, en raison de l'éclairage

[2] Cette alliance entre analyses linguistiques et recherche d'information, envisagées sous l'angle d'une automatisation n'est rendue possible que par l'émergence d'une linguistique formelle et les premières expériences en traduction automatique, autrement dit la naissance du TALN.

historique qu'ils apportent sur un domaine cumulant les difficultés liées à l'étude des corpus et celles liées à l'élaboration de représentations abrégées du contenu des documents[3].

Dans ce chapitre, nous nous intéressons à la recherche d'information intégrant des analyses linguistiques, en tant que celles-ci reposent sur des principes distributionnels tels que nous les avons évoqués dans la partie consacrée au distributionnalisme discontinu et continu. Nous tenterons de donner, dans un premier temps, un cadre à la notion d'information, puis nous nous pencherons sur les principes de l'indexation par unités thématiques. Ceci nous amènera à examiner les principes de l'indexation automatique par descripteurs de contenu extraits des documents. Nous évoquerons donc brièvement les principes généraux de l'indexation et de la recherche de documents, en tant qu'ils reposent sur une conception distributionnaliste[4] de la valeur informative.

Nous soulignerons les insuffisances, constatées de façon unanime, des approches les plus courantes, restant dans le domaine du mot typographique, pour aborder, dans un deuxième temps, les principes d'une Recherche d'Information basée sur une analyse linguistique automatisée. Nous examinerons tout d'abord quelques approches, basées sur un distributionnalisme discontinu, visant à repérer dans les documents des marqueurs thématiques, considérés non plus comme des mots-clés mais comme des unités lexicales complexes. Enfin, nous évoquerons des approches relevant du distributionnalisme continu, visant des applications en RI, notamment par le biais des techniques d'extraction de collocations[5].

---

[3] D'autres indices témoignent de la vitalité de ce domaine émergent dans les années 1960, ainsi que des liens étroits, en France, entre linguistique centrée sur les corpus, linguistique formelle, et recherche d'information : (Gross, 1966 ; Gross, 1967), par exemple, ainsi que la création du Laboratoire d'Automatique Documentaire et Linguistique (LADL).

[4] À savoir : la valeur informative d'un élément (ex. : un mot) dépend de ses contextes d'occurrence (phrase, document).

[5] Les collocations sont des expressions constituées de plusieurs mots, présentant des contraintes proches de celles des mots composés : « Collocations of a given word are statements of the habitual or customary places of that word », (Firth, 1957, p. 181).

2.1. La Recherche d'Information

Nous empruntons à (Bar-Hillel, 1964) une définition du cadre de la recherche d'information, s'inscrivant dans une réflexion critique sur les pratiques du domaine au sein des systèmes d'information existants, examinant l'apport d'une automatisation des procédures standardisées d'archivage (ou d'indexation) et de recherche de documents pertinents. L'auteur a consacré son ouvrage à une réflexion sur les spécificités de l'activité de recherche d'information, tant dans le cadre des systèmes manuels qu'automatiques. Ce faisant, il a entrepris de fournir à l'ensemble du domaine des définitions, des spécifications et des contraintes, tant dans une optique de fixer la terminologie employée que dans celle de recenser les méthodes et les techniques les plus appropriées, à ses yeux. Le cadre de l'ouvrage cité dépasse cependant la simple définition conceptuelle d'un domaine émergent à l'époque de sa parution, la recherche d'information automatisée : Bar-Hillel évoque également les perspectives du domaine, et pose, par exemple, dès les années 1960, la question « Is information retrieval approaching a crisis ? ». L'auteur peut être considéré comme un des théoriciens d'un domaine émergent : la mécanisation (l'automatisation) des systèmes d'information, regroupées sous la dénomination de sciences de l'information (*Information Science*).

L'auteur définit l'objet de la recherche d'information comme visant la réponse à la question posée ci-dessous.

> Assuming that there exists somewhere a body of recorded knowledge – in technical terms, a collection of documents – and assuming that someone has a certain problem for the solution of which this collection might contain pertinent material, how shall he decide whether there are in fact documents in this collection that contain such pertinent material, and, if so, how shall this material be brought to his attention?
>
> (Bar-Hillel, 1964, p. 331)

Autrement dit, la recherche d'information suppose :

- une collection de documents existante, ou en cours de constitution, dans laquelle des connaissances sont enregistrées sous la forme de textes en langue naturelle, principalement[6] ;
- un principe de représentation du contenu (les connaissances) véhiculées par chaque document ;
- un principe d'appariement entre les représentations de contenu existantes et une demande d'information émanant d'un individu ;
- des moyens de présentation du résultat satisfaisant le besoin en information de l'utilisateur.

La recherche d'information est donc le lieu d'une tension entre une représentation individuelle et subjective d'un besoin en information et une représentation collective, à visée objectivante de la réponse à ce besoin. Le nécessaire ajustement de ces deux représentations pose, de façon empirique, la question de la pertinence[7], qui vient de celle, plus fondamentale, de l'association d'un contenu (une somme de connaissances) à un ensemble de formes linguistiques.

2.1.1. Notion d'information

Que recouvre le terme générique d'*information* ? En effet, les théoriciens de la notion d'information, élaborée dans le cadre de l'ingénierie de la transmission des signaux (Shannon, 1948), ainsi que leurs prédécesseurs, n'ont eu de cesse de distinguer l'information véhiculée par les suites de caractères d'un document à transmettre (ex. : un télégramme), de son contenu (ex. : l'annonce d'un événement). Nous aborderons donc deux définitions de la notion d'information, afin de préciser la valeur du terme ainsi que le cadre méthodologique qui en découle, dans le cadre d'une activité de recherche d'information.

[6] La recherche d'information sur des documents multimédias constitue un domaine de recherche à part entière, que nous n'évoquerons pas ici.

[7] Quels critères permettent de garantir qu'un document est pertinent ? Cette pertinence est-elle absolue, ou relative ?

2.1.1.1.Définition quantitative

La définition quantitative de la quantité d'information repose sur l'estimation de la probabilité d'occurrence d'une classe d'événements donnés. La Théorie de l'Information[8], ou, pour reprendre les termes de (Bar-Hillel, 1964)[9] la Théorie de la Transmission des Signaux (*Theory of Signal Transmission*) définit un cadre formel pour la quantification de l'information véhiculée par un signal. L'évaluation de la quantité d'information apportée par un signal est liée à l'adoption d'un processus de codage optimisé pour l'information à véhiculer, dans des conditions où la transmission est susceptible de ne pas être parfaite (ex. : un câble télégraphique). Le signal à transmettre peut consister, par exemple, en un message, composé de caractères pris dans un alphabet.

Considérons l'exemple suivant :

- soit un message *X* à transmettre, composé d'un seul caractère, A ou B. La variable *X*, dans le cas d'une répartition aléatoire, peut donc prendre la valeur A ou B. Dans le cas présent, la probabilité que *X* = A, notée p(A), est la même que celle que *X* = B, notée p(B), c'est-à-dire p(A) = p(B) = ½ = 0,5 ;
- dans ce cas, l'incertitude liée à la composition du message est la même quelque soit le message. Cette incertitude est mesurée par la notion d'entropie *H* calculée sur l'événement *X*, donnée par la formule: *H(X)* = - 0,5 Log2(0,5) = 0,5 = 0,5.

Le même raisonnement peut s'appliquer sur d'autres éléments que les caractères : les syllabes, les mots, ou encore les phrases[10], considérés comme des événements présentant une certaine probabilité d'occurrence. Le dénombrement de ces différents types d'événements permet d'associer à chaque événement $x_1, x_2, \ldots x_n$ (caractère, syllabe, mot, phrase) les

[8] (Shannon, 1948).

[9] Voir notamment (Bar-Hillel, 1964, pp. 288-290), pour une présentation historique de la notion d'information et la nécessité de distinguer entre quantité d'information et contenu associés à un document.

[10] On peut, en effet, envisager de coder l'information au niveau des mots, voire des type de message les plus fréquents (ex. : félicitations pour un heureux événement), ainsi que cela se pratiquait chez les compagnies télégraphiques (Bar-Hillel, 1964, p.278).

probabilités $p_1, p_2, \ldots p_n$, que nous désignerons par p_i. L'entropie H, c'est-à-dire l'incertitude liée à la survenue d'un événement i est donnée par la formule[11] :

$$H(p_i) = - p_i \log 2 \, p_i.$$

Définition 1 : entropie associée à la survenue d'un événement x_i

L'entropie associée à l'ensemble p des probabilités est donnée par la formule :

$$H(p) = - \sum_i p_i \log 2 \, p_i.$$

L'entropie, ainsi que son inverse la néguentropie, usuellement confondue avec la notion de quantité d'information, sont mesurées en *bit*, ou unités binaires. Une diminution de l'entropie associée à des événements est généralement perçue comme décrivant le passage d'un état aléatoire (où tous les événements sont équiprobables) à un état d'ordre relatif. Cette mesure est donc généralement considérée comme caractérisant l'organisation des systèmes (ensembles d'événements).

Cette définition ne s'applique que dans le cadre de l'observation d'événements distincts (des suites de caractères, de mots), quantifiables, formant un signal, dans la perspective de transmettre de façon optimale (rapidité de la transmission, intégrité du signal transmis) ce signal *via* un canal susceptible d'être bruité (une ligne télégraphique). Cette mesure de la réduction de l'incertitude quant à la survenue d'un événement, pris parmi un ensemble d'événements possibles, a cependant connu une forte popularité en dehors du cadre strict de l'ingénierie des télécommunications. En effet, par l'élaboration d'une métrique de la complexité d'un signal, constitué d'événements quantifiables, Shannon a fourni au domaine des sciences humaines, par exemple, les outils quantitatifs qui leur faisaient jusque-là défaut. Ainsi, en psychologie expérimentale, il est possible d'évaluer la complexité d'une expérience (ex. : reconnaître une forme) en dénombrant les événements possibles. Cette quantification permet ainsi de prédire des différences de performance aux différentes expériences, en fonction du nombre de décisions à prendre, par exemple.

[11] Les justifications du recours au logarithme de base 2 se trouvent dans (Shannon, 1948), signalons simplement qu'elle est liée à l'adoption du *bit* comme unité d'information, pouvant prendre deux valeurs (0 ou 1).

Au-delà du domaine des sciences humaines, tous les domaines d'activité manipulant de l'information (émission, réception, stockage, codage), en tant que séquences d'événements possibles, ont repris et développé la notion de quantité d'information associée à un signal. Bar-Hillel note, par ailleurs, que la popularité de la notion d'information est principalement liée à la confusion, entretenue par la plupart des auteurs du domaine de la Transmission des Signaux (Hartley, Shannon & Weaver, ou encore Wiener), entre quantité d'information véhiculée par une séquence d'événements parmi un ensemble d'événements possibles, d'une part, et contenu (représentations sémantiques, pragmatiques) véhiculé par un signal (ou message), d'autre part. Cet auteur propose d'ailleurs une Théorie de l'Information Sémantique afin de distinguer information et contenu[12].

2.1.1.2.Définition fonctionnelle

Nous l'avons vu, d'après Bar-Hillel, la définition quantitative de l'information n'est pas une définition de la fonction informative d'un document, il est donc nécessaire d'envisager une définition fonctionnelle de l'information. Suivant (Bar-Hillel, 1964), on peut affirmer qu'aucune adéquation entre entropie (ou néguentropie) et contenu véhiculé par un message n'est possible : il faudrait, pour cela, énumérer les événements possibles en termes de contenu, ce qui reviendrait à vouloir dresser une liste exhaustive de tous les événements possibles. Ainsi, pour reprendre l'exemple de Bar-Hillel, en se limitant au domaine des télégrammes, il serait nécessaire, pour représenter le contenu d'un message/événement par rapport à l'ensemble des messages/événements possibles, de dénombrer :

- les heureux événements, tels que naissances, mariages, anniversaires, réussite à un examen etc. ;
- les événements malheureux, tels que décès, ruptures, échecs etc. ;
- les événements ni heureux ni malheureux, tels que bonne réception d'un colis, réservation d'un billet de train etc.

De toute évidence, un tel dénombrement est une entreprise utopique : de même que l'ensemble des phrases possibles est un ensemble ouvert, potentiellement infini, l'ensemble des événements du monde possibles ne peut être décrit de façon exhaustive, à moins d'imposer une norme, ne sélectionnant qu'un sous-ensemble fini de ces événements.

[12] Voir le chapitre 15 (Bar-Hillel, 1964).

Le rejet d'une définition quantitative, à visée objective, du contenu véhiculé par un signal (ex. : document) vient également de la prise de conscience de la subjectivité inhérente à tout processus de communication humaine : ainsi, des linguistes cognitivistes tels que Lakoff postulent que la compréhension d'un message, donc son contenu et sa fonction informative pour le destinataire, dépend de la structure cognitive de ce destinataire, et non pas de la seule valeur de vérité du message, au sens logique. Autrement dit, la compréhension d'un signal, dans le cadre d'une communication humaine, n'est pas qu'un simple codage/décodage d'un contenu par le biais d'une langue naturelle, mais plutôt une négociation, un processus d'équilibration intégrant les attentes, la représentation du monde et les connaissances tant du locuteur que du destinataire.

Dans ce cadre conceptuel, deux événements ont la même charge informative s'ils remplissent la même fonction. Cette type de définition fonctionnelle a servi de base, dans le domaine de la linguistique structurale, à l'essor de la phonologie[13], par exemple, dont nous avons montré au chapitre précédent quelle part d'abstraction par rapport aux données observables elle supposait, dans l'optique d'une linguistique de la Langue. Cette définition fonctionnelle peut également servir de base au domaine de la recherche d'information. Si on y ajoute la dimension individuelle, une définition fonctionnelle de l'information peut être exprimée comme suit.

> Deux éléments apportent la même information si, pour un individu donné, à un moment donné, ils remplissent la même fonction par rapport à son besoin en information

Définition 2 : une définition fonctionnelle de l'information

Une fois posée cette définition, reste à définir la notion de fonction. On peut adopter une définition « simple » de la fonction informative : un élément de contenu répondant à un besoin en information. Cette définition n'a de simple que l'apparence, puisqu'elle implique de définir le besoin en information d'un utilisateur de système d'information, effectuant une tâche de recherche à un moment donné, dans un contexte donné. Autrement dit, aucune

[13] Dans cette conception, deux événements (i.e. acoustiques) de nature distincte, observés dans des contextes similaires peuvent être considérés comme deux matérialisations d'une même unité/fonction.

caractérisation absolue n'est possible pour la notion de fonction informative. Ainsi, dans le cadre du filtrage d'information, que nous détaillons plus loin, nous prendrons comme définition approchée d'un besoin en information la caractérisation succincte donnée par les thèmes pris en compte par rapport à un ensemble de documents traitant du domaine financier (ex. : thème 18, stratégie des entreprise, thème 19, cession-acquisition de sociétés). Comme définition approchée de la fonction informative, nous considérerons l'ensemble de phrases (ou de parties de phrases) formant l'ensemble des documents traités, associées à un thème/besoin en information donné.

Cette conception fonctionnelle de l'information se retrouve, par exemple, chez (Michel, 1999), consacré à la mise en œuvre de protocoles d'évaluation d'une application informatique de recherche d'information, pour laquelle la dimension individuelle est primordiale[14]. Cette centration sur l'utilisateur final se traduit, en effet, par une nécessaire remise en cause d'une conception objective de l'information.

> Le projet Profil-Doc (…) part du constat que tous les documents ne sont pas pertinents au même titre pour des utilisateurs différents, même si leur contenu est en relation avec la question posée au système. (…) [L']utilisateur, face à un système en texte intégral qui lui fournit généralement trop d'information, va développer **une stratégie de recherche empirique**. Toutes ces stratégies ont deux caractéristiques : elles portent sur des critères (la forme, le support, le style, le domaine de compétence de l'auteur, …) autres que le contenu du document ; elles sont très fortement individualisées et permettent une personnalisation de la recherche. (…) [C]es propriétés permettront de sélectionner un corpus « personnalisé » suivant les caractéristiques de l'utilisateur, corpus sur lequel portera la question [la requête soumise au système].
>
> (Michel, 1999, p.16)

[14] Le projet Profil-Doc, vise à développer une interface dite de « filtrage » entre une application de recherche d'information largement diffusée (Spirit), et un ensemble d'utilisateurs. Chaque utilisateur est identifié par un profil, spécifiant quelles unités documentaires il intègre à sa stratégie de recherche d'information, qui servira de base à une présélection (filtrage) de documents parmi les réponses fournies par le système.

Dans les termes de (Michel, 1999), nous considérerons des « unités documentaires » au niveau syntaxique, en l'occurrence des parties de phrases, associées à un ou plusieurs thèmes, c'est-à-dire un ou plusieurs besoins en information[15].

Comprise dans les termes présentés ci-dessus, une activité de recherche d'information vise donc à identifier des unités documentaires au niveau syntaxique, en se basant sur une étude sur corpus préalable. Cette étude préalable vise à déterminer le fonctionnement syntaxique, au sens large, des unités documentaires recherchées : types d'unités syntaxiques, choix lexicaux, contraintes de sélection entre unités, lien entre unités documentaires et fonction informative.

2.1.2. Les marqueurs thématiques en Recherche d'Information

Quelque soit l'application, le contexte d'utilisation, la nature des procédures (manuelles, automatiques, semi-automatiques) visant à archiver des documents de façon à ce que des utilisateurs puissent retrouver ceux qui les intéressent, l'objectif central de la RI est de trouver une représentation abrégée du contenu desdits documents, ainsi que des requêtes des utilisateurs, et d'apparier ces deux objets de façon à choisir le document de la base le plus proche de la requête. Autrement dit, toutes les recherches en RI tendent vers le même but : trouver les bons descripteurs de contenu, ou termes associés de façon systématique à un thème donné, jouant le rôle de marqueurs de thème. C'est l'objet de l'extrait ci-dessous, dans lequel les « indices » (*clues*) mentionnés par Bar-Hillel doivent être compris comme un terme générique pour la notion utilisée ici de descripteurs de contenu, ou de marqueurs thématiques.

> The obvious general solution to our main problem, **how to select out of a given collection of documents those documents that are relevant to a given topic** (…) is to assign to each document a clue, or rather a set of clues, and to assign likewise to each topic a set of topic-terms, in such a way that by comparing the set of topic-terms with the

[15] (Michel, 1999) distingue en effet une structure générale dans les documents, de laquelle différentes unités documentaires participent. Ces unités documentaires sont de nature diverse (ex. : éléments typographiques, syntaxiques), leur charge informative dépend de leur fonction, au sens présenté ici : elle dépend d'un utilisateur particulier et de son besoin en information.

> set of clues a decision as to the (probable or possible) relevance of the document can be reached.
>
> (Bar-Hillel, 1964, p.335)

Le principe d'appariement, évoqué plus haut, entre le besoin en information d'un utilisateur du système d'information et les documents archivés susceptibles de satisfaire ce besoin, repose donc sur un appariement entre les indices assignés à chaque document et les termes associés à un thème (*topic-terms*). Cette définition a le mérite de résoudre partiellement la question de la pertinence, mentionnée en introduction au présent chapitre : est considéré comme (probablement ou possiblement) pertinent, par rapport à une requête d'utilisateur, tout document dont les « indices » correspondent aux « topic-terms » contenus dans la requête. Si elle résout – au moins partiellement – la question de la pertinence, cette définition, reprise par l'ensemble des approches dans le domaine, ne résout pas celle du choix des « topic-terms » ni des « indices » associés aux documents. En effet, pour qu'il y ait des termes associés à des thèmes, il faut, d'une part, qu'un ensemble de thèmes (ex. : un thesaurus, une ontologie) ait été identifié et défini comme couvrant l'ensemble des documents archivés. D'autre part, il faut qu'un principe systématique associant à des documents traitant du même thème les mêmes « indices », ou descripteurs de contenu, eux-mêmes associés aux « topic-terms ».

Deux approches dans l'assignation de marqueurs thématiques à des documents, ou processus d'indexation, sont possibles : une approche manuelle, basée sur un langage de description, et une approche automatique, basée sur l'extraction de marqueurs thématiques à partir des documents à indexer[16].

2.1.2.1.Indexation manuelle et marqueurs thématiques

En indexation manuelle, on trouve essentiellement deux types de descripteurs, correspondant à deux types d'indexation :

- l'indexation libre ;

[16] Ces deux approches ne sont pas nécessairement exclusives, cependant, autant l'approche manuelle est susceptible d'utiliser des marqueurs thématiques tirés des documents à indexer, autant l'approche automatique ne peut se substituer à l'opérateur humain dans le processus de description du contenu d'un document par un langage normalisé. En effet, cette opération équivaut, en complexité, à un processus de traduction.

- l'indexation contrôlée.

Dans le premier cas, les descripteurs peuvent être pris dans l'ensemble des mots du lexique d'une langue. Il s'agit habituellement de substantifs, représentant le ou les thèmes principaux abordés dans les documents. L'indexation libre n'est efficace que dans le contexte d'un domaine émergent, pour lequel n'existent pas de dénominations faisant l'unanimité. On le voit, le risque de perte d'information est élevé : des descripteurs pris dans un domaine trop spécialisé, ou inattendu, risquent de ne jamais pouvoir être appariés avec des requêtes d'utilisateurs.

L'indexation contrôlée et l'indexation mixte sont les plus répandues : dans le cas de l'indexation contrôlée, le choix des descripteurs se fait dans un ensemble fermé de termes, ayant fait l'objet d'un consensus, souvent par le biais d'une procédure de standardisation[17] : les langages dits de description de contenu. L'indexation contrôlée n'est pas exempte de difficultés : des descripteurs consensuels ne sont opérationnels que s'ils restent suffisamment discriminants tout en étant génériques, ce qui amène directement à des problèmes ontologiques. L'indexation mixte tente de concilier les avantages des deux techniques, en limitant le recours aux descripteurs libres aux champs les plus subjectifs.

Le processus d'indexation tel que décrit sommairement ci-dessus ne va pas sans rencontrer des difficultés, constatées de façon unanime, qui ont trait à une variation incontournable dans les points de vues adoptés par les opérateurs humains lors de l'indexation. La condensation du contenu grâce à un langage d'indexation pose des problèmes d'ordre pratique[18], mais également des problèmes plus théoriques, ayant trait aux points abordés dans le chapitre précédent, à savoir essentiellement des problèmes de structuration du monde (i.e. les concepts véhiculés par les documents), donc des choix de catégorisation.

> The major feature of the conventional information retrieval process is the replacement of a long and complex linguistic entity, the entire document, by a greatly abbreviated description. The use of such a summary is not solely a consequence of practical

[17] Les termes servant à l'indexation sont souvent tirés des langues naturelles, cependant des systèmes reposant sur des termes non naturels ont également été mis en œuvre (ex. : la Classification Décimale Universelle).

[18] Optimisation du processus d'indexation, choix d'un langage d'indexation, ou encore normalisation et standardisation.

> constraints on the amount of material that can be stored and inspected in searching. *It may also be desirable in principle since the function of the description is to bring out the essential features of the document*[19].
>
> (Spärck Jones & Kay, 1973, p.47)

Les auteurs caractérisent l'activité de recherche d'information, dont font partie les processus d'indexation, comme un moyen de souligner les propriétés essentielles (*essential features*) des documents traités. La question de l'essentiel *versus* l'accidentel est bien un problème de catégorisation, dont nous avons vu qu'il dépendait du modèle adopté, de façon implicite le plus souvent, dans le processus de structuration des classes d'objets (i.e. des classes de documents). En ce sens, l'usage du terme « descripteur » nous paraît trompeur : les éléments choisis pour représenter le contenu d'un document sont bien plus qu'une simple description, ils constituent forcément une prise de décision par rapport à l'appartenance du document à une classe donnée.

2.1.2.2.La variation dans l'indexation humaine

Comme nous l'avons vu plus haut, le processus d'indexation des documents fait appel à des langages d'indexation, plus ou moins proches du langage naturel. Or, la description du contenu d'un document, autrement dit la traduction d'un ensemble de formes d'une langue naturelle vers un ensemble de formes d'un langage contrôlé constitue une analyse de ce contenu. Autrement dit, ce processus correspond à la mise en œuvre d'une visée objectivante, à partir d'un ensemble de formes linguistiques observables, produites dans un contexte particulier, par un individu (ou groupe d'individus) particulier, à destination d'un public particulier (ex. : spécialistes, étudiants).

Nous avons montré, dans le chapitre précédent, quelle tension, entre fidélité aux données et nécessaire abstraction, ce type de processus d'analyse impliquait. Une des conséquences de cette tension est l'extrême variation de l'indexation réalisée par des opérateurs humains. Ce phénomène est mis en évidence par les expériences, relatées dans (Coyaud, 1972), visant à évaluer l'influence de ce qui est dénommé « variation de point de

[19] Italiques ajoutés.

vue chez les indexeurs » sur l'indexation de documents. Coyaud fait remarquer[20] que : « Une des causes essentielles d'échecs, en recherche documentaire, réside dans le fait que l'analyse (humaine) manque de régularité et de cohérence ».

Coyaud évoque les expériences réalisées en indexation[21], visant à comparer les choix d'indexation opérés par des indexeurs humains. L'ensemble de ces expériences se basait sur des documents déjà indexés au préalable, pris parmi un ensemble fermé (quelques centaines de documents), que des indexeurs devaient réindexer. Les dimensions suivantes ont été abordées : variation inter et intra-individuelle, effets de la fréquence d'occurrence sur le choix de mots clés pris comme descripteurs, et comparaison entre procédure manuelle et automatique (statistique). Les résultats de ces expériences peuvent être synthétisés comme suit :

- les décisions de sélection (points de vue) évoluent au cours du temps pour un même opérateur, dans une proportion analogue aux différences observées entre deux opérateurs différents ;
- l'accord entre indexeurs constitue l'exception plutôt que la norme ;
- la fréquence d'occurrence ne semble pas avoir d'influence sur le choix des descripteurs ;
- les différences entre les décisions de sélection opérées par des moyens automatiques (statistiques) et celles opérées par des humains sont comparables à celles constatées entre opérateurs humains.

Coyaud voit dans la variation associée aux indexations humaines un argument en faveur de processus complètement automatisés, si possible basés sur la prise en compte de la dimension linguistique des documents traités. Nous voyons, de notre côté, dans cette variation la tension entre deux modes de représentation de la structure du monde telle que perçue au travers des bases de documents : une conception à visée ontologique, objectivante, selon le modèle scientifique classique, et une conception dans laquelle les catégories ont des frontières perméables (ex. : un document traite *plutôt* d'un thème que d'un autre), où la valeur des

[20] (Coyaud, 1972), p. 133.

[21] Nous renvoyons le lecteur (Coyaud, 1972) pour les références exactes et les détails de chaque expérience.

éléments du système est sensible au contexte, aux attentes, aux effets d'amorce, induisant des « points de vue » changeants.

Autrement dit, nous reconnaissons la part de subjectivité propre à chaque opérateur d'un système d'information comme une donnée à prendre en compte. Notre expérience du sous-domaine du filtrage d'information nous pousse à considérer cette subjectivité comme nécessaire à cette activité de recherche d'information particulière, de ce fait nous la percevons plus comme la manifestation d'une expertise que comme un effet de bord néfaste. Pour cette raison, nous nous démarquons de la vision de Coyaud de l'indexation opérée sur des bases linguistiques comme seule garante d'une objectivité que nous qualifions d'artificielle, pour proposer une conception de la recherche d'information, et plus particulièrement du filtrage d'information, prenant en compte les relations de dépendance existant entre les éléments inclus dans des structures linguistiques particulières.

2.1.2.3. Indexation automatique et sélection de descripteurs de documents

En indexation automatique, le concept de descripteur libre n'est pas applicable : le lexique dont disposent des systèmes automatiques est, par nature, limité. De plus, là où la subjectivité peut être tolérée, dans la mesure où elle reflète l'expérience du domaine des opérateurs humains, une prise d'initiative par un système automatique semble difficilement acceptable, en l'état actuel des techniques. Le domaine de l'indexation automatique se caractérise donc par une volonté de prendre le minimum d'initiatives, donc de risques, ce qui se traduit par le recours exclusif aux « observables », autrement dit les mots présents dans les documents à indexer[22].

À propos du processus d'indexation, nous citons ci-dessous (Spärck Jones & Kay, 1973), qui vise à dresser un bilan de l'interdisciplinarité dans le domaine de la recherche d'information, entre linguistique, et plus particulièrement linguistique de corpus, et indexation automatique de documents. Le contexte historique de parution de l'ouvrage[23] n'enlève rien, à

[22] Ainsi, l'indexation automatique d'un document ne prend généralement pas en compte les relations connues (ex. : synonymie simple) entre les termes d'un document donné et ceux d'autres documents, voire des parties du même document.

[23] Une période sombre pour la linguistique informatique, après la remise du rapport ALPAC au congrès américain, remettant en cause les efforts entrepris dans le domaine de la traduction automatique.

nos yeux, aux remarques faites par les deux auteurs, familiers des grands projets en recherche d'information sur des bases linguistiques[24].

> The conventional view of the documentation process is that it involves "the analysis of each document's content, a formulation of this content in a set of descriptors, and an organization of descriptors such that enquirors can match their search request and not miss any documents relevant to that request [Hutchins, 1967]."
>
> (Spärck Jones & Kay, 1973, p.45)

Nous avons vu plus haut que le choix des descripteurs, en indexation humaine, dépendait du type d'indexation. Dans le cas de procédures automatisées, les descripteurs d'un document donné, c'est-à-dire l'ensemble des termes inscrits dans la base d'indexation, sont choisis uniquement parmi ceux présents dans le document. On le voit, cette situation est propre à l'indexation automatique : en indexation manuelle, il n'existe pas de lien nécessaire entre les termes d'un document et les descripteurs. Tout l'effort porte donc sur le choix de ces descripteurs, à partir des mots typographiques observables, considérée comme une population, en termes statistiques, dont les occurrences vont être considérées comme autant d'événements. Les principes directeurs de l'indexation automatique sont à la croisée de deux disciplines : la statistique, notamment les techniques d'échantillonnage, et l'étude des distributions des événements langagiers. Dans cette optique, le contexte d'occurrence des descripteurs retenus est le document, dont la segmentation est réalisable sur des critères objectifs (ex. : marques de début et de fin de document, marques de paragraphes), contrairement aux délimitations linguistiques, pour lesquelles aucun critère objectif, non dépendant de l'application, du domaine et de l'approche n'est disponible.

Dans les approches les plus répandues, la sélection des descripteurs se traduit essentiellement par l'élimination des mots jugés peu représentatifs du contenu du document. La représentativité d'un terme, dans cette perspective, ne peut être basée que sur sa présence ou son absence au sein d'un document, et plus précisément sa fréquence d'occurrence dans ce document. Le principe de sélection communément admis dans le domaine se fonde sur les

[24] Signalons que Spärck Jones est l'un des organisateurs des conférences d'évaluation TREC (Text REtrieval Conference), que nous présentons plus loin.

recherches d'auteurs tels que Zipf ayant montré, quelque soit la langue, que les mots d'un document peuvent être classés en fonction de leur fréquence d'occurrence qui tend à suivre une loi générale[25]. Ainsi, la sélection des descripteurs associés à un document donné ne prendra en compte qu'une partie de la population des mots des documents : ceux dont la fréquence d'occurrence est comprise entre un seuil maximal, au-dessus duquel les termes sont trop fréquents pour être pertinents (ex. : les mots dits grammaticaux, tels que les déterminants ou les prépositions), et un seuil minimal en dessous duquel on considère n'avoir affaire qu'à des *hapax legomena*, dont le faible taux d'occurrence amène à les considérer comme des accidents[26]. Ce principe de sélection des mots d'un document, en fonction d'une relation supposée entre fréquence d'occurrence et pertinence, est l'objet du passage ci-dessous[27].

> The general assumption behind the extraction of words on a statistical basis, whether these are to serve as entry words to a dictionary or as terms, is that conspicuous words are significant content indicators. *It is not necessary to make any more concerted attempt to discover what a document is about, because a document wears its heart on its sleeve, and any nontrivial word that occurs sufficiently frequently must be a valid content indicator, or it would not be used so often*[28].
>
> (Spärck Jones & Kay, 1973, p.134)

Cet extrait donne la philosophie générale sous-tendant le recours aux approches statistiques en indexation automatique des documents. La difficulté principale, dans ces approches, étant de décider ce qui constitue un mot trivial d'un mot porteur de sens. Soulignons que, tout comme c'était le cas au moment de la parution de l'ouvrage cité ci-

---

[25] Voir, par exemple, (Zipf, 1945), qui a servi de fondement théorique aux approches dominantes en indexation automatique de documents. Pour une discussion des expériences de Zipf, voir (Herdan, 1964), ainsi que (Li, 1992).

[26] Cette généralisation a souvent fait l'objet de critiques, notamment de la part d'auteurs tels que Coyaud, militant pour une approche linguistique de l'indexation des documents.

[27] Signalons au passage combien cette approche se distingue de la définition fonctionnelle de l'information donnée plus haut, centrée sur la perception de cette fonction pour un utilisateur donné. Dans ces approches statistiques, la fonction, donc la pertinence d'un ensemble de mots, est associée à leur fréquence d'occurrence.

[28] Italiques ajoutés.

dessus, la plupart des approches statistiques en indexation automatique restent cantonnées au domaine du mot typographique. Cependant, on comprend tout l'attrait de ces approches, résumé par les auteurs : « looking only at the surface of a document, it is clear that prominent physical features reflect important features of its content, so we need not examine the latter directly ». Les approches statistiques apportent, en effet, une réponse pragmatique et indirecte à une difficulté fondamentale : évaluer le contenu d'un document de façon automatique, de la façon la plus objective possible.

On peut voir une certaine parenté entre l'approche visant à déterminer, de façon automatique, la fonction informative de mots pris comme marqueurs thématiques, à partir des documents à indexer, et l'approche décrite dans le premier chapitre, visant à déterminer la fonction linguistique d'éléments pris dans un échantillon de langue, à partir de leur comportement observable. Dans les deux cas, les approches centrées sur les données linguistiques, ou corpus, visent à répondre à des besoins concrets, tout en abordant nécessairement des questions théoriques primordiales, liés à la généralisation de règles par induction à partir des observables de l'échantillon, au lien entre contenu et formes linguistiques, ou encore à la tension entre une somme de représentations individuelles et une représentation collective objectivante.

2.1.3. Limites des approches basées sur des descripteurs en Recherche d'Information

En raison du coût que représente une indexation manuelle, la plupart des systèmes d'information manipulant des bases hétérogènes de documents adoptent des approches automatisées, partiellement ou complètement. Le degré d'automatisation dépend essentiellement de la taille et de la diversité de la base à indexer, ainsi que de la disponibilité d'opérateurs humains et de leur expertise[29].

[29] Un cas particulier, à cet égard, est l'indexation nécessairement complètement automatisée des documents disponibles sur Internet : la taille, la diversité et la rapidité de mise à jour de cette base documentaire particulière interdisent toute intervention humaine.

2.1.3.1.L'approche « sac de mots »

L'indexation à partir de descripteurs, généralement des mots typographiques isolés, c'est-à-dire des mots simples, présente quelques limites, ayant trait essentiellement au principe de pertinence adopté dans la constitution des index. Les limites de l'indexation automatique par descripteurs tirés des documents sont l'objet du passage ci-dessous.

> Short of comparing the request formulation with the original document, one could think of comparing this formulation with a set of clues obtained from the documents by some mechanical procedure. Such procedures have come to be known as *automatic indexing*. (…) However, the chances that thereby a satisfactory set of clues will be obtained are (…) rather slim. (…) [I]t is (…) rather unlikely that the set obtained thereby will be of a quality commensurate with that obtained by a competent indexer (…). First, there will be serious difficulties as to what is to be regarded as instances of the same word. (…) Second, there arises again the problem of synonyms. Third and most important, *this procedure will yield at its best a set of words and word strings exclusively taken from the document itself*[30].
>
> (Bar-Hillel, 1964, pp.338-339)

Bar-Hillel identifie notamment la variation, tant stylistique (tournures de phrases, voix privilégiées : active, passive) que lexicale (choix des mots) comme limites à une approche automatisée de l'indexation des documents, et considère nécessaire la mise en œuvre d'une théorie de l'information sémantique, autrement dit une théorie du contenu des documents, comme préalable à une automatisation de l'indexation. À cette conception plutôt négative de l'apport des approches automatiques dans le domaine de la recherche d'information, on peut opposer les expériences entreprises par Salton, comparant les performances d'une des premières versions de son système d'indexation automatique par approche vectorielle, SMART, à celles d'opérateurs humains[31]. Les résultats de ces expériences ont eu comme effet de conforter les approches peu théorisées, tenantes d'une position linguistique faible.

---

[30] Italiques ajoutés.

[31] Connues sous le nom de ASLIB Cranfield Research Project, décrites dans (Cleverdon, 1966).

En effet, la sélection de descripteurs de contenu se fonde, le plus souvent, sur une telle position linguistiquement faible : le contexte d'occurrence des éléments retenus est le document et non pas le contexte syntaxique (ex. : phrases, paragraphes). De ce fait, le profil distributionnel des descripteurs de contenu n'inclut aucune information syntaxique, telle que la constituance, par exemple[32]. Pour cette raison, les approches d'indexation automatique sont généralement perçues comme représentant le contenu des documents sous la forme d'un « sac de mots » (*bag of words*).

Bar-Hillel poursuit son analyse des lacunes des approches automatiques en indexation, en critiquant la représentation peu structurée des informations linguistiques qu'elles élaborent.

> If a certain document collection contains both documents dealing with the Export of Cars from France to the USA and the Export of Cars from the USA to France, and if both kinds of documents are indexed, in uniterm or descriptor fashion, by *export*, *cars*, *France*, *USA*, then clearly any request for a list of documents dealing with one topic will be answered by a reference list containing also references to documents dealing with the other topic. (…) False drops of the above-mentioned kind in a request for a reference list of documents dealing with the export of cars from France to the USA can be avoided if the indexing terms are taken to be *export*, *(of) cars*, *(from) France*, *(to) USA* (…).
>
> (Bar-Hillel, 1964, p. 362)

La solution préconisée par Bar-Hillel pour limiter les réponses non désirées passe, principalement, par l'abandon du principe d'indexation par une collection non structurée de mots simples. Cette solution doit être comprise dans le cadre plus général de la théorie de l'information sémantique développée par l'auteur. Nous verrons plus bas que la solution évoquée ci-dessus constitue celle que nous avons adoptée, dans l'optique du filtrage d'information reposant sur une analyse linguistique, bien que nous ne reprenions pas la théorie développée par Bar-Hillel dans son ensemble.

[32] Par ailleurs, les principes de sélection des descripteurs de contenu visent explicitement à éliminer les mots grammaticaux des bases d'index construites, ce qui rend quasiment impossible toute représentation des relations de constituance.

2.1.3.2.Pertinence d'une base de descripteurs figés

Les bases d'index, en raison des volumes documentaires manipulés, ont vocation à être relativement stables. Autrement dit, les descripteurs de contenu, choisis manuellement ou pas, ont vocation à saisir les aspects les moins volatils du contenu informatif des documents. Comme nous l'avons vu plus haut avec (Spärck Jones & Kay, 1973), on retrouve là le problème classique de la métaphysique, qui consiste à distinguer les traits essentiels des objets considérés (i.e., des concepts véhiculés par des documents) de leurs accidents.

En premier lieu, on peut s'interroger sur la pertinence d'une telle représentation figée du contenu des documents, alors que les connaissances évoluant nécessairement, il apparaît inévitable que la valeur des descripteurs choisis à un moment donné, au sein du système que constitue l'ensemble de la base documentaire, doive être remise en cause en fonction de la mise à jour d'une collection de documents, afin de suivre cette évolution. Cette remise en cause n'est possible que dans une perspective métaphysique faible, c'est-à-dire une démarche structurante nécessairement imparfaite et connue comme telle, qui nous paraît être la position dominante en Recherche d'Information[33]. Cette position se traduit d'ailleurs par des choix lexicaux particuliers : on parle rarement, en indexation, d'Ontologie (au singulier), mais bien plutôt d'ontologies (au pluriel), c'est-à-dire de structuration nécessairement locales et imparfaites de concepts.

On peut voir dans la stabilisation d'un espace conceptuel que constitue cette démarche un mouvement partagé par toute démarche posant une abstraction nécessaire par rapport à un ensemble d'observables. On retrouve toute la difficulté, soulignée au chapitre précédent dans le domaine des études linguistiques partant de la Parole, entre point de vue subjectif, inscrit dans un contexte (temporel) et point de vue objectif, atemporel.

2.1.3.3.Prise en compte du point de vue des utilisateurs

La pratique de l'indexation des documents pose, de façon empirique, plusieurs questions fondamentales. La première a trait, d'un côté à la structuration d'un fonds documentaire suivant une hiérarchie de concepts, supposée fixe, première et universelle, de

[33] (Coyaud, 1972, p. 130) : « Le problème de la documentation (…) ne se laisse pas mettre en forme et résoudre par des méthodes mathématiques ou même simplement scientifiques. (…) Lorsqu'on emploie l'expression *Information Science*, à propos des activités documentaires, il ne faut pas oublier qu'il ne s'agit pas d'une science exacte ».

l'autre à une mise à jour en fonction de l'évolution des connaissances. On retrouve, dans ce domaine, les deux positions fondamentalement opposées, évoquées au premier chapitre, entre un point de vue objectif, à visée scientifique et un point de vue subjectif. Les problèmes posés par l'activité d'indexation sont abordés ci-dessous.

> Certaines motivations erronées que l'on aperçoit dans des langues naturelles se retrouvent dans des LD [Langages Documentaires] ; par exemple, dans le LD WRU, le mot « baleine » est codé dans la classe des poissons avec l'infixe Z « simulation », presque comme l'allemand *Walfisch*. *Les classifications ne sont pas nécessairement scientifiques. Au contraire, il y a de bonnes raisons de penser que plus elles sont scientifiques, moins elles risquent d'être efficaces*[34].
>
> (Coyaud, 1972, p.16)

Ce passage illustre, à nos yeux, la tension résultant d'un nécessaire compromis entre plusieurs représentations du monde : celle des indexeurs, à vocation normative et scientifique (point de vue objectif), et celle des utilisateurs (point de vue subjectif). La conclusion que tirent tant Coyaud que Spärck Jones & Kay des manifestations de cette tension, entre les représentations des utilisateurs d'un système d'information et celles des opérateurs de ce système, va dans le sens :

1) d'une automatisation du processus d'indexation,
2) opérée sur des bases linguistiques.

En effet, les auteurs cités voient dans l'adoption d'une description de contenu des documents, plus proche de la langue naturelle, les moyens de dépasser les tensions évoquées plus haut (variation dans l'indexation humaine, limites des indexations automatiques par descripteurs).

La question posée ici peut être reformulée comme celle de la place de l'utilisateur au sein du système d'information. En d'autres termes, on peut comprendre les expériences, menées dans le domaine de la recherche d'information pour aboutir à une plus grande

[34] Italiques ajoutés.

adéquation des systèmes d'information par rapport aux attentes des utilisateurs, comme autant de précurseurs des modèles orientés vers les utilisateurs (*user-oriented models*). Qui plus est, on peut considérer des réalisations telles que le projet Profil-Doc, décrit dans (Michel, 1999), comme des tentatives d'allier une problématique orientée vers les utilisateurs à des modèles basés sur l'usage (*usage-based models*).

Bien que nous souscrivions à une problématique orientée vers les utilisateurs, basée sur l'usage effectif, pour la conception de systèmes d'information, nous nous démarquons des auteurs cités dans la mesure où nous relativisons la portée objectivante d'une telle démarche. Comme nous le verrons plus bas, nous proposons un principe d'appariement, entre un besoin en information exprimé par un utilisateur et une collection de documents, reposant sur une analyse linguistique de ces documents, réalisée de façon automatique. Cette analyse vise à dépasser les limites évoquées plus haut des principes d'indexation par descripteurs limités à des mots typographiques, tirés du stock de mots simples des documents. En ce sens, l'approche que nous proposons suit les conclusions de Coyaud, et de Spärck Jones & Kay. Toutefois, pour le sous-domaine qui nous occupe, à savoir le filtrage d'information, nous ne postulons aucune association régulière, valable pour tous les utilisateurs, entre l'ensemble de formes linguistiques pris comme descripteur de contenu et le contenu lui-même. Nous nous plaçons plutôt dans une optique proche de celle guidant le système Profil-Doc : proposer des solutions afin de représenter la partie linguistique des unités documentaires mises en œuvre dans les stratégies individuelles de recherche d'information.

2.1.4. Recherche d'information basée sur des unités lexicales complexes

Tout l'enjeu de remplacer les langages de description, dont nous avons vu quelles difficultés étaient liées à leur utilisation, par la langue naturelle comme moyen d'indexation et d'appariement entre requête et documents indexés, est celui d'une simplification supposée de l'utilisation des systèmes d'information.

Dans cette approche, le texte, autrement dit une partie de la Langue, est considéré comme un support de l'information. Suivant les recherches amorcées par Harris, poursuivies, entre autres, par Herdan, Biber ou encore Habert, chaque domaine de spécialité (ex. : genre littéraire, domaine d'activité, époque) se caractérise par des contraintes tant au niveau lexical, morphologique, syntaxique, phrastique, que textuel. Autrement dit, est posée une

spécialisation linguistique en fonction des domaines de spécialité, dont les principes sont suffisamment stables pour permettre d'établir des règles générales. Ces règles peuvent être mises à profit dans le cadre de la recherche d'information au sein de bases de documents, en vue d'aboutir à des descripteurs plus pertinents que ceux issus des procédures d'indexation classiques, en ce qu'ils sont basés sur une analyse linguistique, même partielle, des observables et non plus seulement de la prise en compte de propriétés statistiques de ces observables.

2.1.4.1. Analyses linguistiques automatisées et Recherche d'Information, une difficile intégration

Le bilan que constitue (Spärck Jones & Kay, 1973) de l'intégration de techniques issues du domaine émergent du TALN, à l'époque de parution de l'ouvrage, souligne à plusieurs reprises les difficultés rencontrées. Il est intéressant de se pencher, rétrospectivement, sur la conclusion de l'ouvrage, qui représente, encore aujourd'hui, la position dominante sur le sujet outre-Atlantique.

> It is difficult, when considering syntax in information retrieval, to avoid a feeling of puzzlement. Many apparently convincing arguments for its use have been advanced, and many apparently sensible syntactic procedures have been proposed. But insofar as systematic comparative experiments have been carried out, they show that syntactic information contributed little to retrieval performance and may even detract from it. (…) It may be that all the experiments to date have been inadequate. Other possible explanations are (1) that retrieval needs are not properly understood; (2) that the value of the syntactic component of an index description is affected by other system components: it may either be that the correct relationships between different components have not been established, or that other components are defective; and (3) that essentially inadequate or inappropriate methods of handling syntax have been adopted. We are reluctant to consider the possible fourth explanation, namely that an indexing language cannot materially contribute to a good retrieval performance.
>
> (Spärck Jones & Kay, 1973, p. 119)

(Spärck Jones & Kay, 1973) examine les liens entre recherche d'information, et plus particulièrement processus d'indexation, et linguistique, sous l'angle :

1. des méthodes servant à l'identification des unités pertinentes dans les documents à indexer ;
2. de la dérivation d'une description de contenu à partir de ces unités pertinentes ;
3. de la construction et de l'utilisation de classifications et autres formes de structuration des langages d'indexation.

On peut comprendre la démarche exposée tout au long du présent chapitre comme relevant essentiellement du point 1). Le point 2), en revanche, ne nous occupera pas, toutefois le point 3) est abordé dans l'ensemble de notre exposé, par la question portant sur l'indexation en tant que processus de catégorisation. Ainsi que le montre (Spärck Jones & Kay, 1973), l'ambition initiale de l'alliance entre linguistique et recherche d'information était la mise à profit de l'appareil formel (modèles, descriptions linguistiques) développé par la première dans une optique de capitalisation de la connaissance et d'amélioration des performances des systèmes et des applications de la seconde : « We began this survey with two questions : since linguistics and information science are both concerned with the product of linguistic behaviour, namely discourse, we may ask, first, what linguistics can or should be able to offer information science, and second, what information science can offer linguistics », (Spärck Jones & Kay, 1973, p. 195). Dans la conclusion du bilan de cette expérience, les auteurs insistent sur la difficulté de cette alliance.

> Our initial hypothesis was that the information scientist or documentalist would be assisted in his attempts to devise linguistic processing procedures for retrieval if he could exploit the findings of linguists. It is not unreasonable to suppose that while his use of linguistic theory will be influenced by his specific purpose, he needs a substantial general linguistic apparatus. (…) The assumption this whole survey has been intended to examine, in other words, is that the data and objectives of information retrieval do not imply nongeneral, purpose-oriented linguistic theories which are qualitatively different from those that concern ordinary linguists. (…) The most striking fact to emerge from the literature, however, is the difficulty of marrying linguistic techniques and retrieval objectives. The difficulty is indeed so great as to cast doubt on the assertion that general linguistic theories are prerequisites for effective information processing and retrieval. As noted, linguistically very crude procedures seem to work quite well in retrieval, and it is in practice not obvious how more sophisticated ones should be used.
>
> (Spärck Jones & Kay, 1973, p.197)

Les auteurs soulignent la difficulté d'intégrer des analyses linguistiques au processus de recherche d'information, pour aboutir à la conclusion que les approches adoptant un point de vue linguistique faible semblent donner les meilleurs résultats.

> (…) [T]he tempting general conclusion to draw from experience to date is that for the special purpose of document retrieval general linguistic theories are not required. Since comparatively simple approaches like those involving statistically extracted key words, simply coordinated, seem to work as well as ones relying on richer linguistic information, we may conclude that document retrieval systems are necessarily crude. Abbreviated document descriptions are presumed, and ill-designed requests are probable or even certain. Some simplicity in the characterization of information if therefore inevitable, and *it is unlikely that performance for poor requests can be much improved by sophisticated simplicity*[35]. (…) It is more productive to maintain that the difficulty of relating linguistics and information retrieval comes from the fact that linguistic theories are still far from adequate, and that the design of good information retrieval systems is not at all understood. We may then hope that even if simplicity is all that is linguistically needed, it had better be sophisticated simplicity; we should surely be able to do better in providing

[35] Italiques ajoutés.

> document summaries that mere keyword lists, and we may legitimately believe that linguistics should help us here.
>
> (Spärck Jones & Kay, 1973, p.198)

La difficulté principale de l'alliance entre ingénierie linguistique et recherche d'information, dans le bilan que dressent Spärck Jones & Kay, semble provenir de l'incomplétude des modèles linguistiques disponibles, ainsi qu'à leur manque de robustesse. Les passages cités résument toute l'ambivalence des tentatives d'alliance entre linguistique et recherche d'information : entre espoir de meilleures performances et frustration devant l'incomplétude des analyses linguistiques automatiques. Encore aujourd'hui, la conviction générale dans le domaine est celle d'une inutilité des représentations linguistiques de haut niveau, non seulement par le manque de maturité des recherches en linguistique, mais également par les particularités de la recherche d'information : les temps d'analyse des documents doivent être les plus réduits possibles, tout délai de plus de quelques secondes dans la constitution d'une réponse à une requête étant perçu comme intolérable par les utilisateurs. De plus, dans le cas des systèmes d'information interrogeables en langue naturelle, la langue utilisée pour constituer les requêtes est loin d'être celle du locuteur idéal postulé par les linguistiques abstraites. Cette prévalence de la Parole dans ce domaine est l'objet de la remarque : « ill-designed requests are probable or even certain », qui milite, pour les auteurs cités ci-dessus, pour une approche privilégiant la simplicité des analyses linguistiques. La progression rhétorique du passage cité, qui prône une position linguistique faible, dans un premier temps, pour se conclure par une apologie de la « simplicité sophistiquée » (*sophisticated simplicity*), ne peut être comprise que dans le cadre du bilan que représente l'ouvrage entier, commandé et financé par le Comité sur la Linguistique en Documentation de la Fédération Internationale de Documentation (FID)[36].

Un des points évoqués dans le passage cité ci-dessus est le caractère incomplet des spécifications de l'activité de recherche d'information fournies par les professionnels eux-mêmes, lacune à laquelle les auteurs attribuent une partie de l'insuccès des approches intégrant des analyses linguistiques automatisées. Ainsi que nous le verrons plus loin pour le

[36] On comprend aisément qu'une position plus tranchée en défaveur du recours à des analyses linguistiques automatisées dans le domaine de la documentation aurait mis les auteurs en position de porte-à-faux vis-à-vis du commanditaire de ce bilan.

cas particulier du filtrage d'information[37], cette remarque s'applique, encore aujourd'hui, aux tentatives d'alliance entre linguistique et recherche d'information.

2.1.4.2.Un retour à l' « empirisme » ?

Il est intéressant de noter que la position d'une alliance entre linguistique et recherche d'information reposant sur une linguistique forte[38] ne semble avoir été abandonnée qu'au début des années 1990, avec le retour en force des approches surfaciques (*chunking, shallow-parsing*), dans le domaine des approches catégoriques, et statistiques ou probabilistes dans celui des approches non catégoriques. Ce retour de ce que (Habert, 1998) nomme « empirisme » a sonné le glas d'une linguistique forte dans la plupart des domaines d'application, y compris la recherche d'information, ainsi qu'en témoignent aujourd'hui les pratiques effectives : extraction et filtrage d'information par cascades de transducteurs et analyse de surface, prédominance des approches vectorielles en indexation de documents.

La prépondérance d'une linguistique faible dans le domaine applicatif peut être comprise de plusieurs façons :

- la linguistique n'a pas les moyens de fournir un appareillage formel général, pouvant trouver des applications dans différents domaines ;
- les descriptions générales ne sont pas utilisables dans des domaines spécialisés ;
- une linguistique faible est suffisante.

En ce qui concerne le premier point, il paraît difficile de préjuger de la capacité de la linguistique de corpus[39] à fournir des descriptions et des modèles génériques. En effet, le domaine a été marqué, principalement depuis le début des années 1990, par l'émergence de modèles formels alternatifs au générationalisme, pour lesquels le recul fait encore défaut. Cependant, la disponibilité toujours plus importante de corpus annotés, standardisés en plusieurs langues permet d'envisager, à long terme, des avancées dans le domaine de la langue générale.

---

[37] Voir le chapitre III.

[38] Par analogie avec la notion d'IA (Intelligence Artificielle) forte.

[39] Nous opposons linguistique de corpus et linguistique formelle, abstraite, reposant sur des énoncés construits.

Ceci nous amène au deuxième point : à supposer que le domaine des études linguistiques sur corpus soit capable de fournir des descriptions et des modèles génériques (ex. : une grammaire des phrases « normales »), la question reste posée quant à l'utilité de ces objets dans des domaines spécialisés. En effet, la mise au point d'applications (logiciels), en ingénierie linguistique, se caractérise par une centration sur les productions effectives, dans des conditions relativement peu contraintes, autrement dit des phénomènes relevant essentiellement de la Parole. En termes de marché, la valeur ajoutée des systèmes développés (ex. : aide au suivi de la relation-client par filtrage des courriers électroniques) tient plus dans leur capacité à traiter la Parole, c'est-à-dire à pouvoir prendre en compte les spécificités des locuteurs (ex. : violation des contraintes de bonne formation des énoncés, recours à l'implicite, variation des niveaux de langue), que dans leur conformité à une certaine norme.

En somme, bien qu'en termes d'objectifs à long terme, l'élaboration de descriptions et de modèles génériques constitue une visée intéressante, elle n'apparaît pas suffisante, dans le domaine des applications. Qui plus est, cette visée n'apparaît pas forcément nécessaire. L'expérience effective des approches surfaciques, linguistiquement faibles, dans le domaine applicatif amène généralement à des constats tels que : « les erreurs d'étiquetage ou d'attachement ne perturbent que très modérément l'image qui est fournie des fonctionnements syntactico-sémantiques des mots du corpus. Ou pour le dire autrement, *la redondance est suffisante*[40] pour garantir une stabilité correcte des rapprochements [sémantiques] sur la base des comportements [distributionnels] partagés. On peut donc 'composer avec l'imparfait' sans trop de risques » (Habert, 1998, p.159).

Ce constat d'une inutilité des représentations linguistiques de haut niveau dans le domaine applicatif justifie le recours à des descriptions parcellaires, contextuelles, dépendantes d'un domaine de spécialité, telles que les grammaires dites locales. Ces grammaires se caractérisent par un abandon du paradigme déclaratif dominant, et un retour vers une conception plus procédurale de la description des énoncés possibles. Ce type de descriptions se contente, en termes de langage formel, de grammaires beaucoup moins contraintes que celles requises dans le cadre d'une linguistique forte : principalement grammaires dites « context-free » (CFG, *Context Free Grammars*), voire automates ou transducteurs à états finis. Une composante probabiliste est également souvent présente, ce qui assure aux chaînes de Markov cachées, aux automates dits pondérés, ou encore aux PCFG

[40] Italiques ajoutés.

(*Probabilistic Context Free Grammars*) un regain d'intérêt tant dans le domaine applicatif que dans celui du TALN[41].

En effet, le caractère non nécessaire des représentations linguistiques de haut niveau dans le domaine applicatif pose, en retour, la question de leur utilité dans le domaine théorique. Autrement dit, ainsi que nous l'avions évoqué dans le premier chapitre, la question de la validité d'un point de vue uniquement objectivant est posée par le succès des approches centrées sur la Parole, dans le domaine applicatif. Cette remise en cause justifie, pour des auteurs tels que Manning, Abney ou Pereira, le recours à une démarche inductive, dans l'optique de l'élaboration d'une grammaire, tant dans des domaines spécialisés que dans celui de la langue générale.

Nous ne traiterons pas ici de l'automatisation, ni des paramètres de cette automatisation, d'une procédure de construction de grammaire de type inductif : nous nous contenterons de décrire les résultats d'une analyse des corpus guidée par les principes distributionnalistes, s'appuyant aussi bien sur une approche catégorique classique (à base de règles explicites) que non catégorique. Par ailleurs, nous nous concentrerons sur un domaine de spécialité : le domaine financier, et plus particulièrement le sous-domaine des cessions-acquisitions de sociétés.

2.2. Extraction de marqueurs thématiques linguistiques par analyse distributionnelle

Dans cette partie, nous nous penchons sur quelques approches, que nous jugeons représentatives pour le problème qui nous occupe, visant à extraire des documents des éléments linguistiques spécialisés, associés de façon préférentielle à des domaines d'activités précis. Nous aborderons donc, dans un premier temps, le domaine de l'analyse distributionnelle discontinue des corpus visant à en extraire soit des termes, soit des marqueurs thématiques ayant un fondement linguistique (ex. : des structures syntaxiques ayant une valeur thématique particulière). Dans un deuxième temps, nous examinerons les descriptions linguistiques formalisées auxquelles cette analyse permet d'aboutir, des

[41] Voir (Charniak, 1993) pour une présentation d'une approche statistique du TALN et de l'apprentissage de grammaires CFG probabilistes (PCFG).

ressources linguistiques, utilisables par des systèmes informatiques en tant que ressources. Nous examinerons également l'apport d'une analyse non catégorique, notamment par le biais de la notion de collocation.

2.2.1. Analyse distributionnelle discontinue des corpus spécialisés

Ainsi que nous l'avons vu plus haut, les études sur corpus de spécialité, menées dans un cadre distributionnelle, font l'hypothèse d'une spécialisation tant lexicale que syntaxique (i.e. des schémas de phrases en nombre fini), voire pragmatique[42]. Autrement dit, ces études postulent une différence fondamentale entre sous-langages de spécialité et langue générale.

Nous voyons dans cette spécialisation la possibilité de mettre en œuvre des langages formels moins contraints que ceux préconisés pour la description de la langue générale, notamment des grammaires dites « context-free », voire des automates à états finis. Cette spécialisation permet également d'envisager l'induction des régularités constatées en corpus, soit par des approches inductives symboliques[43], statistiques[44] ou encore subsymboliques (réseaux de neurones artificiels), dans un cadre non supervisé[45].

2.2.1.1. Élaboration d'une grammaire d'un domaine de spécialité

La démarche harrissienne vise à aboutir à une grammaire d'un sous-langage de spécialité, en suivant les étapes suivantes.

1. Analyse distributionnelle, visant à établir les contraintes de cooccurrence de certains éléments lexicaux, si possible en interaction avec un expert du domaine. Cette analyse repose sur les principes distributionnels harrissiens évoqués dans le premier chapitre, notamment la mise en relation d'éléments de profil distributionnel similaire, dont la similarité est évaluée grâce à des procédures d'approximation.

[42] La dimension poétique et les jeux de langue, par exemple sont relativement absents des corpus financiers, au profit de la dimension informative.

[43] Voir (Cussens *et al.*, 1997).

[44] Voir notamment (van Zaanen, 2001).

[45] Voir (Finch, 1993).

2. Description de séquences élémentaires, destinées à constituer les phrases-noyaux (*kernel sentences*) du domaine de spécialité.
3. Constitution d'une grammaire du domaine de spécialité, intégrant des règles de formation des phrases-noyaux, ainsi que les opérations transformationnelles valides (ex. : transformation passive, nominalisation).

À chaque étape, la confrontation avec le corpus permet de s'assurer de la conformité de la grammaire élaborée au sous-langage étudié. Quelques travaux fondateurs dans le domaine sont ceux de Harris, notamment (Harris *et al.*, 1989) dans le domaine immunologique, (Sager & Friedman, 1987) dans le domaine médical, ainsi que ceux de (Gross, 1968 ; Gross, 1975).

Le résultat visé de ce processus d'analyse est, pour (Sager, 1987) notamment, ou encore pour (Gross, 1975), ou (Habert, 1998), la constitution de classes sémantiquement homogènes à partir des régularités de construction constatées en corpus. Nous avons vu au premier chapitre quel enjeu représentait la notion de classe pour une science du langage, ainsi que l'influence du modèle classique de la catégorisation hérité d'Aristote. Nous nous démarquons ici des travaux cités ci-dessus : en effet, nous soulignons l'insuffisance des propriétés distributionnelles évoquées ici (notamment contraintes de sous-catégorisation) comme critères de constitution d'une ontologie, au sens où les éléments recensés seraient appelés à appartenir à des classes aux frontières étanches, régies par un principe de hiérarchisation de type taxinomique[46]. Tout au plus faisons nous le constat qu'une partie des éléments qui partagent le même profil distributionnel, tirés de corpus spécialisés, peuvent fournir la base de classes de mots relativement stables et opérationnelles dans une application de recherche d'information.

Signalons que les corpus que nous avons analysés dans le cadre d'une application au filtrage d'information appartiennent au domaine journalistique. Ainsi, dans des phases préliminaires, nous avons étudié des articles du journal *Le Monde*, des dépêches de l'AFP et de AP. Dans un dernier temps, nous avons appliqué les principes définis ci-dessus à un corpus financier, constitué de dépêches, rédigées dans un style journalistique. La prudence par nous

[46] (Gross, 1975) fait d'ailleurs un constat d'échec d'une tentative de constitution de classes sémantiquement homogènes à partir de contraintes distributionnelles. Signalons toutefois que l'ouvrage visait la langue générale, non les langues de spécialité, comme c'est le cas pour (Harris *et al.*, 1989), (Habert, 1998) et (Sager, 1987).

exprimée, quant à la constitution de classes sémantiques à partir de régularités de comportement distributionnel, tient essentiellement aux corpus étudiés, dans lesquels les contraintes fortes du domaine de spécialité (i.e. la finance) se heurtent à des contraintes d'ordre stylistique dans la rédaction des dépêches. En cela, notre objet d'étude se distingue des corpus plus contraints, tels que ceux de l'immunologie (Harris *et al.*, 1989), de la médecine (Habert, 1998), ou encore de l'agronomie (Morin, 1999).

2.2.1.2.Extraction terminologique

Parmi les approches directement inspirées du distributionnalisme classique, menés dans une optique terminologique, les travaux de (Bourigault, 1994), (Bourigault, 2002), de (Bouaud *et al.*, 1997), (Habert, 1998), (Habert & Fabre, 1999) se distinguent par leur fidélité aux principes énoncés par Harris. Ces travaux ont pour vocation d'extraire des unités lexicales complexes, rattachées à des domaines de spécialité (ex. : médecine, agriculture), sur la base de leur comportement linguistique, c'est-à-dire sur la base de leurs contextes syntaxiques d'occurrence. Ainsi, ces travaux font généralement appel à une analyse syntaxique plus ou moins profonde puis à des regroupements d'éléments en fonction des contextes qu'ils partagent. Le rapprochement des unités lexicales peut faire appel à des opérations de normalisation et de généralisation (ex. : réduction des expansions d'un syntagme nominal), des transformations[47] (ex. : *cancer de l'intestin* => *cancer intestinal*), ainsi qu'à des procédures (règles symboliques) ou des indices statistiques[48] (ex. : estimation de la proximité en fonction du nombre de contextes communs).

La particularité des travaux en terminologie est la place laissée à l'émergence spontanée d'éléments linguistiques pertinents, supports de concepts spécialisés. Ainsi, les unités lexicales complexes extraites, bien qu'associées à des domaines de spécialité, n'ont pas vocation à servir de descripteurs dans le cadre de la RI, mais bien plutôt à former la base d'une ontologie du domaine étudié. L'utilisateur (terminologue) est généralement inclus dans la boucle des traitements : il sélectionne les candidats-termes en fonction de leur pertinence. Les travaux tels que (Grefenstette, 1993), ou encore (Daille, 1994 ; 2002) sont comparables dans leur visée, malgré la combinaison d'approches symboliques classiques et statistiques : la

[47] Voir, notamment, (Habert, 1998).

[48] Voir (Bourigault, 2002).

mise au point d'une ontologie, autrement dit la construction d'une hiérarchie de concepts portés par des unités lexicales.

Les travaux évoqués ici restent dans le cadre discontinu de l'analyse distributionnelle : en effet, ils se basent tous sur une conception typographique des mots, plus ou moins corrigée en fonction du problème à traiter (ex. : *des => de les*, *du => de le*). Par ailleurs, les classes d'éléments extraits des corpus n'ont pas, à notre connaissance, vocation à être de nature polycatégorielle.

2.2.1.3.Extraction d'information à partir de schémas conceptuels

Les travaux décrits dans (Riloff, 1994) apparaissent comme les plus féconds pour le problème qui nous occupe. En effet, cet auteur a abordé le problème de la RI intégrant des analyses linguistiques par le biais de l'extraction d'information. Le système mis au point, Autoslog[49], vise à construire ce que l'auteur nomme un dictionnaire de nœuds conceptuels (*conceptual nodes*) pour l'extraction d'information, pouvant être mis à profit pour des tâches telles que la classification automatique de documents. Riloff s'est donc intéressée aux relations entre TALN et RI, par le biais d'une analyse locale pouvant être mise en œuvre aussi bien dans le cadre de l'extraction que du filtrage d'information, qui peut être vu comme une spécialisation de la classification de textes.

Autoslog construit des schémas conceptuels (*case frames*) liés à des événements, à partir de corpus d'apprentissage étiquetés par Circus[50], et annotés manuellement. Rilloff donne l'exemple suivant, tiré des corpus MUC-4 sur les attentats terroristes :

A passerby was hurt when two terrorists attempted to kill the mayor

Exemple 1 : extraction d'information sur une phrase décrivant les conséquences d'un attentat

Dans cette phrase, les informations suivantes, correspondant grossièrement à des rôles casuels doivent être extraites par le système :

- la victime de l'attentat (*a passerby*) ;

[49] (Riloff, 1994).

[50] Voir (Lehnert *et al.*, 1993) pour une présentation de l'analyseur Circus.

- l'agent (*two terrorists*) ;
- la victime visée (*the mayor*).

Autoslog vise à constituer, de façon automatique, des schémas conceptuels (*case frames*) associés à des événements tels que des attentats terroristes, à partir de corpus d'apprentissage, de type journalistique, étiquetés et annotés en fonction des informations à extraire. Les schémas décrits par Riloff constituent, en quelque sorte, une représentation abstraite des événements décrits dans les corpus traités. Ces représentations sont fondées, en partie, sur la notion de *frames conceptuels*, et de *script*[51], et sur les travaux de (Cullingford, 1978) pour une tâche similaire[52].

Les schémas conceptuels utilisés par Autoslog représentent des patrons pour l'extraction d'information, c'est-à-dire un ensemble de séquences pertinentes décrites sous la forme d'une grammaire restreinte, autrement dit une grammaire locale. Autoslog se base, dans la construction de patrons d'extraction à partir de corpus, sur des amorces, autrement dit des termes simples, tels que : *murder*, *bomb*, ou encore *terrorist*[53]. Ces amorces forment la base de patrons syntaxiques, ou phrases-noyaux, tels que **N0 was Ved (by N1)**, ou encore **N0 V N1**. La figure ci-dessous donne un exemple de schéma construit par Autoslog.

[51] Voir (Schank & Abelson, 1977) pour une présentation détaillée. Les *frames* constituent des représentations abstraites, élaborées dans un cadre catégorique et logique, identifiant les agents typiques de situations données (ex. : une vente, un mariage). Les *scripts* sont plus particulièrement centrés sur les successions typiques d'événements de situations données (ex. : lors d'un repas au restaurant, la prise de commande précède le repas, qui précède le règlement de la note et la remise d'un pourboire). Ces deux objets conceptuels sont issus des recherches en Intelligence Artificielle et sont censés constituer des unités cognitives fondamentales.

[52] Dans la pratique actuelle en extraction d'information, ces représentations abstraites sont dénommées des « scénarios d'extraction ».

[53] Les travaux de Riloff portent majoritairement sur les corpus de MUC-3, décrivant des actions terroristes.

Name :	%MURDERED%
Event type :	MURDER
Trigger word :	murder
Activating_conditions:	passive-verb
Slots:	VICTIM <subject> (human)
	PERPETRATOR <prep-phrase, by> (human)
	INSTRUMENT <prep-phrase, with> (weapon)

Figure 1 : un schéma conceptuel pour l'extraction d'information par le système Autoslog

Dans ce schéma conceptuel, l'événement décrit est typé (MURDER), les amorces sont identifiées (murder), ainsi que les conditions d'activation du schéma considéré (voix passive). Ce schéma comporte trois champs (*slots*) instanciés par les éléments extraits des documents traités : les champs VICTIM, PERPETRATOR et INSTRUMENT. Chacun de ces champs correspond à un rôle casuel identifié par un comportement syntaxique typique :

- la position de sujet syntaxique occupée par un syntagme nominal typé en tant que « humain »,
- la position de complément d'agent, repérée par la préposition *by*
- celle de complément circonstanciel, de type « arme », repérée par la préposition *with*.

L'instanciation de ce schéma conceptuel doit être vue comme l'application de la procédure algorithmique ci-dessous[54].

[54] Voir (Riloff, 1994) pour une présentation exhaustive des procédures et des schémas conceptuels considérés.

```
SI l'amorce murder est trouvée

    SI une construction à la voix passive est identifiée

        SI un sujet syntaxique de type humain est identifié

        OU

        SI un complément d'agent de type humain est identifié

        OU

        SI un complément circonstanciel de type arme est identifié

            ALORS construire un schéma conceptuel de type MURDER
```

Procédure 1 : instanciation d'un schéma conceptuel MURDER pour le système Autoslog

L'approche adoptée par Riloff passe donc par la définition d'un schéma conceptuel générique, de nature heuristique, spécifiant des rôles/fonctions assimilés à des places (ex. : sujet grammatical). Le système Autoslog cherche ainsi à remplir les éléments libres (*slots*) de chaque schéma conceptuel avec des éléments extraits d'un document partiellement étiqueté. Cette approche vise à extraire des corpus de MUC-4 non pas des mots-clés, ou amorces, isolés, mais bien des ensembles de mots-clés, structurés par les relations prévues par les différents schémas conceptuels envisagés : agent, victime, victime visée, instrument utilisé, ou encore nombre de blessés. Ces ensembles de mots-clés structurés sont appelés « signatures de pertinence » (*relevancy signatures*) en ce qu'ils sont mieux corrélés avec les thèmes traités dans les corpus de MUC-4 que des mots-clés isolés. Ainsi, Riloff donne l'exemple de *dead*, qui pourrait constituer un descripteur *a priori* valide d'un document traitant d'un attentat. Toutefois, après confrontation avec les corpus MUC-4, *dead* apparaît insuffisamment corrélé avec les descriptions d'attentat, alors que *was found dead*, par exemple, est un bien meilleur marqueur thématique. L'ensemble des travaux de Riloff vise donc à extraire des corpus des unités lexicales complexes, centrées autour de verbes associés à des événements particuliers (i.e. des attentats), pour lesquels une représentation abstraite, inspirée des *frames* et des *scripts* peut être élaborée.

L'approche décrite par Riloff relève d'une application du distributionnalisme classique, dans le sens où les travaux de l'auteur peuvent être vus comme une procédure (un algorithme) de découverte d'ensembles de mots-clés structurés par des relations syntaxiques (ex. : l'agent occupe souvent la place du sujet syntaxique). Par ailleurs, les règles de génération des schémas conceptuels sont explicites, codées par le concepteur du système. L'analyse distributionnelle se limite, dans le cas du système Autoslog, aux contextes positionnels/syntaxiques d'occurrence d'un ensemble d'amorces, contextes traduits sous la forme de grammaires locales, ou patrons pour l'extraction d'information, dépendants d'un domaine de spécialité.

Riloff propose une approche à mi-chemin des approches à base de descripteurs et de celles fondées sur une analyse linguistique en profondeur : « We propose that information extraction techniques can be used to support text classification. This approach represents a compromise between keyword-based and in-depth natural language processing. (…) Information extraction technology is powerful enough to make discriminations that are difficult to make with keyword-based techniques, yet it is more robust and practical than in-depth natural language processing » (Riloff, 1994, p. 4). Le système Autoslog vise principalement à raccourcir le temps de développement de ce que l'auteur nomme un dictionnaire de patrons pour l'extraction d'information, autrement dit un ensemble de grammaires locales : l'auteur avance une réduction de la charge de travail d'un facteur 300[55]. Par ailleurs, l'auteur vise la mise à disposition de techniques d'analyse linguistique automatisée pour des utilisateurs non linguistes, et propose de ce fait un système dit « presse-bouton », où l'utilisateur n'intervient que dans la sélection des données à analyser et dans la validation des patrons d'extraction générés.

La différence essentielle entre Autoslog et les approches exposées ci-dessus, outre le domaine d'application, tient au recours aux amorces, qui guident la construction des patrons d'extraction, ainsi qu'à la recherche d'une forte corrélation thématique pour les unités extraites. Ainsi, les patrons générés, donc les unités lexicales extraites des corpus, visent à une adéquation thématique forte, contrairement à l'extraction terminologique[56]. Par ailleurs, Autoslog et les systèmes dérivés visent à accélérer l'élaboration de patrons syntaxiques sur

[55] De 1500 hommes-heures à seulement 5.

[56] On peut voir l'approche adoptée par Riloff comme cherchant à ne retrouver que les termes associés à des thèmes clairement définis (ex. : attentat, enlèvement).

une base préexistante : le système de Riloff vise essentiellement à apporter une meilleure couverture (i.e. des taux de rappel plus élevés) à un système d'extraction d'information existant. Autoslog génère donc autant de schémas conceptuels que de contextes d'occurrence différents pour les amorces considérées ; aucun regroupement n'est opéré, à notre connaissance, entre contextes d'occurrence proches.

L'apport essentiel des travaux de Riloff en ce qui concerne une approche linguistique du filtrage d'information est la notion de signature thématique, que l'auteur définit comme l'association entre un mot-clé, ou amorce, et un nœud conceptuel[57]. Pour la suite de notre exposé, nous reformulons cette définition des signatures thématiques en : l'association entre un ensemble d'amorces et une structure syntaxique, décrite par une grammaire locale[58].

2.2.1.4. Analyse thématique automatique fondée sur une ontologie sémantique

Les travaux de (Klavans & Kan, 1998) constituent une variante de ceux présentés plus haut, ils font appel à une classification des procès de type ontologique et s'inscrivent, eux aussi, dans un cadre catégorique logique. L'approche de (Klavans & Kan, 1998) dédiée à l'analyse thématique automatique, à partir des structures prédicatives trouvées dans les documents. Les auteurs mettent l'accent sur la limite inhérente aux approches guidées par des descripteurs, généralement des substantifs : bien qu'on puisse savoir de quoi parle un document, on ne peut pas savoir ce qui s'est passé. Les auteurs font appel à une classification des procès inspirée de (Jackendoff, 1993), reposant sur des principes catégoriques et visant à fournir une ontologie sémantique des procès[59].

Ce type d'approche se centre donc sur les prédicats, sous leur forme verbale, associés à leurs compléments typiques (entités nommées spécifiques, substantifs appartenant à un ensemble relativement restreint). L'approche décrite dans (Klavans & Kan, 1998) vise à associer des documents de type journalistique à des profils d'événements, ainsi qu'à un type

---

[57] Par exemple, la signature <dead, $found-dead-passive$> associée à l'ensemble des phrases construites autour de *find* et *dead*, à la voix passive.

[58] Par exemple, les amorces *Thales*, *EADS*, *racheter* et la grammaire locale **N0 V N1**, qui reconnaît l'ensemble des phrases où *Thales* est l'agent d'un événement au cours duquel *EADS* est *rachetée* : *Thales rachète EADS*, ou encore *Le groupe Thales, malgré une conjoncture difficile, s'apprête à racheter son concurrent EADS*.

[59] Le principe d'une telle ontologie est repris par d'autres approches, telles que (Pustejovsky, 1996).

textuel (discussion, rapport, argument). D'après les auteurs, les profils événementiels tirés des documents sur la base des prédicats verbaux peuvent également être utilisés dans une optique de classification en genre textuel : (Biber, 1989), par exemple, propose un système de classification en genre textuel reposant sur 5 dimensions, qui sont autant de paramètres. L'une de ces dimensions a trait aux propriétés verbales : mode et temps, passifs avec et sans agents, formes infinitives. Ces paramètres fournissent également la base d'un processus de classification automatique en genre textuel, dans le cas de (Karlgren & Cutting, 1994).

La particularité de l'approche de (Klavans & Kan, 1998) est la dimension réduite de l'espace de classification : les dépêches journalistiques analysées sont supposées se répartir suivant 8 types d'articles standard (ex. : profils, dépêches, nécrologie, interprétation statistique, ou encore anecdotes). Par ailleurs, les auteurs fondent leur approche essentiellement sur les verbes dits de communication (*say*, *report*), les verbes-supports (*be*), ainsi que les ressources lexicales que sont la caractérisation des contraintes de sélection et de sous-catégorisation des verbes de l'anglais, établie par (Levin, 1993) ainsi que leurs propriétés sémantiques telles que décrites dans le réseau sémantique Wordnet[60].

L'ensemble des classifications automatiques décrites dans (Klavans & Kan, 1998) repose donc sur des ressources lexicales élaborées manuellement, par des études sur corpus traditionnelles, menées dans une perspective lexicographique, reprenant des analyses existantes (Wordnet) ou adaptant des principes d'analyse au domaine particulier des dépêches journalistiques (analyse des 100 verbes les plus fréquents, selon les principes établis par Levin). Les auteurs aboutissent ainsi à une ébauche de typologie textuelle en fonction des prédicats verbaux détectés dans les documents, par exemple :

- verbes de communication (*add*, *say*, *announce*) et éditos, rapports ou bulletins d'opinion ;
- verbes de mouvement (*rise*, *fall*, *decline*) et annonces de bénéfices ;
- verbes d'accord (*agree*, *accept*) et annonces de fusion, de transactions.

L'approche décrite par (Klavans & Kan, 1998) reprend l'hypothèse harrissienne d'une spécialisation linguistique associée à une spécialisation dans le domaine d'activité duquel

[60] (Miller *et al.*, 1990).

émane le corpus étudié. La typologie textuelle ébauchée repose sur des estimateurs de corrélation thème/prédicats verbaux, tout en restant dans une conception relativement classique de l'approche distributionnelle. En effet, la centration sur les prédicats verbaux, donc sur les types d'événements décrits dans les documents analysés, repose sur une caractérisation des contraintes de sélection et de sous-catégorisation des verbes, d'après les principes décrits dans (Levin, 1993). Ces propriétés des verbes de l'anglais forment la base d'une ontologie sémantique en fonction des contraintes mises à jour par le biais d'un cadre méthodologique très contraint. Nous voyons un parallèle entre les études menées par Levin sur les verbes anglais, et celles décrites dans (Gross, 1968 ; 1975 ; 1986), par exemple, suivant des principes distributionnalistes, pour les substantifs, les verbes et les adverbes en français. Par ailleurs, les genres textuels considérés, autrement dit l'ontologie des types d'articles prise en compte, n'ont pas vocation à présenter des frontières floues, ni un gradient d'appartenance catégorielle.

2.2.1.5.LIZARD, un assistant linguistique pour l'extraction de signatures thématiques

L'approche que nous avons suivie et implantée par le biais de l'assistant linguistique LIZARD[61] tente de concilier les avantages des travaux évoqués ci-dessus. Ainsi, nous cherchons à extraire des expressions typiques, associées à des corpus de spécialité, tels que le corpus Firstinvest[62], grâce à une analyse distributionnelle prenant en compte les contextes syntaxiques d'occurrence d'éléments privilégiés : les verbes conjugués. Le but de cette analyse est l'élaboration de ce que Riloff appellerait un dictionnaire de patrons pour l'extraction, et que nous appelons une base de signatures thématiques, décrites sous la forme de grammaires locales. Cette base vise à être utilisée dans le cadre du filtrage d'information, tâche qui s'apparente à la classification automatique de textes pour laquelle Riloff a testé la validité de son approche.

[61] Voir (Balvet, 2002 b).

[62] Voir le chapitre consacré au système CORAIL.

LIZARD est essentiellement un dispositif de recyclage d'étiquettes morpho-syntaxiques, associé au système Intex[63]. En ce sens, LIZARD se rapproche de Zellig, un outil servant à l'extraction terminologique, décrit dans (Habert, 1998).

Les spécificités de notre approche sont :

- l'intégration de l'utilisateur au sein d'un processus interactif[64] ;
- le recours à des procédures d'approximation inspirées de (Harris, 1951) visant à rapprocher des contextes d'occurrence lorsque cela est possible ;
- la focalisation sur des énoncés susceptibles de développer des événements, par le biais d'un prédicat verbal et de ses arguments ;
- l'intégration de connaissances hors-corpus, tirées de ressources lexicales telles que le Dictionnaire Intégral de Memodata, dans l'optique d'apporter plus de généralité aux signatures extraites des corpus ;
- la prise en compte des « signaux faibles » (séquences n'ayant qu'une faible probabilité d'occurrence, mais possédant un fort pouvoir discriminant) ;
- la prise en compte de la variation lexicale (choix lexicaux) et syntaxique (transformations), à l'œuvre dans les corpus spécialisés comme dans d'autres types de corpus.

Nous nous inscrivons dans la continuité des approches décrites plus haut, en reprenant l'hypothèse harrissienne qui fonde l'ensemble des analyses sur corpus spécialisés évoquées ici. Cette hypothèse (Harris, 1988 ; 1990 ; 1991) est celle d'une différence fondamentale dans la nature des contraintes de sélection des entrées lexicales, notamment verbales, entre les textes dits de langue générale et les textes spécialisés. Cette différence peut être interprétée, dans le cadre d'une approche reposant sur la notion de hiérarchie de contraintes telle que décrite dans (Mannning, 2002) notamment, comme une différence de statut des contraintes de

---

[63] Le principe de tels outils, recyclant les étiquettes (information morphosyntaxiques) apportées par des analyseurs morphosyntaxiques est décrit, entre autres, dans (Habert, 1998). D'autres étiqueteurs morphosyntaxiques que Intex sont envisageables, par exemple : Lexter (Bourigault, 1994), ou QTag (Mason, 2000). Le principe d'un tel recyclage est repris, entre autres, par les outils Zellig (Habert, 1998), Caméléon (Séguéla, 2002), Asium (Faure, 2002) et Upery (Bourigault, 2002).

[64] Qui s'oppose à un processus « presse bouton », où l'utilisateur intervient surtout à la fin, pour valider ou corriger les signatures extraites.

sélection, passant de contraintes fortes, ou de haut niveau, dans le cas des textes spécialisés, à des contraintes faibles dans le cas de la langue générale.

Prenons l'exemple du verbe *vendre*. Il est certain qu'en langue générale, la description exhaustive des compléments possibles de ce verbe est difficilement prévisible, elle dépend essentiellement d'une réalité du monde : la classe des objets susceptibles d'être *vendus* n'est pas restreinte. Ainsi, on peut *vendre son âme au diable*, *vendre sa maison*, *ses meubles*, *son corps*, ou encore *des services*. En revanche, dans le domaine financier, bien que la clôture absolue de l'ensemble des compléments de *vendre* soit impossible à réaliser, il n'en reste pas moins que la plupart des objets vendus tombe dans les catégories : société (ou partie de société : filiale, activité, service, branche), capital financier d'une société (notamment : actions, parts, droits), ou encore capital matériel (équipement, machines). Cet état de fait tient autant à des contraintes matérielles liées au monde des sociétés qu'à des contraintes linguistiques, en l'occurrence celles qui s'appliquent dans les langues de spécialité et les jargons. Les unités lexicales complexes que nous cherchons à extraire des textes financiers se rapprochent donc des signatures de pertinence décrites dans (Riloff, 1994), dont nous reprenons l'hypothèse centrale : un ensemble d'amorces structuré par un schéma conceptuel[65] recensant les contraintes de sélection[66] ainsi que de sous-catégorisation[67] constitue une unité d'information plus discriminante que des amorces isolées[68].

Les principes sous-tendant LIZARD sont ceux d'une analyse distributionnelle classique, réalisée dans un cadre catégorique, tout en ayant recours à des procédures d'approximation. Ces procédures ont pour but de généraliser des régularités observées en corpus, induites par des contraintes de sélection et de sous-catégorisation portant sur les prédicats verbaux. Les procédures de généralisation et d'approximation suivies par LIZARD sont présentées ci-dessous, les données textuelles analysées sont issues d'un premier

---

[65] Dans notre cas, un scénario d'extraction d'information, plutôt qu'une représentation abstraite de type *frame* ou *script*.

[66] Les compléments typiques (ex. : *vendre* et *société*, *filiale*, *groupe*).

[67] Les structures syntaxiques typiques (ex. : vendre et les constructions **N0 V N1**, **N0 V N1 Prep N2**).

[68] Riloff donnait l'exemple de *dead*, moins discriminant que *was found dead* pour le domaine des attentats, dans notre cas, *vendre* est moins discriminant que la signature Nom de Société vendre Nom de Société (ex. : *Thales vend EADS*), une instance de la structure **N0 V N1**.

déblayage du corpus, visant à en extraire des segments de phrase comprenant des verbes et leurs compléments habituels[69].

```
POUR CHAQUE mot étiqueté
SI le mot appartient à la classe des éléments généralisables[70]
    Effacer le mot
```

Procédure 2 : approximation, visant à normaliser les contextes distributionnels d'occurrence des prédicats verbaux

Cette procédure a pour résultat ce que nous nommons des « formes schématiques[71] », dans lesquelles, par exemple, seuls les prédicats verbaux et leurs compléments (substantifs) sont gardés. Les contraintes de sélection opérant sur les verbes étudiés sont donc ainsi mises à jour : cette procédure permet d'établir une liste des compléments habituels du verbe *vendre* (ex. : *filiale*, *groupe*, *parts*, *actions*).

```
POUR CHAQUE mot étiqueté
SI le mot appartient à la classe des éléments généralisables
    Généraliser en ne gardant que l'étiquette morphosyntaxique
```

Procédure 3 : généralisation, visant à extraire des schémas de sous-catégorisation, pour l'assistant LIZARD

[69] Dans sa version actuelle, LIZARD n'examine que des groupes verbaux.

[70] Cette classe est paramétrable en fonction du corpus, de l'application, ou encore de l'utilisateur. Elle spécifie quels éléments (classes d'éléments : déterminants, pronoms) sont discriminants, pour chaque phase de généralisation.

[71] Voir le chapitre IV, consacré au système CORAIL pour une présentation plus détaillée de LIZARD et des procédures de généralisation.

Cette procédure a pour résultat des patrons de sous-catégorisation propres à chaque corpus, en fonction des paramètres de généralisation choisis. Cette procédure permet de déterminer, sur le corpus de paramétrage, quelles constructions sont attestées pour chaque verbe étudié (ex. : pour vendre, **N0 V N1**, **N0 V N1 Prep N2**).

À l'issue de ces deux procédures d'approximation et de généralisation, trois vues différentes d'un même corpus sont disponibles :

- le corpus étiqueté d'origine ;
- une vue dans laquelle les contraintes de sélection des verbes sont mises en évidence ;
- une vue dans laquelle les contraintes de sous-catégorisation des verbes sont mises en évidence.

Ces trois vues sont complétées par une quatrième, reposant sur la procédure suivante.

```
POUR CHAQUE schéma de sous-catégorisation
    POUR CHAQUE verbe du corpus
        SI le profil distributionnel du verbe courant
        s'unifie avec le schéma de sous-catégorisation
        courant
            Inclure le verbe et ses compléments dans la
            liste associée au schéma de sous-
            catégorisation courant
```

Procédure 4 : élaboration d'une liste d'entrées lexicales en fonction d'un schéma de sous-catégorisation

Cette procédure, dans laquelle seule l'entrée verbale est conservée - les autres éléments (ex. : déterminants, noms, adjectifs, pronoms) étant représentés par un « + » - sert

de base à l'élaboration d'une base de données lexicale, présentée plus bas, spécifiant pour chaque verbe ses contraintes de sélection et de sous-catégorisation.

2.2.2. Ressources linguistiques issues d'une analyse classique

Les analyses distributionnelles, menées dans un cadre catégorique classique, permettent d'aboutir à des descriptions des régularités observées en corpus. Ces régularités peuvent fournir la base de thesauri et d'ontologies, dans le cadre de la terminologie, ainsi que des bases de données intégrant des descriptions du fonctionnement lexico-grammatical des unités retenues.

2.2.2.1.Thesauri et ontologie(s)

Le domaine des études sur corpus visant des applications concrètes, telles que des systèmes d'ingénierie linguistique, ou encore la pédagogie ou la lexicographie, est riche d'une profusion de travaux visant à constituer de façon semi-automatique des descriptions les plus exhaustives possibles des usages, dans leurs paramètres les plus fins. Ces travaux se caractérisent généralement par le recours à des approches hybrides : statistiques et symboliques (reposant sur des ensembles de règles) ; ils visent à fournir des thesauri, réservés à un domaine de spécialité dans le cas de l'ingénierie, ou encore des « ontologies d'un domaine ». Dans le cas des applications relevant de la pédagogie, la couverture des thesauri constitués vise à être la plus étendue possible, jusqu'à constituer la base d'ouvrages de référence, notamment pour le monde anglo-saxon, tels que le *Longman Dictionary of Contemporary English*, ou encore le thesaurus Roget. La prépondérance des approches sur corpus dans le monde anglo-saxon est à mettre en relation avec la disponibilité de corpus annotés pour l'anglais, de volume et de nature différente, autorisant aussi bien les approches classiques que les approches statistiques ou mixtes.

Les ressources linguistiques ainsi constituées sont toutes le résultat d'une analyse distributionnelle classique, ménageant généralement une forte part d'intervention humaine dans les applications les moins spécialisées, menés dans une perspective lexicographique. Inversement, dans les applications les plus proches de l'ingénierie linguistique, le coût que représente l'intervention humaine tend à être réduit à une phase de validation des ressources constituées. Quelque soit la couverture visée, la structure argumentale, c'est-à-dire les contraintes de sous-catégorisation entre un ensemble de verbes et leurs compléments habituels, constitue généralement la cible de ces études sur corpus.

2.2.2.2. Une base de signatures thématiques sous la forme d'une table du lexique-grammaire

Notre approche des corpus de spécialité, marquée par des objectifs applicatifs immédiats, se concentre sur les contraintes de sélection et de sous-catégorisation des verbes, dont on suppose une association avec un thème informationnel donné (ex. : cession-acquisition de société). Les phrases-noyaux ainsi constituées comprennent :

- des places/fonctions courantes, telles que agent, patient, destinataire. Ces fonctions sont, de façon lâche, associées aux places canoniques des sujets et compléments (respectivement direct, indirect et d'attribution) des verbes[72]. De ce fait, nous utilisons la notation peu marquée suivante : N0, N1, N2, où l'indice (0,1,2 …) symbolise la place au sein de la phrase-noyau. Ainsi, N0 signifie « le premier syntagme nominal, à gauche du verbe ». La phrase-noyau : **N0 acheter N1 pour N2**, décrit ainsi l'ensemble des phrases construites autour du verbe *acheter*, admettant trois syntagmes nominaux, le premier ayant la fonction de sujet grammatical, le deuxième celle de complément direct, le troisième comme complément facultatif, précisant le montant de la transaction.
- dans les cas où les arguments des prédicats (verbaux comme nominaux), les N*i*, constituent une classe suffisamment restreinte, une description sous la forme d'une grammaire locale en est donnée.
- des contraintes de formation précisant les constructions attestées, et les transformations syntaxiques autorisées. Ces contraintes sont en premier lieu tirées des corpus, puis généralisées ou supposées pour les cas non problématiques. Ainsi, par exemple, il peut se trouver que le corpus de référence ne comporte qu'une partie des constructions ou des transformations envisagées, ce qui amène à examiner, en ayant recours à notre intuition linguistique, la validité d'énoncés non disponibles en corpus[73].

---

[72] Par « lâche », nous entendons non catégorique. On peut envisager une probabilité d'association entre la place considérée (sujet, objet syntaxique) et le rôle casuel effectif, toutefois, en l'état actuel, LIZARD n'intègre pas ce type d'information.

[73] Ainsi, la construction semi-figée *mettre la main sur*, observée à l'indicatif (**N0 met la main sur N1**) dans les corpus ne semble pas pouvoir subir la transformation passive : * *la main a été mise sur N1 par N0*, * *N1 a été mis la main sur par N0*.

- des entrées lexicales, dans les cas où les arguments et les entrées verbales sont indissociables, malgré des possibilités d'insertions (ex. : adverbes, adjectifs). On se trouve alors devant des expressions à haut degré de figement, telles que : *mettre la main sur* (synonyme : *acheter*).
- des entrées lexicales, dans le cas de nominalisations disponibles en langue générale, telles que : *achat* (*acheter*), *acquisition* (*acquérir*).

Une fois les phrases-noyaux constituées, reste à choisir un format de représentation, ainsi que le langage formel adéquat. Dans le cadre de nos travaux, le langage formel était imposé par l'application destinée à utiliser les ressources linguistiques constituées par études sur corpus. Il s'agit, en l'occurrence, des variantes de transducteurs à états finis utilisés dans l'ensemble des traitements par le système Intex[74]. Dans ce cadre applicatif, le choix du format de représentation dépend du degré de réutilisabilité souhaité pour les ressources linguistiques considérées. En l'occurrence, la représentation sous forme de transducteurs graphiques Intex, bien qu'utile dans une phase exploratoire, doit être abandonnée, au profit d'une représentation sous forme de tables du lexique-grammaire, telles que décrites dans (Gross, 1975). L'intérêt de ce format de représentation réside dans sa souplesse (peu de contraintes induites par le formalisme) et sa simplicité (du texte Ascii). Une représentation sous forme de tables permet de garantir un degré élevé de réutilisabilité : tant les applications compatibles avec le système Intex que des applications étrangères sont susceptibles d'avoir accès aux ressources linguistiques ainsi constituées. En effet, ainsi que le montre l'extrait ci-dessous, l'adoption de ce type de représentation nous place d'emblée dans le cadre classique de bases de données (i.e. lexico-grammaticales), auxquelles des requêtes sont susceptibles d'être adressées afin de récupérer les informations codées[75].

[74] Voir le chapitre IV consacré à la mise en œuvre industrielle pour plus de détails sur le système Intex, utilisant les dictionnaires électroniques mis au point au LADL, ainsi que (Courtois, 1990), (Courtois & Silberztein, 1990) et (Silberztein, 1993).

[75] Voir (Balvet, 2001) pour une discussion de l'application des tables du lexique-grammaire au domaine de la terminologie.

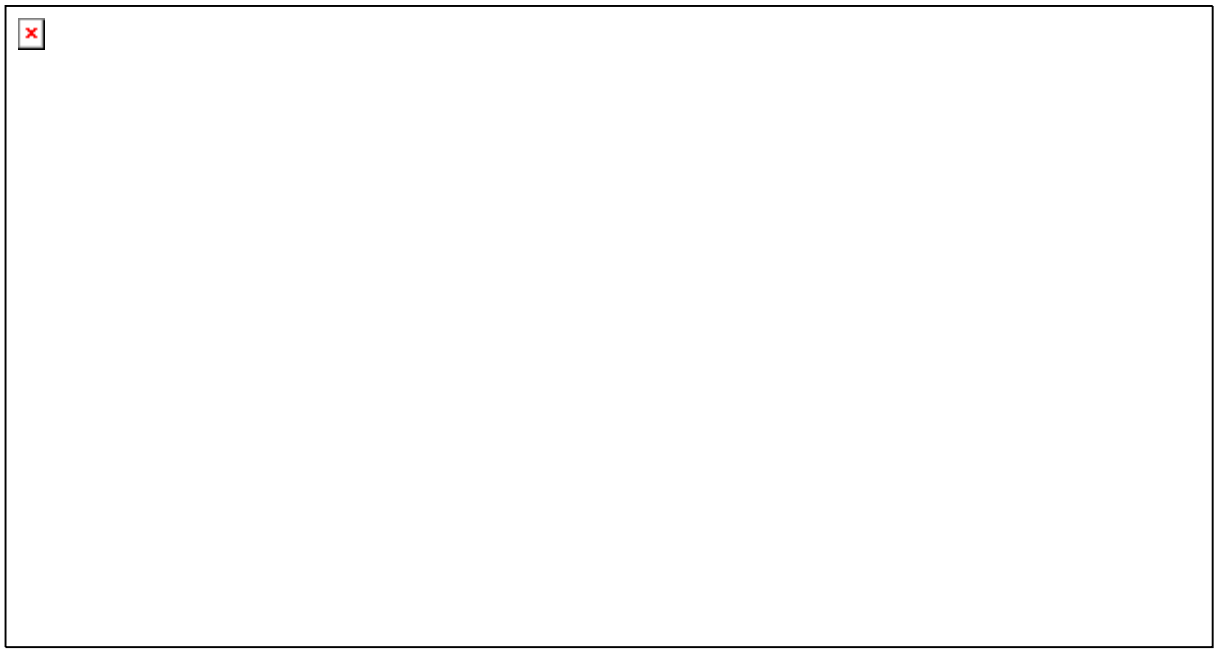

Figure 2 : un extrait d'une base de données lexico-grammaticales du domaine financier

L'extrait ci-dessus représente les paramètres lexicaux et syntaxiques des phrases-noyaux extraites d'un corpus de référence traitant des cessions et acquisitions de sociétés[76]. La table se lit comme suit :

- colonnes A, B, C : spécification du type des trois arguments les plus courants, en l'occurrence, sujet, objet direct et objet indirect ;
- colonne D : spécification de la nature de la particule préverbale, un pronom réflexif (codé par :Refl[77]) décrit par une grammaire locale, ou une chaîne vide (<E>) ;
- colonne E : spécification de l'entrée lexicale, à l'infinitif ;
- colonnes F à J incluse : constructions possibles (constatées sur corpus ainsi que déterminées hors corpus), en l'occurrence construction absolue (**N0 V**), transitive directe (**N0 V N1**), construction transitive indirecte (**N0 V Prep N1**), construction figée (**N0 V Const N1**), construction « maximale » (**N0 V N1 Prep N2**) ;
- colonne K : spécification de l'argument obligatoire, dans le cas d'une construction figée (*la main sur* pour *mettre*) ;

[76] Voir le chapitre IV pour une présentation plus détaillée du corpus financier utilisé.

[77] Pour une présentation des grammaires locales utilisées dans l'ensemble de nos travaux, voir l'annexe II.

- colonne L : complément circonstanciel habituel, décrit par une grammaire locale (:Capital) ;
- colonne M : spécification de la forme nominalisée correspondant à la forme verbal, tirée de ressources lexicales existantes telles que le Dictionnaire Intégral[78] ;
- colonnes N à P : spécification des transformations possibles (constatées sur corpus, ainsi que déterminées hors corpus).

La table ci-dessus peut être considérée comme une base de données lexicales, grâce auxquelles le système Intex permet de générer des grammaires locales sous la forme de transducteurs à états finis. La génération de ces grammaires locales, interprétables par Intex, se fait grâce à des transducteurs particulier : les métagraphes[79], qui permettent de spécifier des contraintes sur les grammaires générées. Ainsi, par exemple, il est possible de ne générer que la grammaire locale correspondant aux entrées verbales pouvant se trouver à la forme passive. La figure ci-dessous donne un aperçu d'un métagraphe.

Figure 3 : automate-patron, générant les grammaires locales correspondant aux constructions figées acceptant la forme active

Le métagraphe ci-dessus se lit de gauche à droite, les parenthèses numérotées indiquent que les séquences de caractères reconnues par les grammaires locales décrites entre parenthèses sont mémorisées. Les états figurant en grisé sont des appels à des sous-grammaires locales[80].

[78] (Dutoit, 2000).

[79] Les transducteurs utilisés par Intex sont généralement appelés « graphes », en raison de leur présentation graphique. D'où la dénomination de « métagraphe » pour des graphes factorisés, ou graphes-patrons.

[80] Pour plus de précision concernant les conventions s'appliquant aux graphes Intex, voir (Silberztein, 1993).

L'opérateur « @ » suivi d'un nom de colonne (de A à Z) fait référence aux colonnes de la table à laquelle le métagraphe est associé. Ainsi, dans le métagraphe ci-dessus, le premier état fait référence à la colonne A, le troisième fait référence à la colonne I, spécifiant quelles entrées sont des constructions figées. La sémantique de l'opérateur « @ » est double : dans les cas où des séquences de caractères autres que « + » ou « - » figurent dans les colonnes de la table, ces séquences sont recopiées dans l'état appelant lors de la compilation. Dans le cas où figurent un « + » ou un « - », l'état appelant constitue une porte logique : tous les appels aux informations de la table, situés après cet état, sont restreints par la contrainte énoncée[81]. En l'occurrence, pour le métagraphe ci-dessus, les appels aux entrées lexicales (@E, sixième état) sont restreints à celles qui vérifient la contrainte spécifiée en I : **N0 V Const N1**, autrement dit seules des constructions figées sont recopiées dans le sixième état à la compilation.

Le recours aux métagraphes permet la spécification de grammaires très génériques (des grammaires-patrons), instanciées par les entrées lexicales contenues dans une table. Ce dispositif fait ainsi l'économie d'une édition manuelle de grammaires locales, pour chaque entrée lexicale considérée. Il permet, de plus, de donner un caractère moins « procédural » aux grammaires locales construites à partir de transducteurs à états finis : une même table peut être associée à différents métagraphes (ex. : un métagraphe pour les formes au passif, un autre pour les nominalisations), des données lexicales éparses peuvent être regroupées dans une base centrale, ce que les transducteurs classiques ne permettent pas de réaliser.

2.2.3. Distributionnalisme probabiliste pour la découverte de signatures thématiques : détection de collocations

Dans la partie précédente, nous avons exposé une méthode distributionnelle, relevant du distributionnalisme classique, permettant de constituer une base de signatures thématiques. Nous explorons ici l'apport d'une approche distributionnelle probabiliste dans le cadre de la constitution de telles bases. Nous examinons, notamment, quelques techniques permettant la détection de groupes de mots présentant un degré de cohésion important : des collocations. Après avoir situé le cadre dans lequel s'inscrit la collocation, nous ferons une présentation générale des principales méthodes de détection de tels groupes de mots montrant une cohésion particulière. Enfin, nous discuterons de quelques collocations extraites du corpus Firstinvest,

[81] Ce principe a été développé dans (Senellart, 1999), ainsi que dans (Silberztein, 1999).

susceptibles de fournir une base pour l'élaboration de ressources linguistiques pour un système de filtrage d'information.

2.2.3.1.Définition

Les collocations sont des séquences constituées de plusieurs mots, pour lesquelles des contraintes de composition sont observables, dans un degré moindre que dans le cas des mots composés. Les collocations regroupent des éléments de nature différente, telles que les expressions semi-figées (ex. : *casser sa pipe*), les expressions idiomatiques, les mots composés et les associations dites habituelles[82]. La notion de collocation trouve son origine dans le domaine de la linguistique anglo-saxonne, dans ses applications à la pédagogie et aux études littéraires (Firth, 1957). Cette notion jouit actuellement d'un regain d'intérêt de la part de la communauté de la recherche d'information intégrant des contraintes d'ordre linguistique. En effet, les collocations, et les techniques de détection automatique employées avec profit, rendent compte de régularités observables dans le domaine de la Parole, ce qui permet d'envisager l'élaboration de grammaires locales de façon automatique, par confrontation avec des exemples positifs tirés des corpus.

Par ailleurs, le recours à des bases de collocations en RI permet d'envisager l'intégration de contraintes compositionnelles et idiomatiques, qui correspond, à nos yeux, à la « simplicité élaborée » prônée par (Spärck Jones & Kay). Cette intégration peut être vue comme un moyen terme entre la position linguistiquement faible dominante dans le domaine et une position linguistiquement plus exigeante, visant des analyses syntaxiques complètes.

2.2.3.2.Quelques techniques d'extraction de collocations

Les principales techniques d'extraction de collocations mettent en œuvre des techniques statistiques, dans le cadre d'approches pauvres en connaissances[83] (*knowledge poor*). En effet, les techniques basées sur des analyses morphosyntaxiques automatiques sont confrontées aux limites des analyseurs disponibles. De plus, le coût entraîné par le temps nécessaire au paramétrage des analyseurs automatiques joue en défaveur de ces approches. Par ailleurs, l'intérêt des approches pauvres en connaissances, ainsi que nous l'avons vu dans

[82] Par exemple, la spécification d'un montant pour une transaction, dans le domaine financier.

[83] Voir (Manning & Schütze, 1999, p. 151) pour une présentation des différentes techniques statistiques d'extraction de collocations.

le cas de l'indexation automatique, est une indépendance relative par rapport aux types de textes traités, en comparaison des approches à base de règles explicites.

Parmi les approches probabilistes pour l'extraction de collocations, nous traiterons essentiellement de celles basées sur des métriques visant à infirmer une hypothèse de cooccurrence entre deux éléments. Autrement dit, nous examinerons plus particulièrement les techniques visant à détecter des associations de mots dans une proportion en contradiction avec une répartition aléatoire[84]. Ces approches font appel à un ensemble de coefficients évaluant la probabilité, ou encore le degré de corrélation entre plusieurs éléments collocationnels, tels que : test du Khi² (coefficient de Pearson), t-test et information mutuelle.

Le test du Khi² et le t-test sont similaires dans leur principe : comparer des valeurs observées (ex. : des fréquences d'occurrence de paires de mots) sur un échantillon à des valeurs théoriques. Dans le cas du Khi², les valeurs comparées sont des effectifs[85], alors que dans le cas du t-test ces valeurs sont des moyennes.

La formule du t-test est la suivante[86], où *moy.* est la moyenne de l'échantillon, σ^2 la variance, N la taille de l'échantillon et μ la moyenne d'une distribution dont on suppose qu'est issu l'échantillon :

$$t = moy. - \mu / \sqrt{(\sigma^2 / N)}.$$

Formule 1 : t-test

Dans le cas du t-test, l'hypothèse nulle, c'est-à-dire l'hypothèse que l'on cherche à infirmer, est la suivante : l'échantillon considéré est pris d'un ensemble de données de distribution μ.

Dans le cas du test du Khi², l'hypothèse nulle est la suivante : les deux séries de mesures considérées (observées et théoriques) ne sont pas corrélées.

Dans les deux cas, des valeurs de référence permettent d'infirmer ou de confirmer l'hypothèse nulle, avec une probabilité d'erreur connue.

---

[84] D'autres approches sont possibles, telles que celle de (Smadja, 1993), basée sur des distances entre éléments collocationnels.

[85] Voir le chapitre IV pour une application du test du Khi² à l'évaluation des performances d'un système de filtrage d'information.

[86] Tirée de (Manning & Schütze, 1999).

Le score d'information mutuelle I[87] entre deux événements *x*, *y* (ex. : des mots), est tiré de la théorie de l'information[88]. Ce score est donné par la formule suivante, où *P(x)* et *P(y)* représentent les probabilités associées aux événements *x* et *y* isolément, et *P(x,y)* la probabilité associé à l'événement *(x,y)*[89] :

$$I(x,y) = \log 2\ P(x,y) / P(x)\ P(y).$$

Formule 2 : score d'information mutuelle

D'après (Manning & Schütze, 1999), le score d'information mutuelle est une mesure grossière de l'information apportée par la survenue d'un événement (un mot) par rapport à un autre. D'après Manninn & Schütze, l'information mutuelle est plus une mesure d'indépendance (lexicale) que de cohésion. De façon générale, les auteurs insistent sur les limites liées à l'utilisation des tests statistiques évoqués plus haut, notamment dans les cas où les éléments étudiés présentent des fréquences d'occurrence basses.

2.2.3.3.Transformation d'un corpus en n-grammes

Dans la plupart des cas, les approches statistiques supposent, dans un premier temps, un découpage des corpus en mots simples, selon une norme revenant généralement à l'adoption de la notion de mot typographique. Dans un deuxième temps, les corpus ainsi découpés, dont on a gardé la structure initiale (l'agencement des mots au sein du texte), sont transformés en n-grammes (généralement des 2grammes) selon le principe de la fenêtre coulissante. L'exemple simplifié ci-dessous illustre les deux premières phases de prétraitement des corpus.

12172. La Fnac lance DigiFnac. Pour répondre à l'offre tout-numérique, la Fnac lance un nouveau service.

[87] (Church & Hanks, 1990).

[88] Voir plus haut.

[89] Dans le cas des collocations, une paire constituée des mots *x* et *y*.

1. Les phrases du corpus sont découpées en mots simples (généralement : suite de caractères comprises entre deux délimiteurs).

12172

.

La

Fnac

lance

DigiFnac

.

Pour

répondre

à

l

'

offre

tout

-

numérique

,

la

Fnac

lance

un

nouveau

service

.

2. Parallèlement, un index des mots du texte est créé, chaque entrée de ce « dictionnaire » est associée à une fréquence d'occurrence[90], et, éventuellement à une position dans le texte.

Entrée	Fréquence d'occurrence
'	1
-	1
,	1
.	3
à	1
DigiFnac	1
Fnac	2
l	1
La	1
la	1
lance	2
nouveau	1
numérique	1
offre	1
Pour	1
…	

3. Enfin, le texte initial est transformé en n-grammes (i.e. 2grammes), autrement dit des groupes de *n* (i.e. 2) mots, constitués à partir du texte grâce à une fenêtre glissante, généralement fixe, partant d'une position *p* dans le texte jusqu'à $p + (n-1)$.

[90] Dans le cas présent, les fréquences d'occurrence n'ont qu'une valeur indicative.

12172	12172
.	.
La	La
Fnac	Fnac
lance	lance
DigiFnac	DigiFnac
.	.
Pour	Pour
répondre	répondre
à	à
l	l
'	'
offre	offre
tout	tout
-	-
numérique	numérique
,	,
la	la
	…

Exemple 2 : étapes principales du prétraitement d'un corpus en vue d'en extraire des collocations

De même que pour la liste des mots du texte, les n-grammes du texte sont indexés et associés à une fréquence d'occurrence. Notons que l'ensemble des étapes détaillées ci-dessus peuvent être adaptées en fonction d'un genre textuel particulier, d'une application, ou encore d'une langue donnée. Ainsi, le découpage des mots peut être plus ou moins fin, jusqu'à inclure des exceptions au principe du mot typographique (ex. : en français, *aujourd'hui*, découpé en *aujourd – ' – hui* ou non, en fonction de l'application). Il en va de même pour les principes d'indexation : les index peuvent contenir, ou non, certains mots, dont les fréquences d'occurrence sont jugées plus ou moins intéressantes (ex. : en français, la préposition *de* est l'un des mots les plus fréquents), ainsi que la ponctuation. La plupart du temps, les tentatives d'extraction de collocations à partir des corpus visent à ne conserver que les mots dits « sémantiquement pleins », au détriment des « mots grammaticaux », repérables par leur comportement distributionnel[91].

Une fois les n-grammes indexés, le comportement distributionnel particulier de certains d'entre eux peut être mis en valeur grâce à des outils statistiques, sélectionnant, par exemple, les paires dont la fréquence d'occurrence effective est supérieure à une fréquence

[91] Une fréquence d'occurrence élevée.

théorique, évaluée par extrapolation d'une loi de distribution donnée (ex. : la loi normale). Nous voyons essentiellement deux types de mesures statistiques pour l'extraction de collocations : les mesures globales, visant à repérer les n-grammes déviants par rapport à l'ensemble du corpus, et les mesures locales, visant à mesurer le degré d'association d'une amorce donnée avec plusieurs candidats au titre de collocation. Ces deux types de mesure correspondent à deux cas de figure dans l'exploration des corpus : dans le premier cas, on cherche dresser la liste de toutes les collocations d'un corpus donnée, dans le second cas, on chercher à distinguer parmi un sous-ensemble de candidats ceux dont la cohésion lexicale est la plus importante. Dans les mesures globales, on trouve généralement des mesures dérivées de la théorie de l'information (ex. : calcul de l'entropie maximale, de l'information mutuelle), qui permettent d'identifier les collocations présentant le plus fort degré d'association, par rapport à l'ensemble des collocations possibles. Dans les mesures locales, on trouve, entre autres, le t-score, le z-score et leurs variantes[92].

En termes linguistiques, la recherche de collocations consiste à isoler les éléments dont les dépendances syntagmatiques sont les plus fortes. Traduits en ces termes, on retrouve les principes de l'analyse distributionnelle harrissienne. Toutefois, là où le distributionnalisme est une analyse systématique, en vue de la délimitation d'unités linguistiques (des paradigmes), la plupart des techniques de repérage de collocations basées sur des approches statistiques, qui constituent à notre connaissance l'écrasante majorité des approches dans ce domaine, limitent l'analyse aux franges les plus cohésives des paires de mots traitées. Ces approches font, par ailleurs, le pari d'une absence de connaissances linguistiques, telles que constituance, ou classement des mots en parties du discours, elles ne cherchent donc pas explicitement à constituer des classes d'éléments linguistiques en tant que telles, mais bien plutôt à isoler des termes d'un domaine spécialisé, ou encore à améliorer le processus d'indexation automatique d'une base de documents. Ces approches relèvent donc plutôt des techniques opératoires en ingénierie linguistique que des outils d'exploration des corpus en vue d'une analyse linguistique. Notamment, le souci de généralisation des régularités constatées en corpus est le plus souvent absent dans ces approches.

[92] Voir (Biber *et al.*, 1998) "T-scores are useful when trying to contrast the use of two words, not for compiling a list of the most important collocates for a single word".

2.2.3.4.Quelques résultats d'une fouille de corpus spécialisé

Nous avons appliqué quelques-unes des mesures évoquées plus haut au corpus financier auquel nous consacrons notre étude. Nous avons, notamment, calculé, pour chaque sous-ensemble des paires de mots possibles, la probabilité associée à la survenue d'une « expansion » en fonction d'une « tête » donnée[93]. Ainsi, par exemple, pour la tête « AOL », les expansions possibles sont données par le tableau ci-dessous.

Tête	Expansion
AOL	)
AOL	.
AOL	dans
AOL	et
AOL	Europe
AOL	France
AOL	pour
AOL	Time

Exemple 3 : expansions associées à la tête « AOL »

Ainsi, pour la tête considérée, chaque expansion a une probabilité égale à 1/8. Cette probabilité permet de calculer, grâce à la formule donnée plus haut, un score d'entropie « conditionnelle » pour chaque expansion d'une tête (voir ci-dessous).

Tête	Expansion	Effectif Tête	Effectif Expansion	Probabilité Expansion|Tête	Entropie Expansion|Tête
AOL	)	8	1	0,125	0,375
AOL	.	8	1	0,125	0,375
AOL	dans	8	1	0,125	0,375
AOL	et	8	1	0,125	0,375
AOL	Europe	8	1	0,125	0,375
AOL	France	8	1	0,125	0,375
AOL	pour	8	1	0,125	0,375
AOL	Time	8	1	0,125	0,375

Exemple 4 : scores d'entropie conditionnelle des expansions de la tête « AOL »

[93] Les termes « tête » et « expansion » désignent respectivement le premier et le deuxième mot d'une paire. Cet emploi ne fait donc pas directement référence à la notion de tête et d'expansion dans le domaine syntaxique.

Le sous-corpus financier considéré, sur lequel nous basons l'ensemble de nos études, comporte un effectif total de 22558 2grammes. Muni des probabilités et des scores d'entropie conditionnelle, il est possible d'évaluer la cohésion lexicale des paires de mots, grâce à des mesures telles que l'information mutuelle, vue plus haut, ou encore une mesure tirée de (Ferret & Grau, 2001). Cette mesure, baptisée cohésion lexicale est mise en œuvre dans le cadre de l'élaboration d'une base de collocations à partir de textes journalistiques, dans un but de segmentation automatique par détection de changement de thème. La cohésion lexicale est donnée par la formule :

$$coh\ (x,y) = \log2\ (N.f(x,y) / f(x).f(y))$$

Formule 3 : cohésion lexicale

Où N représente l'effectif total d'éléments considérés (i.e. 22558), $f(x,y)$ la probabilité d'occurrence d'une paire de mots constituée des mots x et y, $f(x)$ et $f(y)$ la probabilité associée à l'occurrence des mots isolés.

Dans (Ferret & Grau, 2001), la cohésion lexicale est normalisée par l'information mutuelle maximale :

$$Imax = \log2\ N^2\ (Tf - 1).$$

Formule 4 : information maximale

Dans cette estimation de l'information maximale, Tf est la taille de la fenêtre. Dans notre cas, l'information maximale est : $Imax = \log2\ 22558^2 = 28{,}9227031$. Les paires de mots les plus cohésives présente un score de 1,250102.

Le score de cohésion lexicale, calculé pour des 2grammes tirés du sous-corpus financier considéré, permet d'extraire des paires telles que celles présentées ci-dessous.

Lex1	Lex2	Entropie Lex2|Lex1	Coh(x,y) norm
millions	d	0,314493783512482	1,12083119153512
hauteur	de	0,5	1,11960955353711
News	Corp	0	1,11960955353711
M	.	0,5	1,11728908592291
Marie	Messier	0,5	1,11485538089721
dirigé	par	0	1,11485538089721
Pernod	Ricard	0	1,11485538089721
True	North	0	1,11485538089721

Exemple 5 : quelques 2grammes fortement cohésifs

Le tableau présente les 2grammes extraits, associés à un score d'entropie conditionnelle et une mesure de cohésion lexicale. Les 2grammes sont triés par ordre décroissant sur le score de cohésion normalisé. Sur le sous-corpus considéré, la mesure de cohésion telle que tirée de (Ferret & Grau, 2001) permet surtout de détecter des entités nommées, telles que des noms de société (ex. : Pernod Ricard), des noms de personne (ex. : Marie Messier), ou encore des associations habituelles pour le domaine (ex. : [à] hauteur de, dirigé par).

L'utilisation conjointe des scores d'entropie conditionnelle et de cohésion lexicale, projetés sur des 4grammes, par exemple, permet d'étudier des sous-domaines tels que celui des noms propres, commençant par « Jean ».

Lex1	Entropie Lex2|Lex1	Cohésion Lex1|Lex2	Lex2	Entropie Lex3|Lex2	Cohésion Lex2|Lex3	Lex3	Entropie Lex4|Lex3	Cohésion Lex3|Lex4	Lex4
Jean	0	1,1644	-	5,151105E-02	0,9036537	Marie	0,5	1,114855	Messier
Jean	0	1,1644	-	5,151105E-02	0,8233732	Claude	0,5283208	1,034575	Darmon
Jean	0	1,1644	-	5,151105E-02	0,8233732	Claude	0,5283208	1,034575	Darmon
Jean	0	1,1644	-	5,151105E-02	0,8233732	Claude	0,5283208	0,9452001	Cabre
Jean	0	1,1644	-	5,151105E-02	0,8233732	Claude	0,5283208	0,9452001	Decaux
Jean	0	1,1644	-	5,151105E-02	0,7887983	Jacques	0,5283208	0,9452001	Poutrel
Jean	0	1,1644	-	5,151105E-02	0,7887983	Jacques	0,5283208	0,9452001	Bresson
Jean	0	1,1644	-	5,151105E-02	0,7542233	Louis	0,5	0,9654251	Beffa

Exemple 6 : les noms propres construits sur la tête « Jean » (extrait)

Cet extrait permet d'estimer les relations de dépendance entre les différents éléments constituant une famille de noms propres, construits sur la tête « Jean ». Cet exemple limité permet, à nos yeux, d'envisager l'élaboration de grammaires locales de sous-domaines tels que celui des entités nommées, sur la base du comportement distributionnel observable seul[94]. En effet, dans l'exemple ci-dessus, on remarque des différences dans l'incertitude (entropie)

[94] Voir (Charniak, 1993) pour une présentation plus complète de l'induction de grammaires PCFG à partir de corpus.

dans laquelle on se trouve quant à la survenue du mot suivant. Ces différences peuvent être interprétées comme suit :

- en termes de relations paradigmatiques, deux classes d'éléments se dégagent, en l'occurrence les éléments pour lesquels l'entropie conditionnelle est faible (Jean, -, Marie, Claude, Jacques, Louis), par rapport à une classe d'éléments pour lesquels l'entropie conditionnelle est plus élevée (Darmon, Cabre, Decaux, Poutrel, Bresson, Beffa). Ces deux classes peuvent être interprétées comme l'observation des régularités connues quant à la formation des noms propres : certains éléments sont des prénoms, d'autres des noms de famille.
- en termes de relations syntagmatiques, on retrouve une partie des règles de formation des noms propres, notamment des prénoms composés : l'entropie conditionnelle associée au caractère « - » est nulle, pour la tête « Jean ». Une entropie conditionnelle nulle entre une tête et son expansion immédiate est majoritairement associée, dans notre corpus, à des entités nommées (ex. : News Corp, True North, Etats-Unis, Pernod-Ricard). On peut interpréter ce comportement comme la manifestation d'un gradient de compositionnalité des éléments composant une entité nommée. En l'occurrence pour les 4grammes considérés, *Jean* et – sont indissociables, - et les éléments *Marie*, *Claude*, *Jacques*, *Louis* le sont dans une moindre mesure.

Soulignons, toutefois, qu'en raison de la taille modeste du corpus étudié ici (moins de 1 Mégaoctet de texte), il est difficile d'en extraire des collocations par le biais des mesures présentées plus haut. De ce fait, les observations consignées ici sont à prendre comme des perspectives de recherche, dans l'attente de la disponibilité de corpus spécialisés représentatifs, outillés et étiquetés, comme le proposent des auteurs tels que Habert.

2.2.4. Ressources linguistiques issues d'une analyse probabiliste

Les approches distributionnelles probabilistes permettent de constituer des bases de collocation, autrement dit des bases de termes présentant des contraintes de composition. Ces bases peuvent être mises en œuvre dans le domaine de la recherche d'information, afin de dépasser les limites des techniques d'indexation automatiques classiques.

2.2.4.1.Des bases de collocations pour la recherche d'information

Nous l'avons vu, le principe de l'indexation automatique par extraction de descripteurs de contenu, tirés du stock lexical des documents traités, présente des lacunes. La principale d'entre elles est d'oblitérer complètement l'information structurelle donnée par les contraintes d'ordre syntaxique. Ainsi, comme Bar-Hillel le fait remarquer, des documents traitant de thèmes différents, mais présentant un même profil après indexation, seront considérés également pertinents. L'intégration d'une phase de détection de collocations à la procédure d'indexation permet de restaurer une partie de l'information linguistique perdue au cours du processus de sélection des termes descripteurs.

La notion de collocation regroupe, comme nous l'avons vu, des éléments aussi divers que des expressions figées, des tournures idiomatiques ou des termes techniques. Les techniques de détection des collocations permettent également de retrouver des entités nommées (ex. : noms de société, noms propres, toponymes). Or, ces entités nommées constituent des marqueurs thématiques utilisables dans un contexte de RI[95], menée sur des corpus de type journalistique[96].

En tant que séquences particulièrement cohésives sur le plan syntagmatique, les collocations présentent généralement une cohésion thématique forte. Cette cohésion peut fournir la base de systèmes de segmentation thématique, tel que le système ROSA présenté dans (Ferret & Grau, 2002).

2.2.4.2.Des collocations aux grammaires locales probabilistes

Au-delà des applications en RI, il est possible de considérer l'extraction de collocations à partir de textes spécialisés comme une étape préliminaire dans un processus plus général d'induction de grammaires à partir d'exemples positifs. En effet, les techniques d'extraction évoquées ci-dessus tendent à mettre en évidence la cohésion existant entre plusieurs lexèmes, en d'autres termes leurs contraintes compositionnelles, au sens large. Cette cohésion peut traduire

---

[95] Voir (Fourour, 2002).

[96] Ces marqueurs ont une valeur dépendante du contexte historique. Ainsi, dans le courant de l'année 2000, la mention de *Microsoft* ou de *Bill Gates* dans des textes journalistiques, notamment des dépêches, pouvait être associée de façon quasi-catégorique à un thème : la procédure anti-trust menée contre *Microsoft*.

- des contraintes de sélection (ex. : un verbe et ses compléments habituels, des expressions quasi-figées) ;
- des contraintes idiomatiques, privilégiant la cooccurrence de certains termes ;
- des phénomènes de composition, en termes de morphologie compositionnelle.

Toutefois, la constatation d'une certaine cohésion lexicale ne reste qu'une description d'une régularité constatée en corpus, tant qu'aucune procédure de généralisation n'intervient. Le distributionnalisme classique de Harris visait essentiellement, par le recours à des procédures d'approximation et de promotion d'éléments au rang d'unités, à opérer une telle généralisation. Le but poursuivi, détaillé dans (Harris, 1951), est la mise en œuvre d'une analyse en constituants immédiats, en partant des régularités observées en corpus. L'ensemble des travaux dans le domaine de l'induction grammaticale, opérée aussi bien par des procédures statistiques[97] que symboliques, vise une telle analyse en constituants immédiats, à partir de classes construites automatiquement sur corpus. Dans ce domaine, la mise en œuvre de procédures non catégoriques de découverte d'éléments cohésifs peut permettre d'envisager cette cohésion sous la forme d'un continuum. Reprenant la distinction établie par Herdan, les éléments cohésifs les plus fréquents peuvent être associés au domaine grammatical, donc de la Langue, les moins fréquents au domaine lexical, donc de la Parole.

2.3. Conclusion

Nous l'avons vu, le domaine de la recherche d'information partage avec les études sur corpus l'objet d'étude que constituent les productions linguistiques. Ces productions ne sont, cependant, pas envisagées sous l'angle de leur sens, mais de leur contenu informatif. Cette précision permet de contourner le problème de la détermination du sens à partir des seuls observables linguistiques, toutefois la question du contenu reste tout aussi épineuse que celle du sens.

En effet, la détermination du contenu informatif d'un document, tant par des méthodes manuelles classiques (indexation manuelle) qu'automatiques, se heurte au problème de la

[97] Voir les travaux de (Finch, 1993), (McMahon, 1994), et (van Zaanen, 2001).

détermination des éléments informatifs qui relèvent d'un point de vue objectif, par rapport à ceux qui relèvent d'un point de vue individuel et subjectif[98]. Cette différence de points de vue engendre une tension, dont les effets se font sentir, notamment, par une variation inévitable dans le choix de descripteurs de contenu par des indexeurs humains[99], autrement dit, un désaccord profond et inévitable sur des critères de classification.

Nous avons adopté une définition fonctionnelle de l'information, qui permet de concilier ces deux points de vue : la fonction informative d'un document peut être envisagée en termes de valeur au sein d'un système, au sens saussurien. Pour chaque utilisateur d'un système d'information, cette valeur peut être vue comme déterminée :

- par des observables linguistiques, tels que les choix lexicaux et syntaxiques ;
- par un « état cognitif » (ex. : des attentes, une expérience du domaine) propre à chaque utilisateur.

Il est possible d'envisager les observables et l'état cognitif comme deux contraintes, dans un cadre formel proche de celui de la théorie de l'optimalité, évoqué au précédent chapitre dans le domaine de la linguistique de corpus. Par ailleurs, en établissant une analogie avec le domaine linguistique et la distinction entre le plan de la Langue et celui de la Parole, les éléments qui relèvent d'un point de vue collectif sont à chercher du côté des régularités, les éléments relevant d'un point de vue individuel étant à chercher du côté des singularités. En poussant l'analogie, on peut envisager le recours à une approche non catégorique de la valeur informative des documents en RI, basée sur une conception probabiliste du distributionnalisme.

Cette conception fonctionnelle de la valeur informative rapproche plus encore les domaines de la linguistique de corpus et celui de la recherche d'information : on peut envisager l'application de la méthode distributionnelle dans les deux cas, centrée sur les données linguistiques observables. L'ensemble des applications en RI reprennent, souvent implicitement, la conception distributionnelle de la valeur, en partie déterminée par les

[98] Par exemple, une stratégie personnelle de recherche d'information.

[99] Voir les expériences relatées dans (Coyaud, 1972).

contextes d'occurrence possibles : c'est le cas, notamment pour l'indexation automatique, dans laquelle le comportement distributionnel des descripteurs de documents n'est considéré que dans une version simplifiée[100]. Ainsi, le contexte d'occurrence considéré est celui du document tout entier, et non pas une phrase, voire un groupe de mots, d'où découle que la distribution de ces éléments n'est envisagée que sous l'angle de leur occurrence effective.

La RI et la linguistique de corpus sont donc conceptuellement proches : ces deux domaines partagent le même objet d'études, ainsi qu'une partie de la méthode distributionnelle. Cette parenté peut s'expliquer par les origines communes de la linguistique informatique[101] et de la RI : en effet, la naissance du TALN a été provoquée par une volonté de maîtrise de l'information (en tant que contenu) par des organismes gouvernementaux, dans un contexte de guerre froide[102].

Nous avons vu quels espoirs, mais également quelles déceptions étaient attachés à une telle alliance, notamment par le bilan dressé par (Spärck Jones & Kay, 1973), dont les conclusions nous apparaissent toujours valables aujourd'hui : le recours à des représentations linguistiques de haut niveau (ex. : arbres de dépendance syntaxique), dans les phases d'indexation, ne se traduit pas par une augmentation significative des performances des systèmes d'information. Bien au contraire, les approches adoptant la position d'une linguistique faible semblent fournir les meilleurs résultats. La prépondérance de ces approches est manifeste, ce qui pose la question de l'utilité des représentations linguistiques de haut niveau en tant que moyen d'accéder au contenu informatif des documents. L'adoption de ce point de vue linguistique faible en RI est à mettre en parallèle avec la même tendance observée en TALN, soulignée par (Habert, 1998).

---

[100] Les développements récents de la sémantique distributionnelle (*distributional semantics*) dans le domaine de l'IR (DSIR) constituent une tentative, plus aboutie que ce que nous avons présenté ici, d'application des principes distributionnalistes à d'autres domaines que la linguistique de corpus. Voir, à ce sujet, (Rajman *et al.*, 2000).

[101] Que nous considérons comme une branche de la linguistique de corpus.

[102] Outre-Atlantique, les premiers travaux dans le domaine de la linguistique formelle sont, le plus souvent, financés par des organismes dépendant du Ministère de la Défense nord-américain. (Chomsky, 1957), par exemple, a été financé en partie par l'US Army, l'Air Force Office et le Navy Office.

> Le TALN, pratiquement depuis ses origines, a cherché à déterminer la complexité en termes de grammaires formelles, du langage humain (…). Les travaux récents en parsage robuste, surfacique (*shallow parsing*) [Grefenstette 1996], [Roche, 1996] font naître l'hypothèse que, sur le plan syntaxique au moins, le langage articule des fonctionnements réguliers et simples, dominants, avec des zones de complexité, restreintes.
>
> (Habert, 1998, p. 156)

Nous avons vu, au cours du premier chapitre, quelles avancées avaient eu lieu en linguistique de corpus, dans le domaine de l'induction automatique de grammaires à partir des seuls observables linguistiques. L'ensemble des recherches menées dans ce sens tend à remettre en cause la caractérisation formelle du langage naturel établie par Chomsky : les grammaires hors-contexte intégrant une dimension probabiliste, par exemple, sont vues comme généralement suffisantes dans la plupart des applications développées en ingénierie linguistique[103]. Cette prépondérance des approches linguistiquement faibles dans les domaines centrés sur les productions linguistiques effectives pose, de façon générale, la question de la nécessité des approches linguistiquement fortes, basées sur des systèmes de règles explicites.

[103] En somme, le recours à des langages formels plus contraints ne semble nécessaire que dans le cas des énoncés construits par les linguistes eux-mêmes.

CHAPITRE 3

Le filtrage d'information

Ce chapitre est consacré au Filtrage d'Information (désormais FI), une sous-tâche de l'activité de Recherche d'Information (désormais RI). Le FI se caractérise par un contexte de mise en œuvre particulier : une RI en temps contraint, opérée sur un flux d'information[1], à partir d'un besoin en information stabilisé. Le FI est donc essentiellement une situation de diffusion ciblée d'information, dans laquelle l'évaluation de la pertinence se fait document par document, et non pas sur une collection de documents : en conséquence, les documents traités sont soit sélectionnés, soit rejetés, sans aucune autre alternative (ex. : classement d'un ensemble de documents).

Sous la pression du gouvernement fédéral nord-américain, le domaine du FI automatisé s'est essentiellement constitué autour des systèmes développés pour l'indexation automatique (ex. : SMART). Nous tentons d'établir, dans le présent chapitre, que l'activité de FI est loin de constituer une tâche facilement modélisable, malgré le parti pris simpliste de la vision nord-américaine, notamment, de l'automatisation de tâches de RI. Nous posons, en effet, que le FI tel que réalisé par des humains est une tâche cognitive complexe, qui repose sur un ensemble de compétences cognitives[2], l'expertise acquise sur un ou plusieurs domaines, ainsi que le contexte dans lequel est réalisé le filtrage, qui représentent autant de contraintes qu'un processus de catégorisation menant à la décision de sélectionner un document ou non, doit accommoder au mieux. Nous abordons, de ce fait, de manière détournée les problèmes essentiels que sont la modélisation de la compréhension du langage naturel en vue de son automatisation, ainsi que celle de processus de catégorisation complexes, pour aboutir à la question essentielle de la subjectivité nécessaire au processus de filtrage.

---

[1] Par exemple : courrier électronique, dépêches journalistiques actualisées en temps réel.

[2] En l'occurrence des compétences linguistiques, une connaissance du domaine, la faculté de prendre des décisions, et l'interprétation d'un message en fonction d'un contexte.

Nous présentons, dans une première partie, le contexte dans lequel est née la notion de FI, essentiellement attachée au domaine de la documentation (ex. : centres de documentation, bibliothèques), comme l'ensemble des activités de RI. Dans une deuxième partie, nous nous penchons sur les caractéristiques de quelques systèmes de FI. La troisième partie de ce chapitre est consacrée aux problèmes de modélisation de l'expertise humaine que pose l'automatisation du FI, la quatrième partie est, elle, dédiée aux difficultés d'évaluer les performances de systèmes automatiques de FI.

3.1. Aperçu historique de la notion de filtrage d'information

Le FI est né d'un besoin très concret : d'une part réduire la charge de travail des documentalistes, d'autre part, fournir un service personnalisé aux utilisateurs de services de documentation, en leur apportant une information ciblée, en fonction de leurs besoins.

Dans cet aperçu historique de la notion de FI, nous nous appuierons essentiellement sur les écrits fondateurs de Luhn[3], ainsi que sur les actes des conférences d'évaluation américaines TREC, telles que publiées par le NIST. De ce fait, la présente partie a pour but de préciser quelle définition du terme « filtrage d'information » nous adoptons. En effet, le domaine de la recherche d'information subit les influences croisées des différents corps de métier desquels il a émergé : documentation, informatique, ou encore renseignement militaire, qui se traduisent par un certain flou terminologique.

3.1.1. Naissance d'un concept : la veille économique

Dans son article paru en 1958, Luhn pose les bases conceptuelles des systèmes d'information modernes. Il propose un concept que nous traduisons en français par « veille économique » afin de mieux souligner l'aspect stratégique lié à cette activité[4].

[3] Voir (Luhn, 1958). Pour une présentation historique du domaine, voir également (Oard & Marchionini, 1996).

[4] Le terme « Intelligence » en anglais est lié aux activités de renseignement stratégique, qu'on désigne habituellement en français par « veille stratégique ».

3.1.1.1.Les *Business Intelligence Systems*

La notion de « systèmes de veille économique », traduction approchée de « Business Intelligence Systems », définit un cadre pour une activité de gestion de l'information reposant sur les pratiques classiques en documentation (ex. : au sein d'une bibliothèque) où des opérateurs humains définissent des profils pour des utilisateurs individuels, profils servant à la sélection de documents par un système automatique sur la base d'une correspondance exacte (*exact match*). Dans cette conception initiale, chaque profil d'utilisateur, par la description des centres d'intérêt des abonnés au service de diffusion ciblée d'information, est conçu pour identifier un utilisateur unique. De plus, le profil de chaque utilisateur est mis à jour à l'arrivée de tout nouveau document (ex. : commande d'ouvrages). La chaîne de traitement de l'information aboutissant à la confrontation entre les besoins en information (profils) des utilisateurs du système et les informations contenues dans les documents entrants fut dénommée par Luhn « Dissémination Sélective de la Nouvelle Information » (*Selective Dissemination of New Information, SDNI*). Les concepts introduits par Luhn identifient toutes les étapes d'un système d'information moderne, bien que les supports (microfilm, édition sur papier) et les techniques de l'époque supposent des choix d'implantation particuliers.

On le voit, la naissance du concept de filtrage d'information, et de façon plus large celle de recherche d'information, repose sur un besoin concret : assurer une diffusion d'information ciblée dans le cadre d'une activité économique intense, en partant d'une infrastructure documentaire existante (i.e. les centres de documentation, ou bibliothèques classiques).

3.1.1.2.De la *SDNI* à la *SDI*

La notion de Diffusion Sélective d'Information est née des efforts d'un groupement d'intérêts spéciaux[5] (*Special Interest Group*) nord-américain sur la SDNI, abrégée en SDI. Housman, dans son rapport technique délivré en 1969, effectue un recensement des systèmes utilisant la SDI aux États-unis. Il en identifie une soixantaine, neuf d'entre eux totalisent plus de 1000 utilisateurs au moment de l'étude. Ces systèmes suivaient généralement les étapes

[5] Dans l'histoire du développement des nouvelles technologies aux États-unis, les SIG jouent un rôle prépondérant. En identifiant un besoin et des techniques susceptibles d'y répondre, les SIG ont souvent permis d'évaluer la faisabilité d'une approche, tout en quantifiant les retombées économiques par des études de marché.

décrites par Luhn, à l'exception de la mise à jour automatique des profils d'utilisateurs, que seule une infime minorité d'entre eux (4 sur 60) mettait en œuvre.

Ainsi, dès la fin des années 1960, comme l'atteste l'étude de Housman, le besoin de systèmes de diffusion ciblée d'information, prenant en compte les besoins d'utilisateurs individuels, se faisait sentir. Ce besoin, accru par la disponibilité nouvelle d'information textuelle au format électronique, a donné naissance au terme de « filtrage d'information » sur la base de la SDI. Denning, dans son article paru en 1982, est l'un des premiers à utiliser ce terme pour désigner un processus visant à préserver la « bande passante mentale » (*mental bandwidth*) des utilisateurs des systèmes de courrier électronique, un nouveau moyen de communication. Cette réduction du flux d'information avait pour particularité de se baser sur le contenu des messages, et non plus seulement sur des indices tels que l'identité du correspondant, ce qui inaugura la notion de recherche d'information à partir de contenu (*content-based Information Retrieval*), une des branches de la RI actuelle.

On voit là à l'œuvre une deuxième contrainte très pragmatique ayant influencé le développement du filtrage d'information : à la contrainte initiale de maximiser l'information pertinente pour chaque utilisateur, en fonction de son profil, s'est ajoutée celle de minimiser la perte de temps induite par l'information non pertinente introduite par l'augmentation du volume des échanges, due aux nouveaux moyens de communications.

3.1.2. TREC et le filtrage d'information

Les conférences TREC, de même que les conférences MUC pour l'extraction d'information, ont joué un rôle prépondérant dans le développement du domaine de la recherche d'information automatisée. Ces conférences, en regroupant des équipes de différentes nationalités, tant du domaine public que privé, ont eu pour ambition de confronter des approches techniques différentes sur des données normalisées.

3.1.2.1.Une conférence d'évaluation internationale

En 1987, sous l'impulsion, et grâce au soutien financier du ministère de la défense américain (DARPA), était organisée la première conférence d'évaluation de compréhension automatique de messages MUC, précédant les conférences TREC, plus axées sur la fouille de textes (*text retrieval*). La septième et dernière conférence MUC eut lieu en 1998, alors que les conférences TREC en sont à leur neuvième édition, ce qui montre l'importance de

l'engagement d'institutions telles que le ministère américain de la défense dans le domaine de la recherche d'information. Les principes directeurs de ces conférences sont les suivants :

- définir les principaux domaines et sous-domaines de la RI ;
- fournir des données de référence normalisées, dans le but de comparer les performances de systèmes de RI, grâce à des métriques communes ;
- faciliter l'échange entre équipes participantes, issues aussi bien de l'industrie que du domaine public (universités, entités gouvernementales, laboratoires privés).

Le projet TIPSTER, lancé en 1990 sous l'impulsion du DARPA, fut la principale source de financement des conférences MUC, qui a surtout été l'occasion de concrétiser la mise au point de systèmes de sélection de messages grâce aux techniques issues du domaine de l'extraction d'information. TIPSTER mettait l'accent sur le recours à des techniques statistiques pour la présélection des messages (*document detection*), phase considérée comme essentielle et devant précéder toute autre technique plus sophistiquée, TALN notamment. Le DARPA, se basant sur les résultats du projet TIPSTER et l'expérience des conférences MUC, finança et organisa, dès 1992, en collaboration avec le NIST, les conférences TREC, qui reprennent les principes directeurs exposés plus haut.

3.1.2.2.Des débuts hésitants

Sous l'impulsion de la démarche normalisatrice des conférences TREC, le domaine de la fouille de textes s'est spécialisé : des tâches principales et des sous-tâches, organisées en une hiérarchie la plus cohérente possible, ont été définies. Toutefois, tous les sous-domaines de la fouille de textes n'ont pas connu le même développement ; c'est le cas du filtrage d'information, entre autres.

Ainsi, dès Novembre 1991, un atelier sur le filtrage d'information haute performance (*High Performance Information Filtering*), sponsorisé par Bellcore et le SIG sur les systèmes d'information bureautique (*Office Information Systems*), était organisé, au cours duquel plus de quarante publications examinèrent le domaine du filtrage à partir de plusieurs perspectives différentes : de la sélection de l'information à la modélisation de l'utilisateur, en passant par les domaines d'applications, les détails techniques et logiciels ainsi que des considérations sur la confidentialité et des études de cas. Ces publications furent regroupées dans une édition spéciale des *Communications of the ACM* datée de Décembre 1992.

Toutefois, du côté de TREC, le filtrage d'information ne connut que des débuts très hésitants, ne se focalisant que sur un des aspects de l'activité : le filtrage à partir du contenu[6]. Dans les premières éditions de TREC[7], suivant une organisation en tâches et sous-tâches bien établie pour l'indexation et la recherche de documents, par exemple, le filtrage d'information n'était considéré que comme une recherche exploratoire, au même titre que le volet dédié au TALN[8].

3.1.2.3. Une stabilisation tardive

Le filtrage d'information proprement dit n'apparaît qu'à la quatrième édition de TREC. Cette édition est l'occasion de distinguer entre routage et filtrage d'information. Ce dernier est défini comme une tâche de sélection binaire des documents, sur un principe proche de celui établi par Luhn pour la SDI, à la différence que les profils évalués restent fixes après paramétrage sur les corpus d'apprentissage fournis par le NIST. Avec les éditions successives de TREC, le filtrage d'information, initialement confondu avec la tâche de routage d'information, s'est vu lui aussi spécialisé, découpé en tâches principales et sous-tâches. Les dernières conférences TREC ont ainsi abouti à une distinction entre filtrage automatique et filtrage dit interactif (semi-automatique). Le filtrage automatique a, à son tour, été distingué entre filtrage par lots et routage, en fonction de la décision de sélection opérée par les systèmes évalués : binaire pour le filtrage par lot, suivant la définition de Luhn pour la SDI, continue pour le routage (scores de pertinence). Dans la suite du présent exposé, le terme « filtrage d'information » servira à désigner le filtrage par lots tel que défini au cours des conférences TREC, à partir de leur septième édition.

---

[6] D'autres types de filtrage ont été évoqués au cours du développement du domaine, tel que le filtrage collaboratif, ou social, prenant en compte les avis des utilisateurs sur la qualité informative des documents consultés. Il est ainsi envisageable, dans une perspective de diffusion ciblée, de ne prendre en compte que les avis des différents utilisateurs, indépendamment du contenu des documents.

[7] La première édition, considérée comme exploratoire, eut lieu du 4 au 6 Novembre 1992.

[8] Ces deux domaines ont, d'ailleurs, également en commun une naissance remontant aux années 1960, une demande certaine de la part des utilisateurs potentiels, et une mise en œuvre difficile en raison du matériau traité, le langage naturel, éventuellement porteur d'une charge informative.

3.2. Approches pour le filtrage d'information

Cette partie est consacrée à l'étude des approches dominantes en filtrage d'information. Nous examinons les spécificités techniques de quelques systèmes se réclamant du filtrage d'information. Nous distinguons essentiellement entre systèmes basés sur des moteurs classiques d'indexation et de recherche et systèmes visant à reconnaître des séquences de mots-clés dans les documents traités. Nous tenterons d'établir que le premier type de systèmes relève du routage, plus que du FI tel que défini plus haut. Par ailleurs, nous tenterons de montrer que les systèmes appartenant au deuxième type restent limités dans les fonctionnalités de filtrage qu'ils offrent, à moins d'adopter, comme nous le proposons, une approche considérant non plus de simples mots-clés comme descripteurs de thème, mais bien des unités lexicales complexes.

3.2.1. « Filtrage d'information » basé sur un moteur de recherche et d'indexation

Le domaine du FI est largement dominé par les approches reposant sur une adaptation à une tâche de *push* d'un système pensé pour le *pull*. Ces systèmes dominants tirent parti de l'infrastructure commerciale mise en œuvre pour les moteurs de recherche et d'indexation sur lesquels ils reposent, ainsi que sur l'effet de convergence induit par les conférences TREC. Ces systèmes se réclament du FI, or ils sont loin de se conformer à la définition donnée par TREC. L'examen des principes généraux d'indexation automatique, qui sous-tendent les moteurs SMART et toutes leurs variantes (ex. : PRISE, du NIST), nous permettra de trancher entre routage et filtrage d'information.

3.2.1.1. Principes d'indexation automatique

Comme nous l'avons évoqué dans le chapitre II, les principaux moteurs de recherche et d'indexation reposent sur des variantes de l'approche décrite dans (Salton, 1968 ; 1971), connue sous le nom de méthode (ou modèle) vectorielle[9]. Nous l'avons vu, ces systèmes considèrent les documents contenus dans la base à indexer comme des ensembles non

[9] Vector Space Model (ou Method) en anglais.

ordonnés, ou « sacs de mots » typographiques[10], autrement dit les notions d'ordre des constituants, ainsi que la structuration textuelle (ex. : phrases, paragraphes, chapitres) ne sont généralement pas prises en compte.

Par ailleurs, dans cette approche, tous les mots n'ont pas le même statut. En effet, tant les mots très fréquents (ex. : *de*, pour le français) que les *hapax* sont considérés comme peu porteurs d'information. De ce fait, ils sont généralement absents des bases d'indexation. De plus, les différences de casse (majuscules, minuscules) ne sont généralement pas prises en compte afin de réduire le risque de silence, ce qui entraîne l'indexation des entités nommées (ex. : les noms propres) sur les mêmes bases que les autres mots[11].

Une fois les documents débarrassés des mots considérés comme peu porteurs d'information, l'indexation elle-même consiste à construire un vecteur à n dimensions pour chaque document, n étant égal au nombre de mots différents contenus dans le document. Ainsi, chaque document est représenté par un sous-ensemble des mots qu'il contient, considérés comme des descripteurs suffisamment fiables du contenu du document. La base de documents elle-même représente un espace à N dimensions, N étant égal à l'effectif total de mots différents contenus dans la base. Ainsi, chaque document représente un vecteur dans l'espace de la base.

La phase de recherche, initiée par une requête d'utilisateur, consiste à comparer le profil de la requête, dont les mots sont considérés de la même façon que pour la phase d'indexation, avec les profils des documents de la base indexée. Autrement dit, la phase de recherche consiste essentiellement à mesurer la distance entre deux vecteurs dans un espace à N dimensions : celui représentant la requête et celui d'un document de la base, opération répétée pour tous les documents de la base. L'ensemble des documents est ainsi trié en fonction d'une métrique de distance calculée entre le vecteur de la requête et leur vecteur d'indexation.

Les variantes de cette approche reposent sur des algorithmes propriétaires destinés à optimiser les phases d'indexation et de recherche, ou en fixant, de façon plus ou moins

[10] Toute séquence de caractères délimitée par deux séparateurs typographiques : espace, ponctuation.

[11] Un exemple trivial est celui de V. Poutine, président actuel de la Russie, que des moteurs d'indexation classiques considèrent de la même façon que la « poutine », spécialité québécoise, alors qu'un certain nombre d'indices typographiques permettraient de les distinguer.

empirique, des seuils en-dessous desquels les documents ne sont plus considérés comme pertinents. D'autres approches consistent, par exemple, à supposer un espace d'indexation fixe, déterminé par une hiérarchie de concepts considérés comme universels.

3.2.1.2.PRISE, SMART et dérivés

SMART, le moteur de recherche et d'indexation originel de Salton constitue le système duquel découlent, entre autres, PRISE le moteur utilisé par le NIST pour les conférences TREC, ainsi que l'ensemble des systèmes commerciaux les plus répandus du marché. Lors des premières conférences TREC, les données de référence, constituées de corpus textuels variés (ex. : journaux, débats à la chambre des députés, dépêches journalistiques spécialisées), furent indexées grâce à PRISE en vue de simplifier la tâche des relecteurs (*assessors*) humains. Du côté des participants, la plupart des systèmes reposaient sur des variantes de SMART, adaptées en fonction de chaque tâche ou sous-tâche.

Filtrage et routage d'information étaient donc logiquement confondus jusqu'à TREC-4, la tâche de routage consistant en l'adaptation d'un moteur pensé pour le *pull* (recherche d'information dans une base stable de documents) à une tâche de *push* (recherche d'information dans une base non stabilisée de documents). Depuis TREC-4, filtrage et routage se distinguent par la nature de la décision de sélection : binaire pour le filtrage, continue pour le routage. Par ailleurs, la différence essentielle entre les deux tâches est que seuls les documents jugés pertinents sont présentés aux utilisateurs dans le cas du filtrage, alors que l'ensemble de la base, triée selon un score de pertinence, est présentée aux utilisateurs dans le cas du routage. Cependant, dans les faits, l'écrasante majorité des systèmes participant aux tâches de filtrage se contentent de fixer un seuil permettant d'émuler la décision de sélection binaire. Toutes les communications sur le sujet sont ainsi consacrées à la discussion des performances relatives des moteurs d'indexation et de recherche utilisés, d'une part, et des seuils fixés d'autre part.

On le voit, le flou terminologique des débuts de TREC correspond à un flou conceptuel et technique, induit par le recours massif à des moteurs d'indexation et de recherche tels que SMART. Bien qu'au niveau terminologique toute confusion soit désormais impossible entre routage et filtrage d'information, dans les faits la confusion reste réelle. Nous considérons cette confusion persistante comme la marque du peu de maturité du domaine du FI.

3.2.2. Filtrage d'information par reconnaissance de mots-clés

La reconnaissance exacte de mots-clés, ou de séquences de mots-clés, constitue un moyen simple, dont la mise en œuvre informatique est bien maîtrisée, de fournir une décision de sélection binaire pour un document traité par un système de FI. Toutefois, rares sont les systèmes de filtrage industriels basés sur cette technique ; à notre connaissance, aucun système de ce type n'a d'ailleurs participé aux conférences TREC. Ainsi que nous l'avons fait pour les systèmes de routage, nous nous pencherons sur quelques aspects techniques sous-jacents aux systèmes de FI par reconnaissance de mots-clés. Nous tenterons de souligner les limites d'une approche restreinte aux mots-clés, pour aborder la question des expressions typiques d'un domaine de spécialité.

3.2.2.1. Principe des expressions rationnelles

Les expressions rationnelles, qui forment la base des systèmes de FI par reconnaissance de mots-clés, constituent des règles explicites de reconnaissance de caractères ou séquences de caractères. Elles reposent sur la théorie des automates et transducteurs à états finis, leur mise en œuvre informatique est bien maîtrisée[12] et elles présentent des garanties, en termes de maîtrise des temps de traitement, qui en font un outil privilégié dans le cadre d'applications informatiques.

Les expressions rationnelles reposent sur un alphabet de symboles d'entrée et un alphabet de sortie (dans le cas des transducteurs). Cet alphabet comprend aussi bien des caractères atomiques que des opérateurs booléens (i.e. ET, OU, NON), ainsi que des caractères spéciaux. Ces derniers permettent de coder des répétitions (ex. : « * » représente 0 ou plusieurs répétitions d'une même séquence), de spécifier des ensemble (ex. : « . » représente l'ensemble de l'alphabet d'entrée) et des sous-ensembles de caractères à reconnaître (ex. : « [a-z] » représente l'ensemble des caractères alphabétiques en casse minuscule, de « a » jusqu'à « z »).

Les expressions rationnelles permettent ainsi de définir des patrons de recherche, qui peuvent soit servir à une recherche littérale, soit à une recherche étendue grâce aux opérateurs vus plus haut. Les expressions rationnelles sont largement utilisées en programmation, elles

[12] Des bibliothèques informatiques de gestion d'automates et de transducteurs sont disponibles à titre gratuit (ex. : la bibliothèqe *regex* de la GNU Foundation).

forment la base des compilateurs. Elles forment également la base des grammaires formelles et des analyseurs syntaxiques automatiques.

3.2.2.2.SIFT et Infoscope, deux systèmes fondateurs

Historiquement, les premiers systèmes de filtrage d'information par reconnaissance de mots-clés furent dédiés au courrier électronique. Ils ont, depuis, été adaptées à d'autres moyens de communication tels que les serveurs de news, les fils de dépêches et flux d'informations apparentés. L'un des premiers systèmes de ce type, SIFT (T.W. Yan & H. Garcia-Molina,1995), repose sur une définition et une mise à jour complètement manuelle des profils, en fait des listes de mots. SIFT est principalement destiné au filtrage d'information sur les serveurs de news, il fournit une liste ordonnée d'articles, triés selon un taux de pertinence par rapport aux listes servant de profils. La plupart des systèmes de FI à base de mots-clés reprennent les principes de base de SIFT, bien que celui-ci ait essentiellement servi de banc d'essai à son concepteur. Autrement dit, il n'existe pas, à notre connaissance, de version commerciale de SIFT. Cette remarque vaut pour la plupart des systèmes de FI que nous avons rencontré, à l'exception de fonctionnalités très restreintes incluses dans des logiciels grand public, telle que la fonctionnalité de filtrage offerte par Netscape Messenger™.

Infoscope (Stevens, 1992) est proche de SIFT dans le sens où il est également destiné au filtrage des serveurs de news. Cependant, ce système offre une fonctionnalité de paramétrage automatique des profils d'utilisateurs, reposant sur un algorithme d'apprentissage. Le principe de création de profils avec Infoscope est basé sur l'interaction entre le système, qui propose des solutions, et l'utilisateur qui valide, corrige ou refuse ces propositions. Infoscope induit ainsi des règles de sélection binaires à partir des réponses de l'utilisateur, et sur des paramètres simples tels que le temps dédié à la consultation d'un message donné. Infoscope fut conçu dans le but d'éviter à l'utilisateur d'expliciter son expertise et, de façon plus générale, afin de fournir un système convivial capable de s'adapter à chaque utilisateur. De plus, le système conçu par Stevens prenait en compte la structuration informative des documents[13], toujours dans une perspective de centration sur l'utilisateur : Infoscope était ainsi capable de reconstruire l'espace d'information représenté par les serveurs de news de manière à mieux faire ressortir les informations pertinentes, en s'adaptant aux habitudes de chaque utilisateur, ce dont SIFT était incapable. Autrement dit, Infoscope mettait

[13] Des champs réservés : auteur, date, sujet, ainsi que la segmentation en unités textuelles.

en œuvre des fonctionnalités de modélisation de l'utilisateur, qui font partie des techniques destinées à augmenter les performances des systèmes de RI automatique. Cette voie, ainsi que d'autres fonctionnalités telles que le filtrage collaboratif, n'ont que rarement été explorées. Les conférences TREC, de leur côté, s'en sont toujours tenues aux techniques de filtrage par le contenu, indépendamment de conditions d'utilisation réelles : notamment la diversité des besoins en information, les interactions entre utilisateurs et la prise en compte de l'évolution des centres d'intérêt.

On le voit, les systèmes de FI automatique n'ont été mis en place que tardivement et de façon lacunaire, alors même que les bases du domaine étaient posées dès les années 1960 et le besoin, autrement dit le marché, identifié dès l'étude de Housman. Ainsi, les systèmes présentés, SIFT et Infoscope, bien que précurseurs dans les fonctionnalités explorées, n'ont jamais fait l'objet, à notre connaissance, d'une diffusion à grande échelle. De façon générale, la plupart des systèmes de FI existants sont et restent, le plus souvent, des produits de laboratoire, à l'exception des systèmes de routage vus plus haut. Certains auteurs, tels que Oard et Marchionini, avancent que l'une des raisons de cette diffusion défaillante est liée au domaine de prédilection du filtrage d'information : le courrier électronique, les serveurs de news et flux d'information apparentés, pour lesquels l'accès à l'information par chaque utilisateur est intime, fortement subjectif, changeant, donc difficile à contrôler, en termes expérimentaux, ce qui s'accorde mal avec l'impulsion normalisatrice de TREC, par exemple. Nous ajoutons, pour notre part, que le petit volume des données concernées, quelques Mégaoctets, à comparer aux quelques Gigaoctet fournis par TREC à des fins d'évaluation, encourage une approche « artisanale » du filtrage d'information. Cette approche est à mettre en parallèle avec le recours massif du grand public à des moteurs de recherche et d'indexation n'offrant que des fonctionnalités limitées, qui sont cependant jugées suffisantes dans la plupart des cas.

Les systèmes de FI à base de mots-clés, reprenant dans l'ensemble les principes de SIFT, sont limités dans les fonctionnalités de filtrage qu'ils proposent. En effet, dans ce genre de systèmes, la définition des profils reste limitée à la constitution de liste de mots à reconnaître, sur lesquelles des opérations de logique booléenne sont effectuées. Autrement dit, ces systèmes restent dans l'optique « sac de mots » adoptée par les systèmes de routage. Nous posons qu'une approche prenant en compte la structuration du matériau porteur d'information, en l'occurrence le langage naturel, est possible, voire souhaitable.

3.2.3. Filtrage d'information par reconnaissance d'expressions typiques d'un domaine

Dans cette partie, nous proposons une alternative aux systèmes à base de mots-clés. L'approche décrite reste compatible avec la définition retenue du filtrage d'information, elle est basée sur une analyse syntaxique locale, visant à repérer les expressions typiques d'un domaine de spécialité et leurs variantes. Nous définissons en premier lieu la notion de signature thématique, puis nous détaillons la constitution d'un ensemble d'unités lexicales complexes utilisées comme descripteurs thématiques.

3.2.3.1. Notion de signature thématique

Le recours à des termes isolés comme descripteurs thématiques, c'est-à-dire des termes pouvant servir à l'indexation de documents, est limité : ces termes, hors contexte, présentent généralement une forte ambiguïté thématique. Il en va ainsi de « acheter », par exemple. Ce terme pourrait être utilisé comme descripteur de documents traitant de transactions financières, cependant il paraît évident qu'utilisé seul, « acheter » n'est pas un descripteur fiable du domaine, tant ce verbe courant peut apparaître dans nombre de contextes n'ayant rien à voir avec la finance.

Les approches vectorielles tentent de limiter l'ambiguïté thématique en accumulant les descripteurs pour chaque document, cette stratégie n'apporte, toutefois, aucune garantie sur la précision de l'indexation : les vecteurs obtenus restent dépendants des mots trouvés dans les documents. Les méthodes vectorielles se caractérisent par une absence de connaissances sur les objets indexés (ex. : les documents traitant d'un domaine de spécialité), c'est d'ailleurs ce qui fait leur attrait : elles sont indépendantes des documents traités, elles n'ont recours qu'à des propriétés intrinsèques aux objets indexés, en l'occurrence les différences de fréquence d'occurrence des termes. Nous posons que cette approche sans connaissances (*knowledge poor*) n'est pas optimale, notamment pour des applications visant les activités spécialisées. En effet, ces activités, ou domaines de spécialité, se caractérisent généralement par une phraséologie propre, des expressions typiques, ou façons de parler d'un thème donné, dont il est envisageable d'établir un recensement. Ce recensement, établi à partir de textes de spécialité, présente des lacunes, qu'il est possible de combler partiellement grâce à des connaissances générales (ex. : lemmatisation, transformations syntaxiques, termes sémantiquement proches).

Les travaux de Riloff sont une application pratique d'une approche visant à dépasser les limites des systèmes à base de mots-clés grâce à une mise en œuvre raisonnée de techniques issues du TALN, pour des tâches de classification automatique de textes. Riloff définit une notion de signature de pertinence (*relevancy signature)*, basée sur des suites de termes propres à un thème (ex. : les actions terroristes), décrits sous la forme d'une grammaire limitée aux seuls contextes pertinents au regard de la tâche. L'auteur a testé son approche sur les corpus de la campagne d'évaluation MUC-4, dédiée aux actions terroristes. La tâche dévolue aux systèmes d'extraction d'information participant à cette campagne était la mise à jour automatique d'une base de données relationnelles comportant des champs telles que « auteur de l'attentat », « lieu de l'attentat », ou encore « nombre de victimes » à partir de dépêches journalistiques. Dans le cadre de cette tâche d'extraction d'information, Riloff a cherché à montrer la pertinence d'une approche par analyse locale, centrée sur des patrons d'extraction, autrement dit des séquences à reconnaître, construits sur des schémas de sous-catégorisation simplifiés (un verbe et ses compléments typiques). Ainsi, plutôt que de considérer des termes ou listes de termes isolés, comme « bomb », le système Circus cherchait des documents contenant des passages reconnus par une grammaire locale construite autour de l'amorce *bomb* : **was bombed by <perpetrator>**, par exemple. Dans cette grammaire locale, **<perpetrator>** regroupe des syntagmes nominaux (attestés dans les documents ou généralisés) susceptibles de jouer le rôle d'agents.

Nous proposons de reprendre, pour le filtrage d'information, la notion de signature de pertinence introduite par Riloff en extraction. Cette approche, bien qu'adaptée à une tâche d'extraction, à classer dans les activités de *pull*, implique cependant la mise en œuvre d'algorithmes de décision de catégorisation des textes traités en fonction de contraintes multiples, que nous n'avons pas détaillées (ex. : taux de corrélation entre une signature et un thème). Dans le cas du filtrage, autrement dit une tâche de *push*, le temps de traitement doit être le plus limité possible. Cette contrainte implique que la phase de prise de décision soit abandonnée. Nous proposons, pour le FI, la notion de signature thématique, se distinguant de celle de signature de pertinence par une plus grande spécialisation.

3.2.3.2.Des unités lexicales complexes comme descripteurs

Les signatures thématiques sont à considérer comme des unités lexicales complexes pouvant servir de descripteurs thématiques dans le cadre du FI. Autrement dit, nous considérons des unités linguistiques fonctionnellement proches des unités de la terminologie, dans le sens où ces unités sont fortement associées à un domaine de spécialité (ex. : le

domaine financier). Les signatures thématiques sont des grammaires locales, décrivant un ensemble d'expressions associées à un thème. Ces grammaires locales sont décrites sous la forme de transducteurs à états finis, elles sont par ailleurs lacunaires : seuls les éléments pertinents, en termes de thème, sont décrits.

Ces signatures thématiques sont extraites de corpus spécialisés, elles sont centrées sur les cadre de sous-catégorisation des verbes trouvés dans les documents (ex. : <Nom de société> acheter <Nom de société>). Elles permettent une certaine forme de variation par le biais d'un étage transformationnel (ex. : transformation actif/passif). Par ailleurs, elles sont construites à partir de ressources linguistiques à large couverture : les dictionnaires électroniques du LADL et le Dictionnaire Intégral (DI) de la société Memodata, pour les signatures élaborées pour le français. Ces deux ressources sont complémentaires : les dictionnaires du LADL servent essentiellement à l'étiquetage robuste des corpus par le biais de la plate-forme Intex, le DI à la recherche de termes sémantiquement proches de ceux trouvés dans les corpus.

3.3. Problèmes d'évaluation des systèmes de filtrage d'information

Dans cette partie, nous nous penchons sur les problèmes liés à l'évaluation de systèmes automatiques de FI. Nous présentons tout d'abord quelques métriques utilisées en recherche d'information, puis nous nous concentrons sur les métriques développées dans le cadre de TREC pour la tâche de filtrage.

3.3.1. Quelques métriques de la recherche d'information

La majorité des métriques utilisées en RI, développées dans une perspective d'évaluation quantitative des systèmes, supposent la constitution d'un ensemble de réponses de référence, à comparer avec les réponses des systèmes évalués. Ce cadre d'évaluation s'inscrit dans une logique « behavioriste », où seules la correction des réponses fournies est mesurée, indépendamment de la façon dont ces réponses sont élaborées. Par ailleurs, ce type d'évaluation suppose qu'il n'y ait qu'une bonne (ou une mauvaise) réponse possible pour chaque question posée, ce qui n'est pas toujours compatible avec la réalité.

3.3.1.1.Précision et Rappel

La différence observée entre les réponses attendues et les réponses effectives fournit les indices essentiels de bruit et de silence, qui se définissent comme suit :

- le bruit est le nombre de réponses incorrectes fournies par le système évalué ;
- le silence est le nombre de réponses correctes absentes des réponses fournies.

Ces deux indices essentiels vont être utilisés par l'ensemble des métriques citées : précision/rappel, F-mesure et variantes.

La précision et le rappel se définissent comme suit :

- Précision = Réponses correctes / Réponses attendues
- Rappel = Réponses correctes / Réponses fournies.

Le rappel et la précision fournissent des indices relatifs, alors que bruit et rappel sont des indices absolus. Le taux de rappel mesure la capacité des systèmes évalués à couvrir le problème, alors que le taux de précision mesure la qualité des réponses fournies. Les deux indices sont nécessaires à l'évaluation, en effet, un système fournissant en réponse l'ensemble des documents traités aurait fatalement un taux de rappel maximal. Une évaluation ne prenant en compte que le taux de rappel serait incapable de se prononcer sur la proportion de réponses correctes parmi celles fournies.

3.3.1.2.F-mesure, P&R

À partir des taux de précision et de rappel, d'autres mesures ont été développées, qui visent généralement à affecter d'une pondération l'un ou l'autre des deux taux. Ces mesures correspondent à la nécessité de distinguer entre systèmes équilibrés, fournissant des taux de rappel et de précision proches, et systèmes privilégiant l'un ou l'autre de ces taux. En effet, les systèmes équilibrés sont recherchés pour certaines tâches, alors que d'autres tâches mettent l'accent soit sur la qualité (précision) soit sur la complétude (rappel) des réponses fournies. Une de ces mesures, développée dans (Van Rijsbergen, 1979), est la F-mesure, qui se définit comme suit :

- F-mesure = $(\alpha+1)$ * Précision * Rappel / (α*Précision) + Rappel.

Le coefficient α permet de pondérer soit le rappel, soit la précision. Une mesure dérivée de la F-mesure, nommée P&R, fixe α à 1. Il s'agit dans ce cas d'une métrique visant à privilégier les systèmes équilibrés. Ce type de métrique à pondération suppose qu'il soit

possible de déterminer de façon non artificielle un poids, auquel il soit aisé d'associer une interprétation.

3.3.2. Les métriques TREC pour le filtrage d'information

Les métriques développées par TREC s'inspirent de celles présentées plus haut. Comme nous l'avons montré plus haut, le domaine du filtrage d'information a souffert d'un flottement tant terminologique que conceptuel. Ce flou est visible jusque dans les métriques mises en place par les conférences TREC pour l'évaluation des systèmes participant aux tâches de filtrage. Nous passerons sur les trois premières éditions de TREC, où filtrage et routage d'information étaient confondus, pour nous intéresser, dans un premier temps, à la quatrième édition, qui introduisit pour la première fois la notion d'utilité. Nous étudierons, dans un deuxième temps, comment cette notion n'a cessé d'être remaniée à chaque édition de la campagne d'évaluation, pour aboutir à un ensemble de mesures complexes, à partir desquelles il est difficile de départager les systèmes évaluer.

3.3.2.1. Utilité

La notion d'utilité introduite au cours de TREC-4 marque les vrais débuts du filtrage d'information, en tant que tâche distincte du routage. À nouvelle tâche, nouvelle métrique, définie comme suit, pour toute expérience (*run*) R_i, revenant à évaluer la capacité des systèmes de filtrage à trier un ensemble de documents en deux catégories A et B :

- $U_i = u_{ai}A_i + u_{bi}B_i$.

A_i correspond au nombre de documents pertinents trouvés par le système pour l'expérience R_i, et B_i au nombre de documents non pertinents pour cette expérience. Les constantes u_{ai} et u_{bi} correspondent à la valeur d'utilité donnée par un utilisateur pour chaque cas : réception d'un document pertinent, ou non pertinent. Différentes valeurs pour ces constantes sont définies, qui correspondent à autant d'expériences. TREC-4 a fixé trois valeurs pour ces constantes, correspondant à trois scénarios différents : un scénario où on favorise la qualité des réponses[14], un autre où c'est la quantité de réponses qui est

[14] Un poids maximal sur la précision.

recherchée[15] et un dernier scénario dit équilibré, où les poids sur les documents de type A et B sont égaux. Ces trois cas de figure sont synthétisés ci-dessous.

Expérience	Valeur des paramètres	Mesure d'efficacité
R1	$u_{a1} = 1, u_{b1} = -3$	$u_1 = A_1 - 3B_1$
R2	$u_{a1} = 1, u_{b1} = 1$	$u_2 = A_2 - B_2$
R3	$u_{a1} = 3, u_{b1} = -1$	$u_3 = 3A_3 - B_3$

Figure 4 : mesures d'utilité pour trois scénarios d'évaluation

Une particularité de la mesure d'utilité de TREC-4 est qu'il s'agit d'une métrique absolue : on raisonne en nombre de documents pour chaque catégorie, et non pas en proportion de bonnes ou mauvaises réponses. De ce fait, la mesure d'utilité ainsi mise en œuvre n'est pas normalisée entre 0 et 1, comme c'est le cas pour les mesures telles que la précision et le rappel. La comparaison de l'efficacité d'un même système sur plusieurs profils (requêtes) différentes n'est pas aisée, seule la comparaison entre systèmes pour une même expérience est possible. Ainsi, la mesure d'utilité ne permet pas de déterminer de façon globale l'efficacité d'un système donné, mais bien seulement relativement aux autres systèmes sur une même requête. Ce choix est compatible avec la philosophie générale de TREC, qui consiste à départager entre eux des systèmes, sans préjuger de la meilleure façon de résoudre le problème posé (i.e. sélectionner les bons documents). Ce choix sous-entend que la meilleure approche n'est pas connue, mais également que la complexité du problème n'est pas mesurable *a priori*. Cependant, une telle mesure, par son caractère absolu, rend difficile la comparaison avec des systèmes utilisant des données différentes que celles de TREC.

La mesure d'utilité ainsi définie suppose qu'il soit possible de comparer les scores obtenus par les systèmes évalués avec l'utilité effective pour chaque document d'un ensemble de référence, issue d'une pratique réelle (ex. : FI sur un des thèmes de TREC). Or, les

[15] Un poids maximal sur le rappel.

conférences TREC se caractérisent par l'absence de telles données de référence. En effet, étant donnée la quantité de données textuelles fournies par TREC (plusieurs Gigaoctets), il est impossible d'envisager un tri manuel par des experts de chaque domaine couvert par les données. De ce fait, les évaluations TREC se caractérisent également par le recours massif à des techniques d'échantillonnage visant à réduire la masse de données textuelles communiquées à des relecteurs humains. Ces techniques d'échantillonnage sont également mises en œuvre pour la comparaison entre la valeur d'utilité des documents fournis par un système donnée pour une expérience et la valeur d'utilité réelle des documents de référence, impossible à obtenir pour les raisons évoquées. Ainsi, seuls des estimateurs d'utilité réelle sont utilisés pour l'évaluation en filtrage d'information[16].

3.3.2.2.TREC-5, une remise en cause du protocole d'évaluation

Nous l'avons vu, la mesure d'utilité introduite au cours de TREC-4 est loin d'être intuitive, notamment en raison de l'absence de données de référence issues d'une pratique réelle de FI, qui simplifieraient l'évaluation : les seuls points discutables resteraient les pondérations appliquées aux différentes catégories de documents (i.e. pour TREC-4, pertinents/non pertinents). Ainsi, en extraction d'information, les protocoles d'évaluation se basent sur des données triées à la main par des relecteurs humains, autrement dit une référence quasi-absolue (*gold standard*). Il est vrai que, par exemple, la reconnaissance d'entités nommées (ex. : des noms propres), une des sous-tâches de l'extraction d'information, nécessite une expertise moindre de la part des relecteurs que l'évaluation de systèmes de catégorisation de textes tels qu'évalués dans TREC. En effet, les thèmes abordés par TREC sont très variés : de la finance à l'écologie en passant par l'indépendance du Québec ou l'impact des pluies acides sur l'environnement.

Par ailleurs, les techniques d'échantillonnage employées, dans certaines conditions, sont susceptibles de produire des résultats inutilisables pour l'évaluation. Ces conditions sont celles rencontrées au cours de TREC-5 : des corpus très dispersés en termes de thèmes et des effectifs trop restreints pour chaque thème. Ainsi, sur les 49 thèmes évalués pour TREC-5, plus de 30 totalisent moins de 100 documents pertinents (entre 0 et 92 pertinents). Ces effectifs insuffisants ont une conséquence directe sur les indices statistiques employés, tels

[16] Pour un exposé plus complet des techniques d'échantillonnage employées et les mesures de pertinence statistiques des estimateurs d'utilité, voir (Lewis, 1996).

que l'intervalle de confiance calculé pour la mesure d'utilité associée à chaque thème. Dans le cas d'effectifs inférieurs à 100, cet intervalle est inutilisable pour l'évaluation[17]. Lewis, le concepteur du protocole d'évaluation en FI, va même jusqu'à affirmer que, pour la cinquième édition de TREC, étant donnée la dispersion des documents pertinents à travers le corpus, la meilleure stratégie était, dans certains cas, de ne fournir aucun document ; les systèmes adoptant cette stratégie auraient ainsi évité d'être trop pénalisés. Lewis propose d'ailleurs, pour les éditions ultérieures, d'ajuster les données et/ou les métriques servant à l'évaluation afin d'éviter les problèmes rencontrés pour TREC-5. L'auteur envisage même d'avoir recours à des métriques autres que l'utilité. Par ailleurs, les problèmes de représentativité des données fournies pour l'évaluation sont abordés, tant pour le filtrage que pour le routage d'information : la forte dilution des documents pertinents, propre aux corpus de TREC-5, n'a fait que souligner l'inadéquation d'une évaluation reposant sur des données à la fois trop simples et trop complexes. En fournissant des corpus d'apprentissage contenant une forte densité de documents pertinents, avec des effectifs dépassant largement ceux observés au cours d'une pratique réelle, le paramétrage des systèmes en compétition est artificiellement facilité. Par ailleurs, les données de test ne présentant que peu de ressemblance avec celles des corpus d'apprentissage, des performances décevantes sont enregistrées. Lewis mentionne par ailleurs les critiques adressées à l'encontre de la méthode de constitution des données de référence, la méthode dite de *pooling*, que nous aborderons plus bas.

On le voit, cette cinquième édition est l'occasion de difficultés importantes, qui remettent en cause l'ensemble des choix adoptés en matière de protocole d'évaluation. D'ailleurs, Lewis semble ne plus s'investir dans la définition de protocoles d'évaluation après TREC-5, Hull et Robertson prenant la responsabilité des évaluations.

3.3.2.3. Association de l'utilité et d'autres mesures

À la suite des difficultés rencontrées au cours de TREC-5, la notion de filtrage est amendée, afin d'inclure la dimension temporelle et l'interactivité[18] qui lui faisaient jusqu'alors défaut. Les données utilisées pour l'évaluation des systèmes participants sont, à

[17] Pour un exposé plus complet des problèmes rencontrés au cours de TREC-5, voir (Lewis, 1996).

[18] Dans la terminologie TREC, l'interactivité désigne la possibilité de consultation des résultats du filtrage « au fil de l'eau », autrement dit document par document, et non pas à l'issue du tri d'une base de documents en fonction d'un score de pertinence par rapport à une requête d'utilisateur, comme c'est le cas pour le routage.

partir de TREC-6, tirées des archives du FBIS, et non plus de l'ensemble des données servant aussi bien à l'évaluation de moteurs d'indexation et de recherche que de routage. La sixième édition de TREC est également l'occasion, sous l'impulsion de Hull, d'adopter des métriques complémentaires à celle d'utilité, en l'occurrence, précision d'ensemble moyenne (*Average Set Precision*, *ASP*). La notion d'utilité elle-même est redéfinie comme suit, sur la base des éditions précédentes.

	Pertinent	**Non Pertinent**
Document Sélectionné	R+ / A	N+ / B
Document Non Sélectionné	R- / C	N- / D
Utilité (linéaire)	(A*R+) + (B*N+) + (C*R-) + (D*N-)	

Figure 5 : décisions de sélection d'un système de filtrage d'information et mesures d'utilité correspondantes

La mesure d'utilité prend ainsi en compte deux paramètres : décision de sélection et pertinence, et affecte une pondération à chaque document en fonction de l'adéquation de la décision de sélection automatique opérée par chaque système. Les variables R+, R-, N+, N- renvoient au nombre de documents dans chaque catégorie, respectivement : documents sélectionnés/non sélectionnés, pertinents/non pertinents. Les paramètres d'utilité A, B, C, D, qui sont autant de coefficients de pondération, déterminent la valeur relative de chaque catégorie possible. Un paramètre d'utilité positif correspond au gain apporté par chaque document, un paramètre négatif représente le coût entraîné par l'attribution erronée d'un document à une catégorie. De ce fait, pour un profil considéré, plus le score d'utilité est élevé, meilleur est le système. Pour TREC-6, les paramètres suivants ont été testés :

- F1 = (3*R+) – (2*N+)
- F2 = (3*R+) – (N+) – (R-).

De son côté, l'ASP est définie comme suit :

- ASP = Précision * Rappel.

Les deux métriques, utilité et ASP, sont utilisées conjointement afin de fournir des indicateurs de performance pour chaque système. En ce qui concerne l'utilité, les trois scénarios initiaux ont été réduits à deux : le premier (F1) pénalise fortement le bruit dans les réponses fournies (2*N+), le deuxième (F2) pénalise également le silence (R-).

Hull souligne le fait que la mesure d'utilité est peu adaptée à l'évaluation d'un même système sur plusieurs thèmes, puisqu'il s'agit d'une mesure absolue, non normalisée. Par ailleurs, l'auteur fait remarquer que cette mesure ne prend pas en compte le nombre de documents déjà consultés, susceptible de faire décroître la pertinence de chaque nouveau document. Les défauts de l'ASP (et de la F-mesure, très proche) sont également mentionnés : cette mesure ne permet pas de distinguer entre les systèmes qui ne fournissent aucune bonne réponse alors qu'elles existent et ceux qui retournent un nombre quelconque de documents non pertinents. Autrement dit, si on se reporte au tableau ci-dessus, l'ASP ne permet de distinguer entre les cas R- et N+. Malgré les modifications apportées, l'ensemble du protocole d'évaluation repose sur des données qui ne sont pas issues d'une pratique effective de FI, pour les mêmes raisons qu'évoquées plus haut : une forte dispersion thématique, qui nécessiterait le recours d'une batterie d'experts pour chaque domaine abordé par les documents des corpus d'évaluation. Autrement dit, l'essentiel de l'évaluation se fait à partir d'échantillons tirés sur l'ensemble des corpus, pour lesquels des mesures d'utilité sont estimées et non pas des données de référence intégralement vérifiées par des experts du domaine.

Afin de départager les systèmes entre eux, autrement dit de fournir une liste ordonnée des systèmes selon un maximum de bonnes performances, estimées d'après les scores d'utilité obtenus sur l'ensemble des requêtes traitées, des algorithmes de tri, opérant en deux passes, ont été introduits. Pour chaque système, une première passe prend en compte les scores d'utilité F1, F2 … F*n*, obtenus pour une requête donnée. Dans une deuxième passe, une moyenne de ces scores est calculée sur l'ensemble des requêtes. Hull souligne les avantages et les inconvénients de cet algorithme, qui masque les différences entre systèmes en accordant la même importance à toutes les requêtes, indépendamment de leur score d'utilité maximal estimé.

Au cours de TREC-6, le paradigme d'évaluation est passé d'une métrique isolée, l'utilité, à des métriques associées. Cette voie sera poursuivie au cours des éditions ultérieures, avec l'introduction, notamment, de mesures d'utilité non linéaires.

3.3.2.4.Fonctions linéaires / non linéaires d'utilité et métriques associées

Les fonctions non linéaires d'utilité apparaissent au cours de TREC-8, elles sont employées en parallèle aux fonctions linéaires telles que définies plus haut. Les fonctions d'utilité non linéaires reprennent les catégories de documents définies plus haut : documents pertinents et non pertinents, respectivement R+ et N+. Celles testées au cours de TREC-8 sont les suivantes :

- $NF1 = 6*(R+)^{0.5} – N+$
- $NF2 = 6 * (R+)^{0.8} – N+$.

Le principe des fonctions ci-dessus est que l'utilité d'un document pertinent donné dépend de ceux déjà retrouvés par le système. Ainsi, plus un système retrouve de documents pertinents, moins la valeur additionnelle de nouveaux documents pertinents est élevée. Hull et Robertson, les concepteurs du protocole d'évaluation de TREC-8, espèrent que ces fonctions permettront de lisser les différences d'effectif de documents pertinents, donc les différences d'utilité des documents sélectionnés, pour chaque thème.

Par ailleurs, à partir de TREC-7, d'autres pondérations sont affectées aux différentes catégories de documents. Ainsi, pour TREC-7, la fonction d'utilité

$$F2 = (3*R+) – (N+) – (R-)$$

est remplacée par la fonction

$$F3 = (4*R+) – (N+)$$

au motif que le silence est difficile à évaluer pour certains thèmes.

Pour TREC-8, les fonctions linéaires testées sont les suivantes :

$$LF1 = (3*R+) – (2N+)$$

et

LF2 = (3*R+) – (N+).

Afin de faciliter les comparaisons entre systèmes, autrement dit leur classement à l'issue des différentes phases d'évaluation, une fonction de redimensionnement d'utilité (*utility scaling function*) est introduite dès TREC-7, elle précède le calcul de scores moyens d'utilité pour chaque système, sur l'ensemble des thèmes traités, qui fournit un classement global des systèmes. La fonction de redimensionnement vise donc à remplacer l'algorithme de tri en deux passes, expérimenté au cours de TREC-6, elle est définie comme suit :

$$u^*_s(S,T) = \max(u(S,T), U(s)) - U(s) / MaxU(T) - U(s)$$

où u(S,T) et u^*_s (S,T) sont respectivement la mesure d'utilité d'origine et la mesure redimensionnée (*scaled utility*) pour le système S et la requête ou thème (*topic*) T. U(s) est l'utilité associée à la sélection de *s* documents non pertinents et MaxU(T) est le score d'utilité maximal théorique pour le thème T. Le paramètre *s* définit une borne inférieure pour cette fonction de redimensionnement, de son côté MaxU(T) définit la borne supérieure d'utilité. La fonction d'utilité, dans son ensemble, se trouve ainsi bornée et normalisée entre 0 et 1, ce qui rend les comparaisons entre systèmes plus aisées qu'avec l'algorithme de tri vu plus haut. Étant donnée l'importance du paramètre *s*, qui fixe un seuil de performances minimal (i.e. qui permet de distinguer les systèmes les moins performants), plusieurs valeurs ont été mesurées pour TREC-7 et TREC-8, afin d'éviter de fixer ce seuil minimal de façon trop arbitraire. En effet, un seuil inférieur relativement bas permet de mieux séparer les systèmes qui enregistrent de bonnes performances sur des thèmes dont les effectifs de documents pertinents sont bas, ainsi que d'éviter de trop pénaliser les systèmes moins performants. Cette propriété de la fonction de redimensionnement permet d'éviter les écueils rencontrés au cours de TREC-5, où des thèmes généralement trop pauvres en documents pertinents avaient pénalisé l'ensemble des systèmes évalués.

3.3.2.5.Métriques orientées vers la précision

TREC-9 se distingue des précédentes éditions en faisant table rase des métriques non linéaires d'utilité, ainsi que de la méthode de classement des systèmes basée sur un

redimensionnement des scores d'utilité. Cette neuvième édition utilise de nouvelles métriques, dites « orientées vers la précision » (*precision oriented*), ainsi que des métriques adaptées à chaque sous-tâche du FI. L'introduction de ces nouvelles métriques est justifiée par ses auteurs de la façon suivante : en utilisant des métriques basées uniquement sur l'utilité, certains systèmes dont les taux de rappel et de précision sont plus élevés que d'autres systèmes, peuvent se voir moins bien classés que ces derniers. Les inégalités suivantes sont l'illustration de ce phénomène.

Soient deux systèmes de FI, X et Y. Pour ces deux systèmes, il est possible d'observer (U correspond au score d'utilité) :

Précision(X) > Précision(Y)

Rappel(X) > Rappel(Y)

mais

U(X) < U(Y).

Autrement dit, le score d'utilité va à l'encontre de l'intuition qui présuppose qu'un système X, dont les scores de précision et de rappel sont supérieurs à ceux d'un système Y, effectue une recherche d'information de meilleure qualité. Cette observation est valable tant pour les fonctions linéaires que non linéaires d'utilité, dans des conditions différentes : des scores d'utilité négatifs pour les premières, positifs pour les secondes[19].

La mesure principale employée au cours de la neuvième édition, en complément de nouvelles mesures d'utilité linéaire et d'autres mesures basées sur la précision, est la suivante.

- T9P = Nombre de documents pertinents sélectionnés / Max(Cible, Nombre de documents sélectionnés)

[19] Pour une discussion plus détaillée de ce point, voir (Hull & Robertson, 2000).

Avec une cible fixée à 50 pour TREC-9.

Le principe de cette mesure repose sur l'idée de cible, ou but à atteindre (i.e. un effectif de 50) pour chaque système en termes de nombre de documents pertinents, une pénalité est attribuée dans les cas où la cible n'est pas atteinte.

D'autre part, une seule fonction d'utilité linéaire est testée au cours de la neuvième édition de TREC :

- Utility = (2*R+) – N+.

Afin de fournir un intervalle de référence T9U pour les scores d'utilité de chaque système en vue de leur classement, les bornes suivantes sont fixées :

- T9U = (Utility, MinU)
- MinU = -100 pour les thèmes du corpus OHSU (voir annexe), -400 pour le corpus MeSH.

En complément de ces métriques, les mesures d'efficacité suivantes sont utilisées.

- MnT9P, valeur moyenne de T9P sur l'ensemble des thèmes ;
- MacP, moyenne de la précision d'ensemble (mean set precision) ;
- MacR, moyenne du rappel d'ensemble (mean set recall) ;
- MnT9U, valeur moyenne de T9U ;
- MnSU, moyenne normalisée de T9U sur le maximum possible pour chaque thème (i.e. 2*total des documents pertinents) ;
- Zeros, nombre de thèmes pour lesquels aucun document n'est retourné ;
- AveP, précision moyenne non interpolée ;
- P@50, précision à 50 documents.

Les résultats des évaluations basées sur les mesures énumérées plus haut sont consignés dans les actes de TREC-9 ; on trouvera également une discussion de ces résultats dans l'annexe consacrée aux campagnes TREC.

3.4. Problèmes de modélisation d'une tâche complexe : le filtrage d'information

Ainsi que nous l'avons vu précédemment, les différentes éditions de TREC, de la quatrième à la neuvième[20], l'évaluation des systèmes participant aux tâches de filtrage semble avoir posé un problème conceptuel aux responsables successifs : Lewis, puis Hull et enfin Hull et Robertson. En effet, tant les métriques utilisées que les corpus de référence, ou encore les méthodes de constitution de corpus de test statistiquement équilibrés n'ont cessé d'être modifiées. On est ainsi passé, pour l'évaluation de la performance des systèmes, d'une métrique absolue, l'utilité linéaire, pour laquelle plusieurs paramétrages ont été testés, à une métrique relative reprenant les principes des métriques standards que sont les taux de rappel et de précision. En ce qui concerne les corpus utilisés, l'inconstance est là aussi de mise : aucune des éditions de TREC n'a utilisé les mêmes corpus de test afin de ne pas biaiser les résultats[21], ce qui interdit toute étude longitudinale. Le constat qui s'impose, à l'heure où les actes d'une dixième édition de TREC devraient paraître, est l'impossibilité, tant pour le décideur que pour le chercheur s'intéressant au domaine, de choisir *une* approche pour le FI automatique.

D'autre part, aucun système mettant en œuvre une analyse linguistique des données textuelles, même locale, n'a été évalué au cours des cinq éditions de TREC dont nous avons tenté de faire une synthèse, lacune que Lewis souhaitait voir comblée[22]. De même, seuls les techniques de filtrage dites « par le contenu » ont été évaluées au cours de TREC. Après plus de cinq campagnes TREC, le domaine du filtrage d'information, loin de voir ses contours mieux dessinés, semble tout aussi flou qu'au départ.

Dans la suite de notre exposé, nous tenterons de comprendre les raisons de ce que nous percevons comme l'échec des campagnes d'évaluation TREC. Nous insisterons tout d'abord sur la difficulté de constituer une référence indiscutable pour une activité qui revient à

[20] Les actes de la dixième édition ne sont pas encore disponibles.

[21] La méthode de « pooling », utilisée pour créer des données d'apprentissage pour une édition donnée, réutilise une partie des résultats des éditions précédentes.

[22] La thèse de Lewis, soutenue en 1992, porte sur des techniques améliorant l'indexation des documents par la prise en compte de la dimension linguistique, syntaxique notamment.

attribuer de façon automatique une catégorie thématique (ex. : finance, actes de terrorisme) à des objets cognitifs complexes, des textes en langue naturelle. Nous tenterons, ensuite, de mettre en avant la complexité liée au processus même de filtrage, qui consiste à sélectionner des documents en fonction d'un besoin en information. Enfin, nous essaierons de montrer à quel point les campagnes d'évaluation TREC ont une vision simpliste du problème qui nous occupe. À l'heure où se diffusent des initiatives comparables au niveau européen[23], nous jugeons indispensable de faire le point sur ce que nous considérons comme des erreurs tant dans la définition de la tâche que dans le processus d'évaluation lui-même.

3.4.1. Problèmes de constitution d'une référence

La constitution d'une référence, si possible indiscutable, est la première étape logique d'une campagne d'évaluation de systèmes automatiques. Nous aborderons ainsi, dans cette partie, les notions de représentativité qualitative et quantitative, ainsi que l'effort d'explicitation d'une compétence (i.e. filtrer de l'information, c'est à dire décider de la pertinence d'un document) que demande la constitution d'un ensemble de données de référence. Nous serons amené, par ce biais, à déterminer quelles parties de la compétence humaine sont susceptibles de figurer ou pas dans l'ensemble de référence.

3.4.1.1.Représentativité quantitative/qualitative des corpus

Les campagnes TREC mettent l'accent sur les aspects quantitatifs des systèmes évalués. Dans cette logique scientifique visant la reproductibilité des résultats, les organisateurs passent outre les aspects qualitatifs liés au domaine du FI. Nous posons qu'une première cause de l'échec de TREC pour ce domaine, vient justement de cette obsession quantitative.

Historiquement, les campagnes TREC furent principalement mises en place pour évaluer les systèmes d'indexation et de recherche d'information sur des bases documentaires importantes. Plusieurs Gigaoctets de données textuelles constituent ainsi les corpus d'apprentissage et de test fournis aux participants, quelque soit la tâche. Ainsi, les participants à la tâche de filtrage, pour laquelle nous avons vu que la distinction avec le routage n'est que

[23] Les campagnes d'évaluation CLEF, proches de TREC, ou encore les campagnes plus centrées sur la qualité, telles que celles menées dans le cadre du projet Technolangues du Ministère de la Recherche et de la Technologie.

tardive, reçoivent les mêmes données que les participants à d'autres tâches : plusieurs Gigaoctets de textes, regroupant des articles de journaux, spécialisés ou non, des transcriptions de débats politiques, ou encore des dépêches journalistiques, couvrant des domaines aussi divers que la législation nord-américaine, l'impact environnemental des pluies acides, ou encore la baisse des stocks de poisson à la disposition des poissonneries commerciales de la Communauté Européenne.

Des corpus d'une telle ampleur, couvrant des domaines aussi diversifiés, sont bien adaptés à l'évaluation de moteurs d'indexation et de recherche, autrement dit des activités de *pull*, mais pas à celle de systèmes de FI, ou activités de *push*. En effet, le filtrage est avant tout une activité d'experts d'un ou plusieurs domaines, présentant des besoins en information stables, travaillant sur des « petits » volumes de données[24] (quelques Kilooctets par jour), comparés aux Gigaoctets fournis par TREC. Il n'est, de toute évidence, pas possible, ni faisable, ni à notre avis souhaitable de mobiliser l'expertise d'opérateurs humains sur de tels volumes de données. Le remède qui s'impose naturellement est le recours à des techniques d'échantillonnage statistiques, visant à dégrossir le travail de validation humaine des corpus de référence. Autrement dit, TREC vise essentiellement à produire des données de référence quantitativement pertinentes, statistiquement équilibrées afin de ne favoriser aucun système *a priori*. (Lewis, 1996) est d'ailleurs le lieu d'un exposé de haut niveau sur les techniques d'échantillonnage mises en œuvre pour la constitution de corpus de référence pour la tâche de filtrage, dont nous avons vu qu'elles ont été abandonnées dès l'édition suivante.

Ce qui semble faire défaut aux campagnes successives de TREC, ce sont des corpus de référence, issus d'une pratique effective de filtrage d'information par des opérateurs humains. En effet, on ne peut comprendre le recours à une métrique absolue, l'utilité, normalisée et bornée (entre 0 et 1) deux ans seulement après leur introduction, que par l'absence d'un ensemble borné de documents, parmi lesquels un sous-ensemble connu seulement est pertinent. De même, le recours à des estimateurs d'utilité[25] plutôt qu'à des scores d'utilité effectifs mesurés sur le sous-ensemble de documents pertinents, ne peut se comprendre que par cette absence.

---

[24] Nous ne faisons ici que reprendre la définition de la tâche de filtrage telle que définie dans TREC, que nous considérons valide en ce qui concerne le filtrage par le contenu.

[25] Introduits dès (Lewis, 1995).

3.4.1.2.Des données observables : le vocabulaire spécialisé

Une fois soulignée la nécessité de disposer de données de référence indiscutables, se pose la question du contenu de ces données, de leur utilité pour une entreprise normalisatrice telle que TREC, visant à isoler et à contrôler les variables dépendantes dans un cadre expérimental bien défini. TREC, dans cette optique de contrôle de variables, vise logiquement à rendre explicites des compétences humaines, en vue de les formaliser et de les traduire dans un format interprétable par une machine. Cependant, ainsi que l'échec des systèmes-experts en Intelligence Artificielle l'a montré, il semble évident que seule une partie du savoir-faire humain est susceptible d'être ainsi explicité. Les raisons sont essentiellement que les opérateurs humains, lorsqu'ils ont à décider si un document parle d'un thème donné, prennent cette décision en se servant aussi bien de critères objectifs que subjectifs.

Les critères objectifs utilisés en FI sont les données observables dans les corpus, en l'occurrence un ensemble d'expressions typiques pour chaque domaine, ou phraséologie spécialisée. Les approches évaluées dans TREC s'appuient d'ailleurs implicitement sur l'hypothèse que chaque thème peut être associé de façon plus ou moins certaine à un ensemble d'indices linguistiques, en l'occurrence des mots simples dans la plupart des cas, en raison de l'approche « sac de mots » de ces systèmes. Il paraît, en effet, raisonnable de penser qu'on ne parle pas de la même façon selon qu'on décrit l'impact des pluies acides sur l'environnement, ou des opérations boursières, par exemple. Cette hypothèse, qui reste implicite pour la plupart des systèmes basés sur une logique d'indexation, est celle qui guide explicitement les études sur corpus, dont les travaux de Harris constituent un parangon.

En d'autres termes, nous posons que la seule compétence explicitable pour des systèmes automatiques de FI est la décision de sélection d'un document donné à partir d'un ensemble d'indices linguistiques : des mots simples ou composés, des expressions typiques relativement idiomatiques, voire des phrases complètes ou suites de phrases. Par conséquent, un corpus appelé à devenir une référence doit contenir une proportion exploitable d'éléments linguistiques spécialisés, condition que des corpus généralistes sont, à notre avis, peu susceptibles de satisfaire. D'autre part, nous pensons avoir montré la nécessité d'analyser la valeur linguistique des corpus d'évaluation, d'autant plus importante que les approches basées sur une logique d'indexation (vectorielle ou autre), en restant au niveau du mot typographique, ne peuvent avoir accès qu'à une infime partie des compétences explicitables en matière de FI.

3.4.2. Le filtrage d'information, une tâche complexe

3.4.2.1.Subjectivité ou expérience ?

En sus de compétences qu'il est possible de rendre explicites, le FI, ainsi que tout processus de catégorisation et de prise de décision, repose sur un ensemble de compétences que nous nommons implicites, en raison de la difficulté, voire de l'impossibilité de les expliciter. Ces compétences implicites peuvent être vues comme des manifestations d'une certaine subjectivité, voire d'une inconstance de la part des opérateurs humains en FI[26]. Cependant, ces compétences implicites peuvent aussi être vues comme ce qui fait la valeur ajoutée d'un opérateur par rapport à un autre, ce qui lui permet de prendre les bonnes décisions de sélection en ne se basant pas uniquement sur les indices linguistiques objectifs mentionnés plus haut, en d'autres termes : son expérience du domaine.

Nous donnons ici un exemple tiré d'un corpus issu d'une pratique effective du FI, destiné à illustrer notre propos. Le corpus en question nous a été communiqué par la société Firstinvest, propriétaire d'un site Internet offrant des services de diffusion ciblée d'informations financières, sur le modèle de la SDI décrit plus haut. Ce corpus représente environ deux mois d'activité, il traite une vingtaine de thèmes différents. Chaque thème peut être associé à une phraséologie, que nous détaillons dans le chapitre consacré au système CORAIL. Il en va ainsi du thème 19[27] (cessions/acquisitions de société), un thème classique en veille économique. Cependant, cette phraséologie est parfois également partagée avec des documents classés par les opérateurs humains dans d'autres catégories que le thème 19. La dépêche ci-dessous, dans laquelle nous soulignons la phraséologie typique du thème 19, est classée par les experts de Firstinvest dans le thème 18 (accords/partenariats/contrats).

13420. Satellites : l'américain Loral veut se séparer d'Alcatel. Alcatel refuse ce divorce et porte plainte pour violation d'accords. NEW¤2001-04-11 12:05:00.000. L'américain Loral a décidé de mettre fin à sa coopération de dix ans avec Alcatel dans les satellites, rapportent ce matin <i>Les Echos</i>.
L'américain a demandé le 22 février au Français une séparation en bonne et due forme : celle-ci devrait être opérationnelle en février 2002 compte tenu du préavis d'un an prévu dans les accords entre les deux groupes.
Mais Alcatel

[26] Position adoptée par (Coyaud, 1972), entre autres.

[27] Voir la liste des thèmes dans le chapitre IV.

ne l'entend pas de cette oreille : le groupe dirigé par serge Tchuruk affirme vouloir défendre ses intérêts et a attaqué Loral en justice devant le tribunal du district du sud de New York.
La plainte porte sur Loral et sa filiale de construction de satellites Space Systems/Loral (SS/L). Alcatel reproche à ses partenaires d'avoir violé leurs accords et conteste la demande même de divorce.
L'alliance avait été élaborée en 1991 : Alcatel, Aerospatiale et Finmeccanica avaient alors pris 49 % de SS/L et l'année suivante, DASA les avait rejoint. En 1996 et 1997, Loral avait racheté leurs parts, remontant à 100 % du capital de SS/L contre des actions à émettre (Alcatel détient ainsi toujours 3,4 % de Loral).
En dix ans, l'alliance a produit une dizaine de contrats, dont Intelsat7, Intelsat9 et GlobalStar. Elle a aussi permis aux Européens de pénétrer le marché américain et réciproquement.
Reste que la rupture de cette alliance ne remet pas en cause celle dans les services satellites, notamment dans le multimedia où Alcatel a investi 30 millions de dollars dans Cyberstar et Loral 46 millions pour 14 % de SkyBridge.¤4. US

Dans la première partie du titre, *Satellites : l'américain Loral veut se séparer d'Alcatel*, l'expression « se séparer de », prenant comme sujet grammatical (N0) un groupe nominal construit autour d'un nom de société, et comme premier complément[28] (N1) un groupe nominal de même nature, est typique d'une opération de cession de société. Sans contexte et sans connaissances du monde concernant les deux sociétés mentionnées, on peut interpréter cette phrase comme une intention, de la part de *Loral*, de vendre *Alcatel*, qui serait ainsi une filiale, ou une société détenue par *Loral*. Dans les faits, il s'agit bien d'une rupture d'alliance entre *Loral* et *Alcatel*, ainsi que le montre le reste du document, qui sont deux sociétés distinctes. Cette première phrase ne peut donc être comprise avec certitude comme traitant du thème 18 que grâce à des connaissances qui ne figurent pas explicitement dans le document, autrement dit des connaissances sur le monde de la finance.

Dans la dernière phrase, en revanche, le passage souligné correspond bien à une référence au thème 19 : le fait qu'*Alcatel* et *Loral* investissent respectivement dans *Cyberstar* et *SkyBridge* correspond à une opération d'acquisition partielle de société.

Ce document traite donc de plusieurs thèmes, ce qui est courant malgré le soin apporté à leur rédaction par des professionnels. Cette dispersion thématique, qu'on peut également envisager sous l'angle d'une collision de points de vue, s'observe d'ailleurs pour d'autres corpus étudiés, tels que les articles du journal *Le Monde*, ou encore les dépêches de l'AFP. La

[28] Nous adoptons ici une typologie neutre : d'un point de vue distributionnel, *se séparer de* commute avec des verbes à construction transitive directe tels que *vendre*, *acheter* etc...

décision de sélection réalisée par l'opérateur humain doit donc prendre en compte les différents thèmes abordés, réaliser une sorte de pondération de chacun d'eux et aboutir à une prise de décision, autrement dit une prise de risque : classer l'ensemble du document comme relevant du thème 18 plutôt que 19[29].

Cet exemple nous permet d'illustrer l'idée que nous développons en détail plus bas : en situation réelle, le filtrage d'information fait appel, en plus de compétences explicitables, à des connaissances sur le monde, acquises au cours d'une pratique effective, ainsi qu'à un processus de décision capable de faire interagir plusieurs contraintes éventuellement antagonistes.

3.4.2.2.Filtrage d'information et catégorisation

Le FI, autrement dit l'activité consistant à décider, pour un document donné, qu'il traite d'un thème donné, doit être perçu essentiellement comme un problème de catégorisation. Poser le problème en termes de catégorisation nous paraît permettre de mieux saisir la nature des problèmes inhérents à la formalisation de cette tâche pour des systèmes automatiques.

En effet, les tâches de catégorisation, en d'autres termes la reconnaissance de formes (ex. : phonèmes, graphèmes, visage), se caractérisent tout d'abord par une variabilité tant interindividuelle (deux sujets ne voient pas les mêmes formes dans un même signal) qu'intraindividuelle (un même sujet verra plusieurs formes différentes dans un même signal, à des intervalles de temps distincts). Ce phénomène est bien connu dans le domaine de la documentation[30], il a donné lieu à plusieurs stratégies pour l'indexation de documents traditionnelle, visant à cadrer l'espace de catégorisation (ex. : indexation contrôlée).

Cette double variabilité nous paraît fondamentale pour le problème qui nous occupe, en ce qu'il rapproche d'autant le domaine de la linguistique sur corpus et de la rechercher d'information.

---

[29] On peut objecter à cette hypothèse que les experts sont susceptibles, tout simplement, de commettre des erreurs. Nous répondons à cette objection en soulignant le fait que, lorsqu'un document aborde plusieurs thèmes, il n'existe pas de bonne ou mauvaise décision de catégorisation, il n'existe que des réponses violant plus ou moins un ensemble de contraintes antagonistes.

[30] Voir, à ce sujet, (Coyaud, 1972).

3.4.2.3.Décision de sélection binaire et satisfaction de contraintes

Le filtrage d'information est défini comme une tâche où un système (opérateurs humains, logiciel) prend une décision de sélection binaire (oui/non) sur un document, tiré d'un flux dynamique, en comparant le profil informatif de ce document avec les besoins en informations exprimés par une communauté d'utilisateurs. Autrement dit, on attend d'un tel système une réponse définitive, reproductible, instaurant une rupture de continuité dans un processus qui, s'il est pensé en termes de tâche de catégorisation, ne peut satisfaire à ces attentes. Plutôt que de penser le FI comme un processus figé, nous estimons utile d'envisager un fonctionnement dynamique, proche des systèmes à satisfaction de contraintes, dont nous donnons ici une esquisse.

Dans cette vision dynamique, plusieurs objets conceptuels sont requis :

- un ensemble de contraintes ;
- une hiérarchie, ordonnant les contraintes en fonction de leur caractère plus ou moins violable ;
- un processus de satisfaction de contraintes.

Il est possible de reprendre les principes la théorie de l'Optimalité, introduite en linguistique par (Prince & Smolensky, 1993), comme cadre à un tel système à base de contraintes.

Les contraintes d'un système de filtrage dynamique peuvent être distinguées entre :

- contraintes portant sur les observables des documents : essentiellement, les expressions typiques d'un domaine de spécialité, ou signatures thématiques[31] (contrainte ST), ainsi que des principes métaphoriques relativement figés[32] (EM), analysables par les techniques de linguistique de corpus vues plus haut ;

[31] Voir l'annexe II pour une présentation des signatures thématiques du domaine financier, extraites grâce à des procédures distributionnalistes.

[32] Par exemple, celles relevant du domaine notionnel de l'attaque et de la défense, très productif dans les corpus financier : *préparer une offensive contre*, *s'allier à*.

- contraintes portant sur les connaissances du monde, de type encyclopédique (CE)[33].

Une telle hiérarchie de contraintes, mise en œuvre dans le cadre d'un système dynamique de FI, viserait à rendre compte du continuum de certitude chez les opérateurs humains, ainsi que de la variation et la collision de points de vue[34]. On peut faire l'hypothèse que les documents les plus explicites sont ceux pour lesquels les jugements d'appartenance thématique seraient les plus assurés et les mieux partagés par une communauté d'indexeurs. Schématiquement, en accordant aux contraintes liées aux observables : ST et EM, un poids fort, par rapport à celles liées aux connaissances du domaine (CE), il serait possible de prédire une cohérence maximale dans les décisions de sélection relevées chez plusieurs indexeurs pour les documents les plus explicites. À l'inverse, si seules des sociétés peu connues sont mentionnées, et si seules des métaphores figées peu explicites sont employées, on peut s'attendre à ce que la décision de sélection pour un thème donné soit plus difficile.

Par ailleurs, d'autres contraintes peuvent être envisagées : la première phrase d'une dépêche de type journalistique vise généralement à fournir un condensé thématique du document. En d'autres termes, le fait de trouver une signature thématique en première ou en dernière phrase peut être pertinent. On peut traduire cette différence de statut par des contraintes de textualité : titre, développement, conclusion, par exemple.

On pourrait ainsi envisager un processus de catégorisation thématique des documents, ou filtrage d'information, reposant sur un principe d'optimisation de contraintes. Signalons, toutefois, que la détermination d'une telle hiérarchie de contraintes ne peut se baser que sur des situations de filtrage d'information contrôlées, ce qui pose le problème de l'accès à une expertise dans un domaine où la compétition entre experts rend difficile la divulgation de ce type d'information.

---

[33] Par exemple, les liens entre les sociétés-mères et leurs filiales, ou les sociétés dans lesquelles elles ont des participations.

[34] Attribution d'un document à plusieurs thèmes.

3.5. Conclusion

Nous avons présenté une partie de l'activité de filtrage d'information, en nous fondant sur les conférences TREC, visant à structurer l'ensemble du domaine de la recherche d'information. Cette normalisation est principalement effectuée par la comparaison des performances quantitatives de systèmes adoptant des approches différentes pour un ensemble de problèmes, dans un cadre quasi-expérimental. En effet, tant les tâches, que les données et les métriques utilisées dans les évaluations font l'objet d'une standardisation. Ainsi, les conférences TREC définissent le filtrage d'information comme la décision de sélection d'un document pris parmi un flux d'information. Cette décision de sélection est binaire dans le cas du filtrage par lots, qui constitue le cas dans lequel nous nous situons.

Les conférences TREC constituent, par l'ampleur des évaluations menées et la diversité des systèmes testés, un recueil d'expériences capital pour le domaine du FI, notamment dans l'optique d'une adaptation de ce type de campagne d'évaluation à une conception européenne des problèmes de RI. En effet, ainsi que le montrent les publications consacrées aux initiatives comparables tant au plan national qu'européen[35], la conception américaine de l'évaluation montre une préférence envers les évaluations quantitatives de type « boîte noire », alors que la conception européenne, et plus encore française, accorde une préférence aux évaluations dites qualitatives, où la compréhension fine des performances des systèmes évalués est primordiale. Ces deux conceptions se traduisent par une propension à avoir recours à de grands volumes de données hétérogènes, du côté des initiatives américaines. Du côté des évaluations françaises, on observe une tendance marquée vers le recours à des données en quantité plus maîtrisables, issues de pratiques effectives, évaluées par des relecteurs humains. Par ailleurs, ces évaluations ont donné lieu à des réflexions dépassant le cadre de l'évaluation, sur la nature, l'utilité et la représentativité des corpus[36].

L'un des enseignements fondamentaux que nous tirons des évaluations TREC est la nécessité de recourir à des données issues d'une pratique effective. Nous nous plaçons donc dans la continuité de la conception française des évaluations en RI. En effet, nous avons tenté de montrer à quel point les difficultés rencontrées, au cours des éditions successives de TREC

[35] Voir notamment (Landi *et al.*, 1998), (Lespinasse *et al.*, 1999), et (Mariani, 1999).

[36] Voir, par exemple (Habert, 2001).

dans le domaine du FI, à une représentation inadaptée d'un protocole d'évaluation reposant sur des données hétérogènes, non maîtrisables. Ainsi, la succession de métriques, jugées inadaptées quasiment à chaque édition, ainsi que celle des techniques d'échantillonnage tant des corpus de paramétrage que des corpus de test, nous semble principalement due au manque de représentativité des données censées fournir une référence pour l'évaluation. En effet, nous considérons que face à des volumes de plusieurs Gigaoctets de textes hétérogènes, couvrant des thèmes différents à chaque édition, aucune relecture humaine n'est possible. Cette impossibilité d'un contrôle par des experts du domaine nous paraît être la cause principale de l'inconstance constatée dans les protocoles d'évaluation de TREC pour le filtrage d'information.

Signalons que, en raison de cette inconstance, aucune étude longitudinale n'est possible pour les systèmes ayant participé à TREC. En effet, les données de référence et les métriques d'évaluation changeant à chaque édition, il est impossible d'évaluer l'évolution d'un même système au cours du temps. Autrement dit, les campagnes d'évaluation TREC ne semblent pas vouées à s'inscrire dans une durée, tout du moins dans le domaine du FI, ce qui, au regard de l'ampleur des investissements nécessaires, peut paraître surprenant.

Par ailleurs, un des effets de ce type d'évaluation, centré sur les performances chiffrées, est un effet de convergence. Cet effet est visible aussi bien dans le cadre de l'extraction d'information (MUC), que dans celui du FI : la technique la plus efficace, en termes de performances, se répand dans l'ensemble des équipes participantes. Ceci aboutit, au bout de plusieurs éditions, à une uniformité des approches[37] évaluées. Bien que cette uniformité puisse être vue comme l'un des objectifs de ce type de campagnes, visant la diffusion dans le domaine industriel des techniques les plus efficaces en recherche appliquée,

[37] Dans le domaine de l'extraction d'information, les analyses locales et les techniques d'analyse à base de cascades de transducteurs à états finis constituent l'approche dominante aujourd'hui. Dans le domaine du FI, la plupart des systèmes évalués utilisent des moteurs d'indexation et de recherche dérivés du système SMART (Salton, 1971) comme moteurs de filtrage.

il est peu probable qu'une telle uniformité soit souhaitable dans le domaine de la recherche conventionnelle.

CHAPITRE 4

Filtrage d'information par signatures thématiques,

mise en œuvre en milieu industriel

Cette partie est consacrée à la description de CORAIL (Composition de Requêtes assistée par Agents Intelligents Linguistiques), un système de filtrage d'information mis en œuvre dans le cadre du laboratoire de recherche du groupe Thales[1]. En effet, cette plate forme constitue une implantation, dans un contexte industriel, d'une approche linguistique du filtrage d'information. Elle repose sur le principe d'une analyse partielle par cascades de transducteurs à états finis, où le repérage d'expressions typiques d'un domaine permet de sélectionner des documents pertinents parmi un flux d'information dynamique.

Nous insistons sur les aspects techniques du système CORAIL, ainsi que de LIZARD, un assistant linguistique pour l'élaboration de grammaires locales destinées à la Recherche d'Information. Nous détaillerons la chaîne de traitement, de l'acquisition d'un nouveau document à la présentation des filtrats, en passant par le filtrage par reconnaissance de signatures thématiques. Enfin, nous donnerons quelques mesures chiffrées de performance pour le système CORAIL, sur un corpus professionnel.

4.1. Le système CORAIL

Nous donnons ici une présentation du projet CORAIL (Composition de Requêtes par Agents Intelligents Linguistiques), partiellement financé par le secrétariat d'État à l'Industrie suite à l'appel de 1997, « filtrage d'information », lancé conjointement par le ministère de la

[1] Thales Research & Technologies, ex Thomson-CSF.

Recherche et le ministère de l'Industrie, et, mené par Thomson CSF/LCR[2], Informatique CDC/DTA[3] et l'université Paris X/CRIS[4]. Ce projet, d'une durée de deux ans, s'est achevé en Septembre 2000.

4.1.1. Une plate forme industrielle de gestion des documents électroniques : PRIAM

CORAIL s'intègre au sein d'une architecture industrielle de gestion des documents électroniques, PRIAM[5].

4.1.1.1. Architecture fonctionnelle

La figure ci-dessous donne un aperçu de l'architecture de la plate forme PRIAM.

[2] Le Laboratoire Central de Recherches du groupe Thomson-CSF (Thales).

[3] Le département Informatique de la Caisse des Dépôts et Consignations, Direction des Travaux Avancés.

[4] Centre de Recherche en Informatique Spécialisée.

[5] PRIAM, Programme de Recherche en Indexation Automatique, projet interne Thales 1999-2000.

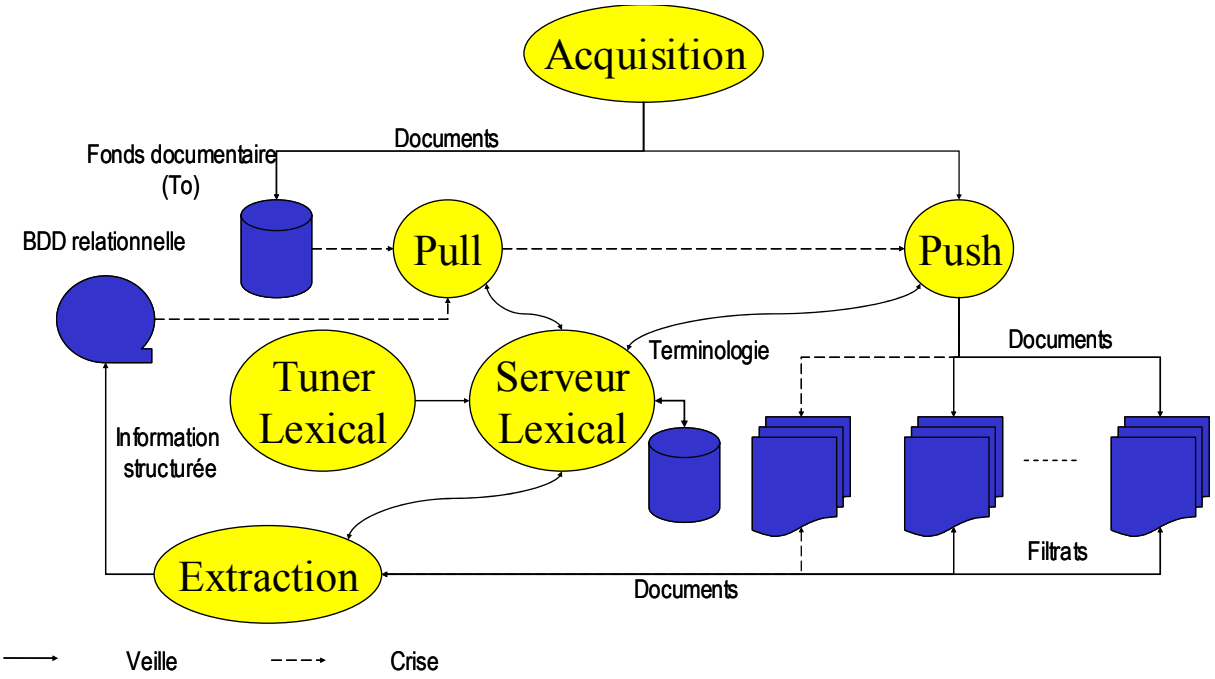

Figure 6 : architecture fonctionnelle de la plate forme PRIAM

PRIAM repose sur une conception distribuée : des agents logiciels prennent en charge chacune des tâches (figurées en jaune), selon une conception centralisée de type tableau-noir. Les agents logiciels sont écrits en Java, le fonctionnement multi-agent est pris en charge par RMI (Remote Method Invocation), passant par un agent central : le superviseur. PRIAM est également distribuée : chaque agent peut s'exécuter sur un hôte différent, en fonction des besoins en ressources. Par ailleurs, les agents sont accessibles via un réseau de type Intranet, sous forme d'applets Java.

Cette plate forme a été conçue de façon à modulariser chaque tâche. Cette modularisation permet de tester des composants logiciels différents pour chaque tâche, et de mesurer leur impact sur les performances globales du système.

Les cinq modules principaux de PRIAM sont :

- le module d'acquisition, c'est-à-dire la collecte de documents à traiter ;
- le module de *push*, qui réalise essentiellement les tâches de filtrage et de routage d'information ;
- le module de *pull*, qui prend en charge l'indexation des documents, grâce à un moteur de recherche et d'indexation du marché ;
- le serveur lexical, qui assure le paramétrage des ressources linguistiques en fonction du domaine ;

- le module d'extraction d'information, qui met à jour des bases de données relationnelles à partir des informations contenues dans les documents[6].

L'agent d'acquisition est connecté, par défaut, sur un fil de dépêches de l'AFP (Agence France Presse), dont le débit[7] permet d'évaluer le respect du traitement en temps réel pour les différents modules. Le module de *push* réunit un agent de filtrage et de routage d'information[8]. Le serveur lexical regroupe l'ensemble des ressources nécessaires aux différents modules : filtres et patrons d'extraction sous forme de cascades de transducteurs, de vecteurs sémantiques, bases de données lexicales etc.

4.1.1.2.Phases de veille, phases de crise

PRIAM a été conçu de manière à offrir des fonctionnalités différentes, en fonction du contexte d'utilisation, en l'occurrence une phase de veille par opposition à une phase de crise. En phase de veille, le système fonctionne en mode ouvert : aucun besoin en information spécifique ne guide les traitements. Ce fonctionnement vise principalement à assister les opérateurs de renseignement, en leur évitant la lecture intégrale de tous les documents, et en leur fournissant des fonctionnalités minimales d'accès au texte. Ainsi, le module de *push* se contente d'indexer les documents, assisté par le module d'extraction.

En phase de crise, le système prend en compte des besoins en information définis en vue d'une prise de décision, soit sous la forme de vecteurs sémantiques[9], soit sous la forme de transducteurs à états finis (filtrage et extraction d'information). Le module de filtrage par cascades de transducteurs, CORAIL, n'est donc sollicité qu'en phase de crise, d'où des contraintes particulières d'utilisation.

---

[6] Pour plus d'informations sur ce module, voir (Poibeau, 2002).

[7] En moyenne, un document par minute, de quelques Ko, représentant quelques paragraphes.

[8] Des essais pour le routage ont été menés avec le logiciel Intuition, développé par Sinequa, qui permet une indexation de documents contrôlée par un espace conceptuel préexistant.

[9] Traités par un module d'indexation développé par la société Sinequa : Intuition, paramétré pour réaliser une tâche de routage.

4.1.1.3.L'alliance filtrage/extraction d'information

Au-delà des contraintes liées aux besoins opérationnels que PRIAM vise à satisfaire, la particularité de cette architecture est l'inter-relation constante entre filtrage, extraction et routage d'information. Ceci vaut d'être souligné, en effet la conception classique des champs d'action de chaque domaine se caractérise plutôt par un découplage de chacun d'eux. Les figures ci-dessous illustrent deux conceptions des liens entre filtrage, extraction et routage d'information.

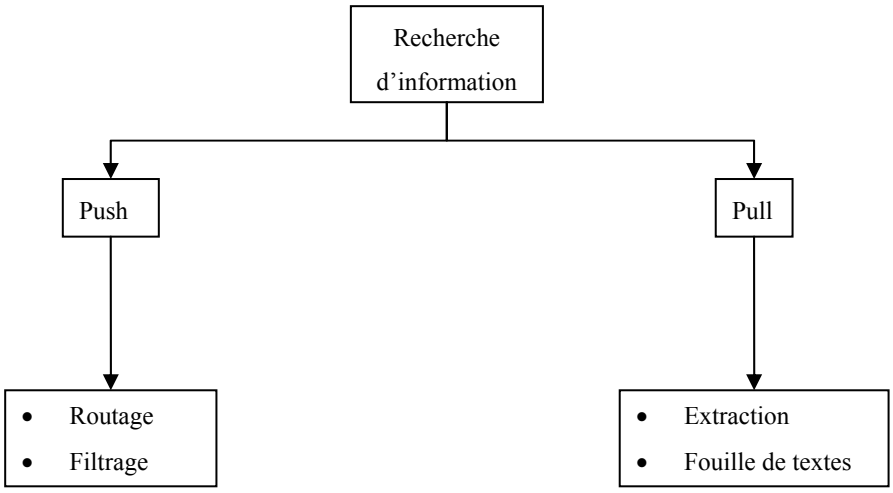

Figure 7 : conception classique des rapports entre activités de *push* et de *pull*

Cette conception classique des rapports entre les deux domaines d'activité principaux de la Recherche d'Information, le *push* et le *pull*, est celle qui guide, notamment, les conférences d'évaluation nord-américaines TREC et MUC. Or, en situation réelle, le découplage de ces deux activités n'a pas lieu d'être. PRIAM met donc en œuvre une conception des rapports entre *push* et *pull* basée sur l'interdépendance entre ces différentes activités.

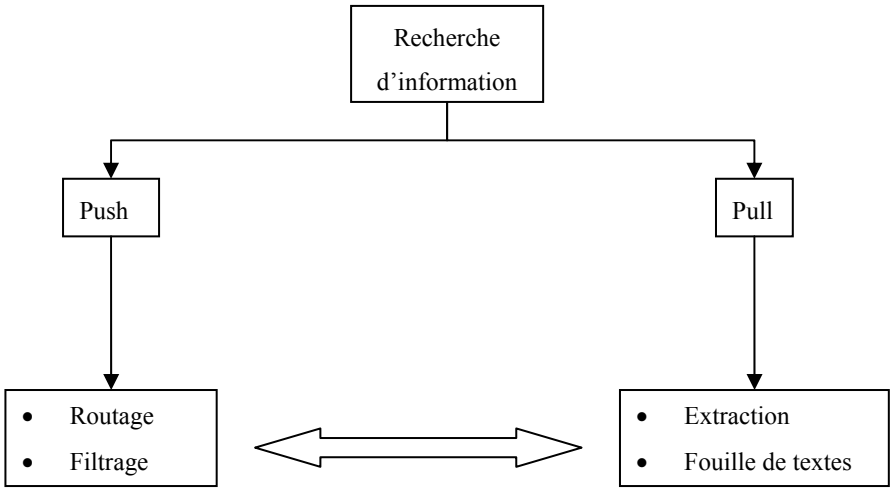

Figure 8 : PRIAM, une interdépendance entre *push* et *pull*

Cette interdépendance n'est pas propre à PRIAM : les travaux de Riloff, notamment, sont l'illustration d'une coopération fructueuse entre *push* et *pull*, en l'occurrence entre filtrage et extraction d'information pour le système Autoslog[10].

4.1.2. TALN et recherche d'information par analyse locale

4.1.2.1.La recherche de la qualité en recherche d'information

Les approches les plus répandues en RI, se basent sur une conception non linguistique de l'information apportée par des documents à traiter[11]. Ces approches adoptent un point de vue sur la langue composant les documents privilégiant les mots individuels, au détriment de la structure (syntaxique, textuelle) d'ensemble. Ces approches, reposant sur des algorithmes statistiques peu dépendants des langues particulières dans lesquelles sont rédigés les documents, ont montré leurs limites :

- en raison du caractère peu intuitif des algorithmes utilisés, il est souvent difficile d'améliorer les performances d'un système donné ;

[10] Voir (Riloff, 1994).

[11] Voir les chapitres II et III.

- de nombreux éléments porteurs d'information sont éliminés au cours des différentes phases d'indexation, ce qui fait baisser d'autant la qualité des résultats ;
- les approches « sac de mots » sont complètement dépendantes des corpus sur lesquels elles opèrent, aucune généralisation n'est possible.

De façon plus générale, on pourrait résumer la philosophie sous-jacente à ces approches non linguistiques comme la recherche du consensus maximal et l'absence de prise de risque : la seule hypothèse guidant ce type d'approche est que le contenu informatif d'un document donné peut-être condensé en une suite de quelques mots, des descripteurs de documents.

Ainsi que nous l'avons évoqué plus haut, les approches linguistiques en recherche d'information se sont développées en parallèle aux approches non linguistiques : l'intuition que des performances acceptables pouvaient être atteintes grâce à une analyse du contenu des documents guidée par des contraintes linguistiques (ex. : ordre des mots, classes de termes, structuration textuelle) est présente dès la naissance de la linguistique informatique[12]. Cependant, après plus de trente ans d'efforts, force est de constater que la percée tant attendue de la recherche d'information de haute qualité, grâce à des techniques linguistiques, n'a pas eu lieu. Ainsi, les conférences TREC, par exemple, n'ont exploré cette voie que lors des cinquième et sixième éditions, suivant l'impulsion donnée par GE Corporate Research & Development, notamment[13], la sixième édition se concluant sur un constat d'échec, en termes de gain en qualité, malgré une démonstration de la viabilité d'une approche mixte quantitative/linguistique[14].

Les échecs rencontrés dans le cadre d'une approche linguistique des problèmes de recherche d'information nous paraissent majoritairement dus à l'adoption d'outils linguistiques informatiques non adaptés à la tâche, principalement dans la profondeur d'analyse mise en œuvre. (Abney, 1996 a.), (Grefenstette, 1996), et (Roche & Schabes, 1997)

---

[12] Voir (Bar-Hillel, 1964), (Coyaud, 1972) et (Spärck Jones & Kay, 1973).

[13] Au cours des éditions ultérieures, GE Corporate Research & Development est l'une des seules équipes à proposer une approche linguistique informatique, en marge des approches quantitatives dominantes.

[14] Voir les conclusions de (Strzalkowski & Lin, 1997).

nous semblent montrer une adaptation nécessaire de la profondeur d'analyse en fonction de la tâche, certaines applications, dont la recherche d'information, pouvant très bien se satisfaire d'analyses partielles et locales.

4.1.2.2.Principes d'une analyse locale

Les analyses locales, telles que le chunking[15], ou l'analyse par grammaires locales et cascades de transducteurs[16] tirent parti de la forte redondance d'information portée par les énoncés analysés. Cette redondance permet de cibler l'analyse aux seuls constituants jugés pertinents pour la tâche. Les analyses locales peuvent être qualifiées d'opportunistes, en ce qu'elles tirent parti de tous les indices disponibles : typographiques (ex. : caractères en majuscule, ponctuation), lexicaux (classes de mots), syntaxiques, sémantiques ou stylistiques. Contrairement aux approches déclaratives dominantes il y a quelques années dans le domaine du TALN, les analyses locales possèdent un caractère plus procédural, en ce qu'elles reposent sur des classes de contextes les plus fermées possibles et des règles d'analyse ordonnées.

Ainsi, dans le domaine de l'étiquetage syntaxique, le travail de Vergne est représentatif de l'efficacité des analyses locales[17] : classé premier au cours de la campagne d'évaluation GRACE, l'étiqueteur syntaxique de l'équipe de l'université de Caen repose sur des ressources lexicographiques très pauvres, ainsi que sur un ensemble de règles d'étiquetage très restreint. Cette approche est en complète opposition avec les approches classiques, reposant sur des lexiques de plusieurs milliers d'entrées et plusieurs centaines de règles déclaratives d'étiquetage : l'étiqueteur de Vergne tire parti de la structuration en propositions reflétée par la ponctuation, afin de délimiter grossièrement les principaux syntagmes. Ce premier découpage est affiné au cours de phases d'analyse ultérieures, en se basant, par exemple, sur des indices morphologiques pour repérer les verbes conjugués et leurs compléments.

Dans le cadre de la recherche d'information, une approche par analyse locale est compatible avec la notion de signatures thématiques[18]. Ces signatures, centrées autour d'un

---

[15] Voir (Abney, 1991).

[16] Voir (Abney, 1996 a.), ou encore (Roche & Schabes, 1997).

[17] Voir (Vergne, 2001).

[18] Voir le chapitre II.

prédicat (réalisé par un verbe ou un nom) et de ses arguments (compléments habituels), constituent la cible à atteindre, des « îlots de certitude ». Dans une telle approche, seuls les passages contenant de tels îlots seront analysés.

4.1.2.3.La technique des cascades de transducteurs

Les transducteurs à états finis constituent un des formalismes grammaticaux les moins contraints de la hiérarchie définie par Chomsky. En ce sens, ils ont été considérés comme inadaptés dans le cadre d'une théorie grammaticale complète, telle qu'envisagée par le générationalisme. En effet, en raison de leur caractère peu contraint, les automates à états finis sont susceptibles de reconnaître et d'engendrer à la fois trop et trop peu d'énoncés, y compris des énoncés jugés non grammaticaux. À cette critique d'ordre formel, une contrainte d'ordre pratique doit être ajoutée : dans l'optique de l'élaboration d'une grammaire de phrases, il est plus difficile de définir un ensemble de grammaires opérationnelles à partir d'automates ou de transducteurs à états finis qu'à partir de formalismes à unification, par exemple[19]. En effet, le mécanisme d'unification permet de propager des contraintes de façon déclarative, tel que l'accord entre déterminant et nom au sein d'un syntagme nominal, là où il est nécessaire de spécifier toutes les possibilités dans les formalismes moins contraints[20].

Toutefois, dans un cadre infra-phrastique, tel que celui qui nous occupe, les transducteurs et automates à états finis, enchaînés en cascades d'ensembles de règles hiérarchisées offrent une simplicité de mise en œuvre supérieure à celle de formalismes déclaratifs. Par ailleurs, dans l'état actuel du prototype CORAIL, le choix du formalisme sous-tendant les analyses linguistiques automatisées est marqué par une priorité accordée à la démonstration de la faisabilité d'un filtrage d'information sur des bases linguistiques.

La technique des cascades de transducteurs suppose d'ordonner les phases de traitement en fonction de la généralité des analyses opérées : des plus génériques aux plus

---

[19] À moins de disposer d'un algorithme traduisant les règles d'une telle grammaire déclarative en transducteurs ou automates à états finis.

[20] En l'occurrence, dans le cadre de la définition d'une grammaire locale restreinte des syntagmes nominaux en français, les quatre possibilités données par le genre (masculin, féminin) et le nombre (singulier, pluriel), doivent être décrites une par une : déterminant masculin singulier + nom masculin singulier, déterminant masculin pluriel + nom masculin pluriel, etc.

spécifiques. Un système de traitement de l'information textuelle générique par analyse locale peut s'appuyer sur les phases d'analyse suivantes, dans l'ordre :

1. reconnaissance des frontières de phrase ;
2. reconnaissance et normalisation des unités lexicales « déviantes » (ex. : *aujourd'hui*) ;
3. reconnaissance et étiquetage des mots simples en parties du discours (ex. : {le,le.Det:ms} {chat,chat.N:ms} {court,courir.V:P3s}) ;
4. reconnaissance et étiquetage des mots composés (ex. : {la,le.Det:fs} {culture,culture.N:fs} {du,de le.PrepDet:ms} {ver à soie,ver à soie.N:ms}) ;
5. réduction des ambiguïtés d'étiquetage ;
6. reconnaissance et étiquetage des expressions figées ;
7. reconnaissance de signatures thématiques (ex. : <FINANCE>{TotalFinaElf,TotalFinaElf.N:+NPropre} {monte,monter} {au,à le.Prepdet:ms} {capital,capital.N:ms} {de,de.PREP} {EADS,EADS.N:+NPropre}</FINANCE>).

Exemple 7 : phases d'analyse d'un moteur de filtrage d'information générique

4.1.3. CORAIL, un système de FI par cascades de transducteurs

4.1.3.1. Intex pour le filtrage d'information

CORAIL est un système de filtrage d'information reposant sur une analyse locale des documents traités, guidée par le principe des signatures thématiques, afin de garantir à la fois une qualité supérieure aux approches quantitatives dominantes, ainsi que des temps de traitement maîtrisés. CORAIL repose sur Intex, un logiciel d'exploration de textes basé sur des cascades de transducteurs à états finis[21]. Le choix d'Intex se justifie par le recours des transducteurs à états finis pour l'ensemble des traitements textuels, ainsi que comme structure de données pour les ressources lexicales électroniques mises en œuvre (i.e. dictionnaires électroniques et grammaires locales disponibles pour le français).

[21] Voir (Silberztein, 1993).

Intex permet de fouiller les textes de façon approfondie : des patrons de recherche peuvent être définis sous forme de grammaires locales utilisant l'ensemble des étiquettes disponibles (environ 40), ce qui en fait un outil particulièrement adapté à une approche du problème du FI par reconnaissance de signatures thématiques.

4.1.3.2.Prétraitements

La phase de « prétraitements » désigne l'ensemble des opérations destinées à normaliser les textes traités. Cette normalisation affecte autant les niveaux les plus bas (ex. : segmentation en phrases) que les plus élevés (ex. : reconnaissance et étiquetage d'expressions figées). Dans le cas du système CORAIL, les différentes phases de normalisation sont les suivantes, illustrées sur un extrait du corpus Firstinvest.

1. segmentation en phrases, grâce à une version modifiée de la grammaire locale Sentence livrée en standard[22]
2. étiquetage des mots dits composés non ambigus, tels que *aujourd'hui* ou *a priori*, grâce à un dictionnaire dédié à ce type de lexèmes
3. normalisation de certaines formes élidées ou contractées, telles que *l'*, *au* (en *le*, *à le*) et délimitation des séquences de chiffres, par une version adaptée de la grammaire locale Replace.

Ces trois premières phases constituent des prétraitement habituels pour tout travail sur corpus, la particularité du logiciel Intex étant de pouvoir définir des grammaires locales et des dictionnaires de normalisation (ex. : étiquetage de *a priori* comme un adverbe), traduits sous la forme de transducteurs à états finis, appliqués de façon séquentielle sur les documents à normaliser. Toutes les étapes de prétraitement sont paramétrables, ainsi que l'ensemble des phases de traitements ultérieures, ce qui permet d'adapter CORAIL à différents types de corpus (ex. : corpus journalistique, littéraire, courrier électronique).

Une fois la normalisation du texte achevée, l'étiquetage des mots simples et composés, ainsi que des expressions figées, le cas échéant, peut avoir lieu. Cet étiquetage repose principalement sur les dictionnaires électroniques mis au point dans le cadre des travaux

[22] Voir l'annexe II.

menés par le LADL[23]. De même que pour les phases précédentes de normalisation, les phases d'étiquetage sont paramétrables : il est possible d'ajouter des dictionnaires spécifiques à un domaine, de sélectionner l'ensemble des dictionnaires appelés par défaut ainsi que l'ordre dans lequel ils sont appliqués : Intex fait appel à un système de priorités, qui permet d'éviter, dès les premières phases, la prolifération d'étiquettes, qui rendent d'autant plus difficiles les traitements ultérieurs. Pour le problème qui nous occupe, les dictionnaires électroniques utilisés sont essentiellement :

- l'ensemble des dictionnaires des mots simples livrés en standard (i.e. les Delaf dans la terminologie Intex) ;
- des ressources (listes/dictionnaires et grammaires locales) développées essentiellement au sein de Thales R&T pour le repérage des entités nommées[24], essentiels au repérage des signatures thématiques du thème 19 ;
- quelques ressources pour l'étiquetage des mots composés (essentiellement les mots composés « grammaticaux »).

En raison des contraintes de temps de traitement réduits, inhérentes à la tâche de filtrage en milieu industriel, le choix des ressources mises en œuvre se fait sur la base de leur intérêt pour la tâche : on ne vise pas à un étiquetage parfait, mais bien plutôt à un étiquetage suffisant[25]. De ce fait, des ressources dont la couverture est imposante, telles que le dictionnaire électronique des noms composés du LADL (i.e. le Delacf), sont délibérément écartées : leur contribution, dans le cadre du FI, n'est pas apparue suffisante au regard de l'augmentation du temps de traitement qu'elles entraînent. Pour cette même raison, dans les premiers essais, la réduction d'ambiguïtés[26] n'était pas mise en œuvre, la stratégie de filtrage par repérage de signatures thématiques suffisant à éviter la plupart des ambiguïtés gênantes[27].

---

[23] Voir (Gross, 1990), (Courtois, 1990), et (Courtois & Silberztein, 1990).

[24] Voir (Poibeau, 2002).

[25] Voir le chapitre II, ainsi que (Habert, 1998) pour une réflexion sur la complétude de l'étiquetage.

[26] Voir (Dister, 2000).

[27] Cette phase de réduction d'ambiguïtés fait désormais partie des traitements appliqués en standard à tous les documents, notamment en raison de l'amélioration des performances globales du logiciel Intex constatée pour les dernières versions.

4.1.3.3.Décision de sélection

La tâche de filtrage d'information, dont nous reprenons la définition donnée par les conférences TREC, repose sur une décision de sélection binaire prise pour chaque document traité, extrait d'un flux dynamique d'information. Contrairement aux approches les plus répandues, CORAIL conditionne la décision de sélection au repérage de séquences-clés, spécifiées au moyen de grammaires locales représentant le besoin en information, autrement dit le profil, de chaque utilisateur. Les grammaires locales utilisées par CORAIL sont typées en fonction de leur statut. On distingue ainsi :

- des primitives, filtres de bas niveau, génériques, dont la coloration thématique est la plus neutre possible, tels qu'une grammaire des dates, par exemple ;
- des filtres proprement dits, dont la coloration thématique est restreinte à un sous-thème (ex. : pour le domaine financier, des grammaires décrivant l'achat d'une société par une autre, ou encore une déclaration d'offre publique d'achat) ;
- des profils d'utilisateurs, représentés par des opérations booléennes (ET, OU, NON) portant sur des filtres.

Les figures ci-dessous montrent la hiérarchisation des ressources pour le filtrage du système CORAIL.

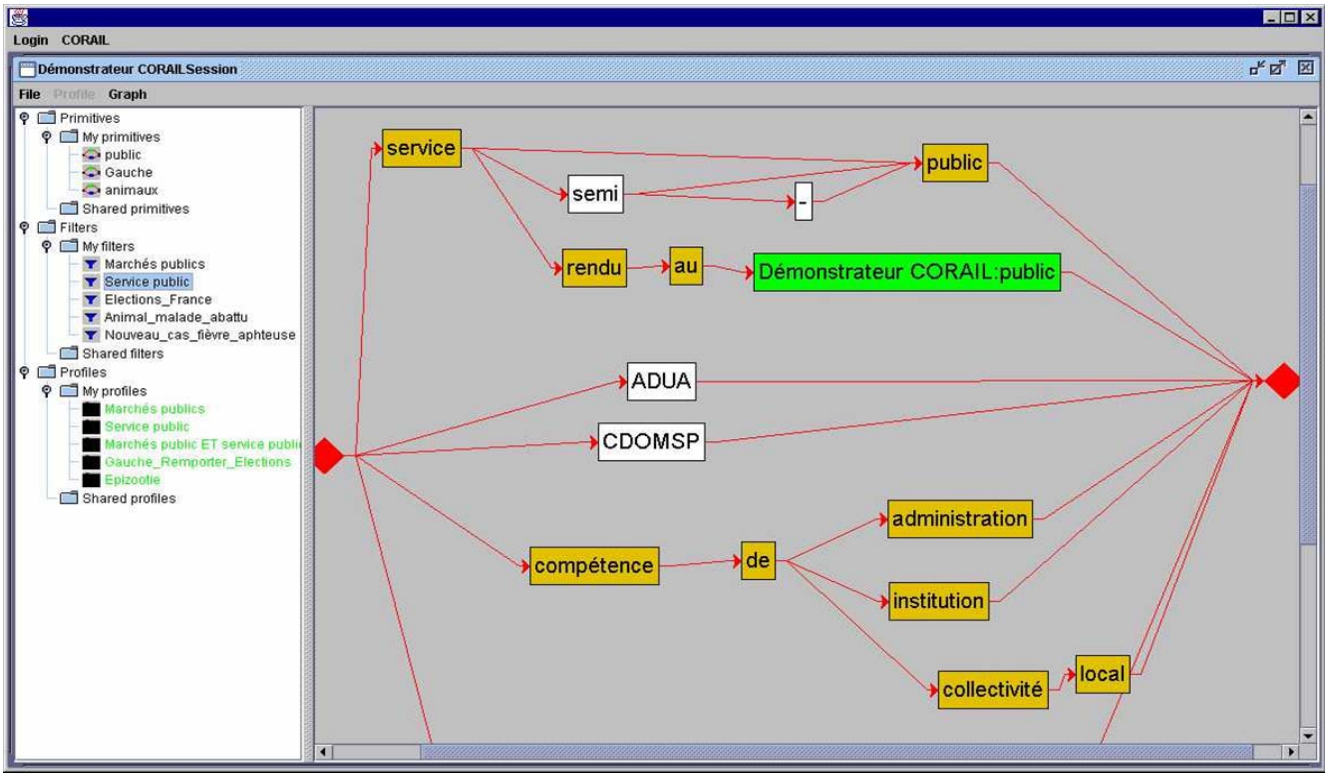

Figure 9 : interface utilisateur du système CORAIL, édition de grammaires locales pour le filtrage d'information

Cette capture d'écran montre la structure de l'interface utilisateur principale de CORAIL : les différentes ressources pour le filtrage sont regroupées dans le cadre de gauche, hiérarchisées en :

- primitives privées (MyPrimitives : public, Gauche, animaux) et partagées (Shared primitives) ;
- filtres privés (My filters : Marchés publics, Service public, Elections_France, Animal_malade_abattu, Nouveau_cas_fièvre_aphteuse) et partagés (Shared filters) ;
- profils privés (My profiles : Marchés_publics OU Service_public OU Marchés_publics ET Service_public OU Gauche_remporter_élections OU Epizootie) et partagés (Shared profiles).

Chaque élément (primitives, filtres, profils) possède des attributs de propriété, spécifiant son caractère partagé ou privé. Cette stratégie permet la réutilisabilité d'éléments jugés suffisamment génériques ou particulièrement stratégiques, tout en garantissant la

confidentialité des données propres à chaque utilisateur (dans la figure ci-dessus, la sous-grammaire nommée « Démonstrateur_Corail : public » est une ressource partagée, décrivant la grammaire locale du concept de public, i.e. *public*, *usager* ou *administré*). Cette gestion des ressources, sur le modèle des systèmes d'exploitation de type Unix, permet également d'augmenter les fonctionnalités de filtrage de CORAIL et d'en faire une plate-forme permettant le filtrage collaboratif par la mise en commun de ressources[28]. On le voit, CORAIL, dans ses objectifs et ses fonctionnalités, se situe dans un cadre applicatif difficilement compatible avec les présupposés des conférences d'évaluation TREC.

Chaque séquence reconnue par un transducteur donné est réécrite en insérant une balise particulière, une étape de post-traitement se charge d'évaluer les conditions de vérité de chaque profil et d'acheminer, par courrier électronique, les documents filtrés aux utilisateurs concernés. Les documents traités se trouvent donc enrichis d'informations apportées par les différentes phases de traitement ; les séquences validant un profil sont mises en évidence grâce à l'insertion de balises de marquage de type HTML, comme le montre la figure ci-dessous.

[28] A notre connaissance, seules des ressources explicites, telles que mises en œuvre ici, permettent le partage que suppose le filtrage collaboratif.

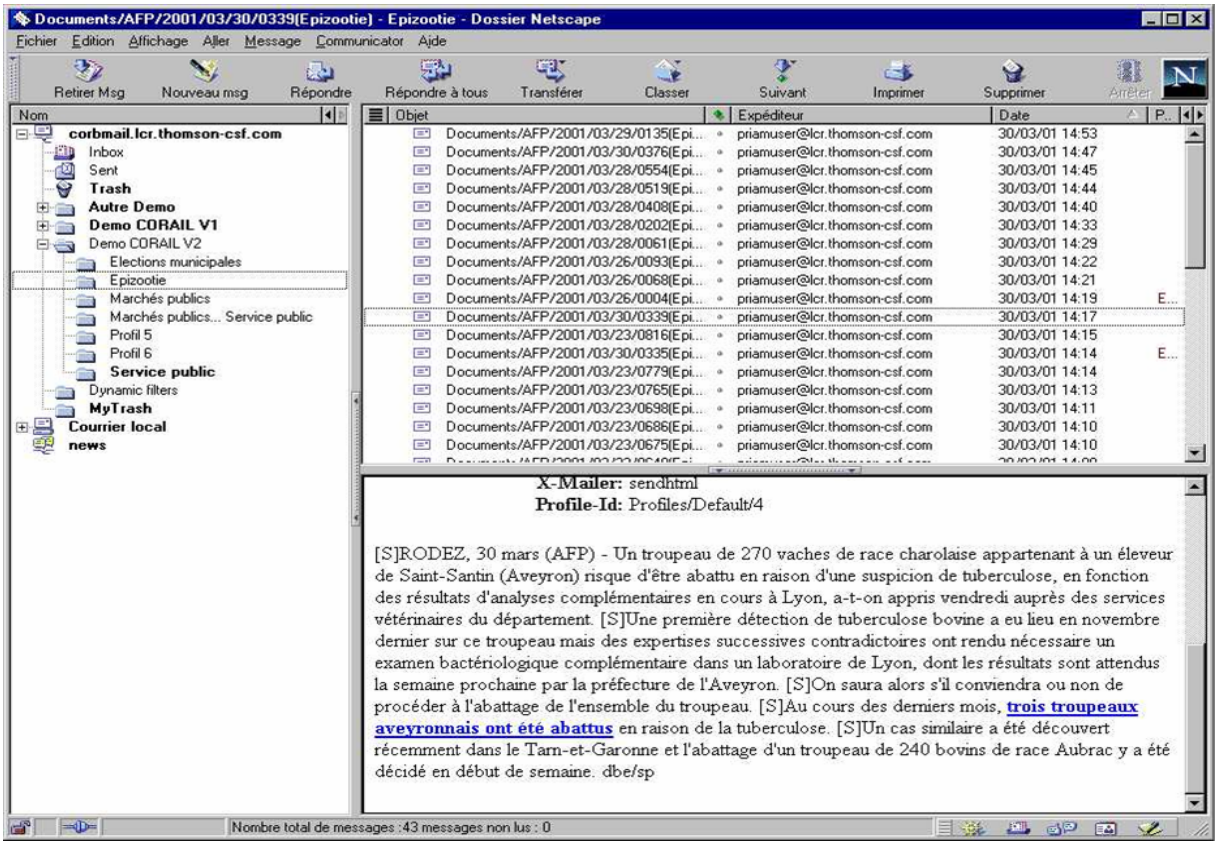

Figure 10 : visualisation des filtrats, acheminés par courrier électronique

L'intégration de l'API Javamail au système CORAIL permet l'acheminement des filtrats par courrier électronique, ainsi que la création à la volée de répertoires correspondant aux différents profils, mis à jour en temps réel (ex. : lors de l'abonnement ou du désabonnement à un profil donné). Les balises de marquage employées sont paramétrables, en l'occurrence, pour cette version de Netscape Messenger, seules des balises de mise en forme de bas niveau (i.e. soulignement et couleur des caractères) sont utilisées, toutefois l'ensemble du jeu d'étiquettes du langage XML, par exemple, peut être intégré.

4.2. LIZARD, un assistant linguistique pour la découverte de signatures thématiques

Cette partie est consacrée à un assistant linguistique, LIZARD (LInguistic wiZARD), de notre conception, destiné aux concepteurs de ressources linguistiques utilisables par un système à base de cascades de transducteurs à états finis. Nous exposons tout d'abord les besoins que vise à satisfaire cet assistant, puis le fonctionnement de cette aide à l'analyse

distributionnelle des corpus spécialisés. Enfin, nous montrons quel type de ressources lexicales LIZARD permet de constituer.

4.2.1. Motivation

4.2.1.1. Automatiser l'analyse distributionnelle des corpus

Le cadre dans lequel nous nous situons, une approche linguistique de la recherche d'information sur des textes de spécialité, présuppose un recours massif aux corpus, dont on tente d'extraire des indices thématiques, non restreints aux termes, mêmes composés. Cette extraction ne peut être menée à bien que par l'étude des observables linguistiques, dans une optique distributionnelle tant discontinue que continue, ainsi que nous l'avons dans les deux premiers chapitres de notre exposé. En effet, on cherche à constituer des classes d'éléments alliant une forme (ou ensemble de formes) et une valeur données, autrement dit des signes, à partir de régularités observées dans la distribution des formes.

Dans le cadre distributionnel discontinu classique, le travail sur corpus demande un investissement certain de la part du concepteur de ressources. Par ailleurs, toute étude à forte composante manuelle, telle que l'analyse des corpus, est sujette à des variations de qualité, liée à la disponibilité de l'opérateur humain (ex. : fatigue, stress). LIZARD vise donc à appliquer de façon systématique différentes phases d'analyse distributionnelle, en vue d'aboutir à des classes d'éléments par rapprochements entre contextes syntaxiques d'occurrence. Les phases d'analyse sont paramétrables, ce qui constitue à nos yeux un prérequis pour ce type d'outils. En effet, ainsi que le travail de Harris l'a montré, tant le domaine de spécialité que l'application visée ou encore la langue traitée peuvent demander des traitements différents. LIZARD se rapproche d'outils mis en œuvre en terminologie, tels que ceux décrits dans (Habert, 1998), ou encore (Bourigault, 2002) : en ce sens, LIZARD est un dispositif de recyclage d'étiquettes (i.e. syntaxiques).

4.2.1.2. Harmoniser et centraliser les ressources lexicales

En fournissant un cadre dans lequel les procédures d'analyse sont appliquées de façon systématique, et en exigeant de la part du concepteur de ressources de rendre explicites une partie de ses méthodes d'analyse de corpus, LIZARD vise également à assurer une harmonisation des ressources lexicales constituées. Ainsi, le format choisi pour ces

ressources, destinées à être utilisées par des systèmes à base de cascades de transducteurs, est proche de celui des tables du lexique-grammaire, tel que défini dans (Gross, 1975).

Ce format est, à nos yeux, suffisamment souple et simple (tableaux de caractères ASCII) pour garantir une certaine réutilisabilité des ressources ainsi constituées[29]. Par ailleurs, ainsi que nous l'avons présenté au chapitre II, ces tables, couplées à des automates-patrons, permettent de factoriser, en quelque sorte, des contraintes générales de construction, et de pallier un des défauts majeurs des grammaires dites locales : leur caractère relativement « procédural », c'est-à-dire dépendant d'un corpus, et d'un contexte particulier d'application.

LIZARD définit un cadre pour l'élaboration de ressources linguistiques, qui permet de centraliser les ressources lexicales extraites de corpus de spécialité. On peut, en effet, envisager l'accumulation de tables/bases de données lexicales particulières au sein d'une même base. La fonctionnalité « lexique-grammaire » du logiciel Intex permet, grâce aux auotmates-patrons, de ne générer que les grammaires locales correspondant à des contraintes définies par le concepteur de ressources.

4.2.2. Fonctionnalités principales

4.2.2.1. Une plate forme multi-agents distribuée

La figure ci-dessous donne une représentation abstraite de LIZARD. Les composants logiciels y sont figurés sous la forme de boîtes rectangulaires, les ressources lexicales produites sous celle d'un cylindre.

[29] Pour plus de précision sur ce point, voir (Balvet, 2001).

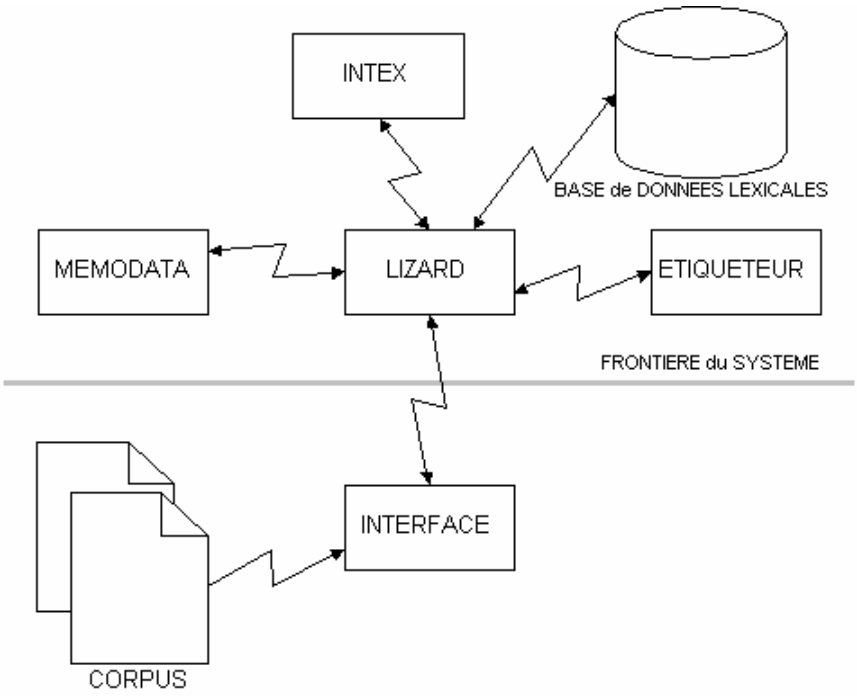

Figure 11 : architecture de l'assistant linguistique LIZARD

Les flèches brisées correspondent aux canaux de communication entre éléments de l'assistant, la plupart des communications sont bidirectionnelles, sauf entre l'interface utilisateur et les données textuelles traitées (corpus). L'orientation des flèches symbolise la rétroaction (feedback) possible ou non entre composants. La ligne grisée figure la frontière visible du système pour l'utilisateur : il n'a accès directement qu'à l'interface d'utilisation.

Les composants retenus sont :

- Intex, pour les opérations liées aux corpus[30], effectuées grâce à des cascades de transducteurs à états finis,
- le Dictionnaire Intégral de Memodata[31], pour les opérations sémantiques telles que le calcul de la distance sémantique entre deux expressions, la comparaison de mots, expressions et phrases.
- un étiqueteur morphosyntaxique générique, en l'occurrence QTag[32], est intégré au système, en concurrence avec Intex, possédant également des fonctionnalités d'étiquetage en parties du discours.

[30] Voir (Silberztein *et al.*, 2001) pour plus de précisions sur la plate-forme Intex.

[31] Les opérations sémantiques sont assurées par le Dictionnaire Intégral (DI), décrit dans (Dutoit, 2000).

L'entrée du système est constituée par des corpus de textes bruts[33], la sortie est une base de données lexico-grammaticales, codant quelques propriétés syntaxiques de surface ainsi que quelques propriétés lexicales (liens sémantiques) d'éléments extraits des corpus : nombre et type de compléments habituels de chaque verbe, transformations possibles, termes sémantiquement reliés. Cette base est le résultat d'une expansion sémantique réalisée par le composant Memodata à partir de schémas de sous-catégorisation rudimentaires extraits des corpus, en interaction avec l'utilisateur.

Dans cette conception modulaire, chaque composant peut être remplacé si l'application le demande : ainsi, on peut envisager de remplacer le DI par Wordnet[34], Intex par d'autres outils d'exploration des textes[35], ou encore d'inclure un nouveau module. Par ailleurs, chaque module peut être aisément transformé en agent logiciel autonome et distribué, en suivant les spécifications de la plate-forme Open Agent Architecture, développée au Stanford Research Institute[36]. En effet, la déclaration d'un agent OAA passe schématiquement par la spécification des services qu'il assure en termes de requêtes et de réponses, les échanges normalisés entre agents OAA étant contrôlés par un agent superviseur. L'intérêt majeur de la plate-forme OAA est la possibilité de faire cohabiter des agents hétérogènes, en l'occurrence, pour LIZARD, les agents Memodata, Interface et Étiqueteur sont écrits en Java, alors que l'agent Intex est développé en C/C++. En mode multi-agent, l'ensemble des échanges entre agents/modules ont lieu sous la forme de requêtes adressées au superviseur central, qui les aiguille vers le bon service. LIZARD peut, ainsi, être transformé en un système multi-agents distribué : les modules gourmands en ressources (tels que Memodata et Intex) peuvent être hébergés sur des serveurs dédiés, pour ne laisser que l'interface utilisateur sur le poste client.

LIZARD fonctionne comme une surcouche au-dessus des composants particuliers intégrés, destinée autant à faciliter la tâche de développeurs experts dans l'élaboration de

---

[32] Voir (Mason, 2000) pour une présentation de cet étiqueteur reprenant le principe des étiqueteurs de type Brill-tagger (Brill, 1992) ainsi que ceux des étiqueteurs statistiques.

[33] Jeux de caractères ASCII, comprenant éventuellement des balises de type HTML.

[34] Voir (Fellbaum, 1998) pour une présentation de ce thesaurus électronique conçu sur des bases psycholinguistiques.

[35] Tels que Cue (Mason, 2000), un outil de gestion des corpus.

[36] Voir (Martin et *al.*, 1999).

ressources linguistiques, que d'utilisateurs non experts. Dans l'état actuel, LIZARD n'offre que les services orientés vers les experts : la fonctionnalité principale de l'assistant est l'extraction de patrons de sous-catégorisation rudimentaires à partir de textes étiquetés et désambiguïsés. Cette extraction repose sur plusieurs phases dites de « généralisation », elle vise à fournir un ensemble d'expressions typiques et non ambiguës d'un domaine de spécialité, en fonction d'un corpus et d'une application particulière visée : les signatures thématiques. Ces phases ont pour but de ne sélectionner que les unités potentiellement intéressantes au regard de l'application visée, de façon paramétrable. Ainsi, la Figure 12 donne un aperçu d'une phase de généralisation visant à ne conserver que la forme lemmatisée des entrées verbales, suivie d'un certain nombre de compléments essentiels[37]. Par ailleurs, les mots mal étiquetés sont conservés tels quels.

4.2.2.2.Extraction de formes schématiques

La fonctionnalité principale de LIZARD est l'extraction d'expressions typiques d'un domaine, que nous appelons signatures thématiques, en plusieurs phases d'analyse distributionnelle, prenant en compte les contextes syntaxiques d'occurrence des candidats signatures thématiques. Cette extraction repose sur les modules Intex et Memodata et vise à produire des bases de données lexicales proches, dans leur format, des tables du lexique-grammaire. Les signatures thématiques recherchées[38] se distinguent des termes (Bourigault, 1993), des unités lexicales complexes (Habert *et al.*, 1997), ou encore des collocations, ou réseaux de collocations (Ferret & Grau, 2001), en ce qu'elles sont centrées autour d'un prédicat et de ses compléments habituels. Cependant, elles se rapprochent de l'ensemble de ces éléments, en ce qu'elles ne valent que pour un domaine, un corpus de spécialité et une application donnés.

Les prédicats autour desquels ces signatures sont construites sont réalisés soit par des verbes pleins à l'actif et au passif (ex. : *racheter la filiale XY*), soit par des formes nominalisées éventuellement associées à des verbes-support (ex. : *se porter acquéreur de la filiale XY*). Le repérage et l'extraction de telles signatures ne nécessite pas d'analyse syntaxique profonde : il est possible d'utiliser la stratégie des « îlots de certitude », connue en extraction d'information, et de limiter l'analyse aux seuls constituants véritablement

---

[37] Principalement des Noms, des Déterminants, des Prépositions, quelques Adverbes.

[38] Par exemple : *Thales rachète sa filiale EADS à Dassault*.

discriminants et par là-même pertinents dans le cadre de l'application visée. LIZARD, ne pouvant extraire directement des signatures thématiques complètes, met en place plusieurs phases, dites de généralisation, visant à permettre le rapprochement d'éléments apparaissant dans des contextes proches. L'ensemble des phases de généralisation sont menées à bien grâce aux fonctionnalités d'extraction de concordances de Intex[39]. Ces phases visent à uniformiser, par exemple, les contextes d'occurrence d'éléments considérés de façon générale comme potentiellement porteurs d'information, tels que les substantifs, les verbes pleins, la plupart des déterminants, des pronoms et des prépositions. Chaque phase de généralisation constitue une vision différente du corpus étudié.

L'extraction de candidats signatures thématiques passe tout d'abord par la phase des « formes schématiques », illustrée par la figure ci-dessous, qui présente à l'utilisateur un corpus partiellement généralisé :

- les verbes conjugués sont figurés sous leur forme canonique (ex. : <racheter> pour le lexème de départ *rachète*), ainsi que les substantifs ;
- les entités nommées sont présentées sous une étiquette unifiée, N+NPropre, regroupant aussi bien les toponymes (ex. : Etats-Unis), les noms propres (ex. : Marcel Dassault) que les noms de société (ex. : Thales, Dassault) ou les noms de produits ;
- la plupart des mots dits grammaticaux ne sont représentés que par leur étiquette de partie du discours (ex. : DET pour déterminant, PRO pour pronom etc.).

Dès cette phase, les éléments tels que les adverbes et syntagmes adverbiaux, les groupes de chiffres, et les commentaires sont éliminés de façon à homogénéiser les différents contextes d'occurrence. De plus, les mots mal étiquetés peuvent être corrigés (ex. : <acter> <rachetées> pour *actions rachetées*, et les contextes non pertinents peuvent être éliminés, afin de faciliter les phases d'analyse ultérieures.

[39] Pour l'exemple d'extraction de signatures thématiques donné ci-dessous, les concordances sont construites de façon à isoler les portions de phrases contenant des verbes conjugués. D'autres concordances peuvent être envisagées.

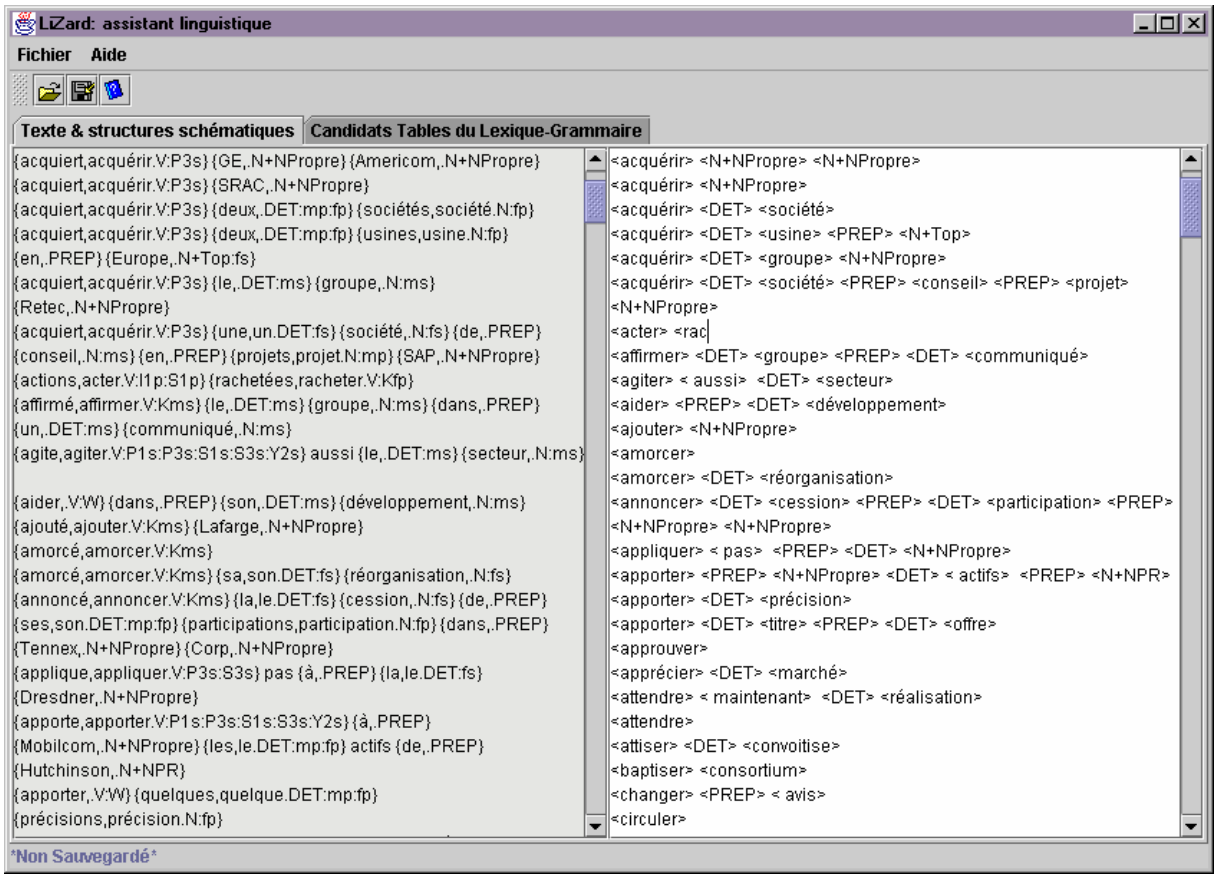

Figure 12 : LIZARD, extraction de formes schématiques

À ce stade, un certain nombre d'observations sur les préférences de sélection des verbes sélectionnés sont possibles : on voit que pour le corpus considéré (un corpus financier), au moins deux constructions sont possibles pour le verbe acquérir : *acquérir* + Nom Propre (un nom de société), et *acquérir* + groupe nominal (Det + (usine, société, groupe, nom propre)).

La stratégie des îlots de certitude, dans le cas d'énoncés tels que *Thales rachète sa filiale EADS à Dassault*, ne retient que les éléments suivants :

- un prédicat verbal, dont la structure de sous-catégorisation attend au moins deux compléments habituels, et dénotant un événement (i.e. une opération financière) considéré comme pertinent pour un thème de veille (i.e. veille économique) ;

- l'agent, le patient et l'objet d'une transaction (respectivement *Thales Dassault* et le syntagme *sa filiale EADS*), en l'occurrence des sociétés, identifiés par leur position dans l'énoncé ;
- la mention facultative d'un montant pour la transaction (ex. : *pour trois milliards d'euros*) ;
- des potentiels d'insertions, éventuellement non bornées, aux frontières des différents syntagmes, ainsi qu'entre la tête et l'extension de ces syntagmes.

Les lexèmes mentionnés ci-dessus, qui constituent des amorces pour la reconnaissance de signatures thématiques, peuvent être rattachés à la structure suivante : **N0 (Insertions) V_Achat (Insertions) N1 (Insertions) PREP N2 (Insertions) (Montant)**, dans laquelle les éléments facultatifs sont figurés entre parenthèses.

4.2.2.3. Passage de formes schématiques à des schémas de sous-catégorisation

La première phase de généralisation est suivie d'une deuxième phase, qui vise à ne produire que des schémas de sous-catégorisation tels que : V + Det + N, V + Prep + Det + N. Les schémas produits sont, dans l'état actuel, dépendants des textes traités. En cela, nous nous rapprochons de (Riloff, 1994).

Dans cette deuxième phase, seules sont présentées les étiquettes de partie du discours des extraits sélectionnés, sous la forme d'une liste qu'il est possible de trier (ex. : tri alphabétique sur le premier champ). Cette liste est destinée à fournir une estimation de la productivité des différents schémas de sous-catégorisation extraits du corpus. La figure ci-dessus donne un aperçu de la liste de schémas de sous-catégorisation générée à partir du corpus étiqueté de départ. La deuxième phase de généralisation sert essentiellement à proposer des candidats-signatures thématiques, qui seront stockées dans la base de données lexico-grammaticales sous une forme proche des tables du lexique-grammaire[40]. La procédure de généralisation concerne les traits morphosyntaxiques associés aux lexèmes : genre, nombre, personne ou encore codes sémantiques (ex. : N+NPropre pour un nom de personne, N+Top pour un toponyme) pour Les mots mal étiquetés ou inconnus du système sont conservés tels que (ex. : *solde*, *avis*).

[40] Une entrée lexicale suivie de traits binaires codant un certain nombre de propriétés syntaxiques et sémantiques, telles que le type des compléments possibles, les transformations valides etc…

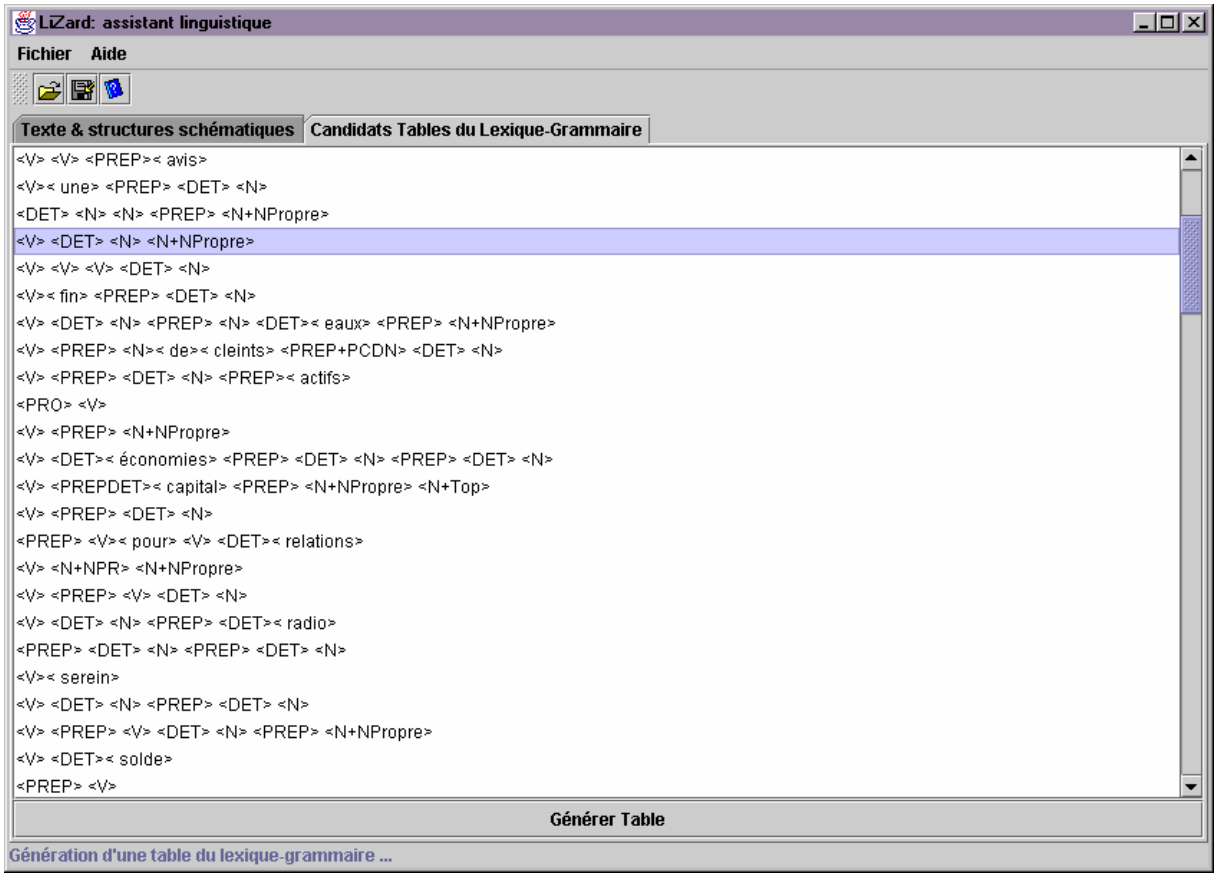

Figure 13 : LIZARD, deuxième phase de généralisation

4.2.2.4.Génération de bases de données lexicales

La phase de génération de bases de données lexicales, à partir des schémas de sous-catégorisation sélectionnés (et éventuellement corrigés) par l'utilisateur constitue la dernière étape du travail sur corpus. Elle vise à regrouper les entrées lexicales, en l'occurrence des verbes, en fonction de propriétés syntaxiques et sémantiques de surface communes. Les entrées sélectionnées seront enregistrées et codées dans un format proche des tables du lexique-grammaire tel que présenté dans (Gross, 1975), exploitable par le logiciel Intex.

La figure ci-dessous présente deux tables correspondant aux deux schémas de sous-catégorisation sélectionnés : V + Prep + NPropre et V + Det + N. En l'état actuel, la validation des tables générées à partir des corpus est réalisée manuellement, toutefois nous

envisageons de l'automatiser en utilisant les fonctions de calcul de distance sémantique de Memodata[41].

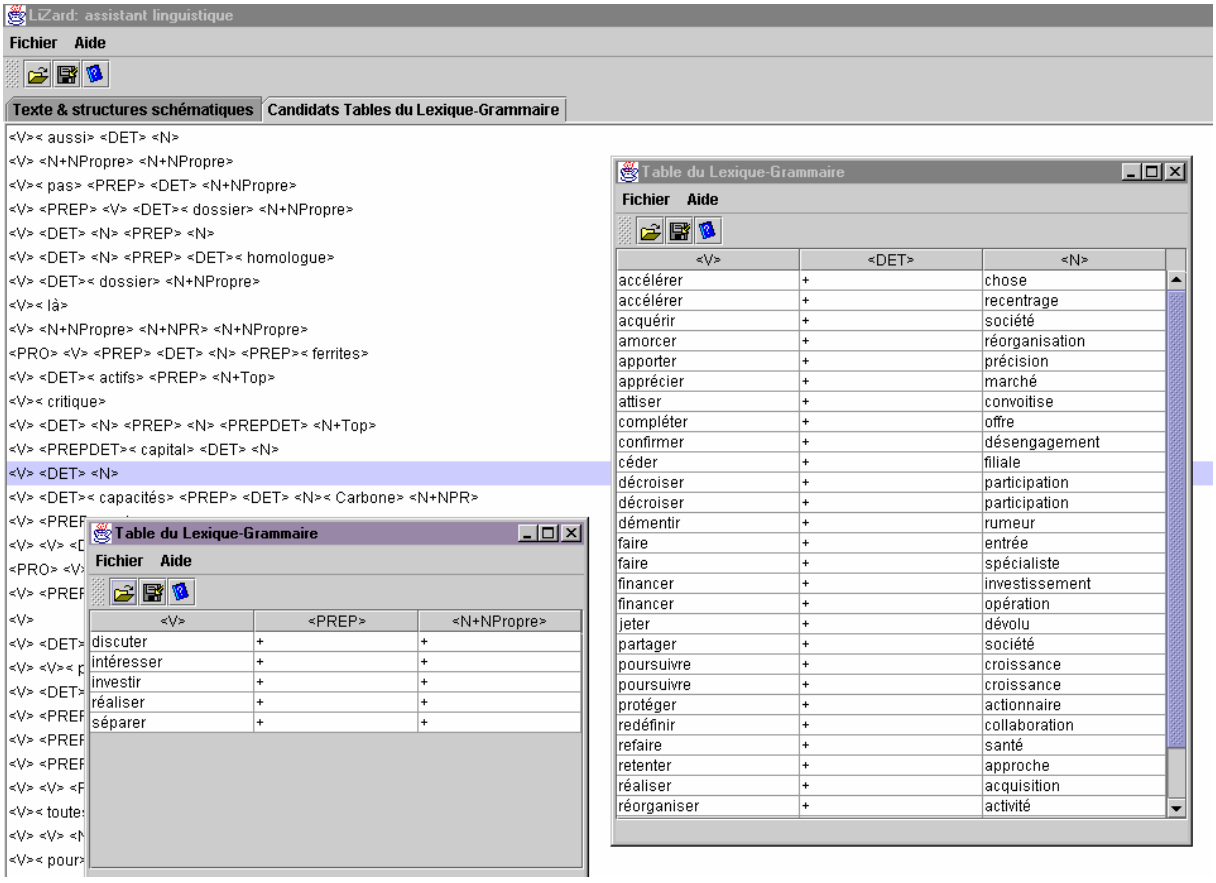

Figure 14 : LIZARD, génération de noyaux de bases de données lexicales

Une fois les tables de signatures thématiques validées, une phase d'expansion, permet de compléter ces ressources lexicales construites sur corpus par l'apport de connaissances hors-corpus. Lors de la phase d'expansion, LIZARD cherche à tout moment à ménager un va-et-vient entre connaissances spécifiques tirées des corpus et connaissances génériques,

[41] Le DI intègre des algorithmes de calcul de distance sémantique qui permettent de trouver, par exemple, que *acheter une société* et *acheter une entreprise* sont plus proches l'un de l'autre que de *acheter des fleurs*. Nous envisageons de mettre en œuvre ces algorithmes afin de proposer à l'utilisateur un regroupement des candidats-signatures, en fonction de leur profil sémantique (ex. : un classement tel que : *racheter DET filiale NPropre* > *racheter DET société NPropre* > *PRO racheter DET conduite*).

établies hors corpus. Ce mouvement est possible grâce à l'intégration des fonctions du réseau sémantique du Dictionnaire Intégral de Memodata, par exemple. En effet, cette ressource lexicale à visée généraliste, décrit la plupart des relations sémantiques et morphosyntaxiques les plus communément admises pour environ 186000 mots-sens[42].

Ainsi, les relations suivantes, entre entrées lexicales, sont codées :

- synonymie ;
- génériques ;
- spécifiques ;
- dérivés.

L'expansion du noyau de signatures thématiques au moyen des fonctions du réseau sémantique est réalisée de façon interactive, en proposant à l'utilisateur des termes sémantiquement proches de ceux trouvés dans la base : synonymes, génériques, spécifiques, ainsi que locutions proches et formes transformées (ex. : formes nominalisées d'une entrée verbale). Ainsi, par exemple, le DI permet de calculer *acheteur* et *achat* à partir de *acheter*, par les relations « personne qui V » et « action de V ». L'algorithme de parcours du réseau lui-même est décrit de façon extensive dans (Dutoit, 2000). (Poibeau, 2002) donne un exemple de paramétrage de cet algorithme pour une tâche d'acquisition de patrons lexicaux utilisés pour l'extraction d'information.

4.2.3. Une base de données lexicales pour la recherche d'information

Le résultat des opérations de fouille de texte et d'expansion des candidats-signatures thématiques est une base de données lexicales, codant le comportement syntaxique de chaque entrée, ainsi qu'un certain nombre d'informations sémantiques (ex. : termes proches). La Figure 15 donne un aperçu d'une base de signatures thématiques extraites d'un corpus financier, destinées à être utilisées par CORAIL[43].

[42] Pour plus de détails sur le Dictionnaire Intégral, voir (Dutoit, 2000).

[43] La table complète pour le thème 19 du corpus Firstinvest se trouve dans l'annexe II.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**N0 =: Nspec**	**N1 =: Nspec**	**N2**	**PPV**	**V**	**N0 V**	**N0 V N1**	**N0 V Prep N1**	**N0 V Const N1**	**N0 V N1 Prep N2**	**Const**	**Complt**	**VN**	**Actif**	**Passif**	**Nominalisation**
2	:Entreprise	:Entreprise	:Entreprise	<E>	acheter	-	+	-	-	+	<E>	<E>	<achat>	+	+	+
3	:Entreprise	:Entreprise	:Entreprise	<E>	acquérir	-	+	-	-	+	<E>	<E>	<acquisition>	+	+	+
4	:Entreprise	:Entreprise	:Entreprise	<E>	augmenter	-	-	-	+	-	<E>	:Capital	<augmentation>	+	+	+
5	:Entreprise	:Entreprise	:Entreprise	<E>	échanger	-	+	-	-	+	<E>	<E>	<échange>	+	+	+
6	:Entreprise	:Entreprise	:Entreprise	:Refl	engager	-	-	+	+	-	<E>	:Capital	<engagement>	+	-	+
7	:Entreprise	:Entreprise	:Entreprise	<E>	entrer	-	-	-	+	-	<E>	:Capital	<entrée>	+	-	+
8	:Entreprise	:Entreprise	:Entreprise	<E>	fusionner	-	-	+	-	-	<E>	<E>	<fusion>	+	+	+
9	:Entreprise	:Entreprise	:Entreprise	<E>	investir	+	-	+	+	-	<E>	:Capital	<investissement>	+	+	+
10	:Entreprise	:Entreprise	:Entreprise	:Refl	marier	-	-	+	-	-	<E>	<E>	<mariage>	+	-	+
11	:Entreprise	:Entreprise	:Entreprise	<E>	mettre	-	-	-	+	-	la main sur	<E>	<E>	+	-	-
12	:Entreprise	:Entreprise	:Entreprise	:Refl	porter	-	-	-	+	-	acquéreur de	<E>	<E>	+	-	-
13	:Entreprise	:Entreprise	:Entreprise	<E>	prendre	-	-	-	+	-	<E>	:Capital	<prise>	+	+	+
14	:Entreprise	:Entreprise	:Entreprise	<E>	racheter	-	+	-	-	+	<E>	<E>	<rachat>	+	+	+
15	:Entreprise	:Entreprise	:Entreprise	<E>	racheter	-	-	-	+	+	<E>	:Capital	<rachat>	+	+	+
16	:Entreprise	:Entreprise	:Entreprise	<E>	recapitaliser	+	+	-	-	-	<E>	<E>	<recapitalisation>	+	+	+

Figure 15 : base de signatures thématiques extraites d'un corpus financier

Cette base est le résultat d'une quinzaine d'heures de travail, elle regroupe environ quatre-vingts entrées lexicales, et représente une partie des contraintes de sélection et de construction associées à chaque entrée (ex. : nombre, type de compléments, transformations autorisées, formes nominalisées). Le format de la base elle-même est libre, bien que les informations contenues doivent être, en l'état actuel, compatibles avec Intex. On peut envisager une représentation XML de ces données, traduites par la suite dans les formats compatibles avec d'autres plateformes[44].

4.3. Mesure des performances du système CORAIL

Cette partie est consacrée à l'évaluation des performances du système CORAIL. Nous détaillons, dans un premier temps, le corpus utilisé, un corpus professionnel issu d'une pratique effective de diffusion ciblée d'information. Dans un deuxième temps, nous donnons quelques mesures de performance de l'approche du FI par signatures thématiques, puis nous complétons l'évaluation quantitative du système CORAIL par des éléments qualitatifs. Les aspects qualitatifs sont, en effet, complémentaires des aspects quantitatifs, ceux que nous

[44] Par exemple, structures de *qualia* dans le cadre du lexique génératif (Pustejovsky, 1996), ou encore structures de traits typés dans le cadre de formalismes grammaticaux à unification.

présentons se basent sur une expérience visant à évaluer l'utilisabilité du système par des utilisateurs « naïfs ».

4.3.1. Un corpus professionnel

Le corpus utilisé pour cette évaluation est issu d'une pratique effective de diffusion ciblée d'information, dans un cadre professionnel.

4.3.1.1. Un corpus financier

Le corpus de référence nous a été communiqué par la société Firstinvest, propriétaire d'un portail financier sur Internet. Les fonctionnalités offertes par ce portail sont classiques :

- alerte sélective (veille) ;
- suivi des opérations financières (archives).

Il s'agit, pour les clients de Firstinvest, de disposer de toutes les informations nécessaires à la prise d'une décision financière (ex. : achat, vente de titres). Les documents, à visée informative, dont le format est proche de dépêches journalistiques (quelques paragraphes, en texte quasi-brut), sont rédigés par des experts financiers, qui leur attribuent une étiquette thématique, prise parmi un ensemble fermé. Le corpus communiqué par Firstinvest représente environ deux mois d'activité de leur portail financier, ce qui représente 2,6 Mégaoctets de texte.

4.3.1.2. Quelques éléments stylistiques

Les documents fournis par Firstinvest sont rédigés dans un style journalistique assez contrôlé. Les dépêches suivent toutes le même format :

- un en-tête d'identification, constitué d'un numéro d'index unique ;
- une phrase de titre ;
- une phrase de sous-titre ;
- des codes de contrôle (spécifiant la date et l'heure à laquelle la dépêche a été diffusée) ;
- le corps de la dépêche ;
- un code de contrôle précisant, lorsque cela est possible, le lieu de rattachement géographique de la dépêche (ex. : FR, pour un document traitant d'opérations ayant eu lieu en France).

Par ailleurs, une structuration légère des documents est effectuée, grâce à des balises de type HTML, identifiant, par exemple, les auteurs de déclarations rapportées (balises), ou encore certains noms d'entreprises (balises), ainsi que les frontières de paragraphe (balises <br>).

13565. Generix séduit le marché¤Le titre affiche une performance positive depuis le début de l'année, les investisseurs semblent convaincus par les objectifs de la société. NEW¤2001-04-19 11:47:00.000. L'éditeur de logiciels de CRM (Gestion de la Relation Clients) Générix attire à nouveau les bonnes grâces des investisseurs. Il se négocie aujourd'hui 22,05 euros, en progression de 4,75 %. Cette semaine aura été bénéfique pour le titre qui voit sa performance depuis le début de l'année repasser dans le vert : + 17 % en quatre mois.
Il faut dire que les décrochages du Nouveau Marché avait provoqué la méfiance du marché sur tout le secteur des éditeurs de logiciels. Cependant, même si elles restent modestes par rapport à d'autres, les performances et perspectives du groupe sont rassurantes.
En effet, en 2000, la croissance des ventes s'est établie à 17 % pour un chiffre d'affaires de 14,5 millions d'euros. De plus, la société est en passe de retrouver une situation d'équilibre : la perte nette 2000 était de 1,8 million d'euros mais au deuxième semestre, le groupe dégageait un bénéfice net de 0,2 million.
Les dirigeants se disent confiants pour l'avenir : ils prévoient un doublement de l'activité tous les deux ans et ont pour ambition d'augmenter le niveau de rentabilité régulièrement. Voilà qui pourrait séduire durablement le marché.¤1. FR

Globalement, le corpus Firstinvest se caractérise par l'emploi majoritaire du mode indicatif : présent et passé composé, le mode conditionnel étant dévolu aux information demandant une confirmation. La voix active semble la plus courante, suivie des formes nominalisées (avec ou sans verbe-support) et de la voix passive. Des contraintes locales semblent toutefois faire préférer telle voix à telle autre : ainsi, les opérations d'achat sont à la voix active ou passive selon que l'opération est valorisée (voix active) ou non (voix passive). Ainsi, les mises en faillites sont au passif (*ISL déclaré en faillite*, *le numéro 1 mondial du marketing sportif, le suisse ISMM Group, a été déclaré en faillite*), alors que les rachats sont majoritairement à l'actif[45].

Par ailleurs, bien qu'on se trouve dans le cadre d'un langage de spécialité, on note un recours massif à des métaphores conventionnelles ayant trait à l'ingestion, l'attaque et les

[45] Ainsi, sur le corpus d'apprentissage constitué des 200 premières dépêches du corpus du thème 19, sur 54 phrases contenant le verbe *racheter*, 8 seulement sont au passif.

alliances : *faire main basse sur*, *mettre la main sur*, *s'allier à*, *lancer une offensive*, *absorber* etc. Le champ notionnel de la compétition est également largement développé dans le corpus, notamment dans le cas du thème 19, bien que la tonalité neutre adoptée pour la rédaction des dépêches se traduise par l'occultation des conséquences logiques d'une telle compétition, i.e. la victoire et la défaite : *les deux groupes étant au coude à coude*, *le groupe français Thales va renforcer ses positions aux Etats-Unis*.

Les dépêches sont généralement structurées de la façon suivante :

- exposition de la nature de l'opération, dans les en-têtes de titre, ainsi que dans une partie du corps de dépêche (ex. : *Ingenico met la main sur IVI-Checkmate*) ;
- exposition des détails de l'opération (montant, partenariats) ;
- motivation de l'opération (ex. : renforcer sa position sur un marché donné, revaloriser une entreprise, se renforcer dans une activité). La motivation des opérations rappelle les frames de Schank & Abelson par le caractère relativement prévisible de l'enchaînement de différentes actions en fonction d'un but donné (ex. : *renforcer sa position => monter au capital d'une entreprise dominante*, inversement *limiter les pertes financières => recentrer son activité => vendre les filiales non stratégiques*).

4.3.1.3.Structuration en thèmes

Les dépêches de Firstinvest couvrent les 21 thèmes suivants.

Thème	Intitulé	Effectif
2	Internet	8
3	Introduction	58
5	Nasdaq	3
6	Vie de la société	367
7	Opération sur le capital	87
8	Résultats	360
10	Téléphone mobile	7
11	UMTS	7
12	Wap	0
13	Produit/service	98
15	Finances perso	0
16	Opérateur	87
18	Accord/partenariat/contrat	218
19	Cession/achat/filiale	303
20	Interview	5
21	Avis	194
22	Rumeur	79
23	Profit warning	16
24	Perspectives/stratégie	283
25	Eclairage	21
26	TNT	4

Figure 16 : tableau synthétique de la répartition en thèmes du corpus Firstinvest

Ainsi que le tableau ci-dessus le montre, le corpus dont nous disposons se caractérise par des effectifs limités, voire nuls dans certains cas. Les effectifs les plus importants sont

ceux liés à des thèmes générateurs d'une intense activité : communication des résultats financiers (thème 8), opérations de cession/acquisitions (thème 19), définition de stratégies (thème 24), et annonces de partenariat (thème 18). On se trouve donc dans une situation de données éparses, contrairement aux bases documentaires de la fouille de textes, où la détection des « signaux faibles », autrement dit des éléments enregistrant des effectifs d'occurrence peu élevés, est primordiale. Notre expérience du domaine nous incline à croire que la détection des signaux faibles fait partie intégrante de la tâche de filtrage d'information et de l'activité de veille en général.

On le voit, le corpus dont nous disposons justifie le recours à une approche linguistique à base de règles d'analyse explicites construites sur corpus par interaction avec un opérateur humain, plutôt qu'une approche à base d'algorithmes d'apprentissage automatique, par exemple, pour lesquels le volume de données d'apprentissage doit être largement supérieur. Par ailleurs, à notre connaissance, la détection de signaux faibles, autrement dits le repérage des *hapax legomena*, est une tâche quasiment impossible pour des algorithmes prenant en compte des seuils de fréquence d'occurrence de certains éléments, alors qu'une approche explicite est intrinsèquement indépendante de la fréquence d'occurrence des éléments recherchés.

4.3.2. Mesure des performances

Nous discutons ici des résultats mesurés au cours d'une évaluation quantitative du système CORAIL.

4.3.2.1. Protocole d'évaluation quantitative

Pour cette évaluation, nous avons suivi un protocole de type « boîte noire », où seule la différence entre le nombre de réponses attendues sur des données de référence et celles observées pour chaque système évalué, est prise en compte. Dans ce type d'évaluation, il est nécessaire de disposer d'un ensemble de données de référence (*gold standard*), réparti en corpus d'apprentissage, ou de paramétrage, et corpus de test. Le principe d'une telle répartition est de fournir aux systèmes évalués un sous-ensemble des données de référence, qui servira au paramétrage, sans limite de temps ou d'itérations (ex. : un système réalisant de l'apprentissage automatique peut subir plusieurs présentations du même corpus d'apprentissage), ainsi qu'un sous-ensemble de test, constitué de données inconnues du système. Le corpus de test sert à vérifier l'adéquation du paramétrage, il est donc nécessaire,

afin d'obtenir des résultats interprétables, que les deux ensembles de données soient comparables (i.e. même domaine).

En sus du corpus de test, nous avons élaboré un corpus de « bruit » à partir de documents tirés d'autres thèmes que le thème évalué. En effet, le corpus Firstinvest nous est parvenu intégralement trié en fonction des thèmes vus plus haut, autrement dit nous ne disposions que d'exemples positifs pour le paramétrage du système. Or, il est intéressant de tester le système avec des documents attribués à d'autres thèmes, dont la phraséologie est *a priori* différente de celle du thème 19. Le corpus de bruit est donc constitué de 50 documents appartenant aux thèmes : 2, 3, 6, 8, 10, 13, 18, 26. Nous avons pris soin d'écarter du corpus de bruit les documents affectés à plusieurs thèmes, dont le thème 19 (ex. : plusieurs documents sont communs entre les thèmes 18 et 19).

Nous avons évalué les performances du système CORAIL sur une tâche de filtrage d'information de la manière suivante : en reprenant la définition de la tâche telle que définie au cours des conférences TREC, nous avons comparé les performances obtenues en élaborant manuellement des filtres sous forme de grammaires locales, puis en intégrant l'assistant linguistique LIZARD. Nous avons constitué une borne inférieure (*baseline*) pour un système de filtrage automatique en mettant au point un système prenant des décisions de sélection binaires de façon aléatoire, indépendamment du contenu des documents. Nous faisons l'hypothèse que le système CORAIL, avec ou sans LIZARD, devrait enregistrer des performances largement supérieures au système RANDOM, qui constitue notre borne inférieure. L'adéquation entre les réponses fournies par CORAIL et la cible (la référence) est mesurée par le test du Khi²[46], qui fournit une estimation de la probabilité de corrélation entre deux séries de données.

4.3.2.2.Indicateurs de performance

Ainsi que nous l'avons vu dans le chapitre consacré aux conférences d'évaluation TREC, le domaine du FI se caractérise par un flottement terminologique et conceptuel, qui se traduit par une absence regrettable de cadre méthodologique stable pour l'évaluation des systèmes automatiques de filtrage. Nous l'avons montré, aucune métrique d'évaluation TREC ne semble faire l'unanimité, essentiellement, selon nous, en raison de l'absence d'un

[46] Voir (Muller, 1973) pour des applications des différents tests de corrélation dans le domaine de la linguistique de corpus.

ensemble de données de référence issu d'une pratique effective du filtrage d'information. Or, nous affirmons que, muni du corpus Firstinvest, qui représente environ deux mois de diffusion sélective d'informations financières, nous nous trouvons dans une situation radicalement différente de celle des conférences TREC : nos corpus d'apprentissage et de test constituent des ensembles bornés, pour lesquels nous connaissons exactement la répartition en thèmes de chaque document. Par ailleurs, le volume de données traité, de l'ordre du Mégaoctet, reste manipulable, contrairement aux volumes titanesques de TREC, qui justifient les méthodes d'échantillonnage (notamment pooling, échantillonnage simple et stratifié) que nous critiquons.

Pour toutes ces raisons, nous choisissons deux métriques de performance standard en recherche d'information : la précision et le rappel. Les scores de silence et de bruit, sur lesquels reposent la précision et le rappel, sont donc calculés simplement en faisant la différence entre les réponses observées et les réponses attendues, l'idéal théorique étant de minimiser les deux taux conjointement (taux de silence et de bruit tendant vers 0%). La figure ci-dessous présente les performances en rappel et précision de deux versions du système CORAIL (avec et sans l'assistant linguistique), par rapport à une borne inférieure.

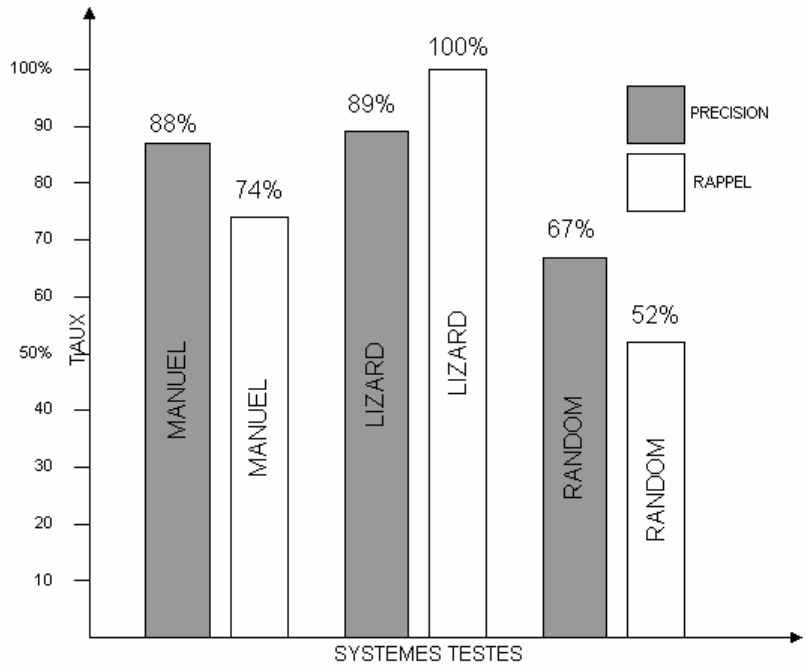

Figure 17 : scores de rappel et de précision pour deux versions du système CORAIL, comparés à un système aléatoire

Le système dénommé « Manuel » repose sur des grammaires locales élaborées manuellement en interne à Thales R&T[47], en se basant sur un recensement des signatures thématiques établi par une entreprise partenaire, E-XML Media[48], sous la supervision d'experts financiers.

Le système dénommé « LIZARD » intègre l'assistant linguistique dans la phase d'élaboration de grammaires pour le filtrage et permet de constituer de façon interactive une base de données lexicales pour le domaine de spécialité visé. Ce système se base également sur le recensement des expressions typiques du domaine financier établi par E-XML Media.

Le système « Random » sert de borne inférieure, il sélectionne les documents indépendamment de leur contenu, de façon aléatoire. Les scores de rappel et de précision donnés ici pour ce système ont été mesurés sur 10 expériences, les réponses variant à chaque essai.

4.3.2.3.Discussion des résultats

La discussion des résultats dégagés de l'évaluation ci-dessus repose sur la mise en œuvre du test du Khi², ou test de Pearson, afin d'évaluer la corrélation des réponses fournies par les différents systèmes et la cible, constituée par les données de référence. Le test du Khi² s'applique dans le cas où des résultats théoriques et des observations effectives sont comparés ; il permet de déterminer la probabilité de corrélation entre les résultats théoriques et les observations. La formule du test de Khi² est la suivante, où *o* représente une valeur observée ou réelle, et *c* une valeur calculée ou théorique :

$$Khi^2 = \sum (o - c)^2 / c.$$

Formule 5 : test du Khi-2

Dans notre cas, nous proposons de considérer les performances d'un système aléatoire idéal comme des valeurs théoriques attendues, nous proposons donc de considérer les réponses fournies par les systèmes évalués comme des observations effectives. En effet, la tâche considérée ici revient à prédire l'issue d'un tirage pouvant donner deux événements,

[47] Voir (Bizouard, 2001) pour plus de détails.

[48] Voir (Amardeilh, 2002).

pour chaque document d'un corpus (corpus de test et corpus de « bruit ») : décision de sélection ou non. Autrement dit, nous considérons que la probabilité théorique associée à chacun des deux événements est égale à ½, pour l'ensemble des documents de chaque corpus. Ainsi, le système théorique présente les performances suivantes :

- Effectif du corpus de test / 2 = 103 / 2 = 56,5 ;
- Effectif du corpus de bruit / 2 = 50 / 2 = 25.

Lors de l'utilisation de tests tels que le Khi-2, on cherche à évaluer la probabilité d'une hypothèse nulle : en l'occurrence que les deux séries de données (observées et théoriques) ne sont pas corrélées. Plus cette hypothèse nulle a une probabilité faible, plus les chances de se tromper en réfutant l'hypothèse nulle sont faibles. Le test du Khi-2 s'applique uniquement sur des effectifs, réels ou théoriques. Le tableau ci-dessous regroupe donc les mesures de performance pour chaque système, exprimées en nombre de documents pertinents retrouvés (*hits*), manqués (*missed*) ou non pertinents (*noise*).

Système	Observés	Théoriques (attendus)	observés - attendus	Khi² = (observés - attendus)²/attendus	Probabilité de l'hypothèse nulle
Manuel					
Hits	76	56,5	19,5	6,7300885	
Missed	27	56,5	-29,5	15,4026549	
Noise	9	25	-16	10,24	
Total				32,3727434	9,34003E-08
Lizard					
Hits	103	56,5	46,5	38,2699115	
Missed	0	56,5	-56,5	56,5	
Noise	13	25	-12	5,76	
Total				100,529912	1,47981E-22
Random					
Hits	53,2	56,5	-3,3	0,19274336	
Missed	49,8	56,5	-6,7	0,79451327	
Noise	24,8	25	-0,2	0,0016	
Total				0,98885664	0,60991949

Figure 18 : résultats du test du Khi-2 pour 3 systèmes de filtrage d'information

Pour chaque système, pour chaque type de réponse, les effectifs observés figurent dans la colonne « observés ». La colonne « théoriques » donne le nombre de documents attendus

dans chaque catégorie de réponse (*hits*, *missed* et *noise*), calculés sur la base du système théorique présenté plus haut, associant une probabilité de sélection de ½ pour chaque document du corpus considéré[49]. Le tableau donne la différence entre effectifs observés et attendus, ainsi que la valeur du Khi² correspondante. La dernière colonne présente la probabilité associée à l'hypothèse nulle, calculée en fonction du nombre *v* (*nu*) de degrés de liberté du tableau de contingence ci-dessus : $v = 2$[50].

Pour $v = 2$, le seuil de pertinence est atteint (probabilité égale à 0,05) pour un score de Khi² supérieur ou égal à 5,991[51]. On le voit, le seul système pour lequel l'hypothèse nulle ne peut être écartée est le système Random (Khi² = 0,98885664). Le test du Khi² permet donc d'infirmer l'hypothèse nulle pour les systèmes :

- Manuel, avec une probabilité d'erreur de $9{,}34.10^{-8}$ (valeur arrondie) ;
- Lizard, avec une probabilité d'erreur de $1{,}48.\ 10^{-22}$.

Le test du Khi² permet donc de compléter l'évaluation classique reposant sur des scores de précision et de rappel, en précisant la probabilité d'erreur associée au rejet de l'hypothèse nulle. Dans le cas des deux variantes du système CORAIL, cette hypothèse nulle peut donc être infirmée.

Une deuxième remarque peut être faite au sujet de ces résultats, elle concerne l'apport d'un assistant linguistique dans le processus d'élaboration de grammaires locales pour le filtrage. Le système Manuel, ainsi qu'on pouvait s'y attendre, enregistre de bonnes performances tant en rappel (74%) qu'en précision (88%) : elles sont largement supérieures aux performances du système Random. La qualité du système Manuel tient au recensement des signatures thématiques du thème 19, établi sur corpus sous la direction d'experts du domaine. Cependant, les manques observés, notamment en rappel, semblent dus à une

[49] Les effectifs attendus font donc toujours référence au même système théorique.

[50] Le nombre de degrés de liberté pour chaque système est donné par la formule : $v = (n - 1)(k - 1)$, où n représente le nombre de colonnes et k le nombre de lignes. Pour chaque système, on a donc $v = (3\text{-}1)(2\text{-}1) = 2$.

[51] Les valeurs de la probabilité associée à l'hypothèse nulle ont été obtenues grâce aux fonctions statistiques intégrées à Excel ™.

application non systématique de principes d'études sur corpus : certaines grammaires sont incomplètes, trop lacunaires car elles ne prennent pas suffisamment en compte la variation syntaxique (diversité des constructions) et lexicale (diversité des choix lexicaux) du corpus. Les performances du système Lizard montrent clairement qu'une meilleure couverture est possible, notamment en intégrant des connaissances génériques tirées du DI de Memodata, sans dégradation de la précision, qui reste comparable à celle du système Manuel. Si on garde à l'esprit les effectifs très limités (101 documents de test, 50 documents de bruit), pour lesquels l'influence individuelle de chaque document est très sensible, on constate que l'approche du filtrage d'information par grammaires locales semble très discriminante et permet de fournir des résultats de haute qualité, sur le corpus considéré tout du moins. Des évaluations menées selon les mêmes principes que ceux exposés ici, sur d'autres types de corpus, permettraient de préciser davantage quelles performances peuvent être attendues d'un système de filtrage d'information basé sur une analyse linguistique de corpus spécialisés.

4.3.3. Questions d'utilisabilité

Au-delà des aspects purement quantitatifs, il est nécessaire de se poser la question de la qualité du système évalué. L'évaluation qualitative de systèmes de recherche d'information automatiques est un domaine de recherche à part entière, c'est pourquoi nous nous limiterons aux expériences réalisées dans le cadre du projet CORAIL, sous la conduite d'une équipe d'ergonomes.

4.3.3.1.Ébauche d'une évaluation ergonomique

Le consortium CORAIL comprenait l'équipe CRIS/Paris X, constituée de deux ergonomes, qui ont dirigé une expérience visant à évaluer l'utilisabilité, abordée sous l'angle ergonomique, du système de filtrage d'information par cascades de transducteurs à états finis, CORAIL[52]. Deux campagnes d'évaluation ont eu lieu, la première sur le site de Thales R&T, la deuxième sur le site de ICDC/DTA. Ces deux campagnes ont concerné des publics d'utilisateurs différents.

L'évaluation menée en collaboration avec le laboratoire Thales R&T visani à analyser l'appropriation du concept de grammaire locale pour le filtrage d'information par des

[52] Les détails des évaluations ergonomiques sont consignés dans (Viard, 2000 a.).

utilisateurs non linguistes. Le protocole d'évaluation comportait, à l'origine, deux volets d'expériences :

- des expériences de compréhension, dans lesquelles la lisibilité des grammaires locales utilisées par le système CORAIL était évaluée ;
- des expériences de production, dans lesquelles les sujets devaient élaborer eux-mêmes des grammaires locales pour le filtrage d'information.

Seule la lisibilité des grammaires locales a pu être évaluée, sur 13 sujets recrutés sur le site de Thales R&T. Aucun des sujets n'était familier des concepts linguistiques de grammaire formelle et d'analyse du langage naturel, la plupart n'étant par ailleurs pas informaticiens[53]. La tâche consistait, après familiarisation avec l'outil et les conventions utilisées[54], à associer, pour chaque phrase d'un corpus d'une dizaine d'énoncés, une grammaire locale. Les grammaires locales présentées comprenaient des leurres, plus ou moins complexes, et étaient élaborées de manière à ménager une gradation dans la complexité de lecture :

- grammaires plates (sans appel à des sous-grammaires) ;
- grammaires plates utilisant la notion de lemme (ex. : toutes les formes conjuguées d'un verbe, toutes les formes d'un substantif) ;
- grammaires à 1, 2 …n sous-niveaux, avec ou sans lemmes.

Pour chaque épreuve étaient mesurées le temps d'exécution, les erreurs commises et leur réparation le cas échéant, ainsi que les commentaires de chaque sujet. Un entretien individuel suivait chaque expérience, permettant aux ergonomes de disposer d'un retour sur les difficultés rencontrées au cours des épreuves, ainsi que d'éléments de nature qualitative sur le système CORAIL.

L'évaluation menée sur le site de ICDC/DTA visait essentiellement à analyser les modes opératoires d'utilisateurs du système de filtrage propriété de ICDC, Exoweb, confrontés au système CORAIL. La différence essentielle entre les deux systèmes étant la délégation *versus* l'autonomie dans le processus de création de filtres : les filtres Exoweb sont

[53] La plupart des sujets utilisaient des outils informatiques, toutefois seul un sujet était informaticien professionnel.

[54] Voir l'annexe consacrée au projet CORAIL.

élaborés par les administrateurs du système, alors que les filtres CORAIL sont élaborés par la communauté d'utilisateurs, en collaboration éventuelle (réutilisation des filtres possible).

4.3.3.2. Quelques résultats

Les conclusions des expériences d'évaluation ergonomique sont les suivantes, elles constituent essentiellement des recommandations dans l'optique de la poursuite du projet CORAIL[55] :

1. évaluation Thales R&T

- le concept de grammaire locale semble présenter peu de difficultés d'assimilation. La présentation graphique, la navigation au sein des sous-grammaires et la sémantique des différents types d'états distingués par leur couleur ont été relativement facilement assimilés, par des utilisateurs n'ayant à leur disposition qu'un manuel communiqué quelques jours avant l'expérience, ainsi que d'une présentation de 20 minutes du système CORAIL par l'ergonome menant l'expérience.
- les différents niveaux d'analyse doivent être distingués, certains ne nécessitant que des connaissances sommaires (ex. : ce que regroupe la classe des substantifs), alors que d'autres supposent de bonnes connaissances en grammaire (ex. : équivalence entre voix active et passive).
- des représentations différentes des relations de dépendance entre constituants ont été observées chez les sujets (ex. : *très exalté* est parfois considéré comme un mot composé, en raison du caractère récurrent de l'association entre un adverbe et un adjectif).

2. évaluation ICDC/DTA

- la nécessité d'un retour a été ressentie par les utilisateurs testés, au cours de l'élaboration d'un filtre. Pour ces sujets, un tel retour peut être fourni par des exemples du langage engendré par la grammaire locale servant de filtre, d'une part, par la mise en relation entre une grammaire locale (ou une partie) et l'ensemble des documents sélectionnés grâce à cette grammaire, d'autre part.

[55] Les recommandations liées à l'interface graphique ne figurent pas ici, pour plus de détails, voir l'annexe consacrée au projet CORAIL.

- des outils d'aide à l'élaboration de grammaires locales ont été demandés par les utilisateurs, notamment en ce qui concerne l'extraction de segments thématiques pertinents, autrement dit des signatures thématiques.
- de même que pour l'évaluation Thales R&T, des représentations concurrentes des objets linguistiques manipulés ont été observées chez les utilisateurs, notamment pour la notion de mot, pour laquelle une hésitation entre une conception typographique et une conception plus linguistique a été observée (les mots composés, les expressions figées et les « groupes de mots » sont-ils des mots ?).

Les résultats de ces expériences visant à évaluer l'utilisabilité d'un système de filtrage d'information reposant sur une analyse linguistique locale militent en faveur de la diffusion de tels outils : la plupart des sujets ont évoqué l'usage qu'ils feraient d'un tel système, en des termes tels que « alléger ma charge de travail », ou encore « ne garder que les messages importants ». Par ailleurs, certains sujets ont fait preuve d'une compétence inattendue en matière d'analyse du langage naturel, notamment en ce qui concerne les concepts de parties du discours (ex. : nom, verbe), de transformation (actif/passif) et d'analyse en constituants immédiats (ex. : groupe verbal). Il est intéressant de noter, pour l'évaluation Thales R&T notamment, que les taux d'erreur sur les différentes tâches ont été particulièrement bas, malgré leur complexité, à tel point qu'un partage entre les tâches n'a pas été possible.

Les expériences réalisées dans le cadre du projet CORAIL semblent donc militer pour la diffusion de systèmes de filtrage d'information visant une haute qualité, d'une part, ainsi que celle de systèmes d'analyse du langage naturel reposant sur des ressources explicites, telles que les cascades de transducteurs à états finis. Cependant, ces expériences soulignent également la nécessité de disposer d'une gamme de fonctionnalités plus ou moins explicitement linguistiques, afin de répondre aux besoins d'une population d'utilisateurs hétérogène : les uns maîtrisant les principaux concepts de l'analyse automatique du langage naturel et adoptant des stratégies analytiques (recensement des éléments pertinents), les autres ayant une vision plus conceptuelle (définition d'un besoin en information en des termes génériques : *agressions entre Israéliens et Palestiniens*, par ex.). Autrement dit, malgré la validation d'une approche du FI par grammaires locales, ces expériences ont également montré la nécessité de mettre en œuvre des interfaces utilisateurs intelligentes, adaptables en fonction du type d'utilisateur (ex. : novice/expert) et de la situation d'utilisation (ex. : phase de veille/phase de crise). Ainsi, une interface conceptuelle apparaît nécessaire, grâce à

laquelle les détails des opérations linguistiques resteraient cachés : l'utilisateur n'aurait, par exemple, qu'à renseigner des champs *Qui ?* (autrement dit l'agent et le patient) *Quoi ?* (l'événement, ex : une attaque terroriste) *Où ?*et éventuellement *Comment ?* (ex. : voiture piégée) pour qu'un filtre à base de grammaires locales soit généré. On est proche d'une conception telle qu'exposée dans (Kalgren *et alii.*, 1994) de systèmes « boîte noire dans une boîte de verre » (*a black box in a glass box*), dans lesquels les objets et la complexité du domaine reste cachée, l'utilisateur n'ayant accès qu'aux niveaux conceptuels les plus élevés.

4.4. Conclusion

Dans cette partie, consacrée au système de filtrage d'information par analyse locale CORAIL, reposant sur des cascades de transducteurs à états finis, nous avons abordé les aspects techniques, opérationnels et ergonomiques de l'implantation du système réalisée au sein du laboratoire Thales R&T, ainsi qu'à la Direction des Travaux Avancés de Informatique CDC, membre du consortium.

Nous avons détaillé le cahier des charges d'une plate forme opérationnelle de gestion électronique des documents, PRIAM, dans laquelle le système CORAIL est intégré. Nous avons montré quelles performances un système de recherche d'information tel que CORAIL était à même de réaliser et quel profondeur d'analyse de la langue naturelle, vue comme support d'information privilégié, était nécessaire.

Les expériences menées sur un corpus professionnel du domaine financier nous ont permis, d'une part, de valider :

- le recours aux grammaires locales, traduites sous forme de transducteurs, pour le filtrage d'information, tant pour la qualité des résultats que pour la maîtrise des temps de traitement ;
- l'approche par signatures thématiques ;
- l'apport d'un assistant linguistique, LIZARD, automatisant certaines étapes de l'analyse des corpus, en termes d'harmonisation et de centralisation des ressources lexicales pour la recherche d'information.

Par ailleurs, les évaluations ergonomiques menées dans le cadre du projet CORAIL ont permis de constater :

- l'appropriation relativement aisée du formalisme des grammaires locales par des utilisateurs non linguistes et non informaticiens, ou à tout le moins peu familiers des problèmes d'analyse automatique des langues naturelles ;
- la bonne lisibilité des grammaires locales présentées sous une forme graphique, par rapport à des expressions régulières, par exemple, plus compacte mais moins immédiatement intelligibles ;
- la diversité des représentations linguistiques des utilisateurs potentiels de systèmes tels que CORAIL ;
- la nécessité de ménager plusieurs niveaux de fonctionnalités linguistiques, en fonction des utilisateurs et du contexte d'utilisation, militant pour le principe de systèmes dits « boîte de verre dans une boîte noire », autrement dit des systèmes où seuls les niveaux conceptuels les plus élevés sont accessibles à l'utilisateur.

Pour notre part, nous insistons sur la nécessité d'offrir à des utilisateurs non spécialistes des fonctionnalités de traitement automatique des langues, dans le cadre d'applications de recherche d'information. En effet, les utilisateurs potentiels de tels systèmes commencent à prendre conscience que les outils les plus utilisés, destinés à des besoins en information peu spécifiques, dans le cadre de situations de veille non stratégiques, i.e. les moteurs d'indexation et de recherche par approche vectorielle, ne sont pas adaptés. Ces utilisateurs se tournent d'ailleurs, parfois, de nouveau vers des approches manuelles, non par conservatisme mais bien plutôt par pragmatisme : seul l'expert humain est à même de leur apporter la qualité qu'ils recherchent. Nous pensons que, bien qu'il soit utopique de vouloir remplacer ces experts, des outils et des approches tels que ceux que nous avons présentés peuvent, au moins, alléger la tâche des experts et assurer une constance dans le niveau de qualité qu'une approche complètement manuelle ne peut pas garantir.

En conclusion, on pourrait avancer que le domaine de la recherche d'information se trouve dans la même situation que celui de la traduction automatique : des utopies originelles, visant à mettre en place des systèmes « presse bouton », où l'ensemble des traitements linguistiques seraient réalisés sans le concours des utilisateurs, on est passé à une conception plus réaliste, où les outils, qu'ils soient proprement linguistiques ou non, sont vus plus comme des aides que comme des experts automatiques. Il nous apparaît, de ce fait, que seul un positionnement des outils d'analyse linguistique automatique dans les termes que nous avons

évoqués, i.e. des assistants pour des tâches complexes, conjugué à un rapprochement avec les besoins effectifs des opérationnels du domaine de la recherche d'information sont à même de voir la généralisation (certains parlaient d'explosion) tant attendue des techniques issues du TALN.

CHAPITRE 5

Conclusion et perspectives

5.1. Un cadre pour une linguistique des corpus

Dans l'ensemble de notre exposé, nous nous sommes efforcé de définir un cadre méthodologique et théorique pour une linguistique centrée sur les productions effectives. La nécessité d'un tel cadre vient du constat :

1. de la prépondérance des approches guidées par les observables dans les domaines applicatifs ;
2. d'une conception empreinte de pragmatisme de la place qu'occupent de telles analyses.

En effet, dans le domaine applicatif, représenté essentiellement par l'ingénierie linguistique, la pédagogie (l'enseignement des langues étrangères) et les approches lexicographiques (terminologie), la prise en compte des productions linguistiques dans leurs paramètres les plus fins, autrement dit la prise en compte de la variation, constitue l'objet central. La description d'un maximum de variantes possibles (ex. : la couverture d'un dictionnaire) est perçue comme fondant la valeur ajoutée des applications développées.

Avec l'avènement du générationalisme, deux linguistiques se dessinent : d'un côté, une linguistique « empirique », de l'autre une linguistique théorique, rationaliste. La seconde s'est fondée en même temps que l'appareil formel sur lequel elle repose. En posant la question de la scientificité d'une linguistique théorique, les tenants du générationalisme ont également contraint la linguistique empirique à prendre position sur la question. En cela, l'un des apports essentiels de la linguistique rationaliste et théorique à l'ensemble du domaine est de nature épistémologique. En posant la question des conditions d'émergence d'une compétence

linguistique, le générationalisme a défini un cadre pour toute théorie linguistique, reposant sur les notions de conditions d'adéquation descriptive, prédictive et explicative de modèles, censés rendre compte de la grammaticalité.

La linguistique empirique a le plus souvent été caractérisée par les tenants d'une linguistique théorique comme une simple méthode de description, arguant du fait qu'elle ne pouvait ni prédire (induire des règles à partir des observables), ni expliquer (fournir les conditions d'émergence d'un système linguistique) la grammaticalité. Qui plus est, l'extrême variation observée dans les productions effectives a été considérée comme fondamentalement incompatible avec l'élaboration d'une théorie linguistique scientifique, reposant sur des principes logiques et catégoriques.

Les récents développements dans le domaine de la linguistique empirique, marqués notamment par l'abandon du principe catégorique au sujet de la grammaticalité, ainsi que la faillite des approches linguistiquement fortes, telles que le générativisme, dans le domaine applicatif, font de la question de la scientificité d'un fondement empirique d'une théorie linguistique une question d'actualité, c'est l'objet du passage ci-dessous.

> All in all, while much still remains to be done, we may well be seeing the beginning of a new version of the Harris program, in which computational models constrained by grammatical considerations define broad classes of possible grammars, and information-theoretic principles specify how those models are fitted to actual linguistic data.
>
> (Pereira, 2000, p. 1250)

Le débat entre fondement empirique et théorique d'une science doit être mis en rapport, dans le cas de la linguistique, avec la disponibilité accrue de données observables depuis le début des années 1990. En effet, le regain d'intérêt pour le programme distributionnel intervient à un moment où, au niveau mondial, des corpus de toute nature (langue générale, littérature, domaines de spécialité, transcriptions de l'oral), dans des langues appartenant à des groupes linguistiques différents, deviennent accessibles[1], rendant, du même

[1] Ce mouvement est d'une telle importance que des organismes supranationaux, tels que l'ELRA (Evaluation and Language Resources Agency) pour l'Union Européenne, ont vu le jour, afin de fédérer et de standardiser les données linguistiques disponibles.

coup, envisageables, voire indispensables des approches guidées par les observables. Nous mettons donc en parallèle le développement d'approches empiriques avec la disponibilité en données linguistiques.

La question des relations entre scientificité et empirie se pose avec d'autant plus d'insistance que d'autres domaines, ayant pour objet les productions linguistiques effectives, adoptent un point de vue linguistique faible, ainsi que des approches non catégoriques.

5.2. Linguistique de corpus et recherche d'information

Nous avons évoqué les liens historiques étroits entre recherche d'information et TALN, et nous avons examiné une application d'un principe d'analyse automatisée, reposant sur une position linguistiquement faible, au problème du filtrage d'information. En effet, nous avons tenté de déterminer la relation entre discrimination thématique et occurrence d'unités lexicales complexes, les signatures thématiques. Dans l'expérience décrite au chapitre 4, une certaine adéquation peut être observée entre les signatures thématiques extraites des corpus et la répartition thématique des documents. Nous avons donc montré quel pouvait être l'apport d'une étude linguistique des corpus dans un domaine applicatif. Toutefois, les bons résultats enregistrés dans l'expérience décrite ne doivent pas occulter le fait que, bien que l'adéquation entre signatures thématiques et thèmes soit bonne, elle n'est pas parfaite.

L'imperfection de l'adéquation signatures/thèmes peut être due à une couverture insuffisante des grammaires locales utilisées. Elle peut également être due à la notion même de signature thématique, telle que nous l'utilisons dans nos expériences : une signature thématique présente dans un document est vue comme caractérisant l'ensemble du domaine thématique du document. Or, bien souvent, les signatures thématiques ne représentent qu'une partie des énoncés présents dans les documents. Il est envisageable que, bien que les signatures thématiques soient de bons marqueurs thématiques, ils ne soient qu'une généralisation utile, qu'une stratégie efficace. En d'autres termes, nous ne prétendons pas avoir décrit l'essence de la compétence des experts financiers, dans le domaine des cessions et acquisitions de société, par les grammaires locales présentées en annexe II. Qui plus est, nous soulignons le caractère irréductible de cette compétence, devant laquelle les approches à base de règles d'analyse explicites sont fondamentalement limitées.

C'est l'objet de la modélisation proposée dans le chapitre 4, basée sur une conception alternative du processus de filtrage d'information, aboutissant à une classification thématique des documents : nous proposons l'esquisse d'un modèle de la décision de sélection à base d'un principe de satisfaction de contraintes hiérarchisées, éventuellement contradictoires, inspiré du modèle OT[2]. Ce modèle, qui intègre les connaissances encyclopédiques nécessaires aux experts financiers[3], constitue une piste à explorer dans l'optique d'une amélioration des systèmes de diffusion ciblée d'information.

Cette proposition de modèle de la décision de sélection constitue la reconnaissance du recours nécessaire à l'expertise humaine, en l'occurrence les connaissances encyclopédiques sur le monde de la finance. Nous sommes conscients de la difficulté de collecter cette expertise, soulignée par Habert dans le passage ci-dessous.

> (…) Harris s'appuyait sur un informateur du domaine et utilisait les catégories d'entités fournies par cet informateur comme point de départ pour déterminer les classes d'opérandes en fonction des opérateurs utilisés. Cependant, une partie des recherches actuelles en TALN qui visent à dégager, à partir d'une analyse syntaxique, les opérateurs et leurs arguments au sein d'un domaine donné, essaient souvent de le faire sans ce recours à un premier dégrossissage conceptuel du domaine. L'économie de ce recours s'explique en partie par la difficulté d'obtenir ce type de renseignements : *on dispose de textes d'un domaine spécialisé, mais pas forcément d'informateurs compétents dans ce domaine*[4]. On rencontre aussi la conviction qu'il suffit de disposer d'un ensemble suffisamment vaste de documents du domaine pour que le retraitement d'analyses syntaxiques fasse émerger les régularités syntactico-sémantiques. La question demeure donc : peut-on induire les schémas d'un domaine sans le recours à une expertise humaine, soit au départ, soit pour valider les regroupements produits automatiquement ?
>
> (Habert, 1998, p. 151)

[2] (Prince & Smolensky, 1993).

[3] Par exemple, les relations entre entreprises-mères et filiales.

[4] Italiques ajoutés.

À la question posée par Habert au sujet du recours à l'expertise humaine, notre expérience dans le domaine du filtrage d'information nous inciterait à répondre négativement. Remarquons cependant que le recours à une expertise extérieure n'est pas synonyme d'objectivité, ni de régularité, ni de validité scientifique : le caractère souvent non tranché des avis d'experts, ainsi que la difficulté d'expliciter toute expertise imposent des limites au type d'approche discuté ici, des analyses linguistiques des corpus spécialisés, intégrant une part d'expertise du métier. Cependant, le recours à l'expertise, ainsi qu'à des corpus issus d'une pratique effective, nous paraissent être le garant d'un compromis acceptable entre visée objectivante et insaisissable essence de la connaissance d'un domaine.

5.3. Linguistique et catégories

Nous avons examiné l'influence de deux conceptions de la structuration des observables linguistiques sur les théories linguistiques développées. Nous avons vu quelles limites étaient attachées au cadre catégorique logique, hérité de la métaphysique aristotélicienne. Ces limites ont essentiellement trait à l'impératif de monocatégorialité : un élément donné ne peut appartenir à plusieurs classes, en vertu des principes de non contradiction et du tiers exclu.

Dans les cas où un élément semble manifester une polycatégorialité, il est nécessaire d'introduire des opérations invisibles, supposant, par exemple, une structure apparente et une structure profonde, qui constituerait, en quelque sorte, la « vraie » nature de l'élément considéré. Ainsi, dans un cadre monocatégoriel, les cas de polycatégorialité apparente sont traités par l'homonymie : la similarité formelle est pensée comme cachant une différence profonde, essentielle.

La contrainte de monocatégorialité est-elle nécessaire à une théorie linguistique ?

En effet, elle implique non seulement le recours à la notion d'homonyme, mais également une complexification des modèles construits, devant ménager un plan surfacique et un plan profond. Cette complexité n'est envisageable que dans l'hypothèse où on recherche des jugements tranchés sur l'appartenance catégorielle d'un élément, ce qui est la marque des approches catégoriques.

En ce qui concerne la grammaticalité, le point de vue catégorique implique qu'un énoncé appartient forcément à l'ensemble des phrases de la Langue, ou non. Or, dans la

pratique effective d'une langue, le jugement de grammaticalité est plus affaire d'opinion, d'attentes, de vision plus ou moins normative sur la langue, que de réelle compétence linguistique. Ainsi que le montre (Manning, 2002)[5], attestabilité et grammaticalité ne sont pas synonymes.

Une linguistique de corpus non catégorique doit donc prendre en compte la variation dans les jugements de grammaticalité. Est ce à dire que tous les énoncés sont perçus de la même façon par les locuteurs ? Non, et c'est l'objet du modèle probabiliste de la grammaticalité proposé par (Manning, 2002), fondé sur la théorie de l'optimalité, appliquée à la syntaxe. Dans ce cadre non catégorique, on peut envisager une grammaticalité graduelle, avec des énoncés violant plus ou moins de contraintes de bonne formation.

Soulignons, par ailleurs, que dans la pratique d'une langue, grammaticalité et intelligibilité ne sont pas nécessairement synonymes : il est possible de comprendre des énoncés agrammaticaux. Une théorie linguistique qui vise à rendre compte uniquement des énoncés grammaticaux ne vise donc pas nécessairement à rendre compte de la compréhension du langage naturel.

Une fois posé le caractère non nécessaire des principes de non contradiction et de tiers exclu pour une théorie linguistique, reste à proposer des principes alternatifs. Le passage ci-dessous peut, à ce titre, fournir des indices quant à la nature des principes recherchés.

> L'organisation cognitive ne résulterait pas d'opérations logiques effectuées sur le réel par un esprit *a priori* rationnel mais d'activités plus primitives telles que les deux processus élémentaires et antagonistes de généralisation (négliger les différences sur la base de la ressemblance), et de la discrimination (ne pas confondre) qui s'appliquent sur les plans perceptif, moteur, comportemental et symbolique.
>
> (Dubois 1991, p. 42)

[5] L'auteur rapporte des cas d'énoncés attestés, tirés de corpus journalistiques, violant des contraintes de bonne formation, tels que : *Steven P. Jobs has reemerged as a high-technology captain of industry, as least as the stock market is concerned*. Manning montre que ce type d'énoncés est trop fréquent pour que l'hypothèse d'une coquille isolée soit retenue.

L'hypothèse proposée par Dubois est celle de l'existence de contraintes de portée plus générale que les principes de non contradiction et de tiers exclu : les contraintes de généralisation et de discrimination. Théories linguistiques catégoriques et non catégoriques peuvent ainsi être conçues comme deux points de vue sur la langue, dont le premier favorise la contrainte de discrimination, alors que le second favorise celle de généralisation.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abney S., 1991. Parsing by chunks, *Principle-Based Parsing*, Berwick R., Abney S., Tenny C. (eds.), Kluwer Academic Publishers.

Abney S., 1996.

a) Partial parsing via finite-state cascades, *Proceedings of the ESSLLI'96 Robust Parsing Workshop*.

b) Statistical methods and linguistics, *The balancing act*, Klavans J., Resnik P. (eds.), MIT Press.

Amardeilh F., 2001. *Extraction d'information : étude de faisabilité appliquée au domaine boursier*, mémoire de DEA, université de Troyes.

Apte C., Damerau F., Weiss S.M., 1994. Automated learning of decision rules for text categorization, *ACM Transactions on information systems*, pp. 233-240.

Arampatzis A., van Bommel P., Koster C.H.A., van der Weide Th.P., 1997. Linguistic Variation in Information Retrieval and Filtering, *Technical Report CSI-R9701*, University of Nijmegen.

Auroux S., 1994. *La révolution technologique de la grammatisation*, Mardaga, Liège.

Baker L.D., McCallum A.K., 1998. Distributional clustering of words for text classification, *SIGIR'98*, ACM, Melbourne.

Balvet A., 2001. Filtrage d'information par analyse partielle, *Actes de la cinquième rencontre des étudiants chercheurs en informatique pour le Traitement Automatique des Langues*, 2-5 juillet 2001, pp. 421-431, Tours.

Balvet A., 2001. Grammaires locales et lexique-grammaire pour le filtrage d'information, Vers une réutilisabilité des ressources linguistiques pour la recherche d'information, *Actes des quatrièmes rencontres Terminologie et Intelligence Artificielle*, 3-4 mai 2001, pp.201-211, Nancy.

Balvet A., 2002.

a) Designing Text Filtering Rules: Interaction between General and Specific Lexical Resources, *LREC Workshop on Using Semantics for Information Retrieval*, 27mai-3 juin 2002, Las Palmas.

b) LIZARD, un assistant linguistique, *Actes de la sixième rencontre des étudiants chercheurs en informatique pour le Traitement Automatique des Langues*, 24-27 juin 2002, pp.425-434, Nancy.

Balvet A., Meunier F., Poibeau T., Viard D., Vichot F., Wolinski F., 2001. Le projet CORAIL : utilisation des grammaires locales pour le Filtrage d'information, pp. 34-43, *REE* n°5 juillet-septembre 2001, EDP Sciences SEE.

Balvet A., Meunier F., Poibeau T., Viard D., Vichot F., Wolinski F., 2001. Filtrage de documents et grammaires locales : le projet CORAIL, *Actes du troisième congrès du Chapitre français de l'ISKO (International Society for Knowledge Organisation) : Filtrage et résumé automatique de l'information sur les réseaux*, 5-6 juillet 2001, Université de Nanterre-Paris X.

Bar-Hillel Y., 1964. *Language and information*, Addison-Wesley publishing company.

Belkin N.J., Bruce Croft W., 1992. Information filtering and information retrieval: two sides of the same coin ?, *Communications of the ACM*, vol.35, n°12.

Bellot P., El-Bèze M., 2000. Classification locale non supervisée pour la recherche documentaire, *TAL*, n° 41, Traitement automatique des langues pour la recherche d'information, Hermès Sciences Publications, Paris.

Benveniste E., 1966. *Problèmes de linguistique générale*, Gallimard.

Besançon R., 2002. *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes, application au calcul de similarités sémantiques dans le cadre du modèle DSIR*, thèse de doctorat, école polytechnique fédérale de Lausanne.

Biber D., 1988. *Variations across speech and writing*, Cambridge University Press.

Biber D., 1989. A typology of english texts, *Language*, n°27, pp. 3-43.

Biber D., 1995. *Dimensions of register variation: a cross-linguistic comparison*, Cambridge University Press.

Biber D., Conrad S. & Reppen R., 1998. *Corpus Linguistics, investigating language structure and use*, Cambridge University Press.

Bizouard S., 2001. *Évaluation d'outils d'acquisition de ressources linguistiques pour l'extraction*, mémoire de DESS, Centre de Recherche en Ingénierie Multilingue.

Bloomfield L., 1926. A set of postulates for the science of language, *Language*, n° 2, pp. 153-164.

Bloomfield L., 1933. *Language*, New York.

Boersma P., Hayes B., 2001. Empirical tests of the gradual learning algorithm, *Linguistic Inquiry*, vol. 32, n° 1, pp. 45-86.

Boons J-P., Guillet A., Leclère C., 1976. *La structure des phrases simples en français, constructions intransitives*, Librairie Droz, Genève

Bouaud J., Habert B., Nazarenko A., Zweigenbaum P., 1997. Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles, *Actes des Ières journées Ingénierie des connaissance*, pp. 207-223.

Bourdeau M., 2000. *Locus logicus, l'ontologie catégoriale dans la philosophie contemporaine*, L'Harmattan, France.

Bourigault D., 1994. *Lexter, un logiciel d'extraction de terminologies, Application à l'acquisition des connaissances à partir des textes*, thèse de doctorat, EHESS.

Bourigault D., 2002. Analyse distributionnelle étendue, *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles*, pp. 75-84, Nancy 24-27 juin.

Brill E., 1992. A simple rule-based part-of-speech tagger, *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento.

Briscoe T., 1997. Automatic extraction of argument structure from corpora, *Proceedings of the 5th conference on Applied Natural Language Processing (ANLP-97)*, Washington DC.

Charniak E., 1993. *Statistical language learning*, MIT Press.

Charniak E., 1997. Statistical techniques for Natural Language Processing, *AI Magazine*, vol. 8, n°4, pp. 33-44.

Chomsky N., 1955. *The logical structure of linguistic theory*, Plenum Press, New York.

Chomsky N., 1957. *Syntactic structures*, Mouton, The Hague.

Chomsky N., 1965. *Aspects of the theory of syntax*, MIT Press.

Church K.W., Hanks P., 1990. Word association norms, mutual information, and lexicography, *Computational Linguistics*, vol. 16, n°1, pp. 22-29, MIT Press.

Cleverdon C.W., Mills J., Keen E.M., 1966. *Factors determining the performance of indexing systems*, Cranfield-ASLIB Research project.

Cohen W., 1996. Learning rules that classify E-mail, *Papers from the AAAI Spring Symposium on Machine Learning in Information Access*.

Comte A., 1996. *Philosophie des sciences*, présentation, choix de textes et notes par J. Grange, Gallimard.

Courtois B., 1990. Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française*, n° 87, Larousse, Paris.

Courtois B., Silberztein M., 1990. Les dictionnaires électroniques du français, *Langue Française*, n° 87, pp. 11-22, Larousse, Paris.

Coyaud M., 1972. *Linguistique et documentation*, collection Langue et langage, Larousse université, Paris.

Croft W. B., Lewis D.D., 1987. An approach to Natural Language Processing for Document Retrieval, *Proceedings of the tenth annual international ACM SIGIR Conference on research and development in Information Retrieval (SIGIR'87)*, pp.26-32, New Orleans.

Cullingford R.E., 1978. *Script application: computer understanding of newspaper stories*, these de doctorat, université de Yale.

Cussens J., Page J., Muggleton S., Srinivasan A., 1997. Using Inductive Logic Programming for Natural Language Processing, *Workshop notes of the ECML/MLnet workshop on empirical learning of Natural Language Processing tasks*, Daelemans W., van den Bosch A. & Weijters A. (eds.), Prague.

Daille B., 1994. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*, thèse de doctorat, université Paris VII.

Daille B., 2002. *Découvertes linguistiques en corpus*, thèse d'habilitation, université de Nantes.

Daille B., Royauté J., Fabre C., 2000. Évaluation d'une plate-forme d'indexation de termes complexes, *TAL*, n° 41, Traitement automatique des langues pour la recherche d'information, Hermès Sciences Publications, Paris.

de Saussure F., 1972. *Cours de linguistique générale*, Payot, Paris.

Déjean H., 1998. *Concepts et algorithmes pour la découverte des structures formelles des langues*, thèse de doctorat, université de Caen.

Denning P.J., 1992. Electronic junk, *Communications of the ACM*, n° 25, vol. 3, pp. 163-165.

Dias G., Guilloré S., Bassano J-C., Pereira Lopes J.G., 2000. Extraction automatique d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire, *TAL*, n° 41, Traitement automatique des langues pour la recherche d'information, Hermès Sciences Publications, Paris.

Dister A., 2000. Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes, *Actes des 5èmes Journées d'Analyse des Données Textuelles (JADT)*, Lausanne.

Dubois D., 1991. *Sémantique et cognition, Catégories, prototypes et typicalité*, éditions du CNRS, Paris.

Dumont J.-P., 1962. *La philosophie antique*, Presses Universitaires de France, Paris.

Dunning T., 1993. Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, n°19, vol. 1, pp. 61-74, MIT Press.

Dutoit D., 2000. *Quelques opérations Texte → Sens et Sens → Texte utilisant une sémantique linguistique universaliste apriorique*, thèse de doctorat, Université de Caen.

Elman J.L., 1990. Finding structure in time, *Cognitive Science* n° 14.

Evert S., Krenn B., 2001. Methods for the qualitative evaluation of lexical association measures, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics Toulouse*, France.

Faloutsos C., Oard D., 1995. A survey of Information Retrieval and Filtering methods, *Technical report CS-TR-3514*, Department of computer science, University of Maryland.

Faure D., 2000. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*, thèse de doctorat, université Paris XI.

Fellbaum, S., 1998. *Wordnet : an electronic lexical database*, MIT Press.

Ferret O., Grau B., 2001. Utiliser des corpus pour amorcer une analyse thématique, *TAL*, n° 42, Linguistique de corpus, Hermès Sciences Publications, Paris.

Finch S.P., 1993. *Finding structure in language*, thèse de doctorat, université d'Edinburgh.

Finkelztein-Landau M., Morin E., 1999. Extracting semantic relationships between terms: supervised *vs.* unsupervised methods, *International Workshop on Ontological Engineering on the Global Information Infrastructure*, pp. 71-80, Dagstuhl Castle.

Firth J., 1957. *Papers in linguistics*, Oxford University Press.

Grefenstette G., 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, MA.

Fourour N., 2002. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français, *Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles*, 2-5 juillet 2001, pp. 265-274, Tours.

Fuchs C., 1980. *Paraphrase et théories du langage; contribution à une histoire des théories linguistiques contemporaines et à la construction d'une théorie énonciative de la paraphrase*, thèse de doctorat d'état, Université Paris VII, France

Fuchs C., 1982. *La paraphrase*, Linguistique Nouvelle, collection dirigée par Guy Serbat. Presses Universitaires de France, Paris.

Fuchs C., 1991. *Les typologies de procès*. Actes et Colloques, Fuchs C. (ed.), Klincksieck, Paris.

Fuchs, C., 1993, *Linguistique et traitement automatique des langues*, Hachette, Paris.

Fuchs C., 1994. *Paraphrase et énonciation*, Ophrys, collection L'Homme dans la langue, Paris.

Galliers J.R., Spärck Jones K., 1993. Evaluating Natural Language Processing systems, *Technical report 291*, Computer laboratory, University of Cambridge.

Gold E.M., 1967. Language identification in the limit, *Information and control*, n°16, pp. 447-474.

Goldsmith J., 2001. Unsupervised learning of the morphology of a natural language, *Computational Linguistics*, vol. 27, n°2, pp. 153-198, MIT Press.

Goujon B., 1999. *Utilisation de l'exploration contextuelle pour l'aide à la veille technologique*, thèse de doctorat, Université Paris IV.

Grefenstette G. 1993. Evaluation techniques for automatic semantic extraction: comparing syntactic and window-based approaches, *Workshop on acquisition of lexical knowledge from text, SIGLEX/ACL*, Columbus.

Grefenstette G. 1996. Light Parsing as Finite-State Filtering, *Workshop on Extended Finite State Models of Language, ECAI'96*, Budapest.

Gross M., 1966. On the equivalence of models of language used in the fields of mechanical translation and information retrieval. Automatic Translation of Languages. W10. Oxford: Pergamon Press, pp. 123-137. Reprinted in Tefko Saracevic ed. *Introduction to Information Science*. New York: R.R. Bowker Company, 1970, pp. 210-218.

Gross M., 1967. Linguistique et documentation automatique. *Revue de l'Enseignement Supérieur* 1-2.

Gross M., 1968. *Grammaire transformationnelle du français*. vol. 1, syntaxe du verbe, Cantilène.

Gross M., 1975. *Méthodes en syntaxe*, Hermann, Paris.

Gross M., 1986. *Grammaire transformationelle du français*. vol. 3, syntaxe de l'adverbe, CERIL, Université Paris 7.

Gross M., 1986. *Grammaire transformationnelle du français*, vol. 2, syntaxe du nom, Cantilène.

Gross M., 1988. Les limites de la phrase figée, *Langages*, n° 90, pp. 7-22, Larousse, Paris.

Gross M., 1990. Le programme d'extension des lexiques électroniques. *Langue Française*, n° 87, pp. 123-127, Larousse, Paris.

Gross M., 1993. Les phrases figées en français. *L'information grammaticale*, Paris.

Guillet A., Leclère C., 1992. La structure des phrases simples en français, constructions transitives locatives, Librairie Droz, Genève.

Habert B., Fabre C., 1999. Elementary dependency trees for identifying corpus-specific semantic classes, *Computer and the humanities*, n° 33, vol. 3, 207-219.

Habert B., Nazarenko A., Salem A., 1997. *Les linguistiques de corpus*, Masson.

Halliday M.A.K., 1961. Categories of the theory of grammar, *Word*, vol. 17, n° 3, pp. 241-292.

Hamelin O., 1985. *Le système d'Aristote*, Librairie philosophique J. Vrin, Paris.

Harman D., 1992. The DARPA TIPSTER project, *ACM SIGIR Forum*, vol. 26, n° 2, pp. 26-28.

Harman D., 1993. Overview of the First Text REtrieval Conference, *TREC-1*, NIST Special Publications, Gaithersburg, MD.

Harman D., 1994. Overview of the third Text REtrieval Conference TREC-3, *TREC-3*, NIST Special Publications, Gaithersburg, MD.

Harman D., 1995. Overview of the fourth Text REtrieval Conference TREC-4, *TREC-4*, NIST Special Publications, Gaithersburg, MD.

Harris Z.S., 1951. *Structural Linguistics*, University of Chicago Press.

Harris Z.S., 1968. *Mathematical Structures of Language*, Interscience Publishers, John Wiley & Sons.

Harris Z.S., 1988. *Language and Information*, Columbia University Press, New York.

Harris Z.S., Gottfried M., Ryckman T., Mattick JR P., Daladier A., Harris T., Harris Z., 1989. The form of information in science, Analysis of immunology sublanguage, *Boston studies in the philosophy of science*, vol. 104, Kluwer Academic Publisher.

Harris, Z.S., 1991. *A theory of language and information : a mathematical approach*, Clarendon, Oxford.

Hayes B.P., 1997. Phonetically driven phonology: the role of Optimality Theory and inductive grounding, *Milwaukee conference on formalism and functionalism in linguistics*.

Herdan G., 1962. *The calculus of linguistic observations*, Janua Linguarum, Mouton & Co., The Hague, the Netherlands.

Herdan G., 1964. *Quantitative linguistics*, Butterworths, London.

Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M., 1997. FASTUS: a cascaded finite-state transducer for extracting information in natural-language text, Roche E. & Schabes Y. (éds.), *Finite state language processing*, pp. 383-406, MIT Press.

Hoenkamp E., Schomaker L., Van Bommel P., Koster C.H.A., Van der Weide Th.P. 1996. PROFILE - A Proactive Information Filter, *Initial Project Plan*, University of Nijmegen.

Housman E.M., 1969. Survey of current systems for selective dissemination of information, *Technical Report*, American Society for Information Science Special Interest Group.

Hull D.A. 1997, The TREC-6 filtering track: description and analysis, *TREC-6*, NIST Special Publications, Gaithersburg, MD.

Hull D.A., 1994. *Information Retrieval using statistical classification*, thèse de doctorat, Stanford university.

Hutchens J.L., 1995. *Natural language grammatical inference*, thèse de doctorat, university of Western Australia.

Illouz G., Jardino M., 2001. Analyse statistique et géométrique de corpus textuels, *TAL*, n° 42, Linguistique de corpus, Hermès Sciences Publications, Paris.

Jackendoff R., 1983. *Semantics and cognition*, MIT University Press.

Jacquemin C., 1997. *Variation terminologique : reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*, thèse d'habilitation à diriger des recherches, université de Nantes.

Jacquemin C., Zweigenbaum P., 2000. Traitement automatique des langues pour l'accès au contenu des documents, *Le document en sciences du traitement de l'information*, Le Maître J., Charlet J., Garbay C. (éds.), pp. 71-109, Cepadues, Toulouse.

Kahane S., 1999. The Meaning Text Theory, Dependency and Valency, *An International Handbook of Contemporary Research*, De Gruyter, Berlin.

Karlgren J., Cutting D., 1994. Recognizing text genres with simple metrics using discriminant analysis, *Fifteenth international conference on Computational Linguistics (COLING'94)*, Kyoto.

Karttunen L., 2000. Applications of finite-state transducers in Natural Language Processing, *Proceedings of CIAA-2000*, *Lecture Notes in Computer Science*, Springer Verlag.

Klavans J., Kan M-N., 1998. Role of verbs in document analysis, *COLING-ACL 1998 Proceedings*, pp. 680-686, Université de Montréal.

Klein D., Manning C.D., 2001. Distributional phrase structure induction, *CoNLL 2001*.

Krenn B., 2000. Empirical implications on lexical association measures, *Rapport de recherche*.

Krenn B., Evert S., 2001. Can we do better than frequency? A case study on extracting PP-verb collocations, *Proceedings of the ACL Workshop on Collocations* Toulouse, France.

Krenn B., Samuelsson C., 1997. *The linguist's guide to statistics, Don't panic*.

Kushmerick N., Johnston E., McGuinness S., 2001. Information extraction by text classification, *IJCAI-01 Workshop on Adaptive Text Extraction and Mining (ATEM 2001)*, Seattle.

Labov W., 1973. The boundaries of words and their meanings, C.-J. Bailey & R. Shuy (eds.), *New Ways of Analyzing Variation in English*, pp. 340-373, Georgetown University Press.

Lakoff G., 1987. *Women, fire and dangerous things*, Chicago University Press.

Landi B., *et al.*, 1998. Amaryllis: An evaluation experiment on search engines in a French-speaking context, *LREC*, pp. 1211-1214, Grenade.

Langacker R.W., 1999. *Grammar and conceptualization*, Cognitive linguistics research vol.14, Dirven R., Langacker R.W. & Taylor J.R. eds., Mouton de Gruyter.

Laporte E., 1988. La reconnaissance des expressions figées lors de l'analyse automatique, *Langage*, n° 90, Larousse, Paris.

Lebart L., Salem A., 1994. *Statistique textuelle*, Dunod, Paris.

Leclère C., 1990. Organisation du lexique-grammaire des verbes français. *Langue Française*, n° 87, Larousse, Paris.

Lee L.J., 1997. *Similarity-based approaches to Natural Language Processing*, Harvard university.

Lehnert W., McCarthy J., Soderland S., Riloff E., Cardie C., Peterson J., Feng F., Dolan C., Goldman S., 1993. UMASS/HUGHES: description of the CIRCUS system used for MUC-5, *Proceedings of the 5th Message Understanding Conference (MUC-5*), pp. 277-291, Morgan Kauffman, San Francisco.

Lespinasse K., Kremer P., Schibler D., Schmitt L., 1999. Évaluation des outils d'accès à l'information textuelle: les expériences américaines (TREC) et française (Amaryllis), *Aupelf-Uref*, John Libbey Eurotext.

Levin B., 1993. *English verb classes and alternations*, University of Chicago Press.

Lewis D., 1996. The TREC-5 filtering track, *TREC*-5, NIST Special Publications, Gaithersburg, MD.

Lewis D., Hill M., 1995. The TREC-4 Filtering Track, *TREC-4*, NIST Special Publications, Gaithersburg, MD.

Lewis D.D., 1991. Evaluating text categorization, *Proceedings of the speech and natural language workshop*, Asilomar, Morgan Kauffman.

Lewis D.D., 1992. *Representation and learning in Information Retrieval*, thèse de doctorat, university of Massachussets.

Lewis D.D., Croft B.W., 1990. Term clustering of syntactic phrases, *Proceedings of the thirteenth annual international ACM SIGIR Conference on research and development in Information Retrieval (SIGIR'90)*, pp. 385-404, Bruxelles.

Lewis D.D., Sparck-Jones K., 1996. Natural Language Processing for Information Retrieval, *Communications of the ACM*, vol.39, n°1, pp. 92-101.

Lewis D.D., Tong R.M., 1992. Text filtering in MUC-3 and MUC-4, *Fourth Message Understanding Conference (MUC-4)*, Morgan Kauffman.

Li W., 1992. Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Transactions on information theory*, vol. 38, n° 6, pp. 1842-1845.

Lin D., 1992. *Obvious abduction*, thèse de doctorat, university of Alberta.

Lin D., 1998.

a) An Information-Theoretic definition of similarity, *Proceedings of International Conference on Machine Learning*, Madison.

b) Extracting collocations from text corpora, *First Workshop on Computational Terminology*, Montreal.

c) Using collocation statistics in information extraction.

Losee R.M., 1996. How part-of-speech tags affect text retrieval and filtering performance, *Rapport de recherche*.

Luhn H.P., 1958. A business intelligence system, *IBM Journal of Research and Development*, vol. 2, n° 4, pp. 314-319.

Malone T.W., Grant K.R., Turbak F.A., Brobst S.A., Cohen M.D., 1987. Intelligent information sharing systems, *Communications of the ACM*, vol. 30, n° 5, pp. 390-402.

Malrieu D., Rastier F., 2001. Genres et variations morphosyntaxiques, *TAL*, n° 42, Linguistique de corpus, Hermès Sciences Publications, Paris.

Manning C.D., 1993. Automatic acquisition of a large subcategorization frame dictionary from corpora, *31st Annual meeting of the Association for Computational Linguistics*, pp. 235-242.

Manning C.D., Schütze H., 1999. *Foundations of statistical natural language processing*, MIT Press.

Manning C.D., 2002. Probabilistic syntax, *Probabilistic Linguistics*, Bod, Hay & Jannedy (eds.), MIT Press.

Manzi S., King M., Douglas S., 1996. Working towards user-oriented evaluation, *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 96)*, Moncton, New-Brunswick, Canada.

Mariani J., 1999. Traitement automatique de la langue française utilisant le paradigme d'évaluation, *Aupelf-Uref*, John Libbey Eurotext, France

Markovitch S., 1989. *Information Filtering: selection mechanisms in learning systems*, thèse de doctorat, university of Michigan.

Martin D., Cheyer A.J., Moran D.B., 1999. The Open Agent Architecture: a framework for building distributed software systems, *Applied Artificial Intelligence*, vol. 13, pp. 91-128.

Martinet, 1985. *Syntaxe générale*, Armand Colin, Paris.

Mason O., 2000. *Programming for Corpus Linguistics*, Edinburgh Textbooks in Empirical Linguistics, Edinburgh University Press.

McEnery T., Wilson A., 1996. *Corpus linguistics*, Edinburgh University Press.

McMahon J.G.J., 1994. *Statistical language processing based on self-organising word classification*, thèse de doctorat, The Queen's university of Belfast.

Mel'Cuk I.A., Clas A., Polguère A., 1995. Introduction à la lexicologie explicative et combinatoire. *AUPELF-UREF, Champs Linguistiques*, collection dirigée par Dominique Willems, Editions Duculot, Louvain-la-Neuve.

Meunier F., Balvet A., Poibeau T., 1999. Projet CORAIL COmposition de Requêtes par des Agents Intelligents Linguistiques, *Linguisticae Investigationes*, XXII, pp. 369-381, John Benjamins B.V, Amsterdam.

Michel C., 1999. *Évaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes*, thèse de doctorat en Sciences de l'Information et de la Communication, Université Lumière Lyon II.

Miles O., 1999. DCG Induction using MDL and Parsed Corpora, *Learning Language in Logic*, Cussens J. (ed.), pp. 63-71, Bled.

Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J., 1990. Introduction to Wordnet: an on-line lexical database, *International journal of lexicography (special issue)*, vol. 3, n°4, pp. 235-313.

Milner J-C., 1985. De l'inutilité des arbres en linguistique, Laboratoire de Linguistique Formelle, Unité de Formation et de Recherches Linguistiques.

Mohri M. 1995. On some Applications of Finite-State Automata Theory to Natural Language Processing, *Natural Language Engineering*, vol. 1, Cambridge University Press.

Mohri M., 1993. *Analyse et représentation par automates de structures syntaxiques composées*, thèse de doctorat, université Paris VII.

Mohri M., 1997. Finite-state transducers in language and speech processing, *Computational Linguistics*, vol.23, n°2, pp. 269-311, MIT Press.

Mohri M., 2001. Language processing with weighted transducers, *Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles*, 2-5 juillet 2001, pp.5-14, Tours.

Morin E., 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*, thèse de doctorat, université de Nantes.

MUC-3, 1991. Proceedings of the Third Message Understanding Conference, *MUC-3*, Morgan Kaufmann, San Mateo CA.

Muller C., 1973. *Initiation aux méthodes de la statistique linguistique*, Collection Unichamp, Champion.

Nasr A., 1996. *Un modèle de reformulation automatique fondé sur la Théorie Sens-Texte. Application aux langues contrôlées*, thèse de doctorat, Université Paris VII.

Naulleau E., 1997. *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*, thèse de doctorat, Université Paris XIII Villetaneuse.

Oard D., 1996. *Adaptive vector space text filtering for monolingual and cross-lingual applications*, thèse de doctorat, university of Maryland.

Oard D.W., Marchionini G., 1996. A Conceptual Framework for Text Filtering, *Technical Report CS-TR-3613*, university of Maryland.

Osborne M., 1999. MDL-based DCG Induction for NP Identification, *Osborne M. & Tjong Kim Sang E. (eds), CoNLL99*, pp. 61- 68, Bergen.

Pedersen T., Kayallp M., Bruce R., 1996. Significant lexical relationships, *Proceedings of the 13th national conference on Artificial Intelligence*, Portland.

Pereira F., 2000. Formal grammar and information theory: together again?, *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, n° 358, pp. 1239-1253, The Royal Society, London.

Pereira F., Tishby N., Lee L., 1993. Distributional clustering of English words, *Proceedings of the 31st annual meeting of the Association for Computational Linguistics, ACL*, pp. 183-190.

Piattelli-Palmerini J., 1979. *Théories du langage, théories de l'apprentissage, le débat entre J. Piaget et N. Chomsky*, Centre Royaumont pour une science de l'homme.

Poibeau T., 1999. Évaluation des systèmes d'extraction d'information: une expérience sur le français, *Aupelf-Uref*, John Libbey Eurotext, France

Poibeau T., 2002, *Extraction d'information à base de connaissances hybrides*, thèse de doctorat, université Paris XIII.

Poibeau T., Balvet A., 2001. Corpus-based lexical acquisition for Information Extraction, *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining (ATEM 2001)*, Seattle.

Popescu-Belis A., 1999. L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures, *Aupelf-Uref*, John Libbey Eurotext, France.

Prince A., Smolensky P., 1993. Optimality Theory, Constraint interaction in generative grammar, *Technical Report*, ROA.

Pustejovsky J., 1996. *The generative lexicon*, MIT Press.

Rajman M., Besançon R., Chappelier J-C., 2000. Le modèle DSIR : une approche de sémantique distributionnelle pour la recherche documentaire, *TAL*, n° 41, Traitement

automatique des langues pour la recherche d'information, Hermès Sciences Publications, Paris.

Ram A., 1991. Interest-based information filtering and extraction in Natural Language Understanding systems, *Bellcore workshop on High-Performance Information Filtering*, Morristown.

Riloff E., 1994. *Information Extraction as a Basis for Portable Text Classification Systems*, thèse de doctorat, université du Massachussets Amherst.

Riloff E., 1995. Little words can make a big difference for text classification, *Proceedings of the 18th annual international conference on research and development in information retrieval (SIGIR'95)*, pp.130-136, Seattle.

Riloff E., 1996. Using learned extraction patterns for text classification, *Connectionist, statistical and symbolic approaches for Natural Language Processing*, Wermter S., Riloff E. & Scheler G. (eds.), pp.275-289, Springer-Verlag, Berlin.

Robertson S., Hull D.A., 2001. The TREC-9 Filtering Track Final Report, *TREC-9*, NIST Special Publications, Gaithersburg, MD.

Robin L., 1973. *La pensée grecque et les origines de l'esprit scientifique*, Albin Michel.

Roche E. 1993. *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*, thèse de doctorat, Université Paris VII.

Roche E. 1993. Une représentation par automate fini des textes et propriétés transformationnelles des verbes, *Linguisticae Investigationes*, XVII, vol. 1, pp. 189-222, John Benjamins B.V, Amsterdam.

Roche E., Schabes Y., 1997. *Finite State Language Processing*, Cambridge, MIT Press.

Rungsawang A., 1997. *Distributional Semantis based Information Retrieval*, thèse de doctorat ENST-Paris.

Sager N., Friedman C., 1987. *Medical language processing: computer management of normative data*, Addison-Wesley.

Salton G., 1968. *Automatic Information Organization and Retrieval*, McGraw-Hill Book Co., New-York.

Salton G., 1971. *The SMART retrieval system*, Prenctice-Hall.

Sapir E., 1921. *Language: an introduction to the study of speech*, Harcourt Brace, New York.

Schulte im Walde S., 1998. *Automatic semantic classification of verbs according to their alternation behaviour*, thèse de doctorat, Institut für Maschinelle Sprachverarbeitung.

Sekine S., Carroll J., Ananiadou S., Tsujii J-I, 1992. Automatic Learning for Semantic Collocation, *3rd Conf. on Applied Natural Language Processing 1992*, Trento.

Senellart J., 1999. *Outils de reconnaissance d'expressions linguistiques complexes dans de grands corpus*, thèse de doctorat, université Paris VII.

Séguéla P., 2002. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*, thèse de doctorat, université Toulouse III.

Shannon C.E., 1948. A mathematical theory of communication, *Bell system technical journal*, n° 27, pp. 379-423, 623-656.

Silberztein M., 1989. *Dictionnaires électroniques et reconnaissance lexicale automatique*, thèse de doctorat, université Paris VII.

Silberztein M., 1990. Le dictionnaire électronique des mots composés. *Langue Française*, n° 87, pp. 71-83, Larousse, Paris.

Silberztein M., 1993. *Le système INTEX, Dictionnaires électroniques et analyse automatique des textes*, Paris, Masson.

Silberztein M., 1999. *Documentation du système INTEX*, LADL, Paris.

Silberztein M., 1999. Traitement des expressions figées avec INTEX, *Linguisticae Investigationes*, XXII, pp. 425-449, Fairon C. (éd.), John Benjamins B.V, Amsterdam.

Slonim N., Tishby N., 2001. The power of word clusters for text classification, *23rd European colloquium on Information Retrieval research*.

Smadja F., 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, vol. 19, n° 1, pp. 143-177.

Soderland S., 1997. *Learning text analysis rules for domain-specific Natural Language Processing*, thèse de doctorat, university of Massachussets Amherst.

Spärck Jones K., 1995. Reflections on TREC, *Information processing and management*, vol. 31, n°3, pp 291-314.

Spärck Jones K., Kay M., 1973. *Linguistics and information science*, Academic Press, New York.

Spärck Jones K., Van Rijsbergen C., 1975. Report on the need for and provision of an ideal information retrieval test collection, *British Library Research and Development Report 5266*, Computer Laboratory, University of Cambridge.

Sta J. D., 1997. *Acquisition terminologique en corpus: aspects linguistiques et statistiques*, thèse de doctorat, université Paris VII.

Stevens C., 1992. Automating the creation of information filters, *Communications of the ACM*, vol. 35, n° 12, p. 48.

Stricker M., 2000. *Réseaux de neurones pour le traitement automatique du langage : conception et réalisation de filtres d'informations*, thèse de doctorat en informatique. Université Paris VI.

Strzalkowski T., Guthrie L., Karlgreen J., Leistensnider J., Lin F., Perez Carballo J., Straszheim T., Wang J., Wilding J. 1996. Natural language information retrieval: TREC-5 report, *TREC-5*, NIST Special Publications, Gaithersburg, MD.

Strzalkowski T., Lin F., Perez Carballo J., 1997. Natural language information retrieval: TREC-6 report, *TREC-6*, NIST Special Publications, Gaithersburg, MD.

Strzalkowski T., Perez Carballo J. 1995. Natural Language Information Retrieval : TREC-4 Report, *TREC-4*, NIST Special Publications, Gaithersburg, MD.

Strzalkowski T., Perez Carballo J., Marinescu M., 1994. Natural language information retrieval: TREC-3 report, *TREC-3*, NIST Special Publications, Gaithersburg, MD.

Tapanainen P., Järvinen T., 1994. Syntactic analysis of natural language using linguistic rules and corpus-based patterns, *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING'94)*, vol. 1, pp. 629-634, Kyoto.

Taylor J.R., 1995. *Linguistic categorization, prototypes in linguistic theory,* second edition, Clarendon Press Oxford.

Trotignon P., 1968. *Aristote, L'Analytique*, Presses Universitaires de France, Paris.

Turenne N., 2000. *Apprentissage statistique pour l'extraction de concepts à partir de textes, application au filtrage d'informations textuelles*, thèse de doctorat, université Louis-Pasteur Strasbourg.

Vergne J., 2001. Analyse syntaxique automatique des langues : du combinatoire au calculatoire, *Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles*, 2-5 juillet 2001, pp.15-29, Tours.

Viard D., 2000.

a) Évaluation ergonomique du système CORAIL, *Rapport de projet*, Consortium CORAIL.

b) Évaluation et recommandations ergonomiques pour le logiciel Intex, *Rapport de projet*, Consortium CORAIL.

Voorhees E., Harman D, 1996. Overview of the fifth Text REtrieval Conference TREC-5, *TREC-5*, NIST Special Publications, Gaithersburg, MD.

Voorhees E., Harman D. 1997. Overview of the sixth Text REtrieval Conference TREC-6, *TREC-6*, NIST Special Publications, Gaithersburg, MD.

Voorhees E., Harman D., 1998. Overview of the Seventh Text REtrieval Conference TREC-7, *TREC-7*, NIST Special Publications, Gaithersburg, MD,

Voorhees E., Harman D., 2001. Overview of the Ninth Text REtrieval Conference, *TREC-9*, NIST Special Publications, Gaithersburg, MD.

Wittgenstein L., 1961. *Tractatus logico philosophicus*, Gallimard.

Yan W.T., Garcia-Molina H., 1995. SIFT-A tool for wide-area Information Dissemination, *Proceedings of the 1995 USENIX Technical Conference*, pp. 177-86.

Yang Y., 1998. An evaluation of statistical approaches to text categorization, *INRT Journal*, Kluwer Academic Publishers.

Yangarber R., 2001. *Scenario customization for Information Extraction*, thèse de doctorat, New York University.

Yangarber R., Grishman R., 2000. Extraction pattern discovery through corpus analysis, LREC 2000.

Zhai C., Tong X, Milic-Frayling N., Evans D. A., 1996. Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report, *TREC-5, NIST Special Publication*, Gaithersburg, MD.

Zipf G.K., 1945. The meaning-frequency relationship of words, *Journal of general psychology*, n°33, pp. 251-256.